Università degli Studi di Milano-Bicocca

QUA_SI Project
PhD Program in Information Society
XXII Ciclo - I
Anno Accademico 2008–2009

CULTURAL HERITAGE ON THE SEMANTIC WEB: FROM REPRESENTATION TO FRUITION

Glauco Mantegari
Ph.D. Dissertation

Advisor: Prof. Stefania Bandini

Contents

1 Introduction **3**
- 1.1 Research Perspective and Objectives . . . 5
- 1.2 Outline of the Dissertation . . . 7

2 1.0 to *n*.0: New Directions in the Development of the Web **11**
- 2.1 The "Web 2.0" . . . 11
- 2.2 Towards the "Web 3.0" . . . 14
- 2.3 Beyond 2.0 and 3.0: The Emerging Key Elements . . . 18
  - 2.3.1 User Generated Content . . . 18
  - 2.3.2 Combining Applications: From APIs to Mashups . . . 22
  - 2.3.3 The Web of Data and the Semantic Web . . . 25
- 2.4 The Enabling Technologies: An Overview . . . 29
  - 2.4.1 XML Technologies: Enabling Syntactic Interoperability . . . 30
  - 2.4.2 APIs and Web Services . . . 31
  - 2.4.3 Mashup Development: Strategies and Techniques . . . 34
  - 2.4.4 Towards Semantics . . . 38

3 The Semantic Web, Cultural Heritage and Archaeology **47**
- 3.1 The Semantic Web in Cultural Heritage . . . 47
  - 3.1.1 ARTISTE - SCULPTEUR - eCHASE . . . 48
  - 3.1.2 MultimediaN E-Culture - CHIP Browser . . . 49
  - 3.1.3 Museum Finland - Culture Sampo . . . 53
  - 3.1.4 Contexta SR . . . 57
- 3.2 The Semantic Web in Archaeology . . . 58
  - 3.2.1 Vbi Erat Lvpa . . . 59
  - 3.2.2 STAR . . . 60
  - 3.2.3 The Port Networks Project . . . 61
  - 3.2.4 Other Projects . . . 63
- 3.3 Around and Beyond the Semantic Web . . . 64
  - 3.3.1 Digital Libraries . . . 65
  - 3.3.2 E-Science and Cyberinfrastructure . . . 67

4 From Representation to Fruition: An Interdisciplinary Research Perspective **73**
- 4.1 Initial Remarks . . . 73
  - 4.1.1 Defining a Research Approach . . . 78
- 4.2 From Representation . . . 80
  - 4.2.1 Core Domain Ontologies: The CIDOC CRM Approach . . . 83
  - 4.2.2 The CIDOC CRM in Use . . . 84
- 4.3 . . . to Fruition: Retrieving Information through Fuzzy Chronologies . . . 92

4.3.1 Archaeological Chronologies: A General Overview 93
4.3.2 Related Work 97
4.3.3 A New Approach Based on Fuzzy Sets 105

5 MANTIC: The Archaeology of Milan on the Semantic Web 111
5.1 The *Milano Antica* Project 111
5.2 MANTIC: a Semantic Portal for *Milano Antica* 113
5.3 Selection of Relevant Data Sources 115
5.3.1 SIRBeC 117
5.3.2 IDRA 120
5.3.3 MANTIC 120
5.4 Conceptual Mapping to the CIDOC CRM 121
5.4.1 SIRBeC Mapping 122
5.4.2 IDRA Mapping 130
5.4.3 MANTIC Mapping 134
5.5 Discussion 137
5.5.1 General Considerations 137
5.5.2 Defining Event-Based Mapping Chains: Issues and Challenges . 142
5.5.3 Representing Actual Instance Data Values 148
5.5.4 Managing Vocabularies 149
5.5.5 Coreference Linking 152

6 ChronoMANTIC: Representing and Retrieving Fuzzy Chronologies 155
6.1 MANTIC Chronologies 155
6.1.1 References to Absolute Chronology 156
6.1.2 References to Historical or Archaeological Periods 158
6.2 Representing Fuzzy Temporal Intervals 159
6.2.1 Determining Fuzzy Temporal Intervals from Label Information . 159
6.2.2 Representing Fuzzy Temporal Intervals for Historical and Archaeological Periods 162
6.3 Evaluation Setting and Criteria 164
6.3.1 An Interface for the Evaluation 165
6.4 Results 168
6.4.1 Defining the Weights for Relevance Measure 168
6.4.2 Analyzing Precision and Recall 168
6.5 Discussion 170
6.5.1 The Model 170
6.5.2 The Retrieval Method 173

7 DemoMANTIC: Design and Development of a Prototype System **175**
7.1 Previous Work: The MANTIC 1.0 System . . . 175
7.1.1 System Architecture . . . 177
7.1.2 Browsing Functionalities . . . 178
7.1.3 Towards a Semantic Web System . . . 178
7.2 Semantic Repositories: General Remarks . . . 180
7.2.1 The Framework Sub-System . . . 181
7.2.2 The Triplestore Sub-System . . . 181
7.2.3 The Reasoner Sub-System . . . 182
7.2.4 Technological Survey . . . 183
7.2.5 Evaluation . . . 185
7.3 The MANTIC 2.0 Architecture . . . 186
7.4 Concluding Remarks . . . 190

8 Concluding Remarks and Future Directions **193**

A Appendix: The Mapping Templates **199**

List of Figures

2.1 A schema representing the difference between the Web of Documents and the Web of Data. . . . 16
2.2 An example of User Generated Content in Google Earth. . . . 21
2.3 A schema representing different types of mashup. . . . 24
2.4 Numbers of mashups available in ProgrammableWeb.com in the semester from July 2008 to January 2009. . . . 25
2.5 Types of mashups available in ProgrammableWeb.com. . . . 26
2.6 The Linking Open Data dataset "cloud". . . . 28
2.7 An example XML structure and its representation as a graph. . . . 30
2.8 Protocol usage by APIs. . . . 32
2.9 Schema of a REST service. . . . 33
2.10 Schema of a SOAP envelope. . . . 33
2.11 The traditional model for Web applications compared to the Ajax model. 35
2.12 An example of the Yahoo Pipes visual editor. . . . 37
2.13 The Semantic Web pile. . . . 39
2.14 URI, URL, and URN. . . . 40
2.15 The structure of an RDF statement. . . . 40
2.16 An example of an RDF graph. . . . 41
2.17 An example showing a scenario where knowledge can be inferred starting from RDFS constructs and an RDF statement. . . . 43

3.1 The MultimediaN E-Culture data cloud. . . . 51
3.2 The architecture of the MultimediaN E-Culture Demonstrator. . . . 52
3.3 The architecture of MuseumFinland on the server side. . . . 55
3.4 The proposed Portus ontology. . . . 63
3.5 A tree-tier framework for the Digital Library Universe. . . . 66
3.6 A general schema of integrated cyberinfrastructure services. . . . 70
3.7 A model of a cyberinfrastructure for archaeology. . . . 71

4.1 Cultural heritage: from representation to fruition. . . . 75
4.2 The basic mapping schema proposed by Kondylakis. . . . 88
4.3 An example of both graph and tabular representation of the mappings to the CIDOC CRM. . . . 89
4.4 CIDOC CRM available serializations. . . . 90
4.5 A representation of the dating and duration properties. . . . 95
4.6 An example of deposition events as meetings. . . . 97
4.7 Determinacy and indeterminacy temporal intervals according to the CIDOC CRM. . . . 98
4.8 A schema of relevant classes and properties for temporal representation in the CIDOC CRM and the Erlangen CRM. . . . 101

4.9 The CIDOC CRM implementation of Allen's relationships between temporal intervals. . . . 102
4.10 CIDOC CRM temporal properties of the `E52 Time-Span` class. . . . 103
4.11 The CIDOC CRM and Erlangen CRM representation of fuzzy temporal intervals. . . . 104
4.12 The period "from the beginning of the 1st century B.C. to the first half of the 1st century A.D." represented as a fuzzy temporal interval. . . . 107
4.13 A fuzzy temporal interval "Pre-Roman Age" and its intersection with another fuzzy temporal interval "Roman Age". . . . 108

5.1 Excavations of the Santa Tecla basilica in the central area of the city in 1961. . . . 112
5.2 Archaeological excavations in Milan: photos posted by a Flickr user. . . . 116
5.3 A polygon related to archaeological features of Milan on Wikimapia . . . 116
5.4 A general schema of the SIRBeC cataloging process. . . . 119
5.5 A graph representing the main activities identified in SIRBeC. . . . 124
5.6 The production of artifacts in SIRBeC and the corresponding graph in CIDOC CRM. . . . 125
5.7 The discovery of artifacts in SIRBeC and the corresponding graph in CIDOC CRM. . . . 127
5.8 A simplified view of the cataloging process in SIRBeC and the corresponding graph in CIDOC CRM. . . . 129
5.9 A graph representing the main activities identified in IDRA. . . . 132
5.10 The archaeological excavation activity in IDRA and the corresponding graph in CIDOC CRM. . . . 133
5.11 The cataloging activity in IDRA and the corresponding graph in CIDOC CRM. . . . 135
5.12 The definition of legal constraints on an archaeological site in IDRA and the corresponding graph in CIDOC CRM. . . . 136
5.13 A graph representing the main activities identified in MANTIC. . . . 137
5.14 The production of structures in MANTIC and the corresponding graph in CIDOC CRM. . . . 138
5.15 The cataloging of sites in MANTIC and the corresponding graph in CIDOC CRM. . . . 139
5.16 The cataloging of structures in MANTIC and the corresponding graph in CIDOC CRM. . . . 140
5.17 Redundant modeling using additional properties and shortcuts. . . . 147
5.18 Coreference graph. . . . 153

6.1 A depiction of the possible references to absolute chronology showing different levels of imprecision. . . . 158

6.2 A schema depicting the basic fuzzy intervals that can be represented with reference to absolute chronology. . . . 160
6.3 A general chronological schema of the ancient history of Milan. . . . 163
6.4 A graphical representation of the transition between the Insubrian and the Roman periods using trapezoidal fuzzy sets. . . . 164
6.5 The selected periods and their representation as trapezoidal fuzzy sets. . 166
6.6 The interface for the evaluation. . . . 167
6.7 Scatter plots of relevance ranking based on intersection confidence and combination of weighted measures. . . . 169
6.8 Precision and recall curves: average recall versus precision for each measure used to rank the results and generalized curve of the results. . . 171
6.9 A graphical representation of the fuzzy interval "Russian Revolution happens". . . . 172

7.1 Use cases of the MANTIC 1.0 system. . . . 176
7.2 General architecture of the MANTIC 1.0 system. . . . 177
7.3 The Mantic 1.0 interface: selection of a value from the typology facet. . . 179
7.4 The Mantic 1.0 interface: selection of a point on the map. . . . 179
7.5 The Mantic 1.0 interface: details of a monument. . . . 180
7.6 The structure of a generic semantic repository. . . . 181
7.7 The analyzed technologies and their models. . . . 184
7.8 General architecture of the MANTIC 2.0 system. . . . 188
7.9 A general schema of the data and control flows in the MANTIC 2.0 system. . . . 189
7.10 Benchmark of the performances of the MANTIC systems. . . . 191

A.1 SIRBeC to CIDOC CRM mapping template: production of artifacts. . . . 199
A.2 SIRBeC to CIDOC CRM mapping template: discovery of artifacts. . . . 200
A.3 SIRBeC to CIDOC CRM mapping template: cataloging of artifacts (part I). . . . 201
A.4 SIRBeC to CIDOC CRM mapping template: cataloging of artifacts (part II). . . . 202
A.5 SIRBeC to CIDOC CRM mapping template: cataloging of artifacts (part III). . . . 203
A.6 SIRBeC to CIDOC CRM mapping template: cataloging of artifacts (part IV). . . . 204
A.7 IDRA to CIDOC CRM mapping template: the archaeological excavation activity. . . . 205
A.8 IDRA to CIDOC CRM mapping template: cataloging of archaeological sites. . . . 206
A.9 IDRA to CIDOC CRM mapping template: the definition of legal constraints on archaeological sites. . . . 207

A.10 MANTIC to CIDOC CRM mapping template: the production of structures. 208
A.11 MANTIC to CIDOC CRM mapping template: cataloging of archaeological sites. 209
A.12 MANTIC to CIDOC CRM mapping template: cataloging of structures. . 210

List of Tables

2.1 A comparison between Web 2.0 and Web 3.0. 17
2.2 Types of User Generated Content. 20
2.3 The Web of documents and the Web of Data. 27

5.1 List of the metadata that have been selected from the SIRBeC-RA card schema. 122
5.2 List of the selected fields from the IDRA database. 131
5.3 The Erlangen CRM implementation of datatype properties. 148
5.4 List of the implemented vocabularies. 151

7.1 Evaluation of the basic dimensions of different semantic repositories. . . 186
7.2 Evaluation of supplementary dimensions of different semantic repositories. 186
7.3 Final evaluation of different semantic repositories. 187

Acknowledgments

There are many people I would like to thank for the support and contributions they gave to my research, which made it possible for me to complete this PhD work.

First of all, I would like to thank Prof. Stefania Bandini, my PhD Advisor, for having stimulated and supported not only this work, but also all the interdisciplinary research initiatives I had the opportunity to carry out during my years at the University of Milano-Bicocca. I would also like to thank Prof. Carla Simone and Prof. Marco Martini for the possibility of working on the "Milano Antica" project they are coordinating at the University, which gave me the opportunity to fruitfully verify my research ideas. Special thanks are due to Dr. Giuseppe Vizzari who has supported in many occasions my research (including this PhD work) during the years, with suggestions and discussions. Even if not directly involved in this work, I moreover acknowledge the support Prof. Roberto Bisiani gave me in my early experiences on new technological applications for archaeology, contributing to the enrichment of my research background and to the refinement of my research perspectives.

My warmest thanks go also to people at the Regione Lombardia who offered me the opportunity to use archaeological data from the SIRBeC system, and in particular Ninfa Cannada Bartoli, Enzo Minervini and Rita Gigante, together with the data co-owners at the University of Milano-Cattolica and the Archaeological Museum of Milan.

I acknowledge the support that Donatella Caporusso of the Archaeological Museum of Milan gave me on different occasions, supplying additional research material and discussing some specific domain issues this work analyzes.

A relevant part of this work, and specifically sections 4.3 and chapter 6, would not have been possible without the research collaboration with the Semantic Computing Research Group at the Helsinki University of Technology. I am extremely grateful for the possibility Prof. Eero Hyvönen gave me in spending an exciting and fruitful research period in his group and for the great support he offered. Special thanks go to Tomi Kauppinen, who not only agreed in doing joint research, but also crucially contributed to it with discussions and ideas and made its model and tools available for experimentation in my case study. My gratitude goes also to Eetu Mäkelä, Panu Paakkarinen and Heini Kuittinen, who actively participated in this research. I would also like to mention other people who made my stay at the University and in Helsinki extremely pleasant, and in particular Tuukka Ruotsalo, Jouni Tuominen, Kim Viljanen, Katariina Nyberg, Tuomas Palonen, and Antti-Jussi Nygård.

This work has greatly benefited from the support of colleagues and friends at the Artificial Intelligence Laboratory, as well as at the Complex Systems and Artificial Research Centre. In particular, Andrea Bonomi and Matteo Palmonari gave precious comments and insights on parts of my work.

Last but not least, I would like to thank my parents, Luciano and Antonia, for their support and understanding throughout all the years I spent at the University.

1 Introduction

In the last few years, the scenario of the Web has evolved dramatically, as new technological elements emerged and a wide range of applications supporting everyday activities became available: terms such as "Web 2.0", "social networks", "Semantic Web", "mashups" and many others are, to different extents, crossing the boundaries of specialized research and becoming familiar even to non technical audiences. The result is that the Web today is such a composite scenario made up of a large number of theoretical, social and technical aspects evolving simultaneously that it seems impossible to completely dominate it.

This is certainly true, at least if we analyze it transversally, i.e. combining the more technological elements with their usage and impact on human communities and their social practices. However, relevant high-level elements can be identified with respect to the modalities through which users interact with the Web today. In fact, a significant shift occurred in less than a decade under the "**Web 2.0**" paradigm, which has been proposed and popularized by O'Reilly (2005) in order to identify new applications which enable users to actively participate in the creation, sharing and aggregation of Web contents, rather than in their mere passive visualization. From this point of view, the era of the "webmaster" which characterized the "1.0" phase is over, since most of the contents today are generated by the users. User Generated Content (UGC) covers a vast spectrum of typologies OECD (see e.g. 2007), including also sophisticated products that were a prerogative of professionals only a few years ago.

On the other hand, this shift occurred in parallel to significant changes in the Web application development cycle, which today is increasingly based on modular approaches that combine publicly accessible applications and data. A key aspect in these new approaches (which are sometimes indicated as "software as services") is of course the diffusion of standards, which makes it possible to integrate different components regardless of the specic architecture and technology each one is developed with. The most evident and practical embodiments of this trend are "mashups" (see e.g Merrill, 2006; Yu et al., 2008), which are Web applications mixing existent services and/or contents. Publicly available Application Programming Interfaces (APIs), as well as the development of Web services, play a key role in this scenario: the Web is becoming a platform (see O'Reilly, 2005) where different and distributed blocks inter-

act dynamically each other in order to accomplish simple to complex tasks.

In parallel to these trends, which mostly characterize the trends around the "Web 2.0" vision, new proposals and practical experimentations in the direction of a supposed "**Web 3.0**" phase exist. In reality, this new phase seems to represent today more a marketing buzzword than a definable concept, and a number of different and alternative definitions exist for characterizing the Web of the near future: "The Web of Data", "The World Wide Database", "The Executable Web", "The Internet of Services", "The Giant Global Graph" and"The Intelligent Web", to name a few. Moreover, 2.0 and 3.0 elements are so strictly inter-connected that they will surely contribute to the future Web in complementary ways (see e.g. Ankolekar et al., 2007; Gruber, 2008). Therefore, at least to a certain extent, these distinctions are meaningless.

The basic idea underlying the 3.0 vision is that the Web is going to increasingly support the execution of complex tasks that are currently considered a human prerogative: a typical example is the real-time planning of a vacation by networked machines, services and data providers in a way equivalent to the expert work of a human travel agent. Interoperability, semantic integration, personalization, ubiquitous applications and natural language processing are only a few of the related keywords in this scenario, which is often considered speculative and not technically feasible at the moment.

However, one of the clearest and central visions in the "Web 3.0" is that of the **Semantic Web**, which in reality relies on a proposal introduced a few years ago in order to indicate an extension of the current Web, in which information is given a well-defined meaning, better enabling computers and people to work in cooperation (Berners-Lee et al., 2001). This apparently simple statement is in reality disruptive, and it raises several issues that are connected for example to the modalities through which information is semantically enriched and represented, as well as how it is processed and exploited. Today's Web contents are in fact mostly produced for human consumption: most of today's Web is a "Web of documents", whose contents often can be hardly processed by machines. Instead, the Semantic Web is a "Web of Data", i.e. it has its primary units in data interconnected by semantic relationships, rather than in documents interconnected by hyperlinks. As such, it is based on a general architecture (known as the "Semantic Web pile" or the "Semantic Web cake") which has been built, and is constantly refined, in order to support the shift from a Web of documents to a Web of data.

Extensive research efforts in the Semantic Web area have produced the basic specifications and technologies for the building blocks of this architecture, for example with the standardization of XML-based languages for metadata and ontologies (such as RDF, RDFS and OWL). On the other hand, a number of application patterns referring to the Semantic Web have been investigated, such as data integration; knowledge engineering with complex ontologies; improved data management; Web service management, coordination and combination; intelligent software agent deployment;

improved search methods; as well as mixtures of these.

Therefore, the Semantic Web, today represents a challenging area of research in many fields. Many contributions highlight the potentialities of **Semantic Web applications in cultural heritage**, thanks to some characteristics such as its high heterogeneity, which makes it difficult to convey cultural heritage information and knowledge to the general public or even to professionals in all its richness. This position, for example, is claimed by Aroyo et al. (2007b, p. 7), who also stress that cultural heritage is "*. . . something that is hard to mOdel. Something open, without clear boundaries. Something where one resource is always related to many others. Something that is thus also an ideal technical challenge for Semantic Web researchers.*".

Since the beginning of the 2000s, Semantic Web technologies and their potentials for the integration and exploitation of digital cultural heritage information have received increasing attention. However, several factors greatly slowed down this research, such as the presence of a gap in the cultural heritage communities between the vision and the practical approaches to "do the Semantic Web", which in turn determines misunderstandings and skepticism. This attitude is also present in specific areas, such as archaeology, which traditionally made early experimentations and usage of computer science tools for research.

On the other hand, the results of notable and recent experiences show that the new vision is increasingly permeating the sector, mostly thanks to the activity of computer science research groups undertaking interdisciplinary research together with cultural heritage professionals. Nevertheless, a large number of research issues still remain open, and there is a general lack of consolidated approaches to "do the Semantic Web" in cultural heritage. Notable exceptions are represented, for example, by the standardization of core domain ontologies (such as the CIDOC CRM) that can be used for the semantic integration of heterogeneous metadata schemata, or the partial results of extensive projects concerning the modalities through which semantic content is created, managed and offered to machines and human users.

1.1 Research Perspective and Objectives

These elements show that research on the Semantic Web in cultural heritage is of primary relevance today, especially if it is carried out with a transversal and interdisciplinary perspective that combines the more technological elements with domain issues and the evaluation of the impact that these solutions should have on the activities and practices of different communities, from professionals to the general public. A path going from the representation of data, information and knowledge in semantic systems, to the fruition of these resources represents an ideal *fil rouge* that not only links most of the projects carried out to date, but could also serve as a guide for the design of a research program.

This PhD work aims to contribute to the scenario of cultural heritage on the Se-

mantic Web in this perspective, and in particular with the investigation of a set of related key elements that take part in the deployment of semantic systems (such as semantic portals; see e.g. Hyvönen, 2009) along a representation → fruition path.

A central aspect is represented by the **in-depth evaluation of relevant and general results obtained to date**, either from a more theoretical point of view and by experimenting them in new contexts. The analysis of the costs and benefits that are connected to the standard solutions represents a necessary condition in order to obtain a clear view of the sustainability of Semantic Web approaches for cultural heritage. In this regard, the problem of representation of cultural heritage data, information and knowledge in contexts of semantic integration of heterogeneous metadata schemata is a fundamental concern, and it is also the area in which notable results of general value have been obtained. Among these the standardization of the CIDOC CRM ontology stands out, and it is the object of vivid debate concerning the model's characteristics, as well as the issues emerging from its actual usage in real application scenarios. Therefore, the analysis and experimentation of the CIDOC CRM is particularly interesting, and new research contributions to the debate are of primary relevance. This investigation should also comprise the definition of a suitable workflow for mapping legacy metadata schemata to the CRM, and the evaluation of the model's available serializations in Semantic Web languages.

Beyond this, there is also a need for more "vertical" research on **domain specific aspects**, in order to provide innovative perspectives for the fruition of cultural heritage. Given the richness and variety of the domain, it is difficult to select a single aspect on which to concentrate. However, since time is a fundamental dimension for the description, study and understanding of cultural heritage it should also constitute a central aspect in Semantic Web applications for the domain. Therefore, the research investigates the temporal dimension with respect to fuzzy temporal intervals and chronologies in archaeology, which represents a complex domain, where a limited amount of related work is available today.

The analysis of the specific characteristics of chronologies in archaeology, and the evaluation of the existing models for representing fuzzy temporal information, necessarily represent the basis for the design of original approaches supporting e.g. innovative and effective methods of information retrieval. In particular, the fuzzy set theory seems to be a suitable framework for this research, since it fits well with the general characteristics of cultural heritage documentation.

On the other hand, the definition of a **case study**, where both the semantic integration of heterogeneous metadata schemata through standard ontologies and the retrieval of fuzzy chronological information are relevant, would enormously contribute to the research with the actual experimentation of the analyzed theoretical aspects. In fact, when moving from the theoretical to the application level, several issues may emerge, as the existing experiences demonstrate with reference to the mapping of legacy metadata schemata to the CIDOC ontology.

This implies either that technical constraints impede the deployment of sound theoretical approaches, or on the contrary, that inconsistencies on the theoretical side emerge as real world scenarios and issues are faced.

Finally, even if this research is only marginally concerned with technical aspects, the actual **development of a prototype system** ideally completes the proposed interdisciplinary research program. In fact, the more practical contribution to the aforementioned problem of "doing the Semantic Web" in cultural heritage necessarily comprises the evaluation of the cost and benefits of deploying a real system. In this regard, and with reference to the scenario of a semantic portal for cultural heritage, two main aspects seem crucial: the creation of a semantic backend based on semantic repositories (such as triplestores); the design and development of highly interactive interfaces based on recent technologies and approaches (such as AJAX and mashups), easing browsing through the repositories. Since several of the involved technologies are not as consolidated as more traditional ones (e.g. relational databases), the evaluation of the performances of the prototype system should also constitute an important aspect.

1.2 Outline of the Dissertation

Taking into consideration the elements introduced so far, this PhD research work is organized as follows.

Chapter 2 provides an overview of the new directions in the development of the Web according to the perspectives of the "Web 2.0" and the "Web 3.0". The analysis highlights the complexity of the current scenario, where complementary visions are co-operating in shaping the Web of the near future. Section 2.3 tries to introduce a set of transversal and emerging key elements, and in particular: User Generated Content, the development of new applications exploiting publicly available APIs, Web services and mashups approaches, and the concepts and constituents behind the visions of the Web of Data and the Semantic Web. Section 2.4 introduces the basic characteristics of the technologies underlying these key elements. Specific attention is paid to the technologies that are involved in the high level Semantic Web architecture (the "Semantic Web pile"), with particular respect to the languages enabling semantic interoperability (RDF, RDFS and OWL) and query on RDF triples (SPARQL).

Chapter 3 analyzes the current scenario of Semantic Web applications in cultural heritage (section 3.1) and in archaeology (section 3.2). This analysis highlights the general elements characterizing the scenario with respect to the problems slowing down or even hindering the adoption of the new approaches, and it offers the discussion of the most relevant projects carried out to date. Section 3.3 extends the analysis scope beyond the strict Semantic Web perspective, and towards the related frameworks of digital libraries, e-Science and cyberinfrastructures in order to complete the state of the art review in a broader perspective.

Chapter 4 introduces the research perspective this work conforms to, taking into

consideration the elements highlighted in the reviews of chapters 2 and 3; an interdisciplinary approach is proposed, along a path linking the representation of cultural heritage data, information and knowledge with their fruition in a semantic portal.

More specifically, a twofold perspective is introduced. On the one hand, the fundamental aspect of representation is analyzed with reference to the ISO core domain ontology CIDOC (section 4.2), highlighting its characteristics, as well as defining the most relevant issues that are connected to its usage and evaluation in real application scenarios.

On the other hand, a specific domain aspect, i.e. the modeling and retrieval of fuzzy temporal intervals and chronologies in archeology is discussed (section 4.3), with reference to existing work on temporal modeling and reasoning, and in particular the CIDOC CRM approach. Thereafter, a new model and method based on the fuzzy set theory are introduced, with the aim of improving the retrieval of relevant information in Semantic Web systems according to fuzzy chronological elements.

Chapter 5 introduces a case study (which is related to the archaeology of the city of Milan in Italy) for the experimentation and evaluation of the research proposals and methods introduced in chapter 4. In particular, section 5.2 sketches the requirements for the creation of a semantic portal in the identified context, while section 5.3 discusses the principal characteristic of the selected data sources. The detailed description of the mapping activity of the data source metadata schemata to the CIDOC CRM according to the proposals introduced in section 4.2 is provided in section 5.4. An in-depth discussion of the elements which emerged from this activity with respect to the structure of the CRM is provided in section 5.5.

Chapter 6 discusses the application of the methods for modeling and retrieving fuzzy chronological information that has been introduced in section 4.3. In particular, the different kinds of chronologies present in the case study are introduced in section 6.1, and their representation with the proposed fuzzy set-based model is described in section 6.2. Thereafter, an evaluation setting, the related criteria, and an interface for evaluation based on an interactive timeline are introduced (section 6.3), and the results of the evaluation of the retrieval method, which involved 12 domain experts, are synthesized in section 6.4. Finally, a discussion on the model and the retrieval method, highlighting its originality with respect to other existing approaches, is provided in section 6.4.

Chapter 7 describes the design and development of a demo semantic portal for the archaeology of Milan, which integrates the elements analyzed in the previous chapters. This portal is based on previous experience (the Mantic 1.0 portal, section 7.1) which was developed using a Web mashup approach, that gave the opportunity to define use cases and provide simple navigation capabilities on a repository based on a relational database. The general characteristics of semantic repositories, together with a technological survey and evaluation of prominent existing solutions, are introduced in section 7.2. The new architecture of the Mantic 2.0 system, which intro-

duces a semantic backend that substitutes the previously present relational database, is described in section 7.3; the discussion and evaluation of this improvement, with particular respect to the system's performances, is provided in section 7.4.

Chapter 8 draws the concluding remarks on the research with respect to both the more general and theoretical elements, and the technological and technical ones; and it suggests some possible directions for future work.

2

1.0 to *n*.0: New Directions in the Development of the Web

In this chapter we provide an overview of the most relevant and complementary visions and approaches that are co-operating in shaping the Web of the future, in order to define the general background of the research. In particular, the elements characterizing the "Web 2.0" (section 2.1) and the "Web 3.0" (section 2.2) are introduced, and key aspects are pointed out (section 2.3), such as User Generated Content (UGC), the development of new applications exploiting publicly available APIs, Web services and mashups approaches, and the visions of the Web of Data and the Semantic Web. Moreover, the basic characteristics of crucial technologies underlying these scenarios are introduced (section 2.4), with specific reference to the high-level architecture of the Semantic Web.

2.1 The "Web 2.0"

The term "Web 2.0" (O'Reilly, 2005) finds its origin in the expression "The Web as a platform", which was coined in 2004 in a conference brainstorming session between representatives of the O'Reilly Publishing Group and MediaLive International. Successively, "Web 2.0" was adopted as the name of a series of Conferences organized by O'Reilly and soon became a general label to describe a new generation of Web applications enabling users to actively participate in the creation, sharing and aggregation of Web contents rather than in its mere passive visualization. Since the first examples of these applications date back approximately to the year 2000 (see e.g. DoubleClick[1] and Akamai[2]), "Web 2.0" is an *a posteriori* definition.

O'Reilly noticed that the bursting of the dot-com bubble at the end of 2001, instead of decreasing the economic and social value of the Web, made it increase enormously, and that the companies that survived the crisis seemed to have some elements in common.

They offer **services** rather than products. The most relevant example of this is Google, whose services can be used with direct or indirect forms of payment, without the traditional forms of licensing and packaging of the software and without the

[1] `http://www.doubleclick.com/`
[2] `http://www.akamai.com/`

necessity to port applications to different platforms. New models of software development and release, which differ substantially from the traditional ones, are related to this approach. In fact, Web services offer functionalities that aim to be useful for particular tasks, rather than complex and multi-purpose solutions. Cconsequently, the software release cycle changes from a complex sequence of scheduled steps and releases to continuous and flexible updates and improvements. For this reason, Web 2.0 applications are often said to be in a stage of "perpetual beta", where new functionalities are quickly added, removed or modified without the deployment of major releases.

In this scenario, **users are co-developers**, because they provide direct or indirect feedback to developers through the usage of the applications themselves. The addition, removal or modification of features can be done on a regular basis as a part of the normal user experience.

The role of users is not only limited to testing. In fact, Web 2.0 applications try to **harness collective intelligence**, i.e. they try to exploit users' behavior in order to make the application more and more effective and useful. Methods such as PageRank[1] by Google, which uses the link structure of the Web rather than the characteristics of documents to provide better search results, is an example of this. Other outstanding examples include Amazon[2], which provides high levels of user engagement whose results are in turn used to suggest relevant contents and to refine search results; collaborative spam filtering, such as that provided by Cloudmark[3]; easy and dynamic content creation such as Wikipedia [4]; collaborative tagging and "folksonomies", which were first introduced by Flickr[5]; and many others. The idea here is that through the "wisdom of the crowds", much value can be added to the applications (in terms of effectiveness) and to user experience: users decide what is useful and relevant and what is not by means of an engaging and natural interaction with the application.

As a natural consequence of these points, **data become the central component of Web 2.0 applications**, and data management represents a crucial competence for 2.0 companies. Successful Internet applications in the 2.0 era are all fed by specialized databases, and in most cases they simply offer personalized views of the company's database. It has been said that "SQL is the new HTML", i.e. the methods for effective and personalized data retrieval are crucial today. The availability of classes of data thus represents a competitive advantage; for this reason it becomes crucial to involve users in populating the databases behind a Web application.

This involvement is possible thanks to significant advancements in the modalities of interaction with Web pages. The availability of **rich and light interfaces** that allow quick and dynamic interactions with multimedia content is a central characteristic in

[1] See: `http://infolab.stanford.edu/~backrub/google.html`
[2] `http://www.amazon.com/`
[3] See: `http://www.cloudmark.com/en/serviceproviders/research.html`
[4] `http://www.wikipedia.org/`
[5] `http://www.flickr.com/`

the Web 2.0 scenario and is enabled by technologies such as AJAX.

Finally, in the Web 2.0 scenario, attention is paid not only to the biggest content providers, but there is the tendency to **embrace the full spectrum of the Web**. The idea here is that the collective power of small Web sites make up the bulk of Internet's contents and as such, it offers a great business potential. This is what C. Anderson refers to as "the long tail" (Anderson, 2006).

To sum up, a Web 2.0 application can be considered as an application in a Web environment which:

- takes the form of a service, rather than a product
- is not limited to a single software product or to a particular machine
- is open and shared
- has in the groups of users and in social interactions fundamental components of its organization
- is open to user contributions in terms of content production; user generated content in turn, increments the value of the application

Given these characteristics, Web 2.0 has been discussed in different ways and subjected to different opinions, from the enthusiast to the skeptic. In fact, several authors question its "disruptive" nature (if compared to the 1.0 era); among these, Tim Berners-Lee when asked in a interview [1] about Web 2.0 and its differences with the 1.0 era, said: " *Web 1.0 was all about connecting people. It was an interactive space, and I think Web 2.0 is of course a piece of jargon, nobody even knows what it means. If Web 2.0 for you is blogs and wikis, then that is people to people. But that was what the Web was supposed to be all along. And in fact, you know, this Web 2.0, quote, it means using the standards which have been produced by all these people working on Web 1.0.*".

In fact, it is clear that the Web is constantly evolving (and is still far from reaching the vision originally defined by T. Berners Lee): from this perspective Web 2.0 can be considered as a natural outcome of the evolutionary process maybe without the need of introducing a clear-cut distinction with a Web 1.0 era. However, it has to be acknowledged that the technological changes that are generally associated with the 2.0 phase marked a big shift in the social impact of the Web because they increased users' participation dramatically. The shift consisted in the passage from a prevalent "read-only" to a "read-write" paradigm. Users became active protagonists in terms of content production, sharing and aggregation mostly because the technical skills that were previously mandatory in order to publish even a single page of text on the Web became optional. The "webmaster" who was a central figure in the 1.0 era, disappeared in favor of a plethora of professional and non-professional profiles, from the

[1] The transcript is available at: `http://www-128.ibm.com/developerworks/podcast/dwi/cm-int082206.txt`

expert in Web programming to the blogger, each one contributing directly at different stages to the development of the Web

The diffusion of Web 2.0 technologies happened through applications, which gave birth to new forms of social interaction mediated by the computer: social networks (such as Facebook[1], MySpace[2], LinkedIn[3] and many others) are probably the best-known examples of this. The dimension of the phenomenon is impressive: statistics provided by Della Valle et al. (2008, p. 22) affirm that MySpace had 185 millions users as of April 2007, making MySpace virtually the 6th largest country in the World.

2.2 Towards the "Web 3.0"

It seems that each new "phase" in the development of the Web requires a software-versioning-like label: in the last few years the idea of a "Web 3.0" has been proposed, first by Markoff (2006). The definition of the Web 3.0 is at least as controversial as that of the Web 2.0, mostly because some of the elements that may contribute to defining it are speculative and not technically feasible at the moment. However, "Web 3.0" represents a good parallel to "Web 2.0", even if it is often considered only as the newest marketing buzzword.

In fact, opinions and predictions about Web 3.0 vary greatly, from those focusing mainly on the technical aspects to those that pay more attention to the increasing pervasiveness of the Web and to its influence on everyday activities. Different perspectives correspond to a number of different labels: "The Web of Data", "The World Wide Database", "The Executable Web", "The Internet of Services", "The Giant Global Graph", and "The Intelligent Web" are only a few examples.

But what are the main characteristics that distinguish this new promised era from the previous one?

A provisional analysis may start from the consideration that the huge diffusion of the Web and the possibility for users to generate content has brought to information overload, making several issues connected to automatic data management and retrieval crucial. Not only different data schemas and formats hinder the possibility to easily integrate applications, but syntactic and semantic differences in datasets make the automatic correlation and retrieval of relevant information a difficult task. A solution that has been proposed to these issues is the idea of a "**Semantic Web**" (see section 2.3.3).

In fact, several authors and technology analysts identify Web 3.0 with the Semantic Web; however, the two do not overlap exactly, and the Semantic Web has to be considered as one of the components of Web 3.0, probably the most relevant one. The development of data formats and query languages that aim to take into consideration

[1] `http://www.facebook.com/`
[2] `http://www.myspace.com/`
[3] `http://www.linkedin.com/`

the semantics of information (such as RDF, OWL, and SPARQL; see section 2.4.4) open new possibilities for the publication, open access and retrieval of structured datasets on the Web. In this vision, data will be automatically inter-linked when formulating a request to Web servers, and links will be conceptually similar to those existing today between Web pages. Mostly referred to as the "Web of Data" (or the "Data Web"), this vision aims to substitute the current Web, which is strongly based on documents and links between them, with a network of resources (i.e. data) that will automatically relate to each other on the basis of semantic relationships (fig. 2.1); the focus here is on the content of documents, rather than on the documents themselves. Similar to this vision are the concepts of "World Wide Database", by N. Spivack[1], which emphasizes the role of the Web as a huge data provider; and "The Giant Global Graph", by Berners-Lee[2], which, for the moment, can be considered a synonym of the Semantic Web. The Semantic Web promises to widen the perspective of the "Web of Data", by enabling the representation and retrieval of the semantics not only for structured documents, but also for semi-structured or unstructured ones.

As a consequence of these evolutions towards data and semantics, a more "**intelligent Web**" would be possible, with applications performing complex tasks that are currently considered a human prerogative. Autonomous agents will crawl the Web and use different kinds of reasoning in order to satisfy complex requests, be they made by a human (even using natural language) or by a machine. A classical scenario which provides an example of this is the planning of a vacation; users will provide high level or general requirements (such as "a warm place" or "a budget of around 2.000 euros"), and computers will be able to process the request by querying different datasets and services, to compose the results, and to return the best detailed alternatives in a reasonable amount of time, as a human travel agent would do. The debate in the Intelligent Web is whether "intelligence" will emerge from systems autonomously learning and reasoning (such as "Cyc"[3]), or from technologies that systematically extract meaning from the existing Web (such as engines exploiting user generated tags).

In order to make this possible, a shift is necessary from the currently available network of separate or loosely integrated applications and content repositories to **full interoperability**. Applications need to be pieced together in order to effectively and efficiently provide functionalities coming from multiple services in a cooperative way. The vision of an "Internet of Services" addresses these requirements and being based on a semantically-enabled infrastructure, it is expected to go much more beyond the effectiveness of the currently available solutions.

Moreover, the effectiveness of applications will be further increased thanks to their customization, in order to satisfy each user preferences and needs. **Personalization**

[1] http://novaspivack.typepad.com/nova_spivacks_weblog/2005/10/towards_a_world.html

[2] http://dig.csail.mit.edu/breadcrumbs/node/215/

[3] http://www.cyc.com/

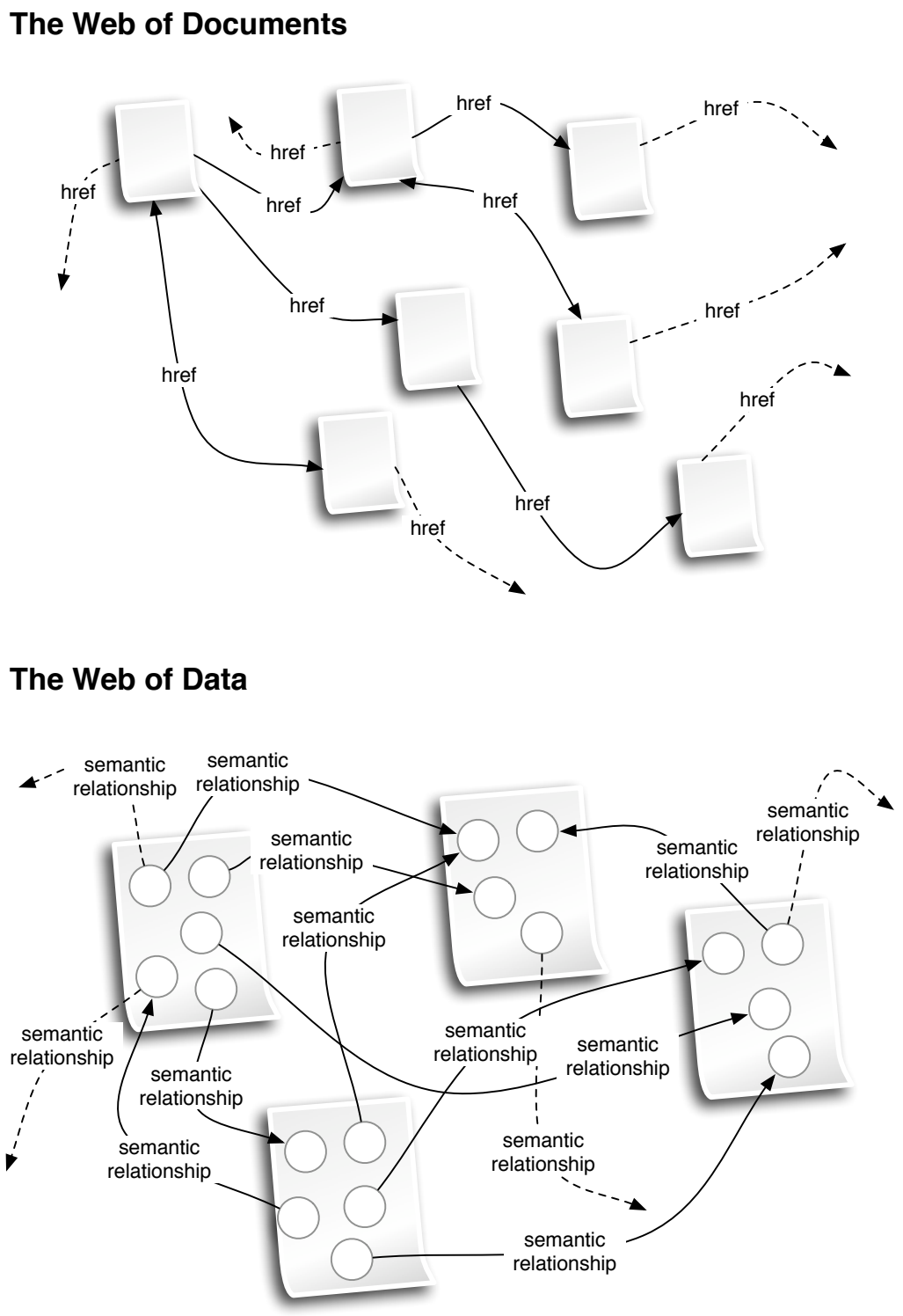


Figure 2.1: *A schema representing the difference between the Web of Documents and the Web of Data.*

is a keyword in the 3.0 scenario, but it involves a number of problems mainly concerning the legal and social issues connected to the ownership and use of personal profiles. In fact, even if languages and tools that address the issues of personal identity management on the Web are today available (such as FOAF[1] and OpenID[2]) and personal information is constantly collected by Web applications such as GMail, the crucial point is how to exploit user profiles, while at the same time respecting his/her privacy. Another related problem concerns the modalities through which personal information will be available, i.e. by direct and consensual input by users, or by indirect and "invisible" profile creation through the analysis of navigation paths, e-mail, etc.

Finally, the transition to a Web that will be more focused on data, on the integration of services and on more server-side rather than client-side computing potentially represents a great opportunity for the development of applications running on a wide range of different devices. Once fast and reliable ubiquitous connectivity will be available, the Web itself will be fully accessible on mobile devices, thus making it possible to design specific services supporting users everytime and everywhere. The "**Ubiquitous Web**" is surely a part of the envisaged new phase of the Web, and the infrastructural elements that are needed to make it possible are rapidly becoming a reality.

Giustini (2007, p. 95) provides a quick comparison between Web 2.0 and Web 3.0 (table 2.1) which may be useful to summarize the main elements that characterize the current scenario.

Table 2.1: *A comparison between Web 2.0 and Web 3.0 (after Giustini, 2007).*

Web 2.0	**Web 3.0**
The document Web	The Web of Data
Abundance of information	Control of information
Controversial	No less controversial
The social Web	The Intelligent Web
The second decade (2000–2009)	The third decade (2010–2020)
Google as a catalyst	Semantic Web companies as catalyst
Wisdom of the crowds	Wisdom of the expert
Mashups, fragmentation	Integration, new tools
Search, search, search	Why search, when you can find?
Google's Pagerank algorithm	Ontologies, semantic systems
Lawless, anarchic	Standards, protocols, rules
Print and digital	Digital above all else

[1] `http://www.foaf-project.org/`
[2] `http://openid.net/`

It has been said that "*in its current state, the Web is often described as being in the Lego phase, with all of its different parts capable of connecting to one another*"(Markoff, 2006); the challenge of the 3.0 era is to automate the creation and management of these connections and to make them exploitable by machines in more "intelligent" ways.

2.3 Beyond 2.0 and 3.0: The Emerging Key Elements

Beyond the debate of the definitions, it is important here to sum up the concrete outcomes of 2.0 and 3.0 phases in order to outline the directions this PhD research takes into consideration. In particular, the following sections discuss three main groups of them: user generated content (section 2.3.1), the integration and combination of applications (section 2.3.2), and the Web of Data and the Semantic Web (section 2.3.3).

2.3.1 User Generated Content

As has been said above, the diffusion of Web 2.0 platforms and applications made users more content providers than simple consumers: User Generated Content (UGC) is becoming a central component of a large number of today's Web applications. The Organisation for Economic Co-operation and Development (OECD) proposes a definition of "User Created Content" (UCC), a synonym of UGC, and "Consumer-generated media", this latter being less diffused than the other two in the literature. In OECD (2007, p. 4) UCC is defined along three main characteristics:

- **Publication requirement**: while theoretically UCC could be made by a user and never actually be published online or elsewhere, we focus here on the work that is published in some context, be it on a publicly accessible Web site or on a page on a social networking site only accessible to a selected group of people (i.e. fellow university students). This is a useful way to exclude email, bilateral instant messages and the like.
- **Creative effort**: This implies that a certain amount of creative effort was put into creating the work or adapting existing works to construct a new one; i.e. users must add their own value to the work. The creative effort behind UCC also often has a collaborative element to it, as is the case with Web sites which users can edit collaboratively. For example, merely copying a portion of a television show and posting it to an online video Web site (an activity frequently seen on the UCC sites) would not be considered UCC. If a user uploads his/her photographs, however, expresses his/her thoughts in a blog, or creates a new music video this could be considered UCC. Yet the minimum amount of creative effort is hard to define and depends on the context.
- **Creation outside of professional routines and practises**: User-created content is generally created outside of professional routines and practices. It often does not

have an institutional or a commercial market context. In the extreme, UCC may be produced by non-professionals without the expectation of profit or remuneration. Motivating factors include: connecting with peers, achieving a certain level of fame, notoriety, or prestige and the desire to express oneself.

The second characteristic mentioned in the list is not always present when evaluating content as user generated, and the characteristic is getting harder to maintain, mostly because there is a trend towards the monetization of UGC (OECD, 2007, p. 9), but also because contents may assume a variety of forms, thus making it difficult to distinguish between professional and non-professional contributions.

In fact, the currently available applications make the creation of even sophisticated contents much easier, and the learning curves are lowering. Several examples of this exist: take, for example the case of georeferenced, three-dimensional textured models of architectural elements. A few years ago, the creation of these models would have required specific, high level competencies in fields such as Computer Assisted Design (CAD), three-dimensional modeling, photogrammetry, Geographical Information Systems (GIS) and Web technologies. On the contrary, excellent results can be obtained today by using free, integrated and easy-to-use applications, such as Google Earth[1] and Google Sketchup[2], and the models can be easily shared on the Web *via* the Google 3d Warehouse[3] repository (fig. 2.2).

If we try to define the types of User Generated Contents, it is becoming extremely difficult to provide an extensive list (fig. 2.2). Not only the scenario is constantly evolving, but there are a large number of platforms, which include and relate different kinds of UGC. A simple example is Flickr, where photographs provided by the users can be annotated by marking sensible areas, and can be rated and commented, not to mention the possibility of georeferencing and displaying them on a map.

Of course, UGC has been the subject of criticism, from those complaining about the low quality of contents if compared to that of professional providers to those complaining about the violation of the privacy when publicly diffusing personal information. However, it is important to notice here that UGC and related applications have evolved so much in the last few years that they are colonizing the territory of professionals and promoting new ways of working. In fact, an increasing number of applications today are offering commercial "pro versions" that in some cases are integrated with traditional professional products or even substitute them: the previously mentioned Google Earth and Google Sketchup, for example, are increasingly adopted by architects when presenting urban scale projects to customers and to citizens who can evaluate them by a direct interaction with the models. In this vein, the term "user" is becoming too generic because it implies a distinction between small-scale publishers who operate for free and the rest of the World, which is constituted by professionals.

[1] `http://earth.google.com/`
[2] `http://sketchup.google.com/`
[3] `http://www.google.com/sketchup/3dwh/`

Table 2.2: *Types of User Generated Content (after OECD, 2007, tab. 3, p. 15)*

Type of Content	Description	Examples
Text, novel and poetry	Original writings or expanding on other texts, novels, poems	Fanfiction.net, Quizilla.com, Writely
Photo/Images	Digital photographs taken by users and posted online; Photos or images created or modified by users	Photos posted on sites such as Ofoto and Flickr; Photo blogging; Remixed images
Music and Audio	Recording and/or editing one's own audio content and publishing, syndicating, and/or distributing it in digital format	Audio mashups, remixes, home-recorded music on bands Web sites or MySpace pages, Podcasting
Video and Film	Recording and/or editing video content and posting it. Includes remixes of existing content, homemade content, and a combination of the two.	Movie trailer remixes; Lip synching videos; Video blogs and videocasting; Posting home videos; Hosting sites include YouTube and Google Video; Current TV
Citizen journalism	Journalistic reporting on current events done by ordinary citizens. Such citizens write news stories, blog posts, and take photos or videos of current events and post them online.	Sites such as OhmyNews, GlobalVoices and NowPublic; Photos and videos of newsworthy events; Blog posts reporting from the site of an event; Cooperative efforts such as CNN Exchange
Educational content	Content created in schools, universities, or with the purpose of educational use Syllabus-sharing sites such as H20; Wikibooks, MIT's OpenCourseWare	
Mobile content	Content that is created on mobile phones or other wireless devices such as text messaging, photos and videos. Generally sent to other users via MMS (Media Messaging Service), emailed, or uploaded to the Internet.	Videos and photos of public events, environments such as natural catastrophes that the traditional media may not be able to access; Text messages used for political organization
Virtual content	Content created within the context of an online virtual environment or integrated into it. Some virtual worlds allow content to be sold. User-created games are also on the rise.	Variety of virtual goods that can be developed and sold on Second Life including clothes, houses, artwork

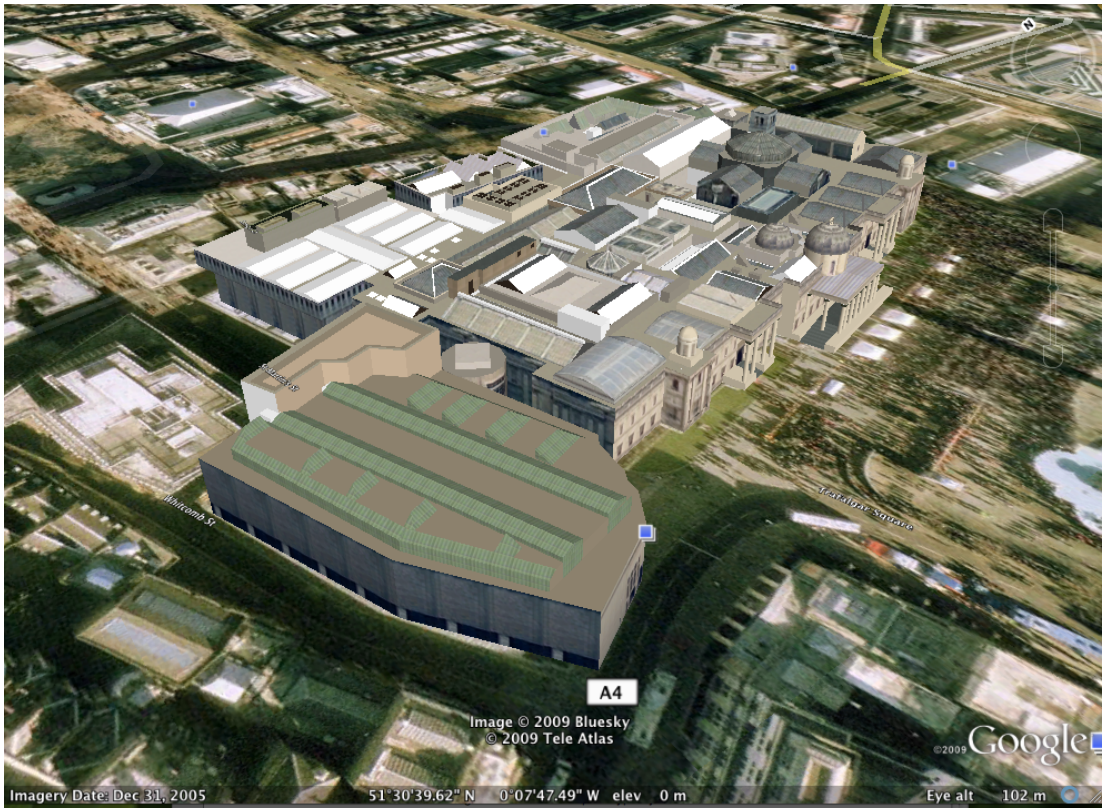


Figure 2.2: *An example of User Generated Content: a georeferenced, three-dimensional textured model in Google Earth (*`http://sketchup.google.com/3dwarehouse/details?mid=340e46987ff58127a1136d3b8e3e0cf`*).*

Consequently alternative expressions have been proposed, such as "Entrepeneurial Generated Content" (EGC).

Beyond these considerations, statistics demonstrate that the proportions of the UGC phenomenon (with the term "user" denoting an articulated typology of contributors, both non-professional and professional) are huge. Those provided by Della Valle et al. (2008, p. 22), for example, affirm that Youtube hosts more than 6 million videos, and Flickr hosts more than 3.5 million photographs, 70% of which are marked with at least one tag. From a more qualitative point of view, data currently present on the Web are 25% original and 75% replicated; 75% personal (i.e. not related to professional activities), among which 95% is unstructured and increasing, and 15% is structured and decreasing. Multimedia digital devices (such as palmtops, smartphones, digital cameras, etc.) are contributing a lot to this scenario and are estimated to double in number by 2010.

2.3.2 Combining Applications: From APIs to Mashups

In contrast to traditional approaches, today specific attention is being paid to the design and development of modular Web applications that combine publicly accessible applications and data. This software development strategy, which is a characteristic of Web 2.0 and is sometimes termed "Software as a Service" (SaaS), aims to provide new ways of creating light and useful applications by remixing available sources instead of releasing software packages that need to be installed locally in the browser. The key concept in this approach is the **loose coupling** between applications, i.e. services can be combined regardless of the specific architecture and technology each one is developed with, provided standard methods allow the services to communicate. This ensures a high level of modularity because each module that contributes to the Web application can be changed and improved without the need to subsequently change the coupled modules. Moreover, the problem of maintenance is dramatically lowered because each service is maintained and improved by its provider, and the development efforts are reduced: Web applications do not need to be fully created from scratch, but the developer can take advantage of services that address specific needs and add customized programming code only when it cannot be sourced from internal or external suppliers, or when better integration between the parts is needed.

These possibilities clearly rely on the technological advances that have invested the Web in the last few years. Key elements have been constituted by:

- The development of **standard formats** for data exchange, thus allowing better interoperability between applications. The definition and diffusion of the eXtensible Markup Language (XML), and the development of languages based on its syntax represent the most important result of this process.

- The release of publicly available **Web Application Programming Interfaces (APIs)**, i.e. methods for requesting specific functionalities and/or data from a Web server in order to use them in external applications. Amazon[1] was one of the notable pioneers of this approach; in 2003, the company released an API which allowed programmers to search into its product database, extract all the desired information, and present it on external Web sites in any desired format, provided any resulting purchases were directed back to Amazon public database.

- The development of standard protocols and architectures for **Web services**. Web services group a number of technologies, and are often considered as an equivalent of APIs. Even if the very basic idea is common to the two, Web services differ from APIs in that they are built according to well defined standard protocols specified by W3C[2] that manage the communication between applications,

[1] `http://www.amazon.com/`

[2] See the W3C Web services glossary at `http://www.w3.org/TR/ws-gloss/`

taking into consideration aspects such as service discovery, service quality, access control, etc.

The overall architecture of the Web is changing according to these trends and towards the vision of "The Web as a platform" suggested by O'Reilly (2005, p. 2), i.e. the idea of a Web where different and distributed blocks interact dynamically with each other in order to accomplish simple to complex tasks.

The assemblage of Web applications by remixing available data and services is referred to as "**mashup**", a term from the musical field that indicates the combination of tracks from two songs (usually belonging to two different genres) into a new one. From the user point of view, a mashup is perceived as an application based on a rich and dynamic graphical interface that does not require any installation and makes the sources of data and services it aggregates completely transparent. Mashups offer some specific characteristics that make them different from traditional integration approaches, and component-based application development. According to Yu et al. (2008), mashups show three main peculiar characteristics:

- They typically serve a specific situational (short-lived) need. They focus mainly on opportunistic integration, which is suitable for non-buisness-critical applications.
- They are composed of the latest, easy-to-use Web technologies: the Web is their natural environment, and mashups are about simplicity, usability, and ease of access.
- They generally have small audiences, so that scalability does not become a crucial issue.

Different kinds of mashups exist; according to the methods by which the aggregation takes place, they can be categorized into three main types (fig. 2.3):

- **Content mashups** are based on the aggregation of contents by a subscription to different and heterogeneous data sources. Contents are processed and integrated in real time in order to be shown through a unified graphical interface or to be further processed by the machine. Examples in this category include mashups ranging from news aggregation, such as AggreGet[1] that integrates popular news sources (e.g. Digg, de.licio.us, stumblebuzz, and others) to FlashEarth[2] that allows the visualization of satellite and aerial images of the Earth coming from different providers (e.g. NASA, OpenLayers, Microsoft Virtual Earth, and Yahoo Maps).

[1] `http://aggreget.com/`
[2] `http://www.flashearth.com/`

- **Service mashups** are based on the integration of available Web services in order to fulfill the functional requirements of a new Web application. This approach makes it possible to develop effective and stable applications easily and quickly, if compared to the traditional Web programming techniques. An example in this category is Hominin[1], which relies on the APIs of an existing Web mapping service (GoogleMaps) and a Simile Timeline in order to show the contents of a local database concerning hominin fossils.
- **Hybrid mashups** combine the two approaches mentioned above. Today this is the most interesting mashup technique, mostly because it shows relevant business potential, as is emphasized by the expression "business mashups" which is sometimes used to identify them. An early example of hybrid mashup was HousingMaps[2] which was released in 2005 in order to provide an effective method for consulting Craiglist's[3] real estate advertisements, by means of a highly interactive interface combining map, search filters on price, and tabular lists.

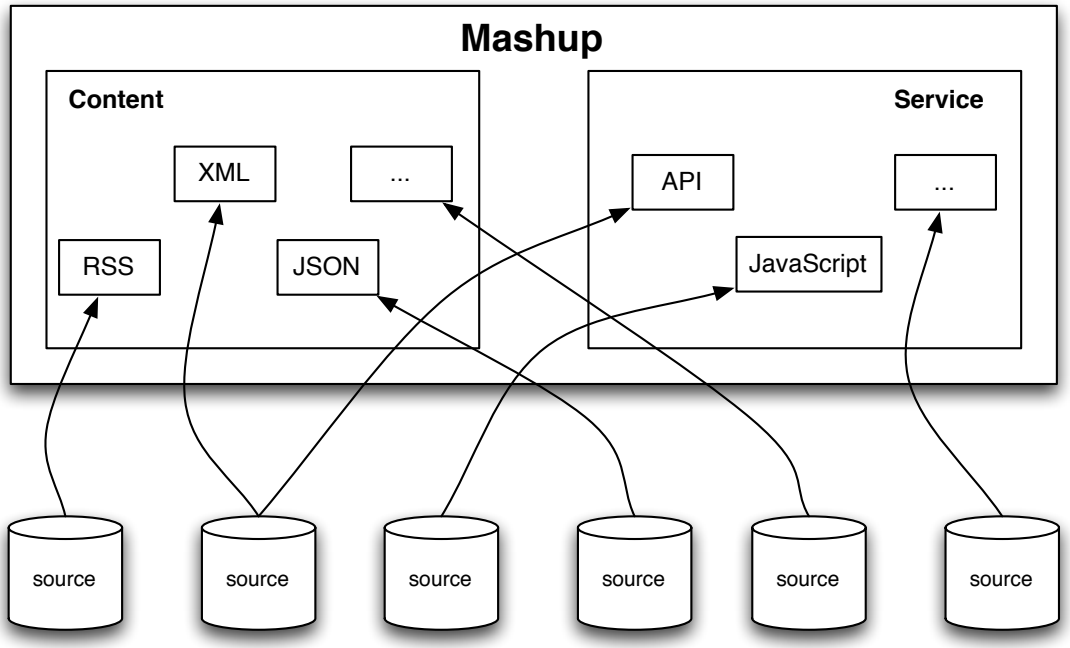


Figure 2.3: *A schema representing different type of mashup.*

Mashups are diffusing rapidly. Statistics provided by ProgrammableWeb[4], a refer-

[1] http://hominin.net/
[2] http://www.housingmaps.com/
[3] http://www.craiglist.org/
[4] http://www.programmableweb.com/mashups/

ence portal for APIs, mashups and the related tools, show a huge increment in mashup creation in the semester going from July 2008 to January 2009 (fig. 2.4). If we take into consideration the types that are available today (fig. 2.5) from the perspective of the application domains we can notice an increasing differentiation, where mapping solutions, which were dominating the scene just a few months ago are today being contrasted by other types, such as those related to shopping and messenging. The business potential related to mashups is being increasingly recognized and the APIs, services and platforms for their creation are the issue for tough competition between big players such as Google, Microsoft, and Yahoo.

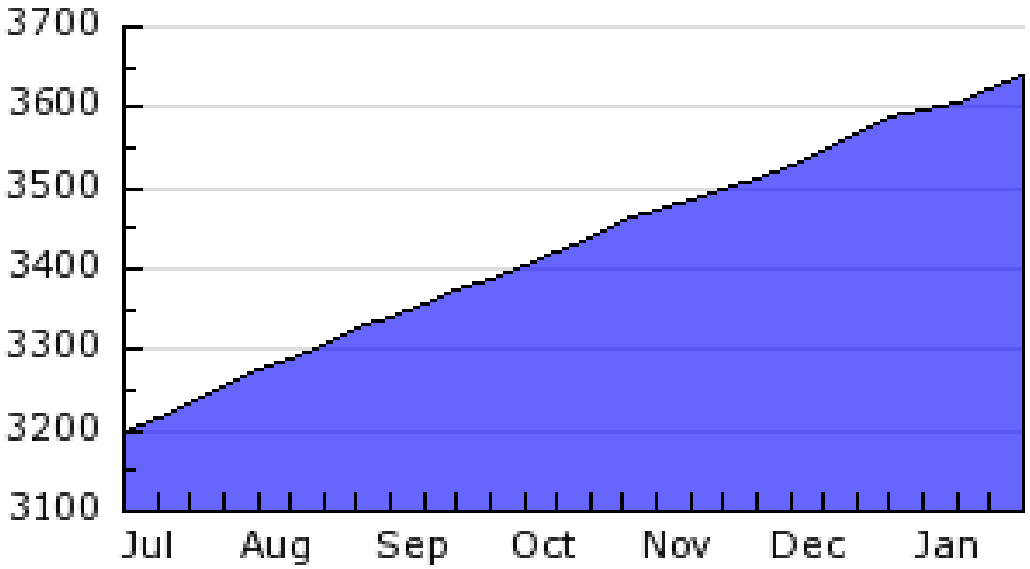


Figure 2.4: *Number of mashups available in ProgrammableWeb.com in the semester from July 2008 to January 2009 (`http://www.programmableweb.com/`).*

In fact, the mashup approach and the related technologies have increased user engagement a lot, much beyond the initial results of Web 2.0. Effective ways to combine sources and create rich user interface are available today; they generally require extremely low technical skills, and the Web is being populated by a huge number of mashups[1].

2.3.3 The Web of Data and the Semantic Web

Web contents today are designed and generated mostly for human consumption and processing; the extensive automatization of their processing would add great value

[1]See e.g. `http://www.programmableWeb.com/`

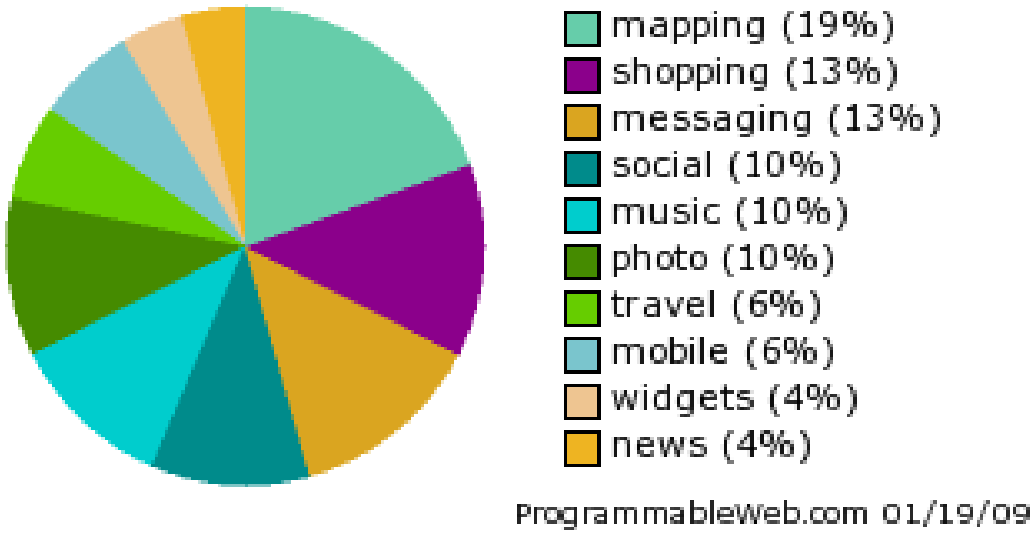


Figure 2.5: *Types of mashups available in ProgrammableWeb.com (`http://www.programmableweb.com`).*

to the current Web, but it also requires machines to become capable of accessing and manipulating information and data in better ways. In fact, even if tasks often require combining data across the Internet, machines are not smart enough to perform this combination. The vision of the "**Web of Data**" is a step in this direction, in that it promotes the idea of linking the contents (i.e. data) embedded in documents (e.g. in HTML pages) in a global information space rather than merely linking one document to another. Documents are currently linked to each other one way or another, but their content, which is what interests us, is undefined for Web applications. The currently available search engines perform queries based on the occurrence of keywords: they show documents containing the entered keywords, without taking into consideration the meaning associated to the keywords within the document; consequently the precision of the results with respect to the user needs is often unsatisfactory. A funny analogy that has been proposed in order to mimic this scenario is that of a parrot that learns some words and repeats them without any understanding of their meaning. The main differences between the current situation and the envisioned "Web of Data" are summarized in table 2.3.

The principles underlying the Web of Data were first defined by Berners-Lee (2007). A concrete outcome and a testbed for this vision is today represented by the "Linking Open Data" project[1], a grassroot community effort funded by the W3C Semantic Web

[1] `http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/`

Table 2.3: *The Web of documents and the Web of Data.*

	The Web of documents	*The Web of data*
Analogy	a global filesystem	a global database
Designed for	human consumption	machine first, humans later
Primary objects	documents	things (or description of things)
Links between	documents (or sub-parts of)	things (including documents)
Degree of structure in objects	fairly low	high
Semantics of contents and links	implicit	explicit

Education and Outreach Working Group[1], which was established in 2007 in order to identify datasets that are publicly available under open licenses, re-publish them in RDF (the Resource Description Framework; see section 2.4.4) and link them with each other. Figure 2.6 represents the currently available "cloud" of interlinked datasets provided by the project, which consists of more than 2 billion RDF triples (i.e. two billion RDF-encoded data) and around 3 million RDF links connecting disparate sources, from bibliographical to musical data.

In order to make a Web of Data possible, the degree of structure in objects must become higher than it is now, and the semantics of contents and links must become explicit. Regarding this, a number of standard languages that make it possible to express semantics at different levels (see section 2.4.4) is emerging, such as the above mentioned RDF. The debate here is whether semantic information would be added to all existing documents on the Web in order to make them understandable for Web applications (the "bottom up approach"), or applications that are able to have a better understanding of the data embedded in documents have to be developed (the "top-down approach"). Moreover, new tools for browsing, crawling, searching and indexing are required. Some of them already exist: in particular, browsers (such as Tabulator[2], Disco[3], the OpenLink RDF browser[4] and the Zitgist browser[5]), and search

[1] `http://www.w3.org/2001/sw/sweo/`
[2] `http://www.w3.org/2005/ajar/tab/`
[3] `http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/.`
[4] `http://demo.openlinksw.com/DAV/JS/rdfbrowser/index.html`
[5] `http://dataviewer.zitgist.com/`

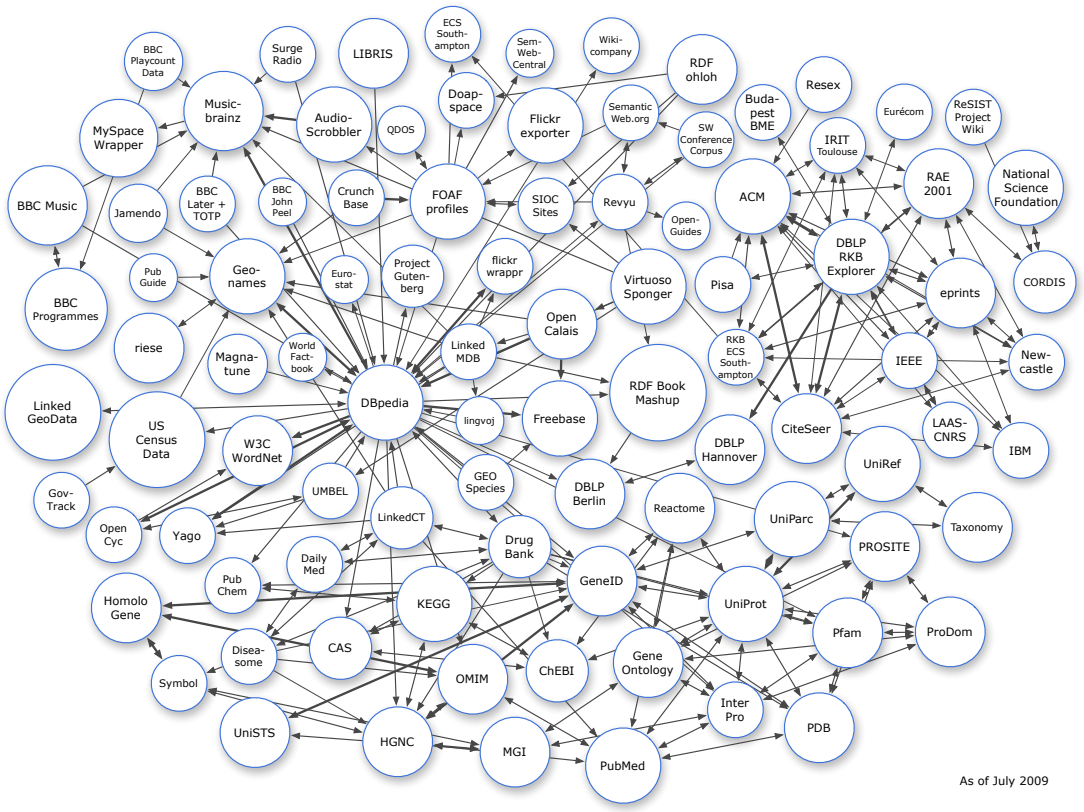


Figure 2.6: *The Linking Open Data dataset "cloud"* (`http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/`).

engines and lookup indexes (such as Falcons[1], Sindice[2], Swoggle[3] and Watson[4]).

The Web of Data is often considered equivalent to the **Semantic Web**. Even if it is hard to find a clear difference between the two in the literature, it seems that the former is a component of the latter, the Semantic Web being a more "conceptual" word addressing a new way of using the Web and a set of issues that go from theoretical and social aspects to the methods and technologies that make it possible to exploit the global information space enabled by the Web of Data. Moreover, it has to be stressed that the idea of the Semantic Web is largely antecedent to that of the Web of Data, and this confirms the distinction suggested here. In fact, the need for semantics in Web documents was proposed by Berners Lee at the very first International World

[1] `http://iws.seu.edu.cn/services/falcons/objectsearch/index.jsp`
[2] `http://sindice.com/`
[3] `http://swoogle.umbc.edu/`
[4] `http://watson.kmi.open.ac.uk/WatsonWUI/`

Wide Web Conference held at CERN in Geneva (Switzerland) in 1994[1], and an explicit Semantic Web was successively exposed in detail: *"The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation."* (Berners-Lee et al., 2001).

This vision has been recently revisited (Shadbolt et al., 2006), starting from the consideration that the first article failed to predict a straightforward transition to the Semantic Web, and argued that it would flourish only when standards will become well established and will provide effective solutions to the growing need for data integration.

But what is the current state of development of the Semantic Web from a general point of view? The scenario is very articulated because, as was said before, the Semantic Web is not limited to the Web of Data. In a talk given to the 2008 Semantic Technology Conference[2], I. Herman from W3C provided a non-exhaustive list of the growing number of application patterns referring to the Semantic Web:

- data integration
- knowledge engineering with complex ontologies
- better data management, archiving, cataloging, digital libraries
- managing, coordinating, combining Web services
- intelligent software agents
- improving search
- mixtures of these

It seems quite clear that the Semantic Web is currently highly related to technological aspects that are being developed or improved in order to make it work: the Web of Data and the related technologies and projects are the most evident result of this activity. On the contrary, the impact that the Semantic Web would have on our use of the Web or in other words, the realization of its visionary promises still needs detailed evaluation.

2.4 The Enabling Technologies: An Overview

The elements that were discussed in the previous section rely upon a large number of technologies that were developed in the last few years. This section introduces the most relevant of them, in order to define the building blocks of the near-future Web.

[1] See: `http://www.w3.org/Talks/WWW94Tim/`.

[2] The presentation is available at: `http://www.w3.org/2008/Talks/0924-Vienna-IH/Slides.pdf`.

In particular, XML technologies (section 2.4.1), APIs and Web Services (section 2.4.2), strategies and techniques for developing mashups (section 2.4.3), and languages and models for expressing semantics (section 2.4.4) are introduced.

2.4.1 XML Technologies: Enabling Syntactic Interoperability

The eXtensible Markup Language (XML) today represents the basic technology upon which most part of the machine-to-machine interaction on the Web is possible. Instead of defining a language, XML declares the characteristics and constraints for the definition of defining new and extensible languages (XML-based languages). XML adopts a tree model whose elements are enclosed in tags delimited by $< >$ and are nested one into the other on the basis of their relative position in the tree (fig.2.7)

```
<person>
   <name>
      <firstName>
               Glauco
      </firstName>
      <familyName>
               Mantegari
      </familyName>
   </name>
   <birth>
      <year>
               1977
      </year>
      <place>
               Treviglio
      </place>
   </birth>
</person>
```

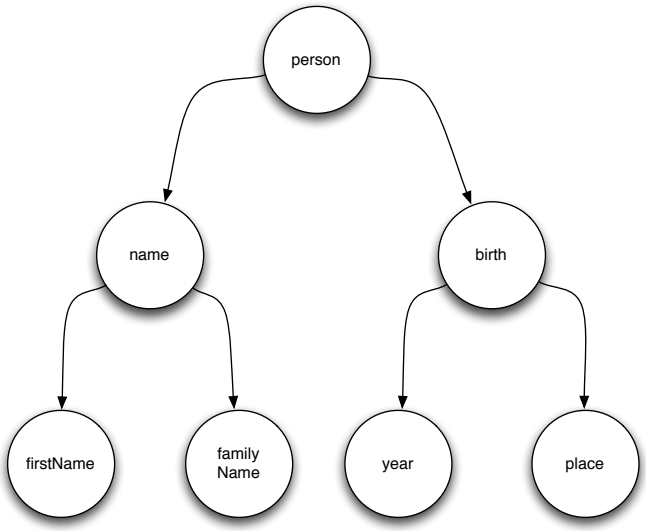


Figure 2.7: *An example XML structure (left) and its representation as a graph (right).*

In order to define a language based on XML, it is necessary to declare its characteristics in a schema, such as the name of the elements, the name of the attributes, how it is possible to combine the different elements in complex tree structures etc. Currently, two schema languages exist:

- the **Documet Type Definition** (DTD), which is the simplest
- the **XML Schema Definition** (XSD), which is more complex, but it offers more possibilities and is based on the XML syntax

With XSD and XML it is also possible to define a new language by integrating other existing languages. In order to avoid inconsistencies that may derive, for example, by equivalent names for different elements of two XML languages to be merged,

the concept of **namespace** has been introduced, which makes it possible to associate different URIs to equivalent names that are present in different XSDs. XML can also be transformed into another language with a different structure and even a different syntax (e.g. HTML, PDF) thanks to the **eXtensible Stylesheet Transformation Language** (XSLT), which is not described here.

Given these very general characteristics, it is important here to stress that XML technologies are the current solution in order to achieve syntactic interoperability. In fact, a plethora of XML-based languages has been developed using the XML technology, many of which are perhaps used unintentionally by the millions of people surfing the Web each day. **XHTML** (eXtensible Hypertext Markup Language), for example, is gradually substituting HTML 4.1 in codifying Web pages, in order to allow a better separation of contents from presentation (typically built using CSS) and the possibility to convey contents in a number of different channels (e.g. traditional, mobile, etc.). Another example is **RSS** (Really Simple Syndication), which makes it possible to subscribe to a Web page and automatically receive its updates without needing to visit the page. In the field of the Geospatial Web, the **KML** (Keyhole Markup Language) has become a *de facto* standard for geographical information exchange between different applications and is the principal format upon which Google Earth is based.

2.4.2 APIs and Web Services

Service providers increasingly expose Application Programming Interfaces (APIs) that can be combined in order to build efficient Web applications. The number of available APIs is constantly growing: ProgrammableWeb.com indexed more than 1.000 of them, as of January 2009, among which the most popular are provided by today's biggest Web companies, such as Google, Flickr, Youtube, Amazon, Microsoft, eBay, and Yahoo.

APIs can be exposed by means of different technologies (fig. 2.8); currently the most consolidated approaches are the REpresentational State Tranfer (REST) and the Service-Oriented Access Protocol (SOAP), two platform-neutral protocols for communicating with remotes services.

REST (whose definition and acronym were first introduced by Fielding, 2000) is currently the most diffused approach. A key factor for its success is its simplicity: REST is based on the HyperText Transfer Protocol (HTTP), and the related methods for managing the retrieval, updating, uploading, and deleting of resources identified by a URL. Web services using REST are usually indicated as "RESTful Web services".

REST is not properly a technology; instead it can be considered a set of principles or a technique which is based on the concept of "resource", and which defines the modalities for accessing resources in order to obtain their representation in the XML format (fig. 2.9). The basic characteristics of REST can be summarized in:

- The communication protocol is based on the client-server paradigm.
- Each request is stateless, i.e. it must contain all the information needed in order

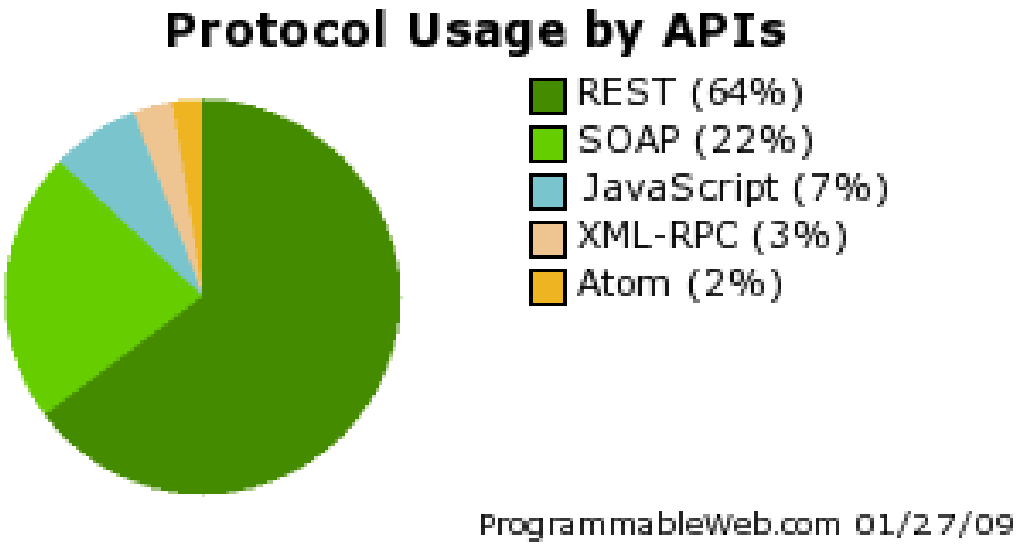


Figure 2.8: *Protocol usage by APIs* (`http://www.programmableweb.com/apis/`).

to be understood and cannot take advantage of any context previously stored on the server side.

- The reply of the server must be classifiable as "cachable" or "not-cachable", in order to increase the efficiency of the network.
- Each resource on the network is accessible through a generic interface, which is typically made by HTTP. The interface must expose a limited and syntactically well-defined dictionary of possible operations.
- Each resource on the network is identified univocally and retrievable through a Uniform Resource Locator (URL).
- The representation of the resources must be interconnected through URLs in order to make the receiving client able to move through different information layers.

The disadvantages of REST is that being based on HTTP, it is limited to applications that use HTTP for communicating with other systems; moreover, REST services are not well specified, and suffer from a limited modularity.

On the contrary **SOAP** is well-defined and its specifications are maintained by the XML Protocol Working Group of the World Wide Web Consortium (Booth et al., 2004). SOAP basically defines the XML format of an "envelope" (fig. 2.10) which is used to exchange XML messages over a number of protocols, ranging from HTML to

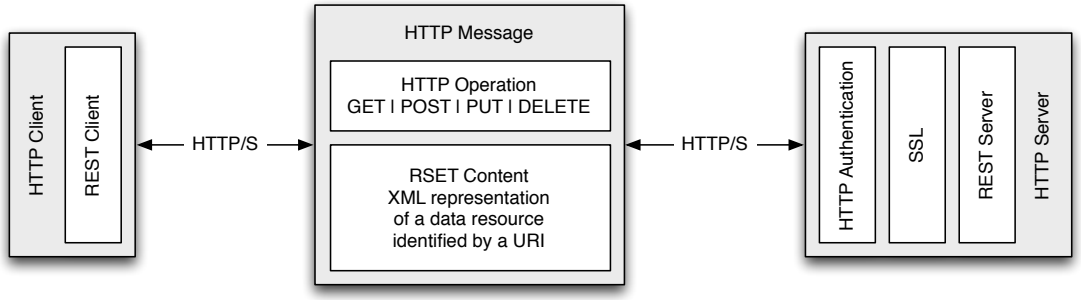


Figure 2.9: *Schema of a REST service (redrawn from Della Valle et al., 2008, fig. 1.4, p. 17).*

the Java Message Service (JMS[1]), the Simple Mail Transfer Protocol (SMTP) and other proprietary protocols. Two key components of the SOAP specification are:

- the use of an XML message to platform-agnostic encoding
- the message structure, which is made of a header (specifying contextual information), and the body (specifying the application payload)

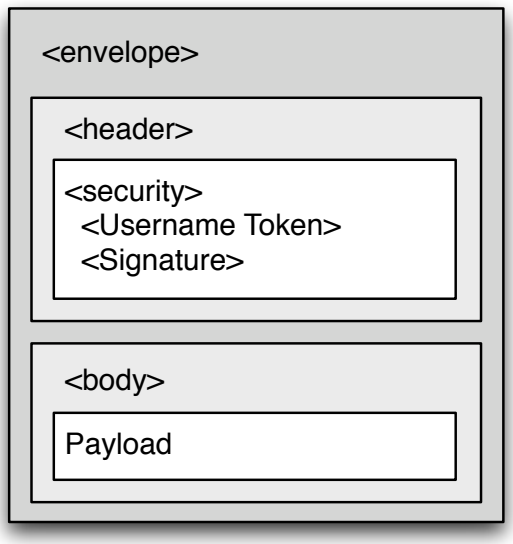


Figure 2.10: *Schema of a SOAP envelope (redrawn from Della Valle et al., 2008, fig. 1.3(a), p. 16).*

[1] `http://java.sun.com/products/jms/`

Several extensions of SOAP exist that address different aspects connected to the communication between machines, such as confidentiality, reliability, etc.; SOAP extensions are generally indicated with the prefix WS- (e.g. WS-Security, WS-Transaction, etc.).

Strictly connected to SOAP is the Web Services Description Language (**WSDL**), which is an XML-based language to describe Web services. It basically takes into account what the service does, how it does it and where it is available. The current specification (Christensen et al., 2001) is a W3C recommendation, and it offers better support for RESTful Web services.

2.4.3 Mashup Development: Strategies and Techniques

The creation of mashups require that content, presentation and/or application functionalities are combined from disparate sources (Yu et al., 2008, p. 44):

- content and presentation typically come in the form of RSS or Atom feeds, various XML formats, HTML, Shock-Wave Flash (SWF) or other graphical elements
- application functionalities are provided by publicly accessible APIs and Web services
- content, functionality and presentation are glued together in different ways, e.g. by using a server-side scripting language (such as PHP, or Ruby) or a more traditional programming language (such as C#, or Java)

From a general point of view, a mashup architecture is composed of three participants that are logically and physically disjoint (Merrill, 2006):

- the API/content providers, which expose their data through Web technologies such as REST, Web services, RSS or Atom feeds, etc.
- the mashup site, which is where the mashup is hosted (i.e. where the mashup logic resides), but not necessarily where the mashup is executed: for example, Web pages can reference scripting APIs, e.g. through browser side JavaScript (e.g. the Google Maps API)
- the client's Web browser, where the application is rendered graphically and where user interaction takes place

A key model for the interaction with the mashup's contents and services is represented by **AJAX** (Asynchronous JavaScript And XML; see (see Garrett, 2005), which makes it possible to dramatically enhance user experience by asynchronously loading and presenting contents, thanks to the combination of different technologies:

- XHTML and CSS for content and style presentation

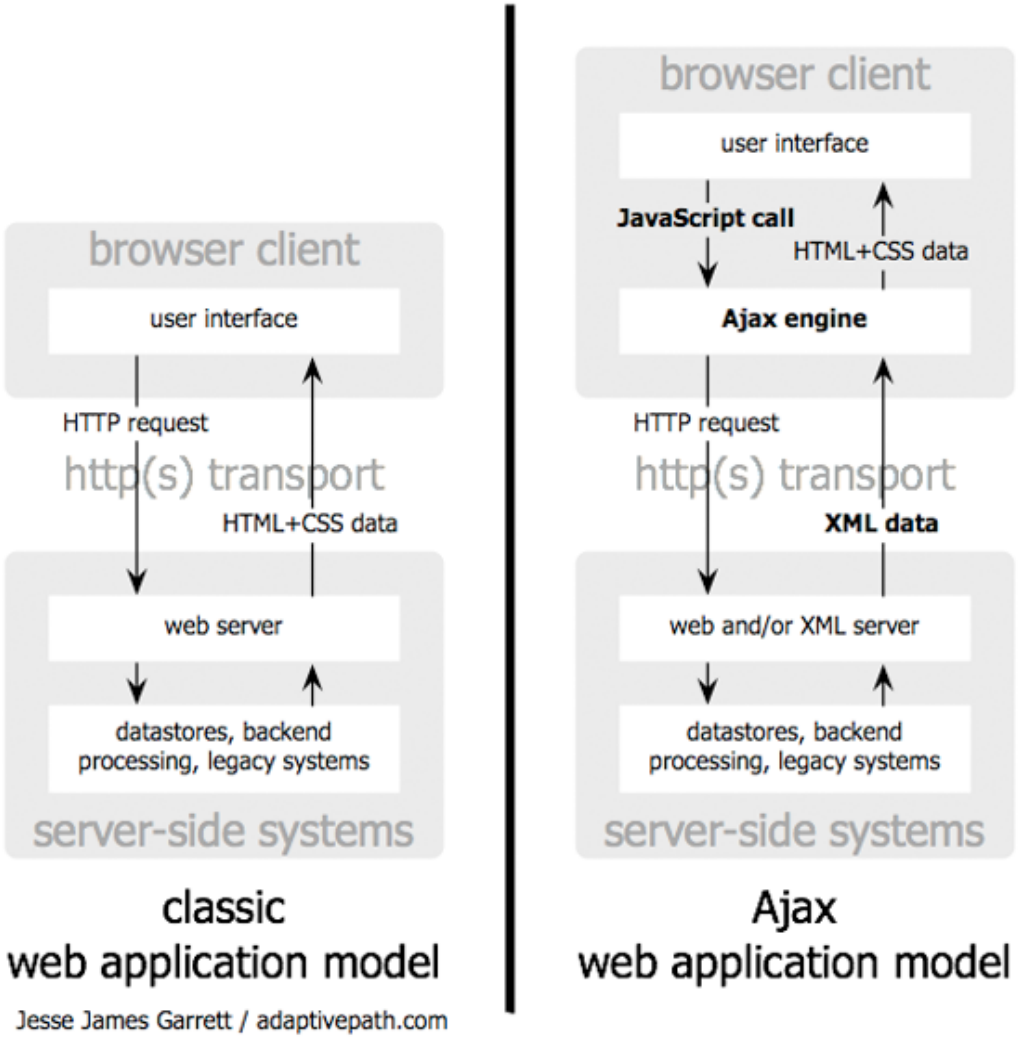


Figure 2.11: *The traditional model for Web applications compared to the Ajax model (Garrett, 2005, fig. 1).*

- DOM (Document Object Model) API for dynamic display and interaction
- asynchronous data exchange, typically in the form of XML data
- browser-side scripting, primarily JavaScript

Figure 2.11 represents the traditional model for Web applications (left) compared to the Ajax model (right).

Given these general characteristics, mashups can be created manually or by using specific tools that assist the developer and ease the integration of the components.

Manual development requires programming skills that may be too demanding for general users. In fact, the integration of data sources requires that the data schema and the format into which data are codified are known. Moreover, the methods and functionalities provided by the APIs need to be known in detail in order to control the application and to obtain the desired behavior. In case the application needs APIs coming from different platforms to interact (which in most cases is the typical situation) manually creating a mashup can become a complex task. On the other hand, developers have complete control over the application, thus making it possible to define tailored solutions for the application domain.

Nevertheless, **tool-assisted mashup development** is increasingly gaining success and it is proving to be a complementary or an alternative approach to manual development. The main advantages here are that general users can deploy effective and professional-looking solutions without having to acquire specific programming skills and competencies; on the other hand programmers can rapidly deploy proofs of concepts, demos or even the basic structure of a complex application without compromising the possibility of successively improving it with manual intervention.

There is a growing number of tools and frameworks that assist in the creation of mashups today. The interest that the business sector is paying to mashups (some people even refers to them as a new software development paradigm) is confirmed by the fact that most part of these tools have been developed by big players in the ICT sector, such as Google, Yahoo, Microsoft, Intel, and IBM. Yu et al. (2008) provide an extensive and up to date analysis of the most popular or representative approaches of end-user mashup tools, i.e Yahoo Pipes[1], Google Mashup Editor (GME)[2], Microsoft Popfly[3], Intel Mash Maker[4], and IBM Quick and Easily Done Wiki (QedWiki)[5]. The general ideas and elements on which these solutions focus are:

- A visual editor for controlling control the application composition and deployment (fig. 2.12).
- Tools for data source selection, e.g. via RSS or Atom feeds, XML sources, or other.
- Tools for data source connection, in case the application needs to relate and combine data coming from different sources in a composite view. These tools generally include methods that aid the development of simple functionalities (such as sorting, filtering, looping, regular expressions, counting, etc.) or advanced ones (such as location or term extraction).

[1] `http://pipes.yahoo.com/`
[2] `http://editor.google.mashups.com/`
[3] `http://www.popfly.ms/`
[4] `http://mashmaker.intel.com/`
[5] `http://services.alphaworks.ibm.com/qedwiki/`

- Tools for integrating externally provisioned services (e.g. Web services) or for the development of useful functionalities (e.g. in JavaScript).
- Tools for defining the layout of the application as a single Web page or as a set of inter-connected Web pages.

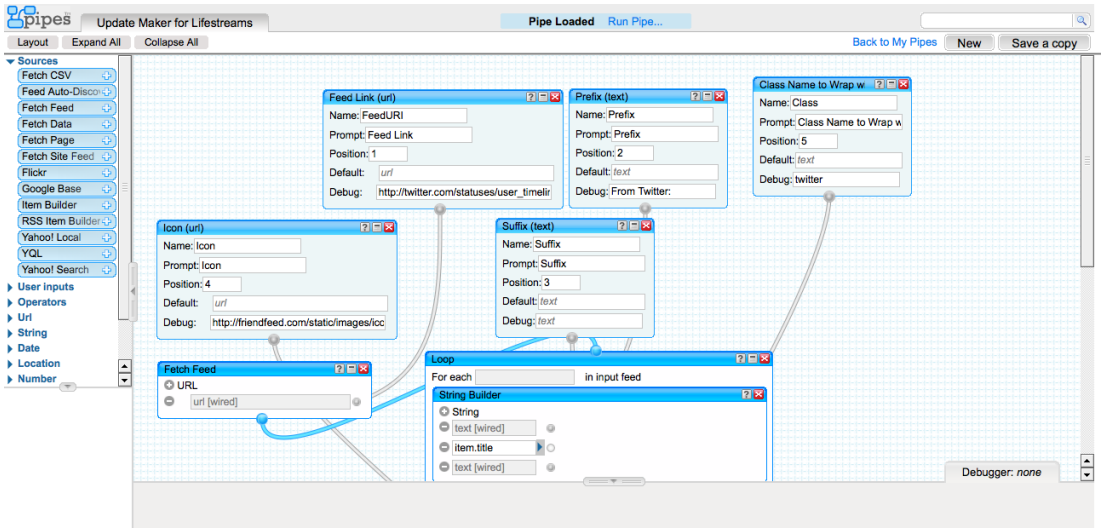


Figure 2.12: *An example of the Yahoo Pipes Visual editor (*`http://pipes.yahoo.com/pipes/pipe.edit?_id=6qrWDqOf3BG8Qhj32h2EvQ`*).*

Each solution differs from the others on the basis of the underlying mashup paradigm, which in turn determines data and service retrieval, combination, and presentation methods. For example, Yahoo Pipes is based on data processing pipelines and on operators interconnecting them, and it exposes output data in RSS, rather than in a user interface; GME is template-based, i.e. it offers the possibility to combine a set of standard modules; Microsoft Popfly is component-based, and operations on each component can be defined in a dedicated XML descriptor; Intel Mash Maker allows data to be retrieved from annotated source Web pages, rather than from structured data sources, such as RSS or Atom feeds; QedWiki, being based on a wiki environment, lets users edit, immediately view and easily share mashups.

These different solutions are useful in order to categorize tool-assisted mashup development approaches. Yu et al. (2008) suggest parameters based on the objects of integration (i.e. the components) and the methods by which they are glued together (i.e. the composition logic).

As has been said before, the development of a mashup solution may require that further manual interventions become necessary in order to accomplish tasks for which none of the existing tools provide a solution. An example of this is the extraction of

contents from Web pages that do not expose their data sources through convenient APIs or Web services. Screen scraping methods become then necessary in order to attempt to extract information from the content provider by parsing the provider's Web pages.

2.4.4 Towards Semantics

The need to express data and information on the Web in meaningful ways in order to allow a better automated processing has been the object of research for several years. The Semantic Web currently relies upon a number of technologies that represent the building blocks of a "pile" (fig. 2.13), which is important to introduce here with particular reference to the languages that currently exist for expressing semantics at different levels.

The base of the pile is made by **URI** (Uniform Resource Identifier), an expression that identifies the mechanism that makes it possible to identify the representation of a resource (such as a Web site address, a document, a file, an e-mail address, etc.) on the Web, which is accessible through a protocol (such as HTTP, FTP, IRC, etc.). More precisely, a URI is a string of characters that can take the form of an URL (Uniform Resource Locator) or a URN (Uniform Resource Name), the former defining the location of the resource in the network, the latter identifying it with a name in a given namespace (fig. 2.14).

Unicode is the standard text encoding format that allows machines to cooperate, because it ensures that text characters are independent from the software platform, the application, and the language used[1].

XML has been introduced in section (2.4.1) and its relevance in the pile is due to the fact that it allows syntactic interoperability, and partial structural interoperability (i.e. the possibility to interpret the structures of documents that have different logical schemas).

RDF (Resource Description Framework) is the W3C model (Manola and Miller, 2004) that deals with a first level of semantic interoperability, i.e. the possibility for machines to exchange information on the basis of the meaning of information rather than its format. In fact, even if it is possible to convert two different XML schemas in specific contexts in order to make them semantically interoperable (e.g. by using XSLT transformation), the operation is not feasible on a global scale, such as that of the Web; at the same time, trying to reach agreements on global schemas is utopistic. Consequently, it is necessary to develop new data models that do not suffer from the limits of XML. Even if it is an XML-based language, RDF offers a metadata data model that completely differs from that of the XML. The basic construct of RDF is the "statement", i.e. a minimal description of a resource; a statement takes the form of a "triple", i.e. an expression which is made with three components (fig. 2.15).

[1] See: `http://www.unicode.org/`

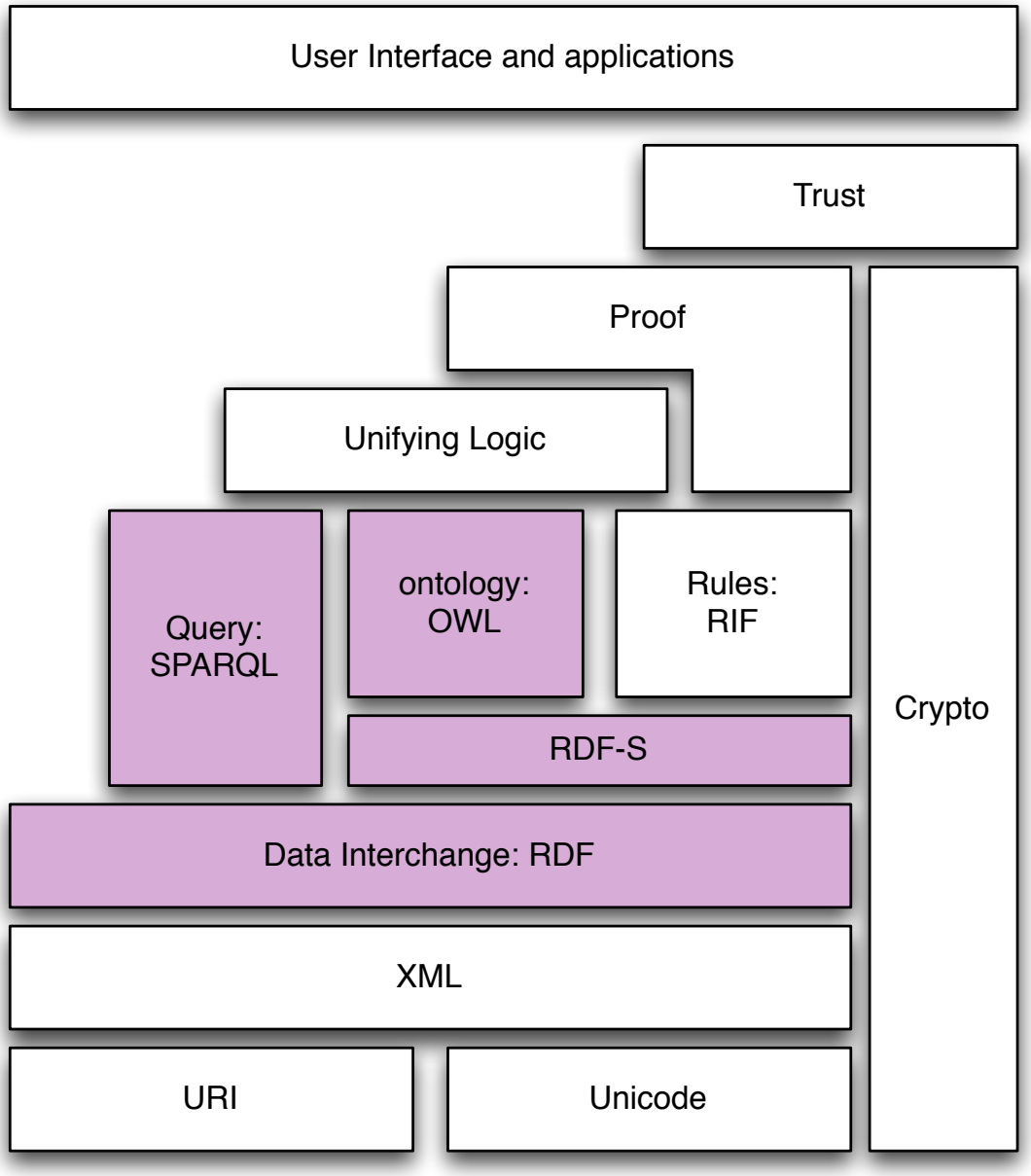


Figure 2.13: *The Semantic Web pile.*

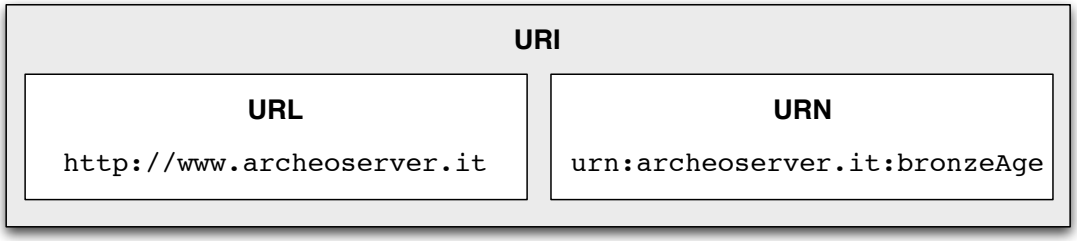


Figure 2.14: *URI, URL, and URN.*

- subject: the identifier of a resource
- predicate: the property or attribute of the subject that has to be expressed
- object: the value of the predicate with reference to the subject

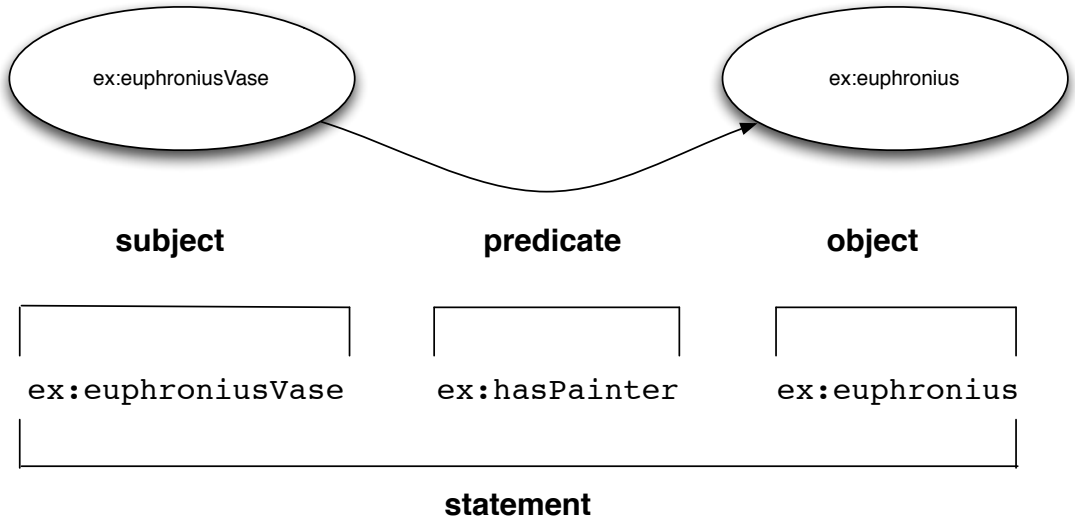


Figure 2.15: *The structure of an RDF statement.*

Each component of a triple is univocally identified by a URI, except for the case where the object is a text string, a number or a date; in this case the object is called a "literal" and does not require a URI.

From a more general point of view, RDF is a labeled, directed multi-graph, which is a model that makes it possible to define a set of nodes and link them through labeled and oriented relationships; in this model, a triple is represented by the connection of two nodes by an arch (fig. 2.16).

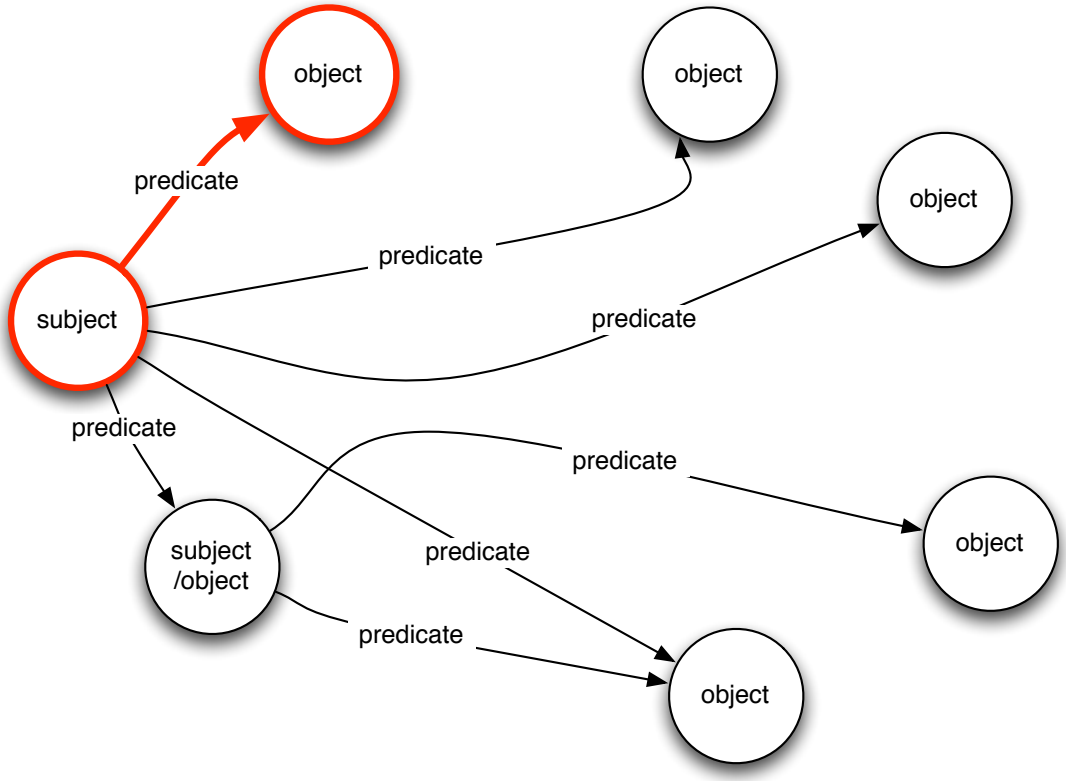


Figure 2.16: *An example of an RDF graph. A triple is highlighted in red.*

RDF is a relational model, but it shows some differences from e.g the model available in RDBMS. In fact, RDF allows greater flexibility and extendibility: new relations can be added dynamically without needing to modify the schema, and the repetition of triples in a RDF graph (e.g. in the case of a merge between different RDF graphs containing a set of identical triples) is not a redundancy. This way RDF makes it possible to easily merge statements coming from different sources, thus overcoming the limits of XML. Nevertheless, it is important to stress here that thanks to its flexibility, RDF can be used to represent relational or tree structures as well. For example, a tuple in a relational database can be mapped directly to a triple, the id corresponding to the subject, the attribute to the predicate, and the value of the attribute to the object.

Two main syntaxes currently exist for RDF:

- Notation3 (N3) syntax has been developed for better readability by humans and for simpler analysis by parsers
- RDF/XML, which is the official syntax, has been developed in order to exploit the number of tools available today for processing XML encoded data

RDF has become the model upon which common vocabularies for shared metadata describing resources are built. The most relevant of these is the Simple Knowledge Organization System (**SKOS**) (Isaac and Summers, 2008), which is an RDF-based extensible family of languages designed for representing any kind of structured vocabulary (such as thesauri, classification schemes, taxonomies, etc.).

RDF files can be made available on a Web server in order to be retrieved and queried; however, other RDF metadata publication strategies exist. Finally, semantical annotation of data in Web sites without needing to develop a fully RDF-based solution is possible through **RDFa** (RDF-in-attributes), which is a series of extensions to XHTML that makes it possible to use RDF metadata directly inside XHTML pages (Adida et al., 2008). This way traditional Web markup is extended with semantic elements and the RDF triples can be extracted thanks to available mappings with RDF.

The main problem that is connected to RDF is that two distinct systems may use different URIs in order to identify the same resource, thus making it impossible to confront and integrate different assertions on the same resource. **Ontologies** provide a solution to this problem and have been the subject of research in computer science for several years in fields such as knowledge representation and knowledge engineering. Without entering into detail, ontologies formally define relationships existing between concepts, i.e. they provide: "*...a formal and explicit specification of a shared conceptualization*" (Studer et al., 1998).

The **RDF Schema** (RDFS) is the first level of ontological language available for the Semantic Web, which adds some logic on top of RDF. More specifically, RDFS is an extension of RDF that makes it possible to express the constructs which describe groups of connected resources and the properties linking the different groups. Unlike XSD or DTD, RDFS does not impose constraints on the structure of a document; rather it gives useful information for the interpretation of the document itself. Using RDFS it is possible to specify:

- classes, i.e. the groups of connected resources (`rdfs:class`); each resource belonging to a class is an instance of that class
- the relationships linking different classes, i.e inheritance between classes (`rdfs:subClassOf`)
- domain and range (`rdfs:domain` and `rdfs:range`), i.e. restrictions on the applicability of certain properties to specific classes or the range of values a property can assume

Several other RDFS constructs are available (Brickley and Guha, 2004), that make it possible, for example, to define hyerarchies of properties or to extend the definition of vocabularies. It is important here to stress that, thanks to RDFS constructs the semantics associated to Web resources can be used by reasoners in order to infer new knowledge starting from existing knowledge or verify if inconsistencies oc-

curr in a given knowledge base. Figure 2.17 shows a simple example of inference. Given two RDFS domain and range specifications (the `ex:hasPainter` is a property of `ex:PaintedGreekVases` that can assume values from the `ex:Painter` class), and an RDF statement (the `ex:EuphroniusVase` was painted by `ex:Euphronius`) a reasoner may be able to infer class memberships that were not stated in the original RDF file, i.e. the `ex:EuphroniusVase` is a member (i.e. an instance) of the `ex:PaintedGreekVases` class and `ex:Euphronius` is a member of the `ex:Painter` class.

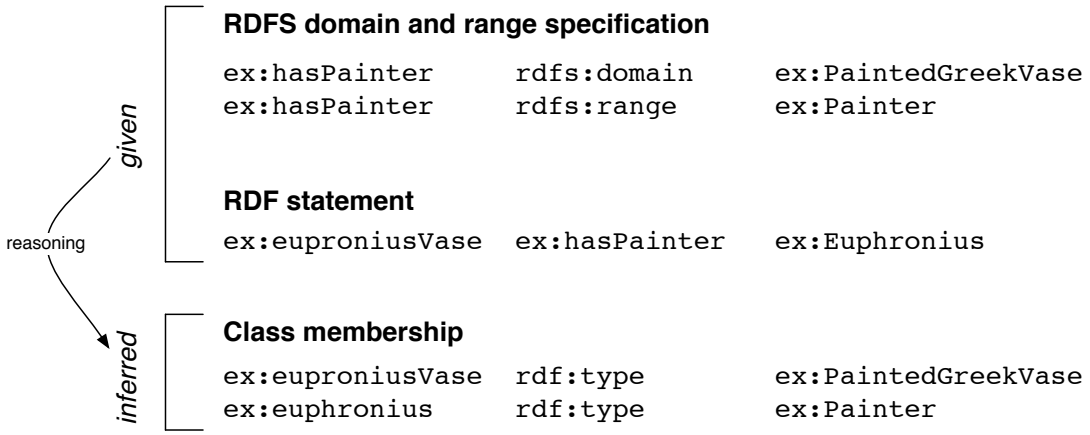


Figure 2.17: *An example showing a scenario where knowledge can be inferred starting from RDFS constructs and an RDF statement.*

OWL (Web Ontology Language) is a family of languages that currently represents the most advanced way of representing knowledge on the Web, by authoring ontologies. It was developed in order to enrich the capabilities of RDFS in being used for automated reasoning. Three main versions of OWL are available today, with increasing levels of expressiveness (Smith et al., 2004):

- OWL Lite is the simplest form that essentially offers basic constructs and simple constraints.
- OWL DL is the most diffused version, because it balances expressivity with computational completeness, decidability and the availability of practical reasoning algorithms. In fact, OWL DL is based on the Description Logics, i.e. the subset of the First Order Logic (FOL) that has been defined in order to be decidable.
- OWL Full offers the best expressivity, but in most cases the aspects of OWL DL are not retained.

Using OWL it is possible to enrich the possibilities of properties characterization a lot, by specifying aspects such as transitivity (`owl:TransitiveProperty`), sym-

metry (`owl:SymmetricProperty`), and inverse property (`owl:inverseOf`). Moreover, it is possible to introduce restrictions on relationships (`owl:Restriction`) such as cardinality or the set of values the property can assume.

SPARQL (Simple Protocol and RDF Query language) is both a query language for RDF (and then also for OWL), and a protocol that enables this kind of request in a Web environment (Proud'hommeaux and Seaborne, 2008). This dual characteristic derives from the fact that RDF resources are generally stored in distributed data sources which may have different schemas: it is then necessary to provide both a query language and a protocol that mediates between the user and the data sources.

Some SPARQL constructs are similar both in syntax and in meaning to those of the SQL (Standard Query Language). This is the case, for example of the `SELECT`, the `FROM` and the `WHERE` clauses, which make it possible respectively to define the variables of a query, the data sources to be queried, and the conditions of the query. Conditions are expressed through the so-called "path expression", that identifies the set of the triples which are necessary in order to resolve the query.

In adition to these, SPARQL offers other constructs that unlike SQL makes it possible to query data sources whose schema is unknown: the `DESCRIBE` clause, for example, makes it possible to obtain a description of the desired resource, while the `ASK` clause determines if a specific statement is present in the data source. The results of a SPARQL query can be shown in the form of an RDF graph by using the `CONSTRUCT` clause, thus allowing a more sophisticated processing.

SPARQL as a protocol acts on the so called "**SPARQL endpoints**", i.e. the access points that data sources expose in order to be able to receive a SPARQL query and to send the results back. It is important to stress that a SPARQL endpoint keeps the incoming query request and the application internal data distinct; this way it is possible to use a preferred technology (such as a RDBMS) for internal data management, provided a translation component between SPARQL and the internal query language (e.g. SQL) is defined. The SPARQL protocol is based on WSDL 2.0 and its specifications currently define the links to the HTTP and SOAP protocols.

Other blocks of the pile are not directly relevant for this research, but are introduced in order to complete the description.

RIF (Rule Interchange Format) is a W3C proposal for the semantic Interchange of rules that may be useful for automated reasoning and are expressed in different languages (such as SWRL, WRL, FOL RuleML, and KIF).

The **Unifying logic** identifies the need for a logic upon which the mediation between the information and knowledge representation of the lower levels and the issues connected to its dissemination to users (upper levels) is built.

The **Proof** block addresses the issues connected to establishing the truth of statements.

The **Trust** block refers to the different levels of trust in data a user can have while retrieving them from the Web.

The **Crypto** block ensures that information is retrieved from affordable sources.

Finally, information is accessed on the Semantic Web through the components of the **User Interface and application** block.

3

The Semantic Web, Cultural Heritage and Archaeology

Semantic Web applications in cultural heritage and archaeology are discussed in this chapter in order to contextualize the specific domain this research takes into consideration. In particular, the issues that are slowing down or even hindering the adoption of the new approaches are highlighted, and the most relevant projects carried out to date are discussed (sections 3.1 and 3.2). Moreover, the frameworks of digital libraries, e-Science and cyberinfrastructures are introduced (section 3.3), in order to extend the analysis perspective towards crucial elements that are emerging around and beyond the Semantic Web vision and technologies, with particular respect to applications to cultural heritage and archaeology.

3.1 The Semantic Web in Cultural Heritage

Since the beginning of the 2000s, Semantic Web technologies and their potentials for the integration and exploitation of digital cultural heritage information have received increasing attention, and today they represent an exciting and dynamic field of interdisciplinary research. Discussions about the issues and advancements in this area took place within the context of workshops, scientific literature, and national and international projects, in a range of contributions varying from the very specific to the more generic, from the more theoretical to the more practical.

It is a common opinion that the diversity and epistemological richness of cultural heritage provides an excellent field for the deployment and experimentation of Semantic Web-based systems. Aroyo et al. (2007a, p. 7) for example, underline the complex modeling challenges emerging from key characteristics of cultural heritage, such as its openness, the lack of clear boundaries, and the network of meaningful relationships that always link one resource to many others.

On the other hand, several factors slowed the diffusion of semantic systems for cultural heritage a lot. Isaksen (2008, p. 5) underlines how many theoretical discussions in the last few years tended to generate a combination of excitement and skepticism, where the potentials of the "new era" are envisaged, but a "veil of mystery" still remains, therefore determining a gap between the vision and the practical approaches to "do the Semantic Web". The outcomes of research initiatives, such as the early DigiCULT Project (DigiCULT Project, 2003), as well as more recent position

papers (such as Parry et al., 2008) provide clear evidence supporting these considerations. Moreover Isaksen (2008) observes that there is still a certain confusion in the "digital cultural heritage" community about what the Semantic Web is; misunderstandings sometimes lead to identifying it with other (even if to a certain extent potentially parallel) things, such as Web 2.0 approaches and technologies. This is certainly true if we consider the broad cultural heritage community that often showed a superficial approach with respect to the deep understanding of the nature and limits of new technologies; however, today there are a number of successful initiatives that show how the proper set of Semantic Web ideas and technologies is permeating the cultural heritage sector, mostly thanks to the activity of computer science research groups undertaking interdisciplinary projects.

Therefore, an outline of the principal characteristics of these projects is considered relevant in order to discuss the current situation.

3.1.1 ARTISTE - SCULPTEUR - eCHASE

ARTISTE was a EU-funded project of the early 2000s that aimed to investigate image retrieval for several major art galleries in Europe (the Louvre, the Victoria and Albert Museum, the Uffizi Gallery, and the London National Gallery) exploiting integrated content and metadata (Addis et al., 2002). Searches through integrated information have been implemented both using image analysis techniques (such as selection by similar color, pattern, texture) and text-based metadata search. The integration and the interoperability between resources have been supported using existing open metadata standards, among which is the RDF Schema.

The achievements in the context of ARTISTE constituted the basis for the **SCULPTEUR** (Semantic and content-based multimedia exploitation for European benefit) project[1] (Addis et al., 2005). Among the principal objectives of the project, the development of new ways of information retrieval and collection exploration with Semantic Web and content analysis techniques represented a central concern. The domain is that of art museums and galleries exhibiting their collections in the form of 2D and 3D digital representations, and the envisaged audience is that of professionals in the art and museums fields.

The project specifically addresses the retrieval of multimedia content, which comes in the form of digital images, 3D models, and videos with associated textual information and metadata. An ontology-driven approach is at the basis of the design of adaptive search modalities and visualization mechanisms for the virtual representations of objects, both in 2D and in 3D. The SCULPTEUR ontology has been developed on the basis of the CIDOC CRM, which has been extended in order to better suit the needs of the project, and has been used as a common model to which each single museum database has been mapped. In addition, an ontology of the system components has

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

been developed in order to dynamically adapt user interfaces to the different typologies of data retrieved.

An online demonstrator[1] allows to search and explore the Victoria and Albert Museum collection by concepts and by content. The concept-based modality makes use of an ontology visualization tool implementing a graph-based approach, and the instance visualization and query browser "mSpace" [2]; content based search extends the ARTISTE approaches in image retrieval based on the characteristics of the 2D and 3D models.

eCHASE (Addis et al., 2006) continued the experience of SCULPTEUR, with particular emphasis on the possibility for third parties to search and browse for the media they require, and to collect, annotate and export groups of relevant objects. The project integrates in the eCHASE system contents coming from different holders through metadata and media import: data is delivered as a set of images with the associated metadata coming in different formats (XML, Excel spreadsheets, EXIF embedded in the images, SQL Server dumps). A workflow-based approach is used for integration, through data cleaning and transformation procedures, and semantic integration of metadata by mapping to a common structure. The common structure is the CRM Core[3] version of the CIDOC CRM, which is a condensed set of metadata elements designed for resource discovery that can be expressed in a Dublin Core compliant form.

A Web client demonstrator has been developed using approaches and technologies previously employed in the SCULPTEUR project, and by providing new functionalities, such as the selection, aggregation, annotation and sharing of groups of media items in order to facilitate content use in external applications, while preserving in the export the established semantic links.

The approaches and experimentations of ARTISTE, SCULPTEUR and eCHASE represent significant experiences in Semantic Web applications for cultural heritage. Many novel ideas emerged during the projects, such as the combination of concept and content based searches on digital images and 3D models. However, at the time of this writing, the projects' demos seem to be unavailable, thus hindering the possibility of a more complete and practical evaluation of the systems.

3.1.2 MultimediaN E-Culture - CHIP Browser

MultimediaN E-Culture[4] is a project promoted by the public-private organization MultimediaN[5], which aims to demonstrate "*...how novel semantic-web and presentation technologies can be deployed to provide better indexing and search support within large virtual collections of cultural-heritage resources.*". The principal outcome of the project is

[1] http://sculpteur.it-innovation.soton.ac.uk/
[2] http://www.ecs.soton.ac.uk/research/projects/292/
[3] http://cidoc.ics.forth.gr/working_editions_cidoc.html
[4] http://e-culture.multimedian.nl/
[5] http://www.multimedian.nl/

a demonstrator system[1] (Schreiber et al., 2006; van Ossenbruggen et al., 2007) bringing together different online repositories of cultural heritage in the Netherlands.

The system is based on application ontologies that have been modeled on the individual databases, as well as on diffused thesauri (the Getty's AAT, TGN, and ULAN), and the lexical resource WordNet. The thesauri have been converted to the RDF/OWL format[2] and the alignment of ontologies has been done by hand. The graphical representation of the MultimediaN E-Culture data cloud (fig. 3.1) shows the inter-linkage between the project's resources, and in particular between the repositories of the cultural Institutions involved (Rijksmuseum, Tropenmuseum, Louvre, The Netherland Institute for Art History, etc.) and the thesauri. As is shown in the figure, thesauri are the means through which concepts, people, and places coming from the different datasets are connected; from the graphical representation, it appears that the overall vision is similar to that of the linked data initiative introduced in section 2.3.3.

Resources are kept distributed, i.e. they are identified by their URIs, while for practical reasons, the associated metadata are stored locally. The interlink of these resources is based on annotations, which make use of a description template derived from the VRA 3.0 Core Categories[3] (a specialization of the Dublin Core metadata specification for art images[4]) and the above mentioned thesauri. Techniques for the automatic semantic annotation based on the existing textual annotations of the single data sources and parsing techniques have been experimented in order to define effective methods for information extraction on metadata associated to the collections; in addition, a manual annotation interface is available.

Users can search for contents by using a dynamic text input field which provides keyword search on the underlying thesauri and taxonomies. In addition, the navigation in the integrated and semantically-enriched content is supported by a generic browser called "\facet" (Hildebrand et al., 2006), which allows to explore data along a number of facets related to wh-views (who what, where, when), production (creator, creation site, date), format (type, material, technique), exhibition (current repository, current site).

From the technical point of view, the architecture of the demonstrator system is based on SWI-Prolog and its Semantic Web libraries (fig. 3.2). Prolog is used as a query language for the searching and clustering algorithms of the application logic module, and a SPARQL-based access has been added later. The query results are returned as Prolog Herbrand terms, which are then elaborated by the presentation generation module in order to provide users with Web documents.

Data from MultimediaN E-Culture, together with ontology mappings from the

[1] `http://e-culture.multimedian.nl/demo/search/`

[2] The MultimediaN E-Culture thesauri are available online at `http://e-culture.multimedian.nl/resources/`

[3] `http://www.vraweb.org/`

[4] An OWL specification of VRA is available at `http://e-culture.multimedian.nl/resources/`

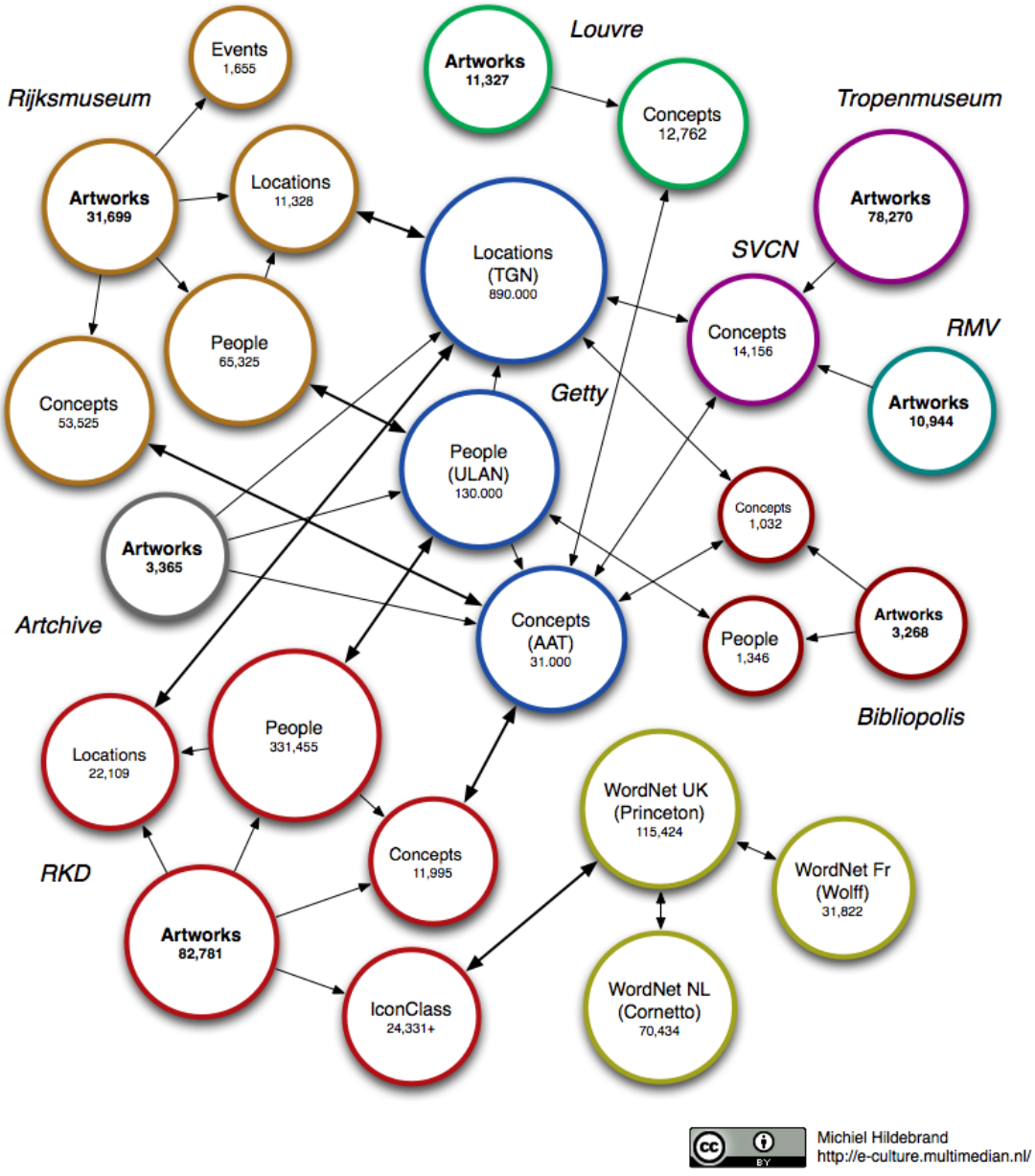


Figure 3.1: *The MultimediaN E-Culture data cloud (`http://e-culture.multimedian.nl/resources/datacloud`.*

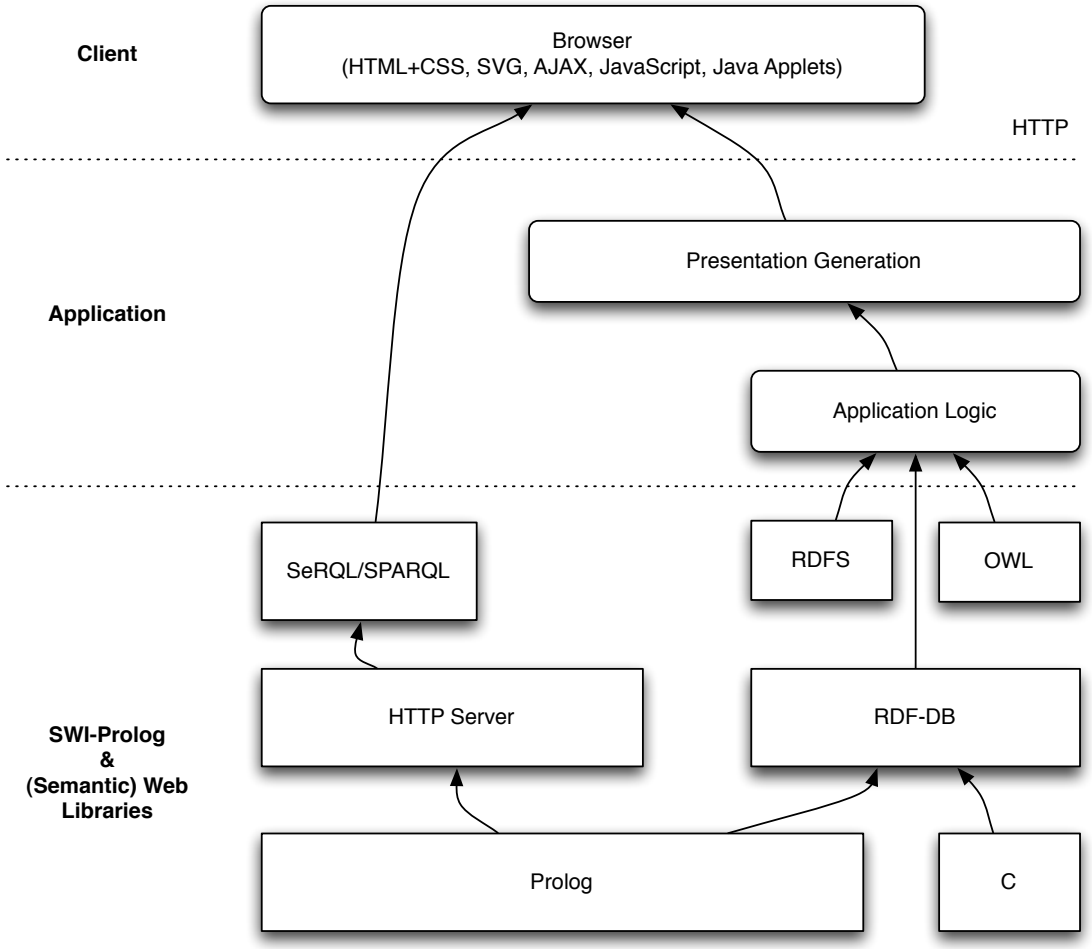


Figure 3.2: *The architecture of the MultimediaN E-Culture Demonstrator (redrawn from Schreiber et al., 2006, fig. 3, p. 4).*

CATCH STITCH project[1], have been used in order to test the **CHIP** (Cultural Heritage Information Personalization) demonstrator browser (Aroyo et al., 2007b). CHIP aims to investigate Semantic Web technologies for the creation of personalized access modalities to digital museum collections, both on the Web and in the museum. The components that have been implemented are:

- an artwork recommender
- a tour wizard
- a mobile tour for PDA devices

From the technical point of view, CHIP makes use of diffused technologies and standard languages, such as the Sesame repository and the SeRQL query language; FOAF profiles of users, and filtering on a person's social network are being exploited in order to refine the recommendation possibilities. On the Web client, available tools and services for the creation of interactive AJAX-based interfaces are used (such as the Simile Exhibit framework) in a Web mashup approach.

The principal and interesting outcomes of the MultimediaN E-Culture project are in the methods for searching and browsing through the dataset, and for presenting data to users that have been developed and tested in the demonstrator system. Even if other generic browsers and services based on the facet paradigm exist (such as Simile Exhibit and Longwell), the creation of a tool in the specific context of cultural heritage information is of primary relevance. On the other hand, the distributed approach in the integration of information coming from Web repositories in the form of URIs represents a relevant experience, which demonstrated e.g. the possibility of inter-linking data sources on the basis of standard and domain specific thesauri. Concerning the CHIP project, an innovative result is the exploration of the possibilities for personalized access to contents based on Semantic Web technologies and standards.

3.1.3 Museum Finland - Culture Sampo

The SeCo (Semantic Computing) research group at the University of Helsinki is one of the major contributors to the development and experimentation of advanced Semantic Web applications for cultural heritage of the last few years. The "**Museum Finland**" project[2] (Hyvönen et al., 2004b, 2005) represents the earlier concrete deployment of Semantic Web tools and services in the museum sector. The aim of Museum Finland has been to integrate information of Finnish museums coming from four databases (using four different schemas and cataloguing systems, and four different database systems), and to present it online through a common Web interface.

The project relies on seven domain ontologies (artifacts, materials, actors, situations, locations, times, collections) that have been semi-automatically created through

[1] http://www.cs.vu.nl/STITCH/
[2] http://www.museosuomi.fi/

processes of analysis and integration of national and local thesauri and data, and collection item data. Additional cultural and commonsense knowledge has been incorporated in order to enrich the representation, and to link collection contents in richer and more meaningful ways. The global knowledge base conforming to the ontological structure is in the RDF format; it is created with a process for transforming local heterogeneous databases requiring as little manual work and technical expertise as possible by the museum staff. This process involves a first syntactic homogenization by transforming the relational database into a shared XML language, and successive stages ending with RDF cards conforming to the global museum ontologies. The integrated repository is then exploited in order to offer a semantic view-based search engine and a semantic linking system, allowing users to explore the collections in rich and highly contextual ways, e.g. with semantic browsing exploiting the relationships between collection data and single objects and providing recommendation links.

From the point of view of the architecture (fig. 3.3), the system is based on a data tier comprising domain knowledge (expressed in the ontologies), and metadata of the data resources (expressed as annotations and RDF cards). A middle layer provides logical rules (in Prolog) between the content and the semantic services, which allows to create semantic recommendations for the browser service, and to support the search engine with hierarchical index-like decomposition of category resources.

This work relies on previous work for the creation of view-based search engines and recommendation systems (Ontogator Hyvönen et al., 2004a) which has been integrated into a new tool developed for the project (OntoViews) in order to enable a multi-facet view-based search. The outcomes of the Finnish Museums project represent a significant advancement in the demonstration of the potential of Semantic Web technologies for integrating and accessing museum collections. Moreover, the release of the developed tools under an open source license[1] enables the reuse of software in the context of new projects dealing with similar topics.

Culture Sampo[2] (Hyvönen et al., 2009a,b) is a follow-up of Museum Finland that considerably extends the approach of the previous project both from the point of view of the domain, the portal functionalities, and the employed technologies. The scope of the system is the cultural heritage of Finland, and the central idea is to create its "semantic memory", combining the possibilities of the Semantic Web and the Web 2.0. The key axes along which the creation of such a memory is made are explained in Hyvönen (2009) and summarized in Hyvönen et al. (2009a, p.1). The integration of a multitude of heterogeneous contents in a global level infrastructure, the possibility to support distributed content production, and the deployment of intelligent search, browsing and visualization techniques for end-users are crucial aspects of the Culture Sampo view.

The envisaged infrastructure is based on a national ontology, which is being devel-

[1] http://www.seco.tkk.fi/software/
[2] http://www.kulttuurisampo.fi/

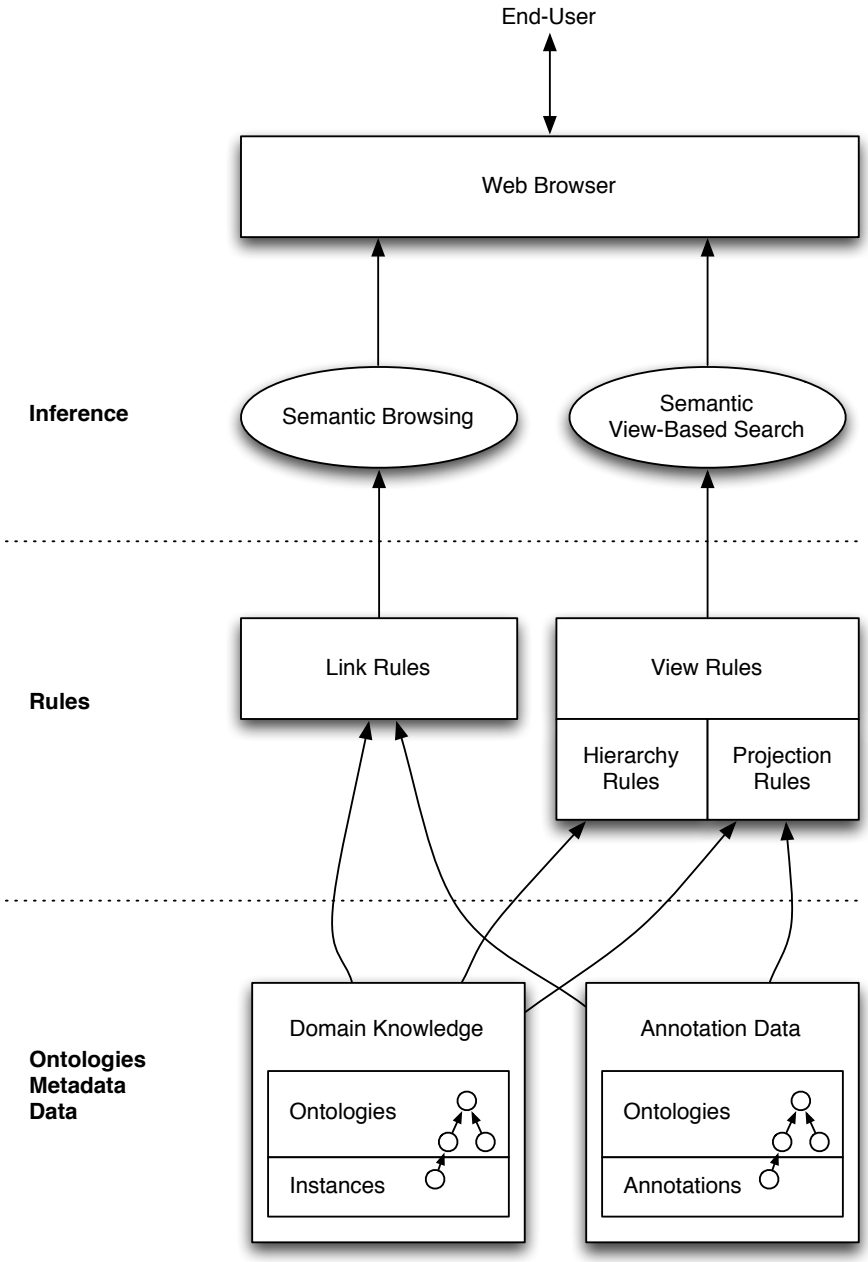


Figure 3.3: *The architecture of MuseumFinland on the server side (redrawn from Hyvönen et al., 2005, fig. 4, p. 233).*

oped in the context of the FinnONTO research project (Hyvönen et al., 2008). This ontology (KOKO) is composed of several domain specific ontologies which are mapped together. With reference to cultural heritage information, KOKO includes locally developed ontologies, as well as ontological versions of international systems such as Iconclass[1] (for art and iconography) and ULAN[2] (Union List of Artists Name).

Culture Sampo aggregates contents from a remarkably high number of institutions, which include 22 museums, archives, and libraries, as well as external contents from the Getty Foundation, Wikipedia, Geonames, and Panoramio. A crucial characteristic of the project is the distributed, and cross-domain content creation, which allows contributors to increment knowledge and maintain the overall ontology and annotation infrastructure. To this respect, several generic tools have been developed, such as a national ontology library service (ONKI[3]), a metadata editor (SAHA[4]), an information extraction tool (POKA[5]), and a semantic content validator (VERA[6]).

The portal provides contents both for machine processing and for human consumption. In particular, a sophisticated interface combining several technologies, services, and mashup techniques according to nine thematic perspectives has been developed. These include map views, relational searches, a collection view, as well as a search and organize facility which allows filtering, aggregation and visualization of contents on the basis of personal interests; moreover, a "skills and processes" perspective allows to find cultural procedural descriptions in the system, and access it through semantic models enriched by multimedia material, such as videos. Finally, specific perspectives concerning respectively Finnish history, the semantic version of the national poem (called "Kalevala"), and the history and cultural evolution of the Karelia region complete the scenario. Machines can access Culture Sampo's contents via an AJAX interface, and the project's knowledge base can be easily integrated in external systems; moreover, a Web widget approach has been proposed (Mäkelä et al., 2007), in order to allow the addition of semantic functionalities to existent systems with only a few lines of code.

Many other aspects of Semantic Web approaches and technologies have been investigated in the Culture Sampo project, such as methods for disambiguating spatiotemporal locations on the basis of ontologies (Kauppinen et al., 2008). Culture Sampo provided many original contributions, from the more theoretical to the more practical and technical ones. On the other hand, this richness underlines the great potentialities, and at the same time, the complexity behind the adoption of a Semantic Web approach for the integration and access to large bodies of cultural heritage information. The availability of the software components developed by the SeCo research

[1] http://www.iconclass.nl/
[2] http://www.getty.edu/research/conducting_research/vocabularies/ulan/
[3] http://www.seco.tkk.fi/services/onki/
[4] http://www.seco.tkk.fi/services/saha/
[5] http://www.seco.tkk.fi/tools/poka/
[6] http://www.seco.tkk.fi/services/vera/

group under the terms of open source licenses should encourage institutions developing similar projects to adopt these components in order to validate the generality of the approaches that have been proposed.

3.1.4 Contexta SR

Contexta SR is a recent project undertaken by the Universidad Técnica Federico Santa Maria (Chile), whose goal is to semantically integrate cultural heritage heterogeneous information using Semantic Web technologies (Astudillo et al., 2008). The project concerns both the technological integration of the systems and tools used by different institutions and the semantic integration of resources through the definition of a common vocabulary.

The Contexta SR approach takes three main modeling dimensions into consideration: ontologies for describing artifacts, ontologies for describing circumstances and social ontologies. From a general perspective, the approach substantially differs from other experiences (such as those mentioned in the previous sections) in that it considers a broad spectrum of services and ontologies, many of which have not been specifically developed for the cultural heritage sector. ISAD(G)[1] provides general guidance for the presentation of archival description, and is the main ontology upon which the project is based; it has been combined with other lightweight and diffused ontologies such as FOAF, the W3C Geo Ontology, the Music Ontology, and the Event Ontology, providing the "semantic glue" between them. In the project's view, this choice allows for easier integration than e.g. that possible with the CIDOC CRM with other Semantic Web repositories.

Physical integration of resources happens through export to XML from the databases of each single institution, and its transformation into RDF through XSLT templates; this operation is performed by an "ingester" component. The RDF graph is then automatically linked to nodes of the central system's graph in a pending repository through filtering, and successively permanently linked or rejected by the manual intervention of experts through an "auditing" component. The final creation of dereferenceable URIs is relevant in this process, since it allows to make the project's resources available for the Linked Data Initiative.

The Contexta SR approach aims to encourage the development of an ecosystem of applications around the core dataset, by providing a platform for the integration of digital repositories on the Web. This would allow to develop "semantic mashups", where several sources in the RDF format are exploited by Web applications, and information is combined in useful ways. The premises of the project seem particularly interesting, and the public availability of a prototype system will in the future allow to evaluate the approach in more detail, and to experiment with the creation of the envisaged semantic mashups.

[1] `http://www.ica.org/en/node/30000/`

3.2 The Semantic Web in Archaeology

Semantic Web applications in the archaeological domain are still very limited in number if compared to the experiences carried out in the broader scenario of cultural heritage. Nonetheless, several potential benefits are generally acknowledged concerning the Semantic Web and archaeological research, for example for what concerns e-publication of data and documents (see e.g. Richards, 2006) and their accessibility by machines for automatic processing, and by humans for scholarly research.

Many factors negatively influence the development of new projects in this area. At a general level, several contributions agree on the fact that the scarcity of economic resources that traditionally characterizes archaeology hinders the investment in a promising but still not consolidated framework such as that of the Semantic Web. A few years ago Ross (2003, p. 10) stressed that "*...the heritage sector is likely to be left behind because the financial rewards for creating the mark-up necessary to make the Semantic Web a reality are only evident to the commercial sector.*". Today the attitude and possibilities of the cultural heritage community seems to have changed and a more dynamic situation is emerging, even if the gap with the commercial sector is still present; on the contrary, archaeology seems not to be following the cultural heritage trend and really runs the risk of being left behind.

Beyond this Isaksen (2008) underlines other factors that negatively influence the situation, such as the tendency of archaeologists to concentrate on recording and analysis rather than dissemination of results or even raw data. To this regard the attitude twoards protectionism on data sometimes represents a crucial problem, as previous personal experiences have stressed (Mantegari et al., 2006). Moreover, the lack of technical skills in the archaeological computing community, which is traditionally focused on relational databases and spatial data management (through GIS technologies) rather than on the (Semantic) Web and the related technologies is mentioned by Isaksen (2008) as a key factor.

Finally it has to be acknowledged that archaeological data are of a complex nature, and their inherent uncertainty makes the definition of models of representation an extremely difficult task. This characteristic shows its influence in a number of aspects ranging from the representation itself to the modalities through which data are combined, retrieved, and displayed. For example, standard models such as the CIDOC CRM have been defined mainly for the cultural heritage and museum communities, and they do not comprise specific aspects related to archaeological knowledge that would require the model's extension. On the other hand, the complexity of scientific inter-relationships between different resources makes the development of effective interfaces for accessing data a complex challenge.

As was proposed with reference to cultural heritage, the outline of the principal characteristics of existing Semantic Web applications for archaeology is considered a relevant basis for the discussion of the current scenario; it is provided in the following

sections.

3.2.1 Vbi Erat Lvpa

The Vbi Erat Lvpa Project[1] (Doerr et al., 2004b) represents one of the first experiences in the use of Semantic Web technologies for the integration and exploitation of archaeological data coming from heterogeneous sources, such as structured databases and corpora. The project has been sponsored by the European Union Program "Culture 2000 - multiannual cooperation agreements in the field of cultural heritage"[2], and is based on previous experiences in the creation of a local database (1994), and a test version of a Web site (2001).

The specific application domain regards archaeological finds carrying iconographic and epigraphic elements of the Roman era. The project deals with several data sources that present complementary information, which needs to be reconciled in an integrated repository in order to be cross-retrieved. The integration is made through the use of a domain model (the CIDOC CRM), and the creation of a mapping workflow, which consists of rules for formatting source data, and an algorithm for transforming these data according to the model. Crucial in this workflow are issues of data quality, that go beyond the inconsistencies e.g. in spelling that are often present in cultural heritage documentation, and concern the presence of multiple records without a common identifier for the same object. For example, the same object is present in two different databases with different identifiers, and the possibility to integrate the specific information contained in each single database requires the detection of an equivalence relation between the objects' identifiers. This problem, also referred to as the object identity problem, is naturally linked to the complementary nature of the data sources the project deals with. The fact that most identifiers do not match between resources, made it necessary to define a system for semi-automatic coreference detection and correction, which is based on data cleaning procedures.

From the technical point of view, a few details are described in the project's documentation. The central repository is based on RDF, and data transfer from satellite data sources to the central repository makes use of XML. An existent integration/migration tool, a suite for RDF management, and a semantic Web portal generator are combined in order to obtained the desired mapping and RDF storage functionalities (Doerr et al., 2004b, p. 5). The online portal offers an interface for accessing information in traditional ways, through simple hyperlinks, drop-down lists, etc.; this choice hinders the richness of the underlying information integration environment and makes the interaction poor if compared to the current possibilities offered by mashup technologies. Beyond the functionalities described in (Doerr et al., 2004b) which include a cross domain search service, and a service for the statistical evaluation of data, since 2002 the system has been developed towards a multifunctional

[1] http://www.ubi-erat-lupa.org/

[2] http://europa.eu/legislation_summaries/culture/l29006_en.htm

Web tool for various user groups (Forschungsgesellschaft Wiener Stadtarchäologie, 2005). It currently provides a set of online functionalities ranging from data input, access methods for educational purposes, and an administration system (L.O.D.A.S.) for small and medium scale Web database projects.

3.2.2 STAR

Semantic Technologies for Archaeological Resources (STAR) is a project undertaken by English Heritage and the Royal School of Library and Information Technology (Denmark), whose general aim is the application of Semantic Web technologies for the integration and exploitation of digital archive databases, vocabularies and the associated grey literature in archaeology. More specifically, the objectives of the project are[1]:

- Open up the grey literature to scholarly research by investigating the combination of linguistic and KOS-based methods in the digital archaeology domain.
- Develop new methods for enhancing linkages between digital archive database resources and to associated grey literature, exploiting the potential of a high level, core ontology.
- Apply multi-concept query generalization techniques to archaeology cross-domain research.
- Design and implement a demonstrator search system, in collaboration with English Heritage.
- Evaluate the demonstrator with a view to cost/benefit issues and application more widely in the archaeological domain.
- Engage with the archaeological community to inform research and to disseminate outcomes.

One of the most relevant activities in the STAR program has been the creation of an extension of the CIDOC CRM (v. 4.2[2]) for the description of archaeological datasets, with particular respect to the archaeological excavation. The extension (named "CRM-EH") was created in 2004 (Cripps et al., 2004) by the English Heritage Research and Standards Group in collaboration with CIDOC; it provides 125 sub-clasess and 4 properties, which are based on the concept of "context" and the events that modify it, such as the "context stuff" removal during an excavation[3].

[1] http://hypermedia.research.glam.ac.uk/kos/STAR/

[2] http://hypermedia.research.glam.ac.uk/media/files/documents/2008-04-01/CIDOC_v4.2_extensions_eh_.rdf

[3] A diagram of the model is available at http://cidoc.ics.forth.gr/docs/AppendixA_DiagramV9.pdf.

The development of specific tools easing the mapping of datasets to the CRM-EH ontology is currently being pursued. A mapping/extraction tool is in fact considered essential within the context of the project, and a bespoke utility application assisting users in the process of data mapping, cleansing and extraction has been created (Binding et al., 2008). This application allows the mapping of RDF entities to database columns, constructing structured SQL queries, cleansing data, and output the result to RDF files. A prototype search/browse client application allows to test cross-searching and exploring of integrated data extracted from the previously separated databases. The exploration makes use of Boolean full-text search operators, by which users can retrieve a set of results and use them as entry points to the structured data.

Moreover, a pilot set of Semantic Web services based upon SKOS providing archaeological thesauri are available[1], together with demonstrators such as a thesaurus client browser built on a previous Web interface (FACET project[2]).

The applications make use of data coming from several datasets and aggregated to RDF following the CRM-EH ontology. A case study on STAR has been published by the AHRC ICT Methods Network[3], which summarizes the approach and the results obtained until 2008, with particular respect to data extraction on three archaeological datasets.

The creation of the CRM-EH extension represents an important step forward for archaeologists, mostly because it broadens the scope of the CIDOC CRM beyond the community of museums, as it has been observed by Isaksen (2008, p. 8). On the other hand, the evaluation of the tools developed in the context of the STAR project is still restricted to the available demonstrators that provide limited capabilities; a full release of the datasources and the new versions of the software will possibly provide very interesting contributions in the field of Semantic Web applications for the archaeological research.

3.2.3 The Port Networks Project

The development of a Linked Data system supporting scholarly research on Roman ports in the Western Mediterranean is the focus of the "Port Networks" project. The project aims to allow domain experts to translate their data into a common structure; this will happen with the definition of a procedure and the related technology enabling archaeological data providers to (Isaksen et al., 2009, p. 1):

- develop a common conceptual structure (domain ontology) capable of reflecting a level of inquiry relevant at an inter-site scale
- cope with overlapping categorization systems

[1] http://hypermedia.research.glam.ac.uk/kos/terminology_services/
[2] http://www.comp.glam.ac.uk/~FACET/
[3] http://www.methodsnetwork.ac.uk/resources/casestudy13.html

- map local relational database schemas to the concepts represented in the domain ontology
- map locally used terminology with canonical (i.e. universal) identifiers
- export data to a centralized repository for use as a communal knowledgebase
- export data in a format suitable for local hosting in order to promote distributed data connectivity

The Portus Network represents the case study of a doctoral work on the transparent negotiation and sharing of local application terminologies, instances and ontologies (the TRANSLATION framework), which aims to develop "*...a process by which resource providers are able to publish their data to the Semantic Web that keeps its semantic origins as transparent as possible via an explicit ontology.*"(after Isaksen, 2008, p. 1)

Being in an initial phase of development, the project's overall results cannot be evaluated at the moment; however, some preliminary elements can be introduced.

From the point of view of the general architectural principles, the Portus Network proposes the combination of the data centralization/data distribution approaches that emerged in the last few years, through the development of a centralized vocabulary of canonical concept URIs and a centralized triplestore, coupled with widely distributed source data. The distributed nature of data and their fair stability (major updates of the datasets normally occur once a year, over the course of archaeological excavation seasons) suggested moreover to avoid implementing dynamic and real-time mapping, therefore improving the system's performances and stability.

The procedure and tools that have been developed in order to support domain experts in exporting proprietary data to the common ontology show interesting approaches and results. The Portus ontology (fig. 3.4) is based on a "classification" layer and an "instance data" layer. The former allows to link independent datasets through canonical URIs, which provide a vocabulary of concepts that may be common to any instance data; it also makes reuse of existing vocabularies, such as SKOS and HEML. Canonical URIs are managed by a classification service that allows to use SKOS predicates in order to define different kinds of matching between types in different schemata.

The mapping procedures of proprietary data to RDF statements are based on a "Data Inspector Wizard" that guides users in creating a connection to a data source, and in establishing ontology-to-database column matching. The result is an XML configuration file specific to the dataset, which can be modified any time using the wizard. This file is then used in conjunction with the database by a "Data Importer" tool that automatically generates RDF for the data source. The RDF becomes available to data providers, who can autonomously publish it in their Web sites, and it is also uploaded into the project's central triplestore in order to enhance performance, security and querying functionalities.

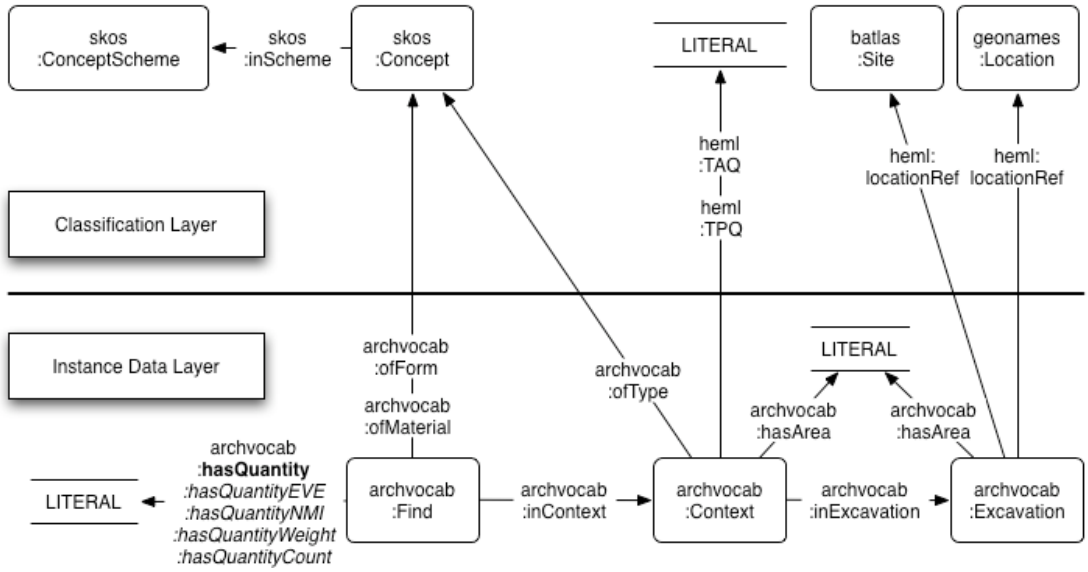


Figure 3.4: *The proposed Portus ontology (Isaksen et al., 2009, fig. 1, p. 3).*

Until now, the main results of the project are in the development of a mapping procedure and tools that have been tested successfully on domain experts, therefore easing the creation of ontology-compliant Linked Data. The evaluation of the impact this network of linked data will have on the study of Roman ports, and the possible generalization of procedures and tools for easily mapping proprietary data to a common ontological structure will be two of the most interesting results of the project in the long term.

3.2.4 Other Projects

Several other projects are being defined in the context of Semantic Web applications for archaeology; however, they are in an initial stage or have provided minor results to date if compared to the experiences described in the previous sections.

Archaeology Platform (@PL) (Eckkrammer et al., 2008) deals with the integration of excavation data using the CIDOC CRM and proposes the creation of a peer-to-peer server framework in order to support it. The project, carried out by the Vienna University of Applied Science, will release the code of the applications with an open-source license, but it cannot be further evaluated at the moment, due to the lack of a concrete implementation.

VERA (Virtual Environments for Research in Archaeology) from the University of Reading (UK)[1], is a project based, similarly to archaeology Platform, on the creation

[1] `http://vera.rdg.ac.uk/`

of a peer-to-peer server framework. The aim is to share data between the IADB (Integrated Archaeological Database) from York University and the Vindolanda database from Oxford University, using RDF technology. In addition, SOA techniques are investigated in order to exploit the IADB.

The **COINS** (Combat On-line Illegal Numismatic Sales) project (Hermon, 2007) is not directly related to the archaeological research needs, but it makes use of archaeological data on coins for creating a repository supporting the fight against the illegal trade and theft of coins. The project integrates data from Fitzwilliams and the Archaeological Superintendency of Rome collections through the mapping of their databases to the CIDOC CRM. A specific ontology has been developed, using the CIDOC CRM concepts and classes (Felicetti and D'Andrea, 2008a), as well as a multilingual thesaurus (Felicetti and D'Andrea, 2008c) which has been implemented in SKOS; the mapping templates are also available for public evaluation (Felicetti and D'Andrea, 2008b). The project's Web site[1] offers an online version of the management tool[2], as well as the possibility to download several tools developed in the course of the project[3].

The **Perseus-Arachne** project (Kummer, 2007; Babeu et al., 2007) deals with the syntactic and semantic integration of a digital library (Perseus) and a database of archaeological objects (Arachne) which mainly provide data and information concerning the Classical period. In order to do this the CIDOC CRM was used as a common model for metadata sharing, and issues related to mapping sources to this ontology as well as the identification of named entities have been faced; moreover, methods for indexing and navigating through the integrated content have been explored, using existing Semantic Web tools, such as the Longwell browser.

Other works include contributions on several topics such as the translation to RDF of existing data and the development of interfaces for easily performing querying (Byrne, 2006), the creation of ontologies for archaeology (Missikoff, 2004; Fernández González et al., 2007), the application of principles of edition philology to archaeological texts and archives (Holmen et al., 2004) as well as more general perspectives on Knowledge Management (McAuley and Carswell, 2008).

3.3 Around and Beyond the Semantic Web

The ideas, approaches and technologies of the Semantic Web show a transversal intersection with high-level research frameworks into which the semantic integration of heterogeneous information coming from multiple sources represents a central concern. Today terms such as "digital libraries", "e-Science" and "cyberinfrastructure" represent more than simple buzzwords, even if it is not always simple to define their

[1] http://www.coins-project.eu/

[2] http://www.coins-project.eu/COINS-MT/bin-debug/main.html#

[3] http://www.coins-project.eu/index.php?Itemid=53&option=com_content

boundaries and their relationships with the elements introduced in the previous sections.

Nevertheless, our review cannot disregard their relevance and their potential impact in the scenario this work takes into consideration. An introduction to these frameworks will be provided in the next sections in order to broaden our perspective towards the elements that are emerging around and beyond the Semantic Web, cultural heritage, and archaeology.

3.3.1 Digital Libraries

The term "**digital library**" (DL) is generally used in order to denote different initiatives undertaken by institutions hosting cultural heritage material with the aim of offering useful digital data and services to their audience. The origins of digital libraries are quite recent, and their development has been crucially stimulated during the last few years by the funding opportunities connected to the activities of the European Union (Unit of Technology Enhanced Learning: Cultural Heritage), and the National Science Foundation and other agencies in the USA (in the context of the "Digital Libraries Initiatives").

More precisely, the term was introduced in the early 1990s (Fox et al., 1995; Borgman, 1999), and the concept it represents has been the object of a number of competing visions concerning its scopes and functionalities (an overview is provided by Candela et al., 2007, pp. 14–16). A recent "digital library manifesto" states that a DL is "*an organization, which might be virtual, that comprehensively collects, manages and preserves for the long term rich digital content, and offers to its user communities specialized functionality on that content, of measurable quality and according to codified policies.*"(Candela et al., 2007, p. 17). Therefore, a digital library cannot be considered a single application or system, but part of a universe, where several components, concepts and actors are involved. Regarding this, the manifesto proposes a three-tier conceptual framework (fig. 3.5) where the digital library is the higher level of abstraction, on top of the technological components providing the required functionalities (DL System - DLS), and the methods and tools for the creation, management and extension of the DLS (DL Management System).

In fact, specific software systems and repositories for digital libraries are available today, such as Fedora[1] and DSpace[2] (which are currently joining their efforts in the DuraSpace initiative[3]), as well as Eprints[4], dLibra[5], and several others. The creation of a digital library can therefore rely on these solutions, some of which are incorporating Semantic Web technologies, such as Fedora with a SPARQL support.

[1] `http://www.fedora-commons.org/`
[2] `http://www.dspace.org/`
[3] `http://duraspace.org/`
[4] `http://www.eprints.org/`
[5] `http://dlibra.psnc.pl/`

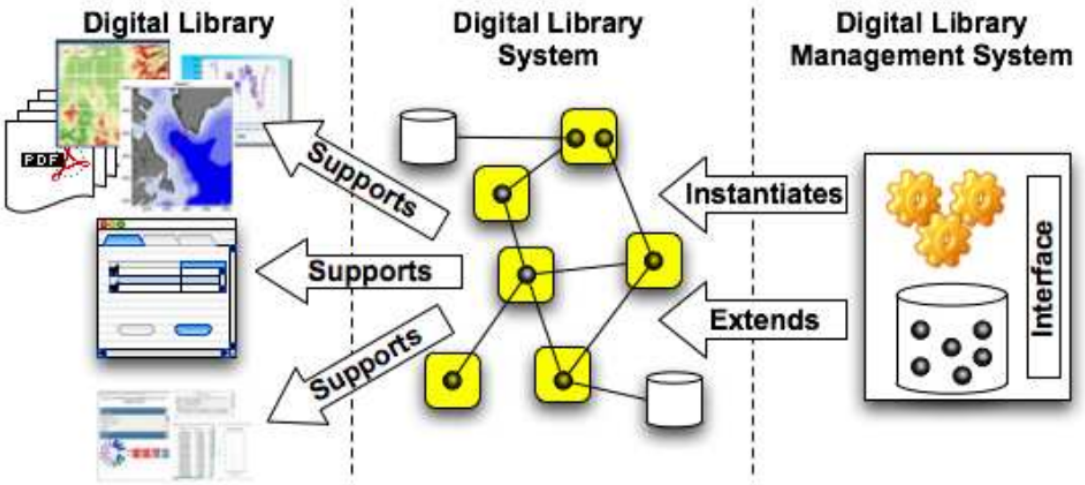


Figure 3.5: *A tree-tier framework for the Digital Library Universe (Candela et al., 2007, fig. 1.2, p. 17).*

In fact, digital libraries have often been considered as specialized search engine portals built around metadata collections, basically offering search and access capabilities on digital material, and the available software offer ready-to-use solutions for these requirements. However, if these can be considered fundamental characteristics of DLs, the search and access paradigm is nowadays considered limiting with respect to the potential and to the advantages digital libraries can show. For example, (Lagoze et al., 2005) points out that since traditional libraries cannot be considered only well organized warehouses of books, so digital libraries must offer much more than searchable collections of digital contents. The specific nature of new services takes different forms in the context of different projects, ranging from improved searching capabilities across federated repositories to the possibilities of users contributing knowledge to the library through annotations, reviews, and the like.

From a more general point of view, new information models for digital libraries are envisaged today "*...expressing the expanding web of inter-relationships and layers of knowledge that extend among selected primary resources.*" (Lagoze et al., 2005). Most recent contributions in the field of digital libraries agree on this perspective, and they stress the importance of creating networks where semantic links between units of information are established and exploited. While traditional libraries deal with documents as the minimal unit of information, digital libraries should focus on contents, in the vein of the "Web of data" vision introduced in section 2.2. However, if the interoperability between resource descriptions in different systems and searches on metadata repositories are achievable today (thanks moreover to the widespread acceptance of metadata

standards such as the Dublin Core[1] and the OAI-Protocol for Metadata Harvesting[2]), semantic integration mostly remains an unmet, even if exciting promise.

The relationship between the new vision of digital libraries and research on the Semantic Web is evident today, to the point that specific scientific events are dedicated to these two converging worlds[3]. Recent contributions in this field, such as Soergel (2009) propose the creation of unified systems, where standards from the DLs and the Semantic Web are harmonized; others envisage the development of "Semantic Digital Libraries", which can be exploited and can integrate a number of existing approaches and technologies (Kruk, 2009).

Even if in most cases they do not directly fit into the digital library framework, the documentation of the projects that have been introduced in the previous sections makes references to it. From another point of view these projects show significant point of contact with the semantic digital library proposals, especially when Semantic Web technologies are used to integrate and retrieve data from semi-structured or unstructured documents (e.g. in part of the "Perseus-Arachne" project).

Finally, it has to be noticed that big scale projects for the creation of digital libraries dealing with cultural heritage material are under development since a few years ago; however, they have not been introduced in the previous section because their adoption of Semantic Web approaches and technologies are in most cases absent or in a preliminary stage. Among the most relevant initiatives are BRICKS[4], DELOS[5], and "Europeana"[6]. The latter, in particular, is an extremely big and ambitious project of the European Union that provides an increasing multilingual repository of images, sounds and texts. To date 4.6 million digital objects, coming from a vast body of museums, galleries, archives libraries and audiovisual collections in Europe, including major Institutions such as the British Library and the Louvre, are present. Europeana is moreover experimenting semantic search modalities on the repository, but the results are still preliminary; however, it has to be noticed that semantic search engine is based on software developed in the context of the previously discussed MultimediaN E-Culture project.

3.3.2 E-Science and Cyberinfrastructure

Semantic Web and digital library applications are increasingly considered as fundamental parts of programs dealing with the development of innovative forms of support to scientific and scholarly research. New visions and trends emerged in the last few years around the terms of "**e-Science**", "**e-Research**", and "**e-Scolarship**", which,

[1] http://dublincore.org/
[2] http://www.openarchives.org/pmh/
[3] For example: http://www.icsd-conference.org/
[4] http://www.brickscommunity.org/
[5] http://www.delos.info/
[6] http://www.europeana.eu/

from a general point of view, denote scenarios of collaboration between scientists on a global scale that rely on digital scientific publishing, and the possibilities of cross-retrieving and analyzing primary data over a network. This vision is in reality rooted in the very origin of the World Wide Web, which was conceived as a technology for sharing scientific documentation in electronic formats over a network, and the proposals of a Semantic Web and a Web of data can be seen as further advancements in this direction. The relationship between the Semantic Web and e-Science has been analyzed in different contributions, such as Goble et al. (2006), who analyze the potential benefits of a closer collaboration between the two communities, supporting a scenario of dramatic reduction in the time needed to create new scientific results.

With respect to the creation of digital scholarship in the Humanities and Social Sciences, specific and extensive proposals and research frameworks have been recently introduced. ACLS (2006, p. 7) provides for example a list of different visions of digital scholarship that are identifiable in recent practice:

1. building a digital collection of information for further study and analysis
2. creating appropriate tools for collection building
3. creating appropriate tools for the analysis and study of collections
4. using digital collections and analytical tools to generate new intellectual products
5. creating authoring tools for these new intellectual products, either in traditional form or in digital form

While collection and tool-building is obviously a necessary and enabling condition for defining innovative scholarly services, the generation of new intellectual products is acknowledged as the ultimate objective of digital scholarship. Some contributions in the cultural heritage domain are embracing this perspective; Kummer (2007, p. 10) for example, when discussing the "Perseus-Arachne" vision of digital scholarship states that "*A new form of work environment could support scientists by offering a tool that supports a complex workflow starting from targeted information search to compiling and arranging thoughts and ideas to argumentation chains and online publishing.*". Some early contributions in archaeology show how the discipline is interested in these perspectives: for example, the idea of using the Web and the hypertext principle in order to express the structure of archaeological theories has been investigated by Gardin and Roux (2004) as the most recent development of Gardin's logicist program launched in the late 1970s in France. Moreover, Bogdanovic et al. (2004) proposed a scenario where archaeological knowledge building is made in a collective, interactive and distributed fashion over the Internet.

With reference to more recent projects, such as those introduced in the previous sections, elements that go in the direction of e-Scolarship/e-Science can be easily

found, even if Perseus-Arachne is the only project that explicitly situates its activity in an e-Scholarship framework. For example, the analysis of the STAR project as a case study is made in the context of the AHRC activities for e-Science[1]; from another point of view, projects such as "Port Networks" have been designed in order to support scientific research.

In the context of electronic publishing, Richards (2006) stresses how the distinction between publication and online archives has become blurred, and how the adoption of a Semantic Web approach would greatly improve the support to research through e.g. access to relevant and updated data and information. Moreover, the e-journal "Internet Archaeology"[2] is experimenting the possibility of making "*...the underlying data available via a digital archive in such a way so that readers are provided with the opportunity to "drill down" seamlessly from the publication into the archive to test interpretations and develop their own conclusions*[3].

Many researches in the context of e-Science, e-Research, e-Scolarship, and the like are designed and carried out under a "**cyberinfrastructure**" framework, which has been proposed by the US National Science Foundation (Atkins et al., 2003), and is generally indicated in Europe with the synonym **e-Infrastructure**[4]. The term basically refers to the possibility of combining middleware services and global high-speed research networks in order to allow scientists to set up secure, controlled environments for collaborative sharing of distributed resources for their research (Hey and Trefethen, 2005, p. 818). From a technological perspective, the creation of a cyberinfrastructure goes hand in hand with developments in the areas of service-oriented architectures and distributed/grid computing (Foster, 2005). In particular, the combination of services and the computing capabilities of machines participating in the infrastructure would offer several innovative opportunities to scientific research, such as the possibility of obtaining huge storage and computing power[5], or to work collaboratively and in real time on the same research materials and data from remote locations on the globe.

Atkins et al. (2003, p. 13) have proposed a general schema of the facilities and services that can be provided by a cyberinfrastructure layer in an integrated way (fig. 3.6).

In the last few years, proposals going in the direction of a cyberinfrastructure have been elaborated for the cultural heritage and the archaeological research (e.g. Kinthigh, 2006). Barceló et al. (2004) even if not mentioning directly a cyberinfrastructure, introduce the idea of "tele-archaeology", which is conceived as the use of telecommunications to provide archaeological information and services. More specifically, the proposal is based on a telescience perspective, in which scientists collabo-

[1] `http://www.methodsnetwork.ac.uk/escience/e-science.html`
[2] `http://intarch.ac.uk/`
[3] `http://intarch.ac.uk/leap/`
[4] `http://cordis.europa.eu/fp7/ict/e-infrastructure/`
[5] See, for example, the "Globus Alliance" association: `http://www.globus.org`.

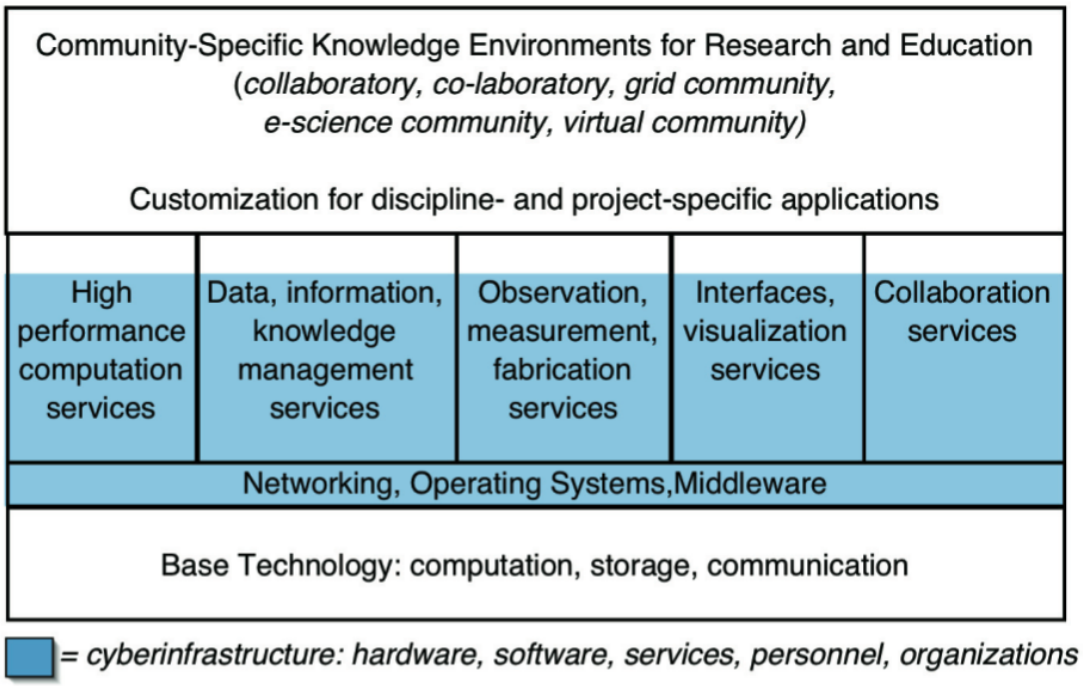


Figure 3.6: *A general schema of integrated cyberinfrastructure services (Atkins et al., 2003, fig. 2.1, p. 13).*

rating from different locations create, manipulate and aggregate knowledge and do archaeology "at a distance" (for example between different research centers and different archaeological sites). Snow et al. (2006) have recently designed a general logical model for an archaeological cyberinfrastructure, which integrates digital library middleware, document and image search capabilities, GIS analytical kits, visualization tools, and content management (fig. 3.7). The motivation for this proposal is the need for archaeologists to find meaningful links between different archaeological items, on the basis of the possibility to access distributed data sources. It is evident how the ideas and technologies of the Semantic Web should play a crucial role in this scenario.

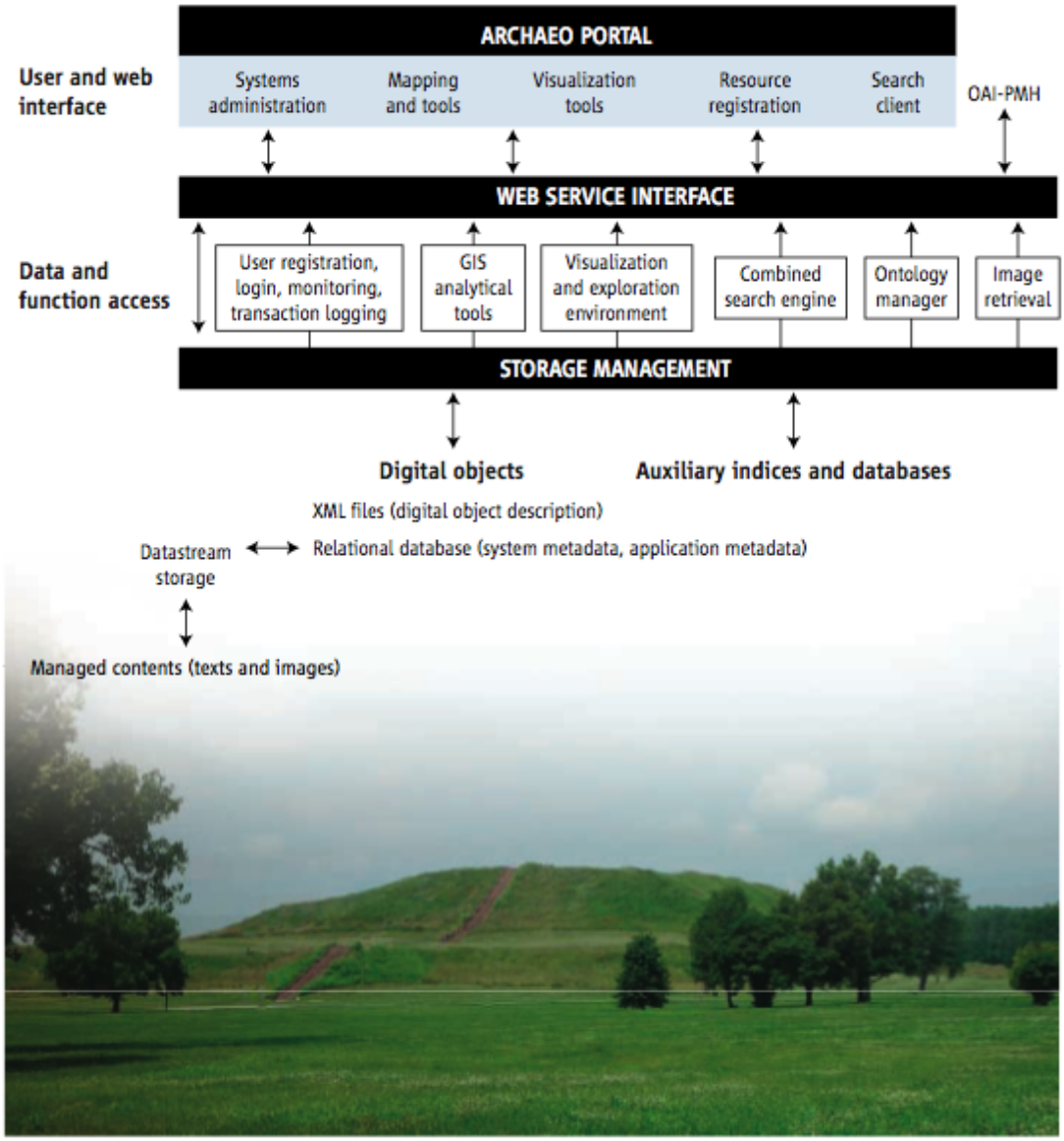


Figure 3.7: *A model of a cyberinfrastructure for archaeology (Snow et al., 2006, p. 958).*

4

From Representation to Fruition: An Interdisciplinary Research Perspective

The overviews and discussions of the previous chapters highlight the necessity of interdisciplinary research coupling the more technological elements with the analysis of crucial and specific issues of the domain under exam. This chapter introduces a transversal and interdisciplinary research proposal (section 4.1) that is characterized by a twofold perspective.
On the one hand, the fundamental aspect of representation of domain data, information and knowledge is analyzed with reference to the ISO standard and core domain ontology CIDOC CRM (section 4.2). In particular, the identification of the most relevant issues that are connected to its usage in real application scenarios, the definition of a workflow for mapping heterogeneous metadata schemata to the CRM, and the evaluation of the available serialization of the model in Semantic Web languages (mainly RDFS and OWL) are set as the methodological key points for novel experimentation.
On the other hand, a more "vertical" contribution is proposed, with the analysis of the modeling and retrieval of fuzzy temporal intervals (section 4.3). In particular, this analysis takes into consideration archaeological chronologies, and it discusses the related work mainly with reference to temporal representation in the CIDOC CRM. Thereafter, a new approach and a model based on the fuzzy set theory are introduced, together with methods for reasoning on the fuzzy temporal intervals that can be used for improving information retrieval in Semantic Web systems for archaeology and cultural heritage.

4.1 Initial Remarks

The review provided in the previous chapter highlights the potentialities and the complexity of the current scenario of Semantic Web applications for cultural heritage, together with the related frameworks of digital libraries, e-Science and cyberinfrastructure. At the same time, it emphasizes the heterogeneity of the research and application approaches, as well as the general lack of consolidated solutions and best practices to "do the Semantic Web" in cultural heritage: in fact, most of the contributions (with significant exceptions that are slowly emerging) did not cross the boundaries of the single projects. On the other hand, high level proposals, such as that of cyberinfrastructures, while of great interest also require massive resources and large scale agreements between different Institutions in order to be successfully deployed, and therefore they run the risk of remaining unaccomplished visions (at least in the short to medium term).

A number of open research challenges are present today, both for what concerns the evaluation of the few general results obtained in new contexts and the definition and experimentation of novel approaches.

The design of a research perspective in the context of the Information Society should couple the technological elements that characterize the scenario with the analysis of crucial and specific aspects and issues of the domain under exam.

Figure 4.1 represents a basic schema on which this analysis should be based. First of all, it is important to stress that the notion of "cultural heritage" in itself is very complex to define. In fact, cultural heritage can be intended as a qualification given to existing things with respect to their significance according to the cultural elements of a particular social group. Therefore, cultural heritage is a social construct rather than something objective and fixed; consequently, it can assume different forms and characteristics in different contexts, including different typologies that range from material items (e.g. monuments, paintings, archaeological artifacts, etc.) to immaterial ones (e.g. oral narrations, music, poetry, etc.).

However, according to the perspectives of this research, no further investigations on the concept will be provided here, and a general and commonsense notion is assumed. With respect to the scenario we are investigating, a crucial path is the one starting from the models that are used for representing data, information and knowledge connected to cultural heritage and the activities performed on it, to the fruition of these data, information and knowledge by different audiences (which are on a transversal axis from professionals to the general public). This, in turn, means that most of our fruition of the cultural heritage coincides in fact with the fruition of its representations, in terms of data, information and knowledge.

Naturally, the development of models of representation and methods of fruition requires domain-expert work. If, moreover, representations need to be "understood" and processed by machines, and fruition should be supported by ICTs, an interdisciplinary perspective is needed, coupling domain knowledge with technology-related knowledge. The path from representation to fruition can be considered as a simple yet powerful reference for the design of relevant interdisciplinary research in the context of Semantic Web applications for cultural heritage, according to the interdisciplinary analysis perspectives connected to studies on the Information Society.

The path from representation to fruition represents a *fil rouge* linking most part of the projects that have been discussed in the previous chapter, regardless of the specific perspectives and results obtained within each single project. In his recent position paper, Hyvönen (2009) provides a detailed proposal concerning the identification of the components of semantic portals for cultural heritage, which further confirms the validity of this perspective. According to this proposal, three principal components can be identified.

The **content model** is directly related to the metadata, the ontology, and the logic levels of the Semantic Web pile (fig. 2.13). Its importance is crucial, since the design

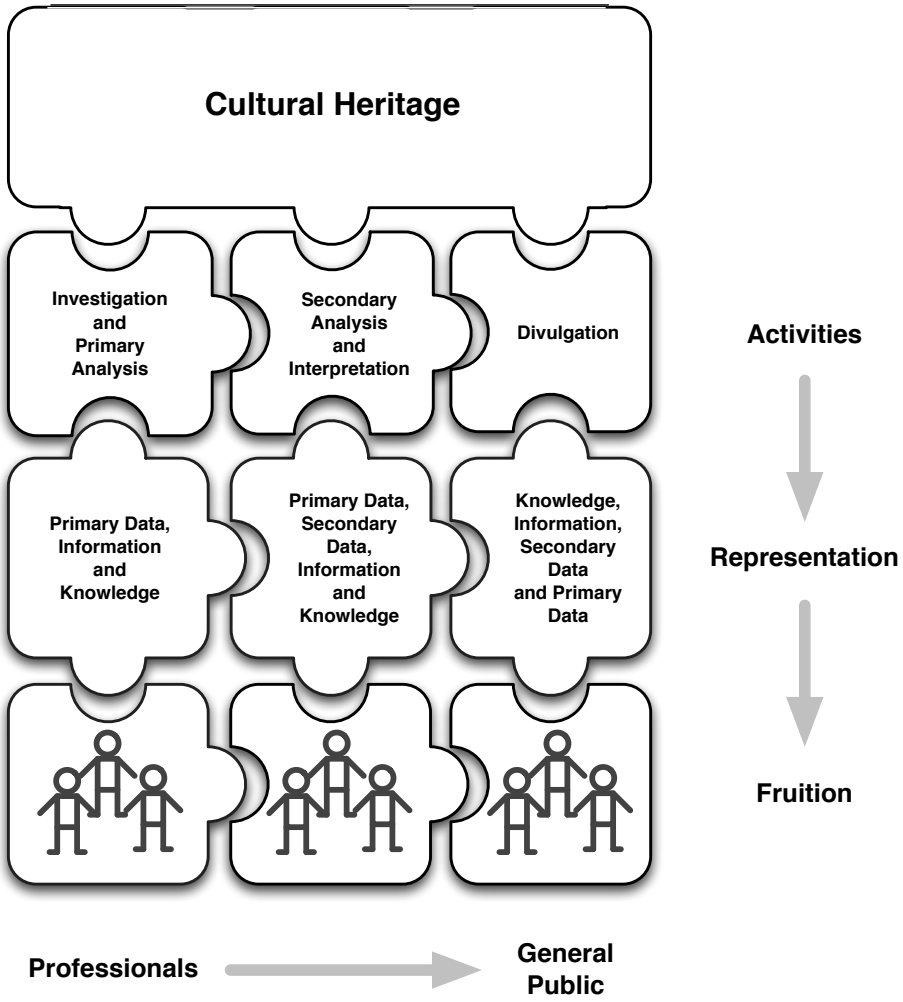


Figure 4.1: *Cultural heritage: from representation to fruition.*

and deployment choices made when designing this component directly influence the possibility to develop and enable innovative functionalities, such as the integration with external systems and the creation of advanced methods of information retrieval based e.g. on semantic relations.

The metadata and the ontology levels are the ones from which important results of general value have been obtained, from the development of schemata based on the Dublin Core standard (such as VRA) to the definition of domain ontologies facilitating semantic interoperability, such as the CIDOC CRM. On the contrary, the logical level has been explored less and rules for deriving new facts and knowledge from repositories are generally absent, with few notable exceptions such as the "Culture Sampo" project.

The methods and tools for both the creation and the harvesting of contents (ontologies, terminologies and semantic annotations) constitute the **content creation system**. Here the analyzed experiences show a high heterogeneity, which is directly related to the specific contexts into which the single systems are created.

On the one hand, content harvesting from legacy data sources, which usually come in the form of relational databases, is predominant. Different approaches for mapping legacy data to common metadata or ontological structures have been experimented. One crucial concern seems to be the creation of assisted procedures relying on specific software, easing the creation of these mappings for domain experts who generally have limited technical background knowledge (see e.g. Isaksen et al., 2009).

On the other hand, methods and tools for the creation of new content, as well as for enabling semantic annotations have been proposed and tested, together with procedures for e.g. ontology alignment. To this regard, the "Culture Sampo" experiences stand out, with specific applications having been developed and released under open source licenses.

Semantic Services are built on top of contents in order to support users in "*...finding and learning the right information based on her conceptual view to culture and the context of using the system.*" (Hyvönen, 2009, p. 11). The range of possible services is vast, covering potentially different audiences (from general users to specialists) with different needs (from unified access to general and distributed information to sophisticated information retrieval). In particular, the following macro-categories are identified:

- Semantic search aims to improve the traditional forms of information retrieval (in terms of precision and recall) by finding the concepts related to the documents at the metadata and ontology levels, instead of finding occurrences of (key) words in documents. A key aspect of semantic search is mapping the literal search words of users with the ontological concepts used by machines. To this respect, however, a potential problem is connected to queries ending with no results. In order to avoid this situation, the faceted paradigm can be used, providing interdependent and multiple categories, whose selection guide in querying controlled terms that are related to the actual contents present in the

repository. This approach can be extended towards semantically-enabled facets, which can be dynamically projected from a system's underlying ontologies in the user interface.

- Semantic autocompletion can be coupled to faceted search in order to provide a more keyword-style search, instead of facets with predefined categories. In this case, the system tries to guess the concepts the user is searching for in real time, i.e. while the user is typing, by using the ontologies and reasoning on them. Semantic autocompletion is particularly useful for finding meaningful keywords in large search vocabularies.

- Semantic browsing tries to exploit the linked-data approach, by enabling browsing on hyperlinks that relate different RDF resources. Moreover, logical rules can be defined in order to exploit semantic relationships for providing recommendations of contents that are related to the content the user is actually visualizing.

- Relational search makes it possible to detect and explore serendipitous associations between instances, following paths of semantic connections. A particularly effective implementation of this idea is the Culture Sampo application, which given two personal names of artists from the ULAN vocabulary, is able to provide the chain of relationships through which these artists are linked, exploiting semantic relationships such as "teacherOf", "knows" and "studentOf".

- Personalization and context awareness can be enabled by exploiting user profiles and the context of usage in order to make systems adapt to different personal information needs and interests. In addition, the detection of location information (in terms of space and time) can be used in order to provide context sensitive services on mobile devices.

- Interactive visualization tools and mashups are increasingly integrated in new applications in order to develop highly interactive modalities of access to information. The most typical approaches in cultural heritage make use of interactive map services, but more sophisticated and ready to use frameworks for data rich interactive Web interfaces (such as Simile Exhibit[1]) are also experimented.

- Cross-portal reuse of content exploits open Semantic Web standards in order to integrate existing contents into a new portal. This integration can be implemented in two main forms. On the one hand, different triple stores can be merged, and services for end-users can be created on the extended knowledge base. On the other hand, the triple stores can be kept separate and Web services (based e.g. on REST) can be created in order to expose contents for re-use in different semantic portals.

[1] http://www.simile-widgets.org/exhibit/

These elements apparently involve the use of a potentially high number of heterogeneous technologies. However, the principal problems in developing semantic portals for cultural heritage are not connected to the basic enabling technologies, which are in most part available today. In fact, since its inception, the Semantic Web relies on a reasonably well defined development roadmap (Berners-Lee, 1998), and its high-level technological elements (see section 2.4.4) are almost at a sufficiently mature stage of development, especially for what concerns the languages for (meta)data interchange, and for representing and querying ontologies. In addition, the body of software tools and services for managing semantic repositories and publishing their contents on the Web is ever increasing; effective solutions, even free of license costs, are available, thus reducing the required technical and financial resources. From this perspective, the main difficulty still resides in the possibility of combining different technologies into new systems, and in the need to develop *ad hoc* solutions for all the situations where a specific application is not available. However, this situation is rapidly changing, thanks to the definition and the adoption of technological standards that allow to simplify the integration of different applications, as the approaches introduced in chapter 2 and the outcomes of the projects described chapter 3 demonstrate. Moreover, specific tools are available, and others are under development and will be likely available in the short-to-medium term.

This of course does not imply that technological and technical problems are absent or solved. Nevertheless, a true interdisciplinary research, where domain knowledge guides the design of technological systems is deemed fundamental, in order to provide a significant contribution to the scenario of the Semantic Web and cultural heritage. If we read Hyvönen's proposal from this point of view, it is easy to see that every single point it introduces should be necessarily based on this principle, for example when defining the criteria for semantic browsing and search and evaluating the relevance of the results they provide.

4.1.1 Defining a Research Approach

It is evident that the design of interdisciplinary research according to the transversal perspective from representation to fruition requires the identification of the key principles and facing of crucial problems, coupling domain analysis with the evaluation of the more technological possibilities and constraints.

With reference to the discussion provided so far (and in particular to the aforementioned lack of consolidated solutions) a central aspect is represented by the **evaluation of the general results obtained to date** by their application in new contexts. In fact, the analysis of the costs and the benefits that are connected to the adoption of standard solutions represents a necessary condition for obtaining an objective view over the actual sustainability of Semantic Web approaches for cultural heritage. Taking into consideration the domain problems (sections 3.1 and 3.2), which slow down or even hinder the adoption of the new and promising approaches, the results of this activity is

intended to provide an original contribution. The value of this evaluation is moreover relevant to the broader scenario of the Information Society, since a detailed discussion of the issues that are related to the standard approaches gives light to the concrete reasons why the new technologies still show limited impact on the actual practices of the cultural heritage communities.

On the other hand, **domain specific aspects** should be identified, and novel proposals and experimentations in light of the Semantic Web framework need to be designed and experimented on. This more "vertical" and in-depth research perspective aims to complete a higher-level and general evaluation with a specific contribution which can constitute the basis for more extensive domain-related work.

The identification of a **case study**, where the more theoretical aspects can be practically applied would greatly support these research perspectives. Most of the projects introduced in chapter 3 demonstrate that in the path from the theoretical level to the application one, several issues may emerge, both practical and methodological ones. Beyond the presence of technical constraints that make it difficult in some cases to actually respect the theoretical approaches in their entirety, these issues also contribute to highlight potential theoretical problems that can emerge only when the theory is actually applied to real world scenarios.

From this point of view, even if the scope of this research is only marginally concerned with technical aspects, the **actual development of a prototype semantic portal** is considered relevant. This is not only a natural aspect connected to the investigation of a case study in the Semantic Web scenario, but because it would also allow for the evaluation of the feasibility of the new perspectives of Web system development, such as those introduced in chapter 2.

According to these aspects and principles, the specific research perspective adopted here is defined by the following building blocks:

- The representation of data, information and knowledge represents a fundamental aspect in the context of Semantic Web applications for cultural heritage. In fact, this area is the one in which most research has been done and general results have been achieved. In particular, the CIDOC Conceptual Reference Model (CRM) stands out as the most relevant standard, and it has been increasingly adopted for e.g. the semantic integration of heterogeneous metadata schemata. At the same time, the complexity of the CRM and the issues connected to its actual adoption in Semantic Web-related projects represent an ideal case for novel experimentation and evaluation according to the perspectives introduced at the beginning of this section. Therefore, a specific investigation of the model and the related issues is of primary relevance, and will be provided in section 4.2 with respect to:
  - the characteristics of the CRM
  - the issues connected to its use in real application scenarios

  - the definition of a suitable mapping workflow for the integration of heterogeneous legacy metadata schemata using the CRM
  - the evaluation of the currently available serializations of the model in Semantic Web related languages

- Since time represents a fundamental dimension for the description, study and understanding of cultural heritage, it should also constitute a central aspect of Semantic Web applications in this domain. More specifically, fuzzy temporal intervals and fuzzy chronologies are of primary relevance (especially in the context of archaeology). Therefore, this aspect represents an ideal domain aspect which is investigated in section 4.3 according to:

  - the analysis of the specific characteristics of chronologies in archaeology
  - the evaluation of the existing models for representing fuzzy temporal information, with particular respect to rich and standard ones, such as the CIDOC CRM
  - the proposal of an original approach for modeling and retrieval fuzzy chronologies, based on the fuzzy sets theory

- A case study where the actual use of the CIDOC CRM is analyzed with reference to the integration of heterogeneous data sources is provided in chapter 5. The case study concerns the archaeology of the city of Milan (Italy), and it offers an ideal context for the detailed evaluation and discussion of the elements introduced in section 4.2.

- The same case study is particularly interesting for the application of the proposals concerning the modeling and retrieval of fuzzy chronologies proposed in section 4.3. Chapter 6 offers the analysis and discussion of these proposals, together with an evaluation of the retrieval according to domain expert evaluation.

- The development of a demo portal according to the perspectives defined in chapters 2 and 3 should be made on the identified case study. Chapter 7 offers a discussion of the demo development experience with reference to the creation of a semantic backend and dynamic mashup-style interfaces on top of it.

4.2 From Representation . . .

The creation of models of data, information and knowledge representation constitutes a traditional concern of cultural heritage documentation: classification systems, list of terms, catalogs, thesauri, and more generally schemata of domain concepts and their relationships have been created and used long before the introduction of computers

in the sector; better, they have often represented crucial aspects in the theoretical debate. Therefore, it is natural that in the context of Semantic Web applications for cultural heritage, the implementation of these resources in machine processable formats constitutes a key area of research and experimentation. On the other hand, the possibilities connected to the creation of formalized frameworks and domain models of knowledge using knowledge representation languages (such as the ones connected to the Semantic Web) offer the possibility to improve research on the documentation and analysis methods of cultural heritage. The resulting scenario covers a wide spectrum, both for what concerns the scopes and approaches and the obtained results.

For example, taxonomies and **thesauri** have been defined in order improve and homogenize the classification processes, and to perform information retrieval based on e.g. query expansion on narrower/broader term relationships, synonymity, etc. Basic hierarchical structures are at the basis of e.g. Iconclass[1], which is a taxonomy of art and iconography related terms. Recent and more sophisticated approaches are instead based on richer structures, which make it possible to define composite sets of relationships between concepts and terms, which can assume the form of lightweight ontologies rather than simple thesauri. Transformations of existing cultural thesauri into the SKOS format has been done in many cases; however, Hyvönen (2009, pp. 5–6) highlights the number of potential issues connected to this process, and particularly the possible semantic ambiguities in the translation of Broader Term (BT) relationships. Different approaches combining simple taxonomies with richer thesauri structures have been experimented on, as the projects introduced in the previous chapter show. At the same time, general and *de facto* standard thesauri have been used (such as Getty's Art and Architecture Thesaurus), together with specific and customized ones (such as the SKOS versions of the English Heritage thesauri used in the STAR project; see May, 2009).

Metadata schemata have also been created and aligned in order to facilitate the syntactic integration and cross retrieval of distributed contents, for example on the basis of equivalences between the elements contained in different metadata sets. On the one hand, different local or national schemata have been developed (such as the ICCD standards for the Italian cultural heritage cataloging[2]) and they have been implemented in digital systems. On the other hand, high-level standards that have been elaborated mostly in the context of archival and digital libraries initiatives have been used. Among these are, for example, Dublin Core, TEI[3], as well as more specific ones, such as VRA (which specializes the Dublin Core set for the description of artwork).

However, as the vision of the Semantic Web began to diffuse widely, research work increasingly embraced the development and use of more complex and expressive models of domain knowledge, such as **ontologies**.

[1] http://www.iconclass.nl/

[2] http://www.iccd.beniculturali.it/

[3] http://www.tei-c.org/index.xml

To this respect, Doerr and Iorizzo (2008) consider the Semantic Web as the most recent and relevant step towards the creation of a network where knowledge will be integrated, managed, shared, and exploited on a global scale. The creation of such a "global knowledge network" represents an old dream that has permeated decades of research in key areas of computer science, as well as in interdisciplinary projects. Starting from Artificial Intelligence research in the 1960s and 1970s, huge efforts have been dedicated to this topic, and particularly to the problem of integrating and exploiting heterogeneous information in meaningful ways. The achievements in the field of semantic networks and database integration have been crucial, as well as experiences towards common data schemata standardization and, more recently, explicit knowledge representation models of domain discourses. In particular, research in knowledge representation, such as the investigation of ontological structures, have given important contributions to the development of the Semantic Web vision, through e.g. advancements in the methods for merging ontological structures with information systems.

On the other hand, in his position paper Ross (2003) stressed the role of ontologies (the "jewels of the Semantic Web"), while at the same time highlighting the lack of suitable models for the cultural heritage sector. A short list of the key issues provide a good idea of the scenario that was present at that time. Since 2003, this scenario has changed a lot, as the projects introduced in the previous chapter demonstrate.

The development of domain ontologies easing the semantic integration of heterogeneous and specific resources (e.g. in "Port Networks") co-exist with experiences making use of several integrated ontologies (e.g. in "Culture Sampo"). The use of foundational ontologies (such as DOLCE[1]) and top ontologies (such as IEEE-SUO[2]) has also been experimented on. In addition, relevant contributions investigating both general and specific aspects of knowledge representation, ontologies and the Semantic Web in cultural heritage are present beyond the reference provided so far, such as Zhang et al. (2004); Missikoff (2004); Signore et al. (2005), as well as others that are constantly appearing in the context of International Conferences[3].

However, the most notable result in this research area, as well as one of the few general results in the context of Semantic Web applications for cultural heritage, has been the development and standardization of the CIDOC CRM core ontology, which represents today a fundamental reference, which is however still open to research challenges.

[1] http://www.loa-cnr.it/DOLCE.html

[2] http://suo.ieee.org/

[3] See for example "Museums & the Web" (http://www.archimuse.com/conferences/mw.html), and "Computer Applications and Quantitative Methods in Archaeology" (http://www.leidenuniv.nl/caa/).

4.2.1 Core Domain Ontologies: The CIDOC CRM Approach

The CIDOC Conceptual Reference Model[1] is a formal ontology for cultural heritage documentation that was created by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). Initially based on a entity-relationship model, it has been developed with the object-oriented methodology since 1996; the latest version (5.0.1, March 2009) consists of 90 classes and 148 properties, which are defined in Crofts et al. (2009). The work on the standardization of the CRM culminated in an important result in December 2006, when the CRM was accepted as the ISO 21127 standard[2].

The model is organized around an event-centric structure. Event-modeling is considered an essential part of the complex knowledge required for historical and cultural documentation, and a major aspect for cultural-historical analysis; therefore, the model assumes that an event-centric perspective could provide a more accurate view of the past or current life history of a cultural object (see Doerr and Kritsotaki, 2006). This choice makes the CIDOC approach substantially different from the more traditional cultural heritage documentation, which is usually object-centric, i.e. it focuses on the description of the features of the single objects. In the CIDOC CRM, instead, the basic idea is that the historical and cultural contexts can be represented by things, people and ideas meeting in space and time: according to this view, cultural objects are documented with reference to the meeting(s) they where present at (for details, see section 4.3.1.1 and Doerr et al., 2004b).

However, it is important to notice that the event-centric perspective does not hinder the possibility of representing a number of crucial and common aspects characterizing cultural heritage documentation. For example, the model provides classes and relationships for[3]:

- acquisition, collection, institution, deaccession and disposal information
- attribute assignment, description, appellation information
- object name and classification information
- object production information, material and technique information
- measurement information
- image information, objects and carriers
- mark and inscription information

[1] http://cidoc.ics.forth.gr/

[2] http://www.iso.org/iso/catalogue_detail.htm?csnumber=34424

[3] A more extensive list organized around functional units and their graphical representations can be found at http://cidoc.ics.forth.gr/cidoc_graphical_representation_v_5_1/graphical_representaion_5_0_1.html.

- location information
- spatio-temporal relationships
- part and component information, taxonomic discourse

Thanks to its rich structure, the CRM aims to facilitate the understanding of cultural heritage documentation by providing a common and extensible semantic framework that any cultural heritage information can be mapped to. The model supports a set of specific functionalities, such as (Crofts et al., 2009, p. i):

- inform developers of information systems as a guide to good practice in conceptual modeling
- serve as a formal language for domain experts and IT developers to formulate requiremenets for information systems
- serve as a formal language for the identification of common information contents in different data formats
- support associative queries against integrated resources by providing a global model

From the point of view of the Semantic Web, the model provides the "semantic glue" that is needed in order to facilitate the integration, mediation and interchange of heterogeneous cultural information. As a high-level domain model, it does not take into consideration the implementation and application-related details, such as the technical aspects and the terminology that are specific to different contexts.

The CRM is extensible, and the development of classes and properties for the needs of more specialized communities and applications is encouraged. Being a conceptual specification, the model can be implemented using a variety of knowledge representation languages; however, its use as an ontology for the semantic integration of heterogeneous sources on the Web has favored the definition of different serialization using Semantic Web-related languages, such as RDFS and OWL (see section 4.2.2.2).

The general characteristics of the model and its official standard structure make the CRM an important reference for the development of Semantic Web portals for cultural heritage. At the same time, its usage in real application scenarios has highlighted a set of crucial elements concerning the definition of mappings from legacy data and heterogeneous data schemata to the model that need to be briefly introduced and evaluated accurately in a concrete case study, such as that provided in chapter 5.

4.2.2 The CIDOC CRM in Use

At a general level, most of the existing experiences show that different degrees of intellectual and domain-expert work are required in order to define mappings between legacy data schemata and the CIDOC CRM.

The explicit representation of events, for example, may be complex where this information is only implicitly modeled in the legacy data schemata (such as in the COINS project Felicetti and D'Andrea, 2008a), while it becomes more straightforward where the schemata already adopt an event-centric perspective (such as in the CRM-EH project; see Binding et al., 2008). Moreover, implicit or explicit event information can show considering variable degrees of detail, depending on the very specific nature of the adopted cataloging and documentation methodology, its objectives, granularities and scopes. For example, even if an event such as the discovery of an artifact can be identified, specific information about the actors who were present or participated in it, as well as the time spans during which it happened, or even the causal connection with other activities that determined the discovery may not be available. Consequently, it may become difficult to select the most suitable classes and properties in the CRM according to the specific typology of event that is present in the data sources.

Beyond these general considerations, several issues and difficulties seem to be commonly associated to the definition of mapping chains to the CIDOC CRM, such as:

- Issues in **data quality** (concerning both the schema and the instance levels; see e.g. Kummer, 2007, pp. 48–50) may hinder the possibility of defining appropriate mappings, for example when information of a different nature is contained within free-text database fields.
- **Different chains for equivalent metadata** may result from the mapping activity performed on the same data schemata by different domain experts. Even if the CIDOC CRM is accurately described in the definition document, it is still open for interpretation, for example for what concerns the meaning of a certain class or property as is described in the respective scope note. From this point of view, the generality of the model, which is a necessary characteristic for preserving the maximum flexibility possible, also constitutes a drawback. On the other hand, the interpretation of legacy data schemata, if not adequately documented, may also in some cases be open to subjective interpretation.
- **Identical chains for different metadata**, on the contrary, may come out in some cases, requiring the introduction of additional "assertion chains" (Nussbaumer and Haslhofer, 2007, p. 11) in order to solve ambiguities in the mapping subgraphs.
- The creation of **virtual entities** (Binding et al., 2008, p. 284) is required in a number of cases, such as when events that are only implicitly present in the original schemata need to be made explicit in the mapping chain. This moreover contributes to a relevant increase in the number of triples constituting the chains, and complicates the mapping schemata.

In addition to the issues connected to the creation of mapping chains, the fact that the model does not address more practical aspects that are related for example to implementation-specific storage and processing aspects, makes other difficulties come out, such as:

- **Representing actual instance data values**, for which the CRM has little intrinsic provision (Binding et al., 2008, p. 284). While the absence of details on the treatment of instance data values is coherent with the high-level conceptual nature of the model, it requires that suitable strategies are defined, depending on the constructs of the actual language used for implementing the model.
- **Managing vocabularies**, for the same reasons, is demanded by the application level. However, vocabularies are a crucial component of cultural heritage documentation (e.g. in the form of typological lists), and their implementation and alignment with the model (through the `E55 Type` class) represent a central concern. The use of available and standard lists or the development of new ones is moreover an alternative that needs to be carefully analyzed.

Finally, the problem of **coreference linking**, i.e. the identification of different possible resource identifiers that refer to the same world entities, also requires that homogeneous and accurate mappings of different legacy schemata to the CRM have been created. In fact, if the same concept in different schemata has been modeled differently with respect to the CRM's classes and properties, the semiautomatic co-reference detection that is discussed for example in Doerr and Iorizzo (2008); Kummer (2007); Babeu et al. (2007) would result in logical inconsistencies.

The points that have been summarized here represent the most problematic elements that emerge from the review of the literature concerning the existent experiences, and provide a good guide for experimentation of the CIDOC CRM in novel contexts. Mapping existing data schemata to the CRM can then easily become a complex activity, for which a suitable workflow needs to be defined.

4.2.2.1 Defining a Mapping Workflow

The creation of mapping templates specifying the correspondence of legacy data schemata to the CIDOC CRM structure can be done in several different ways. The experimented approaches vary a lot, from automatic procedures with low manual intervention to completely manual work. A general methodology and the related software tools allowing to support it are not available to date, even if interesting work has been done in this area, such as the development of the Archive Mapper for Archaeology (AMA[1]).

Naturally, two main stages are required in order to define the mapping templates:

[1] `http://www.epoch-net.org/index.php?option=com_content\&task=view\&id=222\&Itemid=338`

1. the analysis of every single data schemata, with reference to the underlying semantics of e.g. database fields

2. the alignment of the analyzed data schemata to the CRM concepts and relations

Taking into consideration the documentation provided by existing projects (Cripps et al., 2004; Kummer, 2007; Binding et al., 2008; Felicetti and D'Andrea, 2008b), a basic process can be defined. It consists of a sequence of successive operations performed on every single data source, which can be iterated in order to refine or even change the representation choices according to the results of intermediate evaluations. More specifically, the stages characterizing a basic mapping workflow can be summarized as follows:

1. detailed analysis of the data schema

2. eventual selection of relevant subsets of the data structure with reference to the scope and goals of the case study

3. domain-expert analysis of the semantics underlying the schema

4. identification of the events and processes that are implicitly or explicitly represented in the data schema

5. grouping of data according to the identified events and processes

6. mapping of each database field (attribute-value pairs) to the CRM classes and properties (subject-object-predicate triples) with respect to the blocks of identified events

7. evaluation of the results and eventual re-iteration of some step of the sequence until the most satisfactory result is obtained

Specific attention has naturally to be paid to the identification of the events and processes (point 4), since they represent key aspects for the correct interpretation and usage of the CRM.

The documentation of the mapping activity should be organized according to the event-centric perspective of the CRM in the vein of similar works such as Felicetti and D'Andrea (2008a,b). This approach differs substantially from the traditional "object-centric" documentation and may seem confusing at a first sight; however, it shows the advantage of highlighting the nature and the principles of the CRM, making it possible to better control the mapping definition and to highlight its structure.

To this respect a specific and formal proposal for a mapping language for information integration has been been suggested by Kondylakis et al. (2006). This language is based on the basic mapping schema depicted in figure 4.2, and it has been used in relevant CIDOC CRM-related works, such as Kummer (2007).

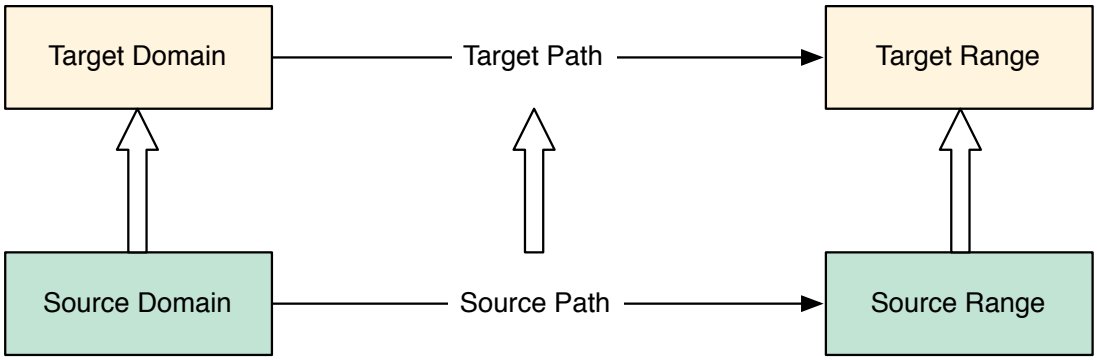


Figure 4.2: *The basic mapping schema proposed by Kondylakis et al. (redrawn from 2006, fig. 1, p. 1).*

However, a more intuitive approach can be defined in order to provide an easily accessible and readable documentation, which can be coupled with more formal schemata. Figure 4.3 represents the suggested approach, which makes reference to the mapping activity defined in the case study presented in chapter 5. More specifically, the sub-graphs created with the CRM's classes and properties shown on top of the figure can be used in order to introduce and effectively represent the relevant triples in a compact way, while the table at the bottom expands the graph representation according to the single triples constituting the mapping chains on e.g. a database field. Some conventions have been used to increase readability:

- the classes representing the activities are marked with a red outline
- the correspondences between general CRM classes and actual domain concepts are indicated using additional labels (in orange) whenever deemed necessary
- only direct properties have been used

The subgraphs of the mapping template will be shown in chapter 5, while the tables with the mapping chains are present in appendix A.

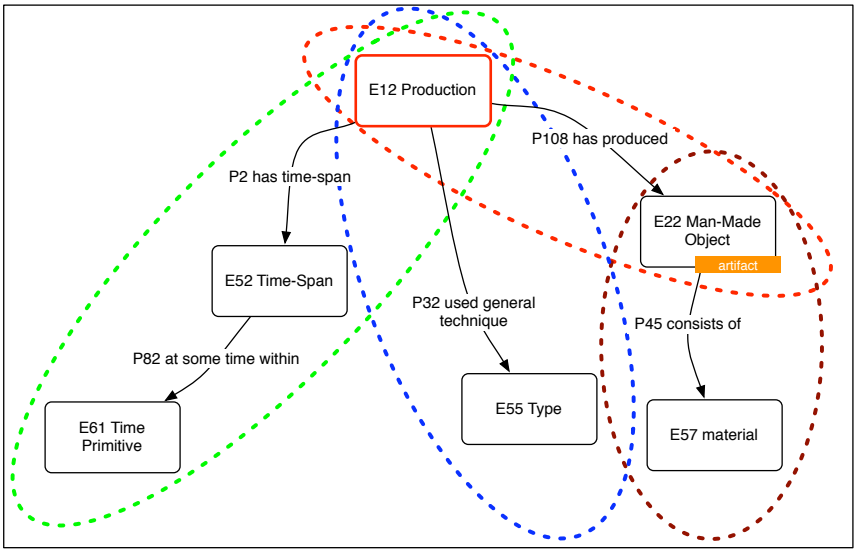
4.2.2.2 Evaluation of the Available Serializations

Different serializations of the CIDOC in RDFS and OWL are available (fig. 4.4).

The first partial RDFS serialization of the CIDOC CRM (v. 3.2) was released by the CIDOC CRM Special Interest Group in 2001, while the first proposal for an OWL serialization is dated to 2006.

The RDFS implementation of version 4.2 constitutes the basis for:

- The first OWL translation by the CIDOC CRM Special Interest Group.



Production of the artefact

table	field	CIDOC CRM representation	meaning
implicit information		E12 Production — P108 has produced → E22 Man-Made Object (artifact)	
Sub_dt	DTZG DTZS	E12 Production — P2 has time-span → E52 Time-Span	Chronology (century and part of century)
		E52 Time-Span — P82 at some time within → E61 Time Primitive	
Sub_mtc	MTC	E12 Production — P32 used general technique → E55 Type	Technique
		E22 Man-Made Object (artifact) — P45 consists of → E57 Material	Material

Figure 4.3: *An example of both graph and tabular representation of the mappings to the CIDOC CRM.*

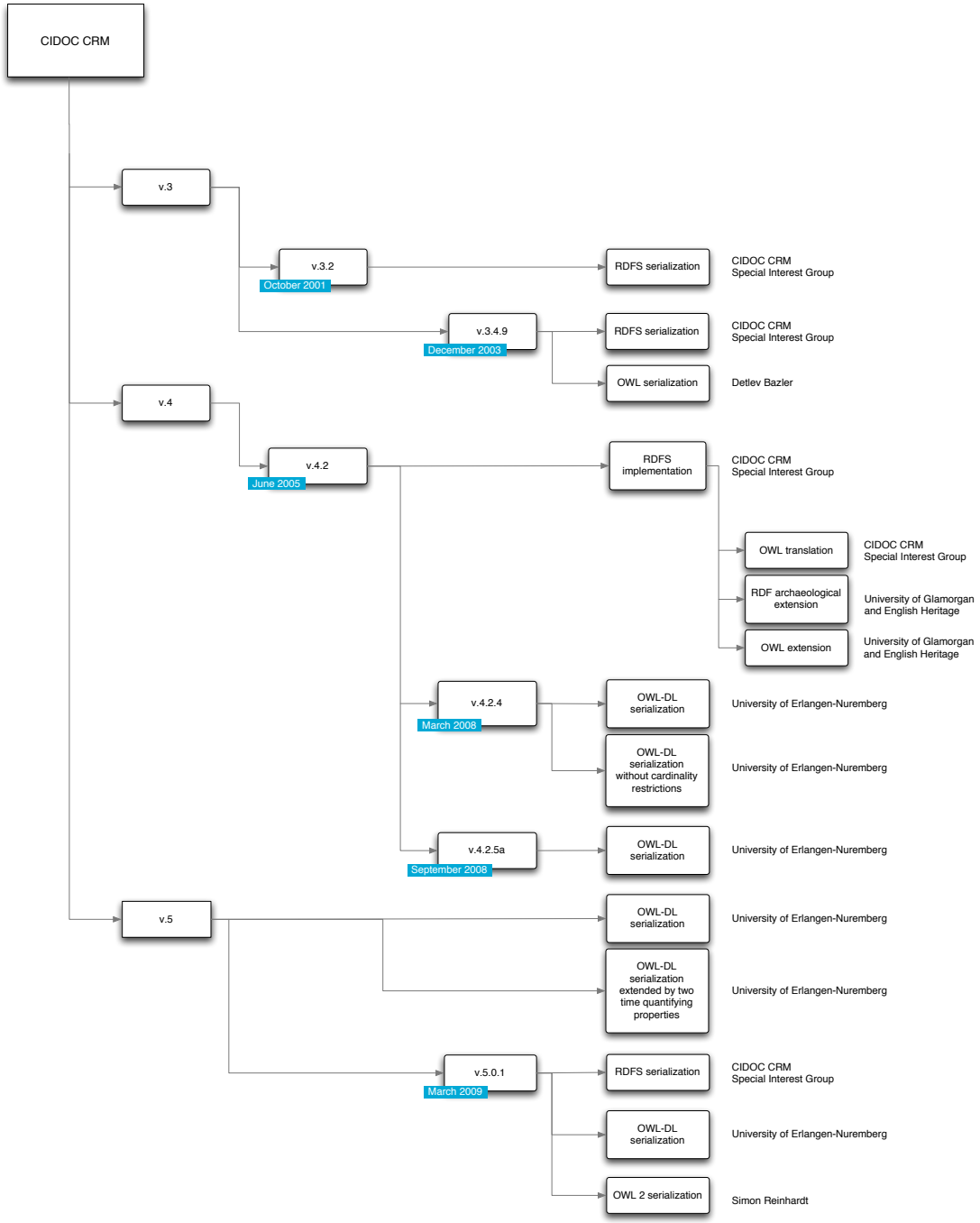


Figure 4.4: *CIDOC CRM available serializations.*

- The CRM-EH extension, which has been developed by the University of Glamorgan and English Heritage in order to model the archaeological research process in detail (Binding et al., 2008). CRM-EH comprises 125 extension sub-classes and 4 extension sub-properties[1]. The general guidelines of conceptual modeling, as well as the processes of data mapping and data extraction are well documented in Cripps et al. (2004); May (2006a,b); Binding et al. (2008) and a graphical schema representing classes, properties, and notes on the usage of them is provided[2].

- The extension to the CIDOC CRM v4.2 properties with OWL statements (InverseOf, TransitiveProperty, SymmetricProperty etc.)[3](see Binding et al., 2008, note 1, p.1). Work done by the University of Glamorgan and English Heritage implements the OWL constructs that are relevant for their STAR project[4] and cannot be considered a full translation of the CRM in OWL.

Full serializations in OWL DL have been made from version 4.2.4 especially by the University of Erlangen Nurenberg, Chair of Computer Science 8 (Artificial Intelligence), under the name of "Erlangen CRM"[5]. The motivation for this work is the need to take full advantage of the CIDOC CRM's benefits by using a modern knowledge representation language, as stated in Goerz et al. (2008, p.1). The implementation of version 4.2.4 is documented in Goerz et al. (2008) which provides an overview of the motivations of the work and the principles that guided the implementation, and in Oischinger et al. (2008), which offers a short documentation. Several issues emerged from the adoption of OWL-DL with respect to the CRM specifications, such as:

- whatever was underspecified or unspecified in the CRM document had been left open in the OWL implementation as well

- some features could not be implemented or had not been implemented for certain reasons

- the `E55 Type` class could not be represented as a metaclass, as described in the CRM document

- shortcuts have not been included

[1] `http://hypermedia.research.glam.ac.uk/media/files/documents/2008-04-01/CIDOC_v4.2_extensions_eh_.rdf`

[2] `http://cidoc.ics.forth.gr/docs/AppendixA_DiagramV9.pdf`

[3] `http://hypermedia.research.glam.ac.uk/media/files/documents/2008-04-01/CIDOC_v4.2_extensions_glam__.rdf`

[4] Semantic Technologies for Archaeology Resources. See `http://hypermedia.research.glam.ac.uk/kos/STAR/`

[5] `http://www8.informatik.uni-erlangen.de/IMMD8/Services/cidoc-crm/index.html`

- for the sake of decidability, data type properties could not be defined as inverse functional
- the definition of new properties in the way familiar to logic programming could not be expressed in OWL DL

Since v. 4.2.4. the Erlangen CRM has been implemented according to the successive new versions of the CIDOC specifications. Version 5 introduces two additional time quantifying properties (see section 4.3.2.1 for details). Moreover, the proposal for encoding v. 5.0.1 in OWL 2[1] was proposed in May 2009.

Given this scenario, and taking into consideration the accurate work done by the the team at the University of Erlangen-Nurenberg, the Erlangen CRM serialization in OWL DL of the CIDOC CRM v.5.0.1 seems to be currently the most interesting choice.

4.3 . . . to Fruition: Retrieving Information through Fuzzy Chronologies

Time is a fundamental dimension for the description, study and understanding of cultural processes. Cultural heritage, as a product of cultural processes, naturally inherits this characteristic. The documentation of cultural heritage usually pays specific attention to the acquisition of temporal information, which may relate to a variety of aspects, such as the period of production of artworks, the architectural phases in the building of historical structures, etc. Instances of temporal intervals are used for the annotation of cultural objects and also for querying datasets containing these objects.

Therefore, methods for the retrieval of cultural heritage information should pay specific attention to improving the possibility of matching query and annotation intervals, for example by examining their similarity or closeness. Checking whether two time intervals have something in common allows for answering queries like "find all paintings that were painted around the middle of the the 16th century".

However, representing time in Semantic Web ontologies is not straight-forward, because the question of when a certain time was or will be is often uncertain, subjective or vague (Nagypál and Motik, 2003). For example, it may not be known exactly when a given archaeological artifact was manufactured (uncertainty), when the "Middle Ages" in a certain area was according to opinions of different historians (subjectivity), or when the spring starts (vagueness, imprecision). In addition, transitions between different phases, such as historical periods, are usually complex processes which are not identiable by clear cut dates, even if conventional calendric markers are mostly used in order to simplify historical sequences. All these elements are at the basis of imprecise temporal representations, for example in the cultural heritage, historical and archaeological contexts.

[1]OWL 2 is currently a W3C Recommendation (approved 27 October 2009). See `http://www.w3.org/TR/2009/REC-owl2-overview-20091027/`

To this respect, archaeology is the discipline within the cultural heritage scenario that offers the most composite and challenging scenario, and provides a challenging domain for original research.

The material items that archaeologist recover during their research activities are, in fact, the physical results of actions and events that happened in the past, involving specific actors in given places and times. The possibility to "read" this information, and to discretize and characterize past events with respect to their temporal properties (such as dating, duration, and the temporal relationships with other events) are fundamental tasks of the archaeological research process. In particular, the creation of linear and ordered sequences of events (**chronologies**), and the detection and analysis of discontinuities and changes in these sequences are crucial, since they represent the basis for the study of the dynamics and the trajectories of cultural processes, and the conditions into which they took place. Therefore, chronologies and the temporal intervals that constitute them should be a central element in the research on matching query and annotation intervals that are connected to archaeological items.

4.3.1 Archaeological Chronologies: A General Overview

Archaeological chronologies are traditionally divided into two main types: relative and absolute.

Relative chronology represents the backbone of chronological reasoning. The term "relative" refers to the fact that events are chronologically defined relative to other data through qualitative temporal relations, such as precedence, succession, contemporaneity, etc; relative chronologies are therefore based on the inter-dependence of the data being studied (Lucas, 2005, p. 3). Actual calendric dates associated to the events and the sequence may in some cases remain unknown, while generally they are established by successive analysis methods and comparison with other data.

Seriation and stratigraphy are probably the most known methods for the creation of relative chronologies. In particular, stratigraphy is of primary relevance in the context of the archaeological excavation; it provides a set of principles for the analysis of the physical relations between different archaeological layers (or "stratigraphic units") and for defining relative chronological sequences from them (see e.g. Harris, 1979). The method relies on a theory, whose basic axiom is the "law of superposition", which states that layers in a stratification are physically ordered in a time sequence, with the oldest on the bottom and the youngest on the top. Since the archaeological stratification represents the material result of actions and events that took place on an archaeological site, the discretization of stratigraphical units coincides with the discretization of events, which together define the chronological sequence of the site.

Absolute chronology is based instead on the direct temporal characterization of events with reference to a scale of measurement, which is typically the calendar; absolute chronologies are therefore based on a framework that is independent of the data being studied (Lucas, 2005, p. 3). The most straightforward methods for obtaining ab-

solute chronologies are related to the direct presence of dates on objects belonging to the context being studied, such as coins, or historical documents. However, the development of absolute dating techniques that are based, for example, on the analysis of the changes through time of the chemical or physical properties of the material archaeological items are made of (e.g. radiocarbon decay analysis and thermoluminescence) (Aitken, 1990) offers the possibility to establish the antiquity of an item with respect to the calendar.

In fact, the distinction between absolute and relative chronologies is not correct. Most of the absolute chronological attributions are in fact **estimates**, rather than "absolute" elements, and the levels of precision and confidence of these estimates can vary a lot[1]. For this reason, alternative definitions have been coined and can be found in the literature, such as "chronometric", or "scientific". The latter, however suggests the idea that relative methods are not scientific, which is evidently not true, since these methods rely on scientific theories that are supported by empirical evidence.

However, the traditional distinction between relative and absolute will be maintained here in its commonsense meaning, i.e. to refer to qualitative and quantitative chronological characterizations respectively.

There is not a single approach possible to building chronologies, and the combination of both relative and absolute methods is usually employed, depending on the specific nature of the context being studied and the aims of the research. A vast body of literature concerning general discussions on chronologies and dating methods is available; on the contrary, there is a surprisingly lack of research on the epistemological foundations of chronology building in archaeology, and on the definition of formal frameworks of analysis.

Notable exceptions are Buck and Millard (2004), and Doerr et al. (2004a). This latter, in particular addresses the scenario of chronological reasoning (in the vein of Gardin, 1990) through the analysis of the chronological consequences that are supported by primary evidence and background knowledge, and proposes a sub-division of relative chronologies in three groups:

- by event order, where direct evidence about the temporal ordering of multiple events is available (e.g. stratigraphy)
- by event inclusion, where direct evidence of inclusion relations between two events is available (e.g. super-process composed of known subprocesses)
- by temporal distances, where primary evidence of temporal distance and duration is available (e.g. the estimate of the change rate of stylistic or technological skills)

[1]More in general, an "absolute" attribution is simply impossible, since time is theoretically subdivisible into infinite units.

In addition, categorical dating is introduced, for all the situations where dating is not obtained by observing particular items and factual relations (as in absolute and relative chronologies), but by attributing an object to a class whose chronology is known (such as in typological analysis).

4.3.1.1 Temporal Fuzziness and Archaeological Chronologies

As Lucas (2005, p. 32) says: "*. . . the temporal location of archaeological entities in an absolute framework (usually calendar years) remains fuzzy, and requires effort, expense and a suitable context that normally means that a single date, or a handful of dates, stand as proxies for the whole site or phase of site — or even a whole artifact type. Moreover, excepting where historical dating can be used, our absolute dates may only be good to within half a century or so*".

In fact, **dating** in archaeology means to provide an estimate of the temporal location of a past event. Barceló (2009) explains that dating is obtained through comparison with a given reference event (or "event 0"), and in particular by calculating the distance from a reference event along a previously defined measurement scale, such as the calendar; with reference to figure 4.5, dating means therefore to provide an estimate of *distance(reference,A)*. On the contrary, duration is obtained through comparison with the event that follows in the sequence, and in particular by subtracting the dating of the twos (fig. 4.5, *duration(A)*).

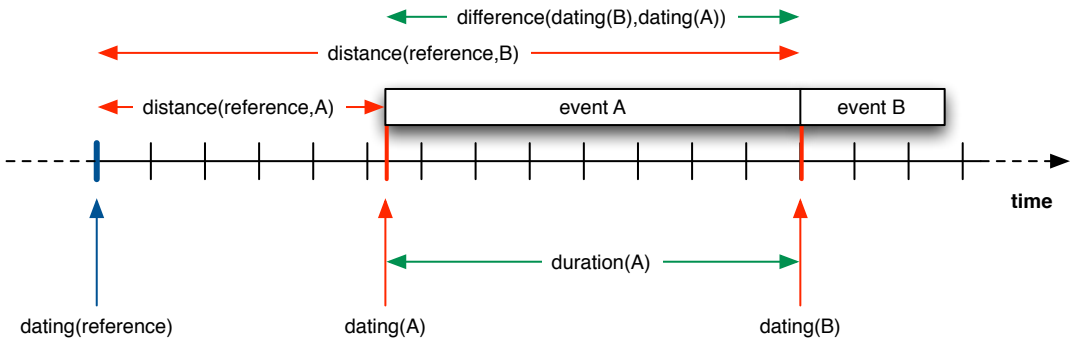


Figure 4.5: *A representation of the dating and duration properties.*

Methods for improving the precision of the *distance(reference,A)* calculation are the object of continuous research, especially in archaeometry. For example, in the context of radiocarbon decay analysis, statistical correction methods (calibration methods) based on measurements obtained with other absolute dating techniques (typically dendrochronology) have been introduced. In fact, the duration of radiocarbon years varies according to the stochastic nature of the radiocarbon decay process, requiring the correction of the raw results which can show a significant divergence from the final estimate. However, the process of calibration has in turn generated new issues (for an overview and discussion see e.g. Barceló, 2009), such as the fact that

calibrated dates are inevitably connected to a non-uniform temporal scale where different areas show different precision levels. Additional elements of imprecision can also depend on measurement errors induced by the contamination of the samples by human and/or natural phenomena, or data processing errors.

The inevitable imprecision of the dating techniques obviously determines that archaeological chronologies are inherently fuzzy with respect to their absolute temporal properties. In fact, when grouping single elements and combining them in higher level chronological categories, several issues concerning the interpolation of single fuzzy temporal intervals emerge. With reference to radiocarbon dating, for example, Barceló (2009) shows the complexity of combining dated intervals with different probabilities, which inevitably determines that the period resulting from this combination should be non-uniformally distributed in a certain temporal interval (for a more extensive discussion, see Buck and Millard, 2004).

In fact, the formal model proposed by Doerr et al. (2004a) strongly takes into consideration these aspects, through the definition of the concepts of determinacy and indeterminacy intervals. According to this model, the goal of chronological reasoning should be to minimize these intervals, therefore improving the precision of the dating and duration estimates of the temporal intervals during which past events happened.

More specifically, the model is based on the concept of "event", which is conceived as the meeting of living or dead items, which brings about a change of state, at any scale. This view, referred to as "events as meetings", implies that meetings happen within a spatiotemporal kind of coherence volume. With reference to fig. 4.5, dating an archaeological artifact means to provide a temporal approximation of the coherence volume of some event(s) in which that object was present (Doerr et al., 2004a, p. 2–3); the dated event is usually the artifact's production. This perspective is extremely accurate from the epistemological point of view, even if it may seem contra-intuitive at first sight. For example, affirming that a statue is chronologically attributed to the "2nd century B.C." means in reality to affirm that the production event of the statue happened sometime within that temporal interval and in a specific location (spatiotemporal coherence volume), where agents (the artist) and the artifact (the statue) were present.

Figure 4.6 shows a more articulated example of the "events as meetings" view, where two events (the building of a house and a successive volcano eruption destroying it), happened within a spatio-temporal coherence volume, and involved participants (people and physical items). These latter are moreover represented coherently with their state before, during, or after the events.

The related formal model is based on a ET (Event/Time) structure providing the mathematical foundation of the "events as meetings" view, together with actual temporal characterizations. The model assumes that "*. . . the true temporal extent of an event cannot be observed, but that it is possible for a suitable observer to identify dates that are definitely before or after the true endpoint of an event. Semantic relationships between events*

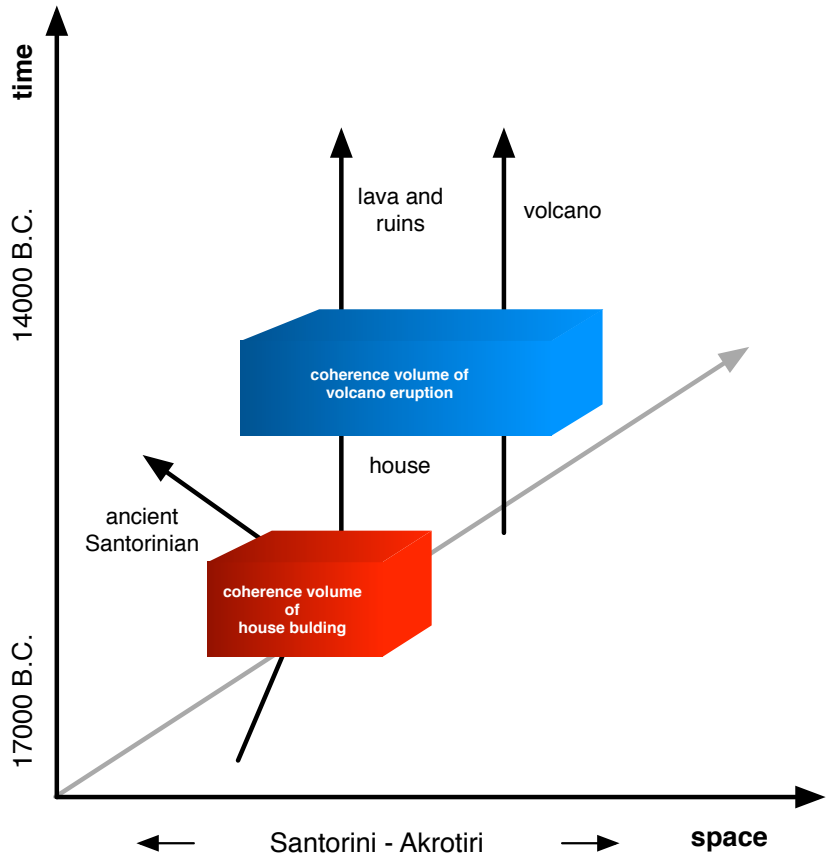


Figure 4.6: *An example of deposition events as meetings (redrawn from Doerr et al., 2004a, fig. 2 p.3).*

and absolute dating give rise to sets of "temporal consequences" that relate or approximate the endpoints of the true intervals of the events under consideration." (Doerr et al., 2004a, p. 4). On the basis of this, different kinds of determinacy and indeterminacy temporal intervals can be identified, as illustrated in figure 4.7.

4.3.2 Related Work

In the context of Semantic Web applications for cultural heritage and archaeology, while there is general agreement on the importance of the temporal and chronological dimensions for retrieving relevant information, there are few contributions specifically facing this aspect.

Heterogeneous aspects have been investigated, from modeling proposals to the visualization of information according to interfaces based on temporal metaphors.

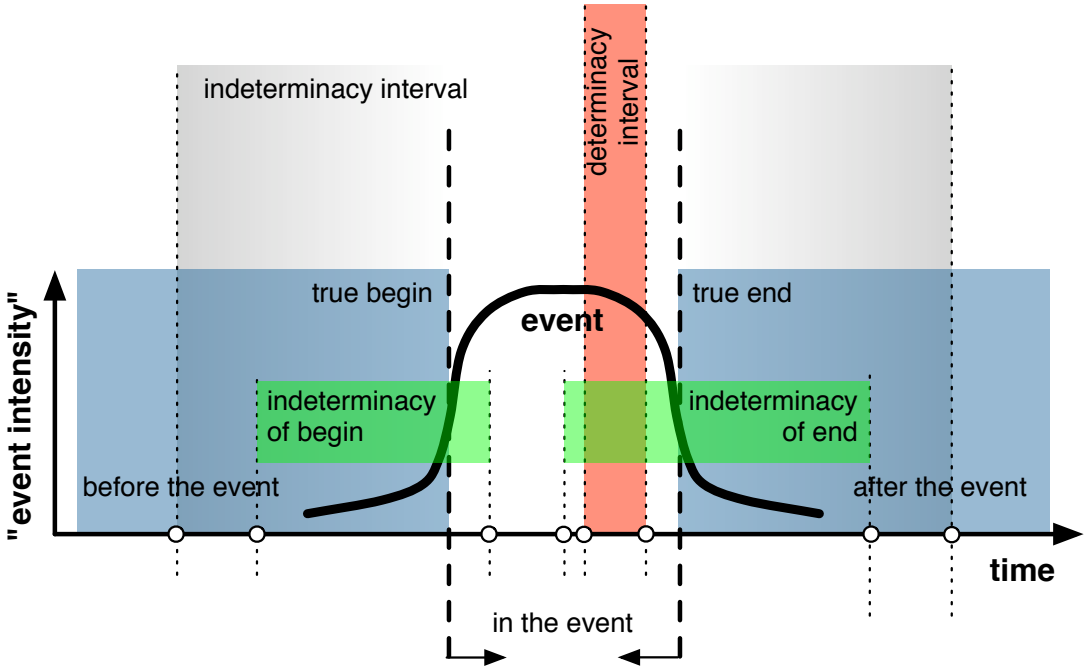


Figure 4.7: *Determinacy and ideterminacy temporal intervals according to the CIDOC CRM (Doerr et al., 2004a, redrawn[fig. 4 p. 4).*

This latter group includes work on managing and visualizing historical collections on interactive timelines, such as those provided by "MultimediaN E-Culture" (Schreiber et al., 2006) and "Culture Sampo" (Hyvönen, 2009), which provide tools such as interactive timelines for visualizing events and perform simple retrieval on temporal intervals. More specific contributions from these projects dealing with temporal ontologies (such as Kauppinen et al., 2008) offer interesting perspectives and results, but they do not analyze temporal fuzziness.

Johnson (2008) reviews the approach adopted in the TimeMap-ECAI[1] project, and provides an updated sets of proposals and advancements towards the integration of Web 2.0 techniques and tools for Web-based cultural atlases. In particular, it introduces some considerations about new modalities for developing interactive timelines concerning historical events. Central in these considerations is the development of historical gazetteers (Mostern and Johnson, 2008), modeled as a database of naming events, where causal relationships (such as IsCausedBy, IsDependantOn, IsPartOf, IsRelatedTo) are introduced, rather than only temporal ones.

Nagypál and Motik (2003) describe the approach adopted in the VICODI project[2]

[1] http://www.ecai.org/
[2] http://www.icodi.org/

for the development of an ontology of European history which can be used for the semantical indexing of historical documents. Key in the design of this ontology are the problems conected to the inherently imprecision of historical accounts with respect to temporal information. Starting from the consideration that "*. . . debate and disagreement over temporal specifications in history is rather the norm than an exception*" (Nagypál and Motik, 2003, p. 908), the cases of uncertainty, subjectivity and vagueness are taken into consideration, and a temporal model taking them into account is introduced. This model is based on fuzzy sets that have been integrated in the RDF(S)-based ontology model KAON[1]. Moreover, a mechanism is introduced to evaluate whether e.g. a crisp temporal relationship *intersects* holds between two fuzzy temporal intervals. From the point of view of the modeling principles and choices, the approach proposed by this work is full of interesting elements; however, no practical implementation for evaluation in the form e.g. of a publicly accessible application is available.

Accary–Barbier and Calabretto (2008) introduce a research project in the context of digital libraries dealing with the possibility of comparing different temporal models of knowledge in archaeological documentation that may emerge from fuzzy or even contrasting chronologies. The model proposed in this work considers the domain of relative chronologies and is based on Allen's well-known temporal interval algebra (Allen, 1983). A constraint propagation algorithm is implemented in order to detect inconsistencies concerning different opinions on chronological sequences.

Beyond these experiences in the domains of cultural heritage and archaeology, the more general literature concerning temporal modeling and reasoning is extremely vast, and is out of the scope of the present research. The general properties of time ontologies have been analyzed e.g. in (Vila, 1994), while specific contributions have investigated the possibility of incorporating temporal reasoning into RDF (Gutierrez et al., 2007). The scientific literature testifies to the richness of the debate about a number of theoretical aspects, such as the definition of basic time primitives (intervals vs points), or other properties for time, e.g. whether it is discrete or dense, bounded or unbounded, and what type of precedence the time ontology allows: linear, branching, parallel, or circular.

The implementation of the theoretical model introduced in the previous section in the CIDOC CRM deserves a more extensive discussion that is introduced in the next section.

4.3.2.1 The CIDOC Model for Temporal Information

Being the CIDOC model an event-based one, it naturally deserves particular attention to temporal representation, as is clearly shown in a general schema of the most relevant classes and properties (fig. 4.8). The CIDOC classes and properties for temporal representation implement on the logical level the theoretical model supporting

[1] `http://kaon.semanticweb.org/`

chronological reasoning that has been introduced in the previous section.

The most general class is `E2 Temporal Entity`, which is defined as being abstract, i.e. no instance of it can be created. It is used as a superclass for all the classes that have a temporal component and in particular:

- `E4 Period`, which "*...comprises sets of coherent phenomena or cultural manifestations bounded in time and space*" (Crofts et al., 2009, p. 3), and therefore is linked through `P7 took place at` to `E53 Place`
- `E3 Condition State`, which "*...comprises the states of the objects characterized by a certain condition over a time-span*" (Crofts et al., 2009, p. 3)

`E3 Condition State` is not further specialized, while `E4 Period` is the superclass of several classes (such as `E5 Event`) that will be introduced in more detail in chapter 5.

The CIDOC basic temporal relations connected to `E2 Temporal Entity` implement the relationships between temporal intervals identified by Allen (1983) and are schematized in figure 4.9 together with graphical representations of their meaning.

Specialized classes are in turn characterized by specific properties, such as the ones of `E4 Period` that make it possible to define spatio-temporal relations between different instances of `E4`. Two of them are of direct interest for archaeological chronologies:

- `P9 consists of` can be used to express the internal subdivisions of a period (e.g. the Italian Bronze Age, and its sub-periods
- `P10 falls within` can be used to describe a period that falls within the geographical area and the time span of another period. Differently from `P9`, no logical connections between two periods is assumed here; therefore `P10` is suitable for the representation of the relations between periods belonging to chronologies built according to different perspectives, such an archaeological and an art-history ones

It is important to stress that `E2 Temporal Entity` and `E4 Period` do not directly comprise temporal extents "*...in the sense of Galilean physics, having a beginning, an end, and a duration*" (Crofts et al., 2009, p. 21). In fact, temporal extents are represented by `E52 Time-Span` class, which is linked to `E4 Period` by the `P4 has time-span` property.

The basic assumption behind `E52` instances is that they represent approximations of the real world time spans in which temporal entities take place. This assumption fits well with the fuzziness of archaeological chronologies where time spans (temporal intervals) represent approximations of the temporal extent of an event, as discussed above. Consequently, time spans should not be defined by fixed start and end points,

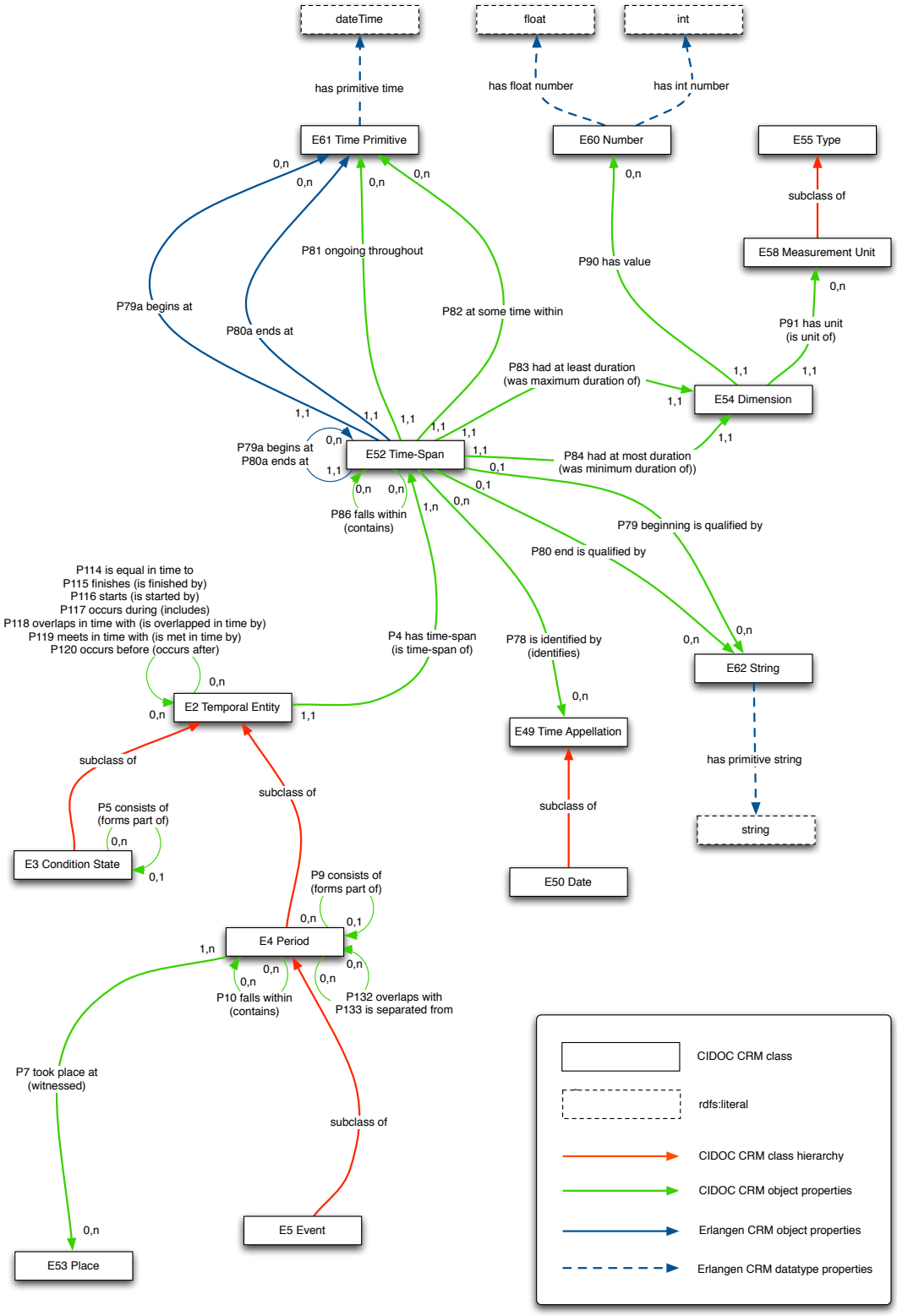


Figure 4.8: *A schema of relevant classes and properties for temporal representation in the CIDOC CRM and the Erlangen CRM.*

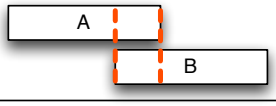
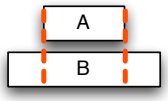
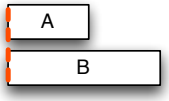
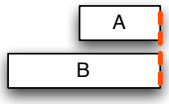
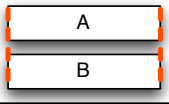
Property	Inverse property	Graphical representation
P120 occurs before	occurs after	
P119 meets in time with	met-by in time	
P118 overlaps in time with	overlapped-by in time by	
P117 occurs during	includes	
P116 starts	started-by	
P115 finishes	finished-by	
P114 is equal in time to	-	

Figure 4.9: *The CIDOC CRM implementation of Allen's relationships between temporal intervals.*

and a set of properties is introduced in order to quantitatively characterize time spans, taking into consideration their fuzziness (fig. 4.10).

In particular:

- `P79 beginning is qualified by` and `P80 end is qualified by` make it possible to express the beginning and the end of a time span in some way, with an instance of `E62 String`
- `P81 ongoing throughout` and `P82 at some time within` describe respectively the minimum period of time covered by a time span (i.e. its inner boundary), and the maximum period of time within which a time span falls (i.e. its outer boundary
- `P83 had at least duration` and `P84 had at most duration` describe

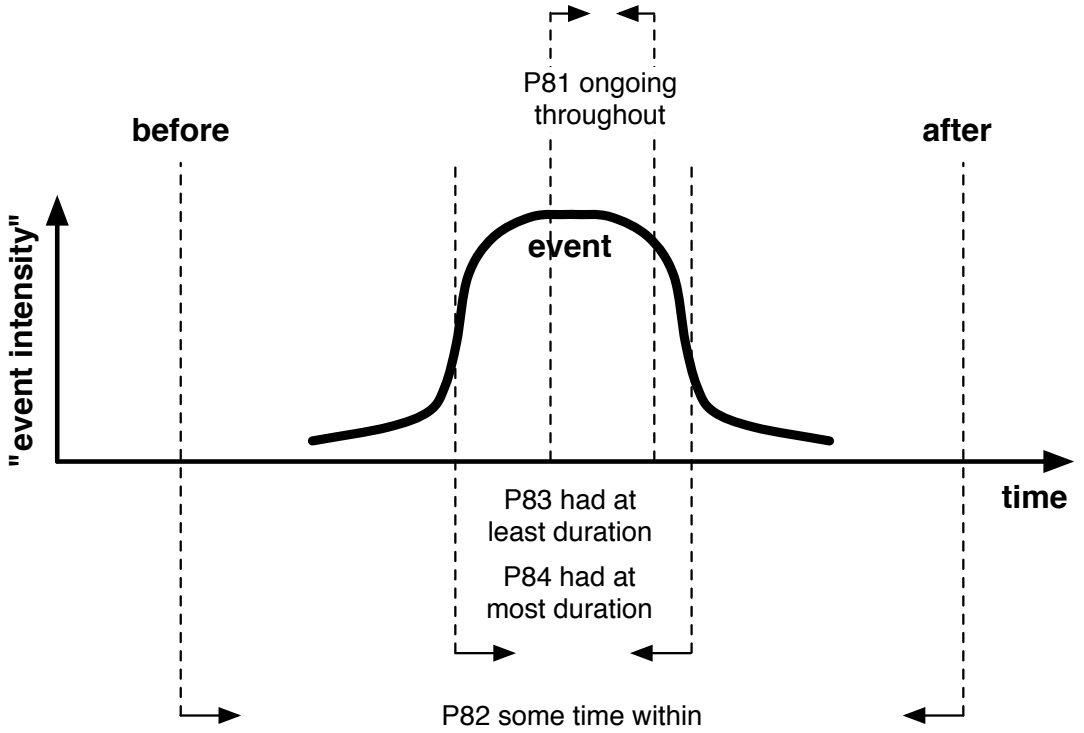


Figure 4.10: *CIDOC CRM temporal properties of the* `E52 Time-Span` *class.*

respectively the minimum and the maximum lenght of time covered by a time span (i.e. the estimates of a time span duration)

Therefore, the model basically offers two different ways of representing events with respect to time:

- through qualitative temporal relationships between events, by using Allen's time operators (`P114–P120`)
- through quantitative approximations represented by time spans and the related properties and classes

The first case assumes that events have crisp boundaries, since Allen's model (and the related temporal algebra) is based on crisp temporal intervals; the second case, instead, allows to take into consideration fuzzy intervals.

The need to preserve the two modalities and the way they are related have been discussed. In fact, in real application scenarios some issues can emerge, such as those discussed on the CIDOC CRM Special Interest Group mailing list[1].

[1]See, for example, `http://lists.ics.forth.gr/pipermail/crm-sig/2009-August/`

Starting from these debates, extensions to the standard model have been proposed, such as the Erlangen CRM, which introduces two additional properties (`P79a begins at` and `P80a ends at`) for representing fuzzy time spans. If we consider the situation represented in figure 4.11, where a time span A has a fuzzy beginning and a fuzzy end, it can be represented in two different ways according to the two models:

- the CIDOC model makes use of the `P81` and `P82` properties, using a theoretically sound and compact representation
- the Erlangen approach makes it possible to represent both the temporal relationships of *A* with *A fuzzy begin* and *A fuzzy end*, and the actual most suitable date-pairs (using `E61` and the `xsd:dateTime` datatype) for *A fuzzy begin* and *A fuzzy end*

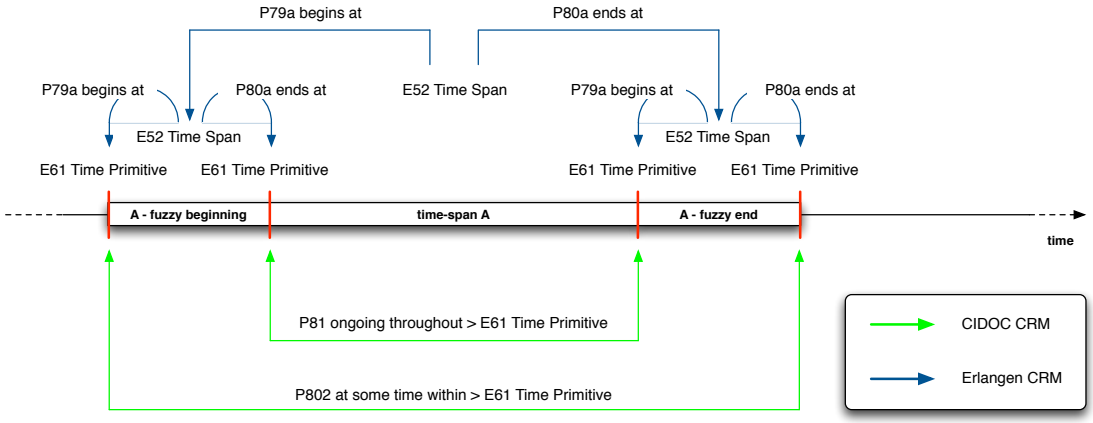


Figure 4.11: *The CIDOC CRM and Erlangen CRM representation of fuzzy temporal intervals.*

Both representations are based on well-advised considerations. The CIDOC CRM approach is theoretically correct and coherent with the global approach of the model. Its possibilities for supporting chronological reasoning are shown, for example, in Holmen and Ore (2010), with a few simple cases integrating deduction rules based on Allen operators with fuzzy time spans. On the other hand, the Erlangen model enables the deployment of simpler chronological reasoning mechanism dealing with dates, while complicating the model.

These examples, beyond the more theoretical and model related issues connected to fuzzy temporal representation, bring us to consider that different application scenarios justify the possibility of experimenting new approaches for representation. In

`001386.html`, `http://lists.ics.forth.gr/pipermail/crm-sig/2009-August/001387.html`, and `http://lists.ics.forth.gr/pipermail/crm-sig/2009-August/001392.html`

particular, considerations about the objectives and efficiency of chronological reasoning may suggest to couple the complex and rich representation of the CIDOC CRM, with simpler models enabling the efficient and effective retrieval of relevant information, such as the one introduced in the next section.

4.3.3 A New Approach Based on Fuzzy Sets

This section introduces a new approach for the representation of fuzzy temporal intervals which can be used for improving cultural heritage information retrieval on the Semantic Web. The work described here is presented in Kauppinen et al. (2009), and represents one of the results of a personal research experience made in collaboration with the SeCo Research Group at the Helsinki University of Technology[1].

4.3.3.1 Motivation

The CIDOC and the Erlangen models provide a sound and rich approach to supporting chronological reasoning, enabling the possibility to draw chronological consequences for validating existing scientific interpretations or produce new ones (as is shown e.g. in Holmen and Ore, 2010). However, in contexts where the objective is not to support automatic interpretation, but to improve the retrieval of meaningful information for fruition in and beyond the community of experts, these models can be too rich, and hinder the possibility of deploying effective and efficient retrieval methods. If we are dealing with the common scenario where temporal intervals related to chronologies are still fuzzy, but can be defined reasonably by a set of approximation dates, a lighter ontology should be developed, and it can be integrated with the standard CIDOC representation in an actual application.

Therefore, here we propose an approach that adapts the ideas exposed in Nagypál and Motik (2003) and in Visser (2004) of using fuzzy sets for the representation of temporal annotations and queries. Fuzzy sets allow the representation of imprecise temporal intervals such as "around the middle of the 2nd century B.C." which can be either used as query or annotation intervals. Visser (2004) has identified the following different types of boundaries for periods:

1. exact boundaries
2. persistent boundaries
3. unknown boundaries
4. fuzzy boundaries

[1] `http://www.seco.tkk.fi/`. The model and the reasoning engine using it, which are presented in this work, have been developed by Tomi Kauppinen, of the SeCo Research Group.

.

Our interest is in types 1 and 4. A model is introduced, and a method for matching query and annotation intervals based on their weighted mutual overlapping and closeness is proposed. In this regard our method aims to extend existent and similar works that have produced relevant results. In particular, Nagypál and Motik (2003) introduced a mechanism to evaluate whether e.g. a crisp temporal relationship *intersects* holds between two fuzzy temporal intervals. The result is a value explicating the level of this confidence. We evaluated its usability also for relevance calculation, and we add a combined measure (overlaps and closeness together) in order to improve precision and recall. Visser (2004) proposed to calculate the overlap between two fuzzy temporal intervals, but did not provide any evaluation results confirming the usability of the overlap relation. Also, closeness was neither considered nor tested in their study.

The approach and the model introduced in this section has been applied to a real case study, and the results have been evaluated, as discussed in chapter 6.

4.3.3.2 Modeling Fuzzy Temporal Intervals

Whereas in traditional set theory an item x either belongs or not to a given set A, in fuzzy set theory (Zadeh, 1965; Zimmermann, 1996), a membership grade μ having a value in range [0,1] is introduced, expressing how much x belongs to A.

According to this theory, a representation of fuzzy temporal intervals that are related to historical and archaeological chronologies can be introduced. Figure 4.12 shows, for example, a depiction of a possible trapezoidal fuzzy set representation for the period "from the beginning of the 1st century B.C. to the first half of the 1st century A.D.".

Following Visser (2004) a fuzzy temporal interval T representing the imprecision of a time period can be defined using a quadruple $< T_{fuzzybegin}, T_{begin}, T_{end}, T_{fuzzyend} >$, where:

- $T_{fuzzybegin}$ expresses the earliest beginning of the period
- T_{begin} expresses the date at which the period has begun for sure
- T_{end} expresses the earliest end of the period
- $T_{fuzzyend}$ expresses the date at which the period has ended for sure

The actual choices concerning the dates that are used in order to define this quadruple are of course a choice of a domain expert, and they can vary according to the opinions of different domain experts. With reference to figure 4.12, the personal choices for representing the period are:

$$\begin{aligned} T_{fuzzybegin} &= 105 \;\; B.C. \\ T_{begin} &= 95 \;\; B.C. \\ T_{end} &= 43 \;\; A.D. \\ T_{fuzzyend} &= 57 \;\; A.D. \end{aligned} \tag{4.1}$$

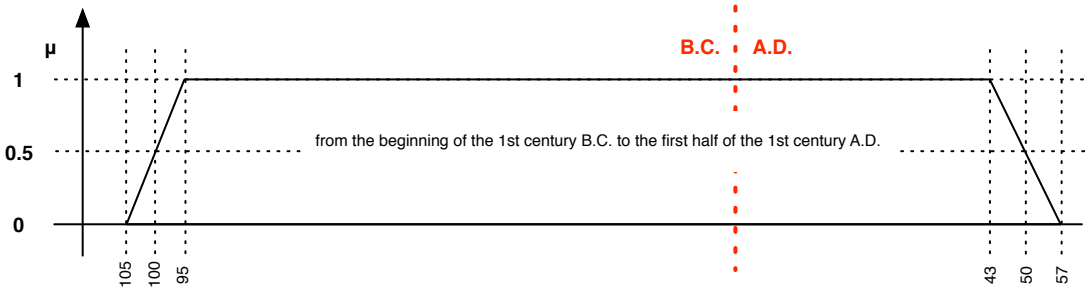


Figure 4.12: *The period "from the beginning of the 1st century B.C. to the first half of the 1st century A.D." represented as a fuzzy temporal interval.*

Since the indication "beginning" is more precise than "first half" the actual temporal distance separating $T_{fuzzybegin}$ from T_{begin} (10 years) is less than the one separating $T_{fuzzyend}$ from T_{end}.

4.3.3.3 Enabling Temporal Reasoning on Fuzzy Temporal Intervals

In order to enable reasoning on the fuzzy temporal intervals modeled with this approach, a proportional **overlap function** has been introduced, expressing the intersection of two intervals:

$$o_t : A, Q \rightarrow p \in [0, 1] \tag{4.2}$$

This function tells how much two fuzzy intervals A and Q overlap, and is represented in terms of temporal overlaps $o_t = overlaps(A, Q) = |A \cap Q|/|A|$. More specifically Q is intended here as a possible query interval, while A is a possible annotation interval; therefore the function answers to the question "How much does a given query interval Q *overlap* with an annotation interval A?".

Moreover, an **overlappedBy** function:

$$o_b : A, Q \rightarrow p \in [0, 1] \tag{4.3}$$

has been introduced for expressing how much a query interval Q is *overlapped by* an annotation interval A", and is represented by $o_b = overlappedBy(A, Q) = |A \cap Q|/|Q|$.

The universe X in our case is defined to be infinite, which means that $|A|$ i.e. the cardinality[1] of A is defined by $|A| = \int_x \mu_A(x)dx$. Hence we get:

$$o_t = overlaps(A, Q) = |A \cap Q|/|A| = \frac{\int_x \mu_{A \cap Q}(x)dx}{\int_x \mu_A(x)dx} \tag{4.4}$$

Calculating o_t (or o_b) intuitively amounts to computing and dividing the integral areas of the two membership functions in the formula.

For example, the situation depicted in figure 4.13 is based on a fuzzy temporal interval Q="Roman Age" that intersects with another fuzzy temporal interval A="Pre-Roman Age". The two intervals are modeled as follows[2]:

$$\begin{aligned} A_{fuzzybegin} &= 510 \;\; B.C. \\ A_{begin} &= 490 \;\; B.C. \\ A_{end} &= 222 \;\; B.C. \\ A_{fuzzyend} &= 89 \;\; B.C. \end{aligned} \tag{4.5}$$

$$\begin{aligned} Q_{fuzzybegin} &= 222 \;\; B.C. \\ Q_{begin} &= 89 \;\; B.C. \\ Q_{end} &= 452 \;\; A.D. \\ Q_{fuzzyend} &= 569 \;\; A.D. \end{aligned} \tag{4.6}$$

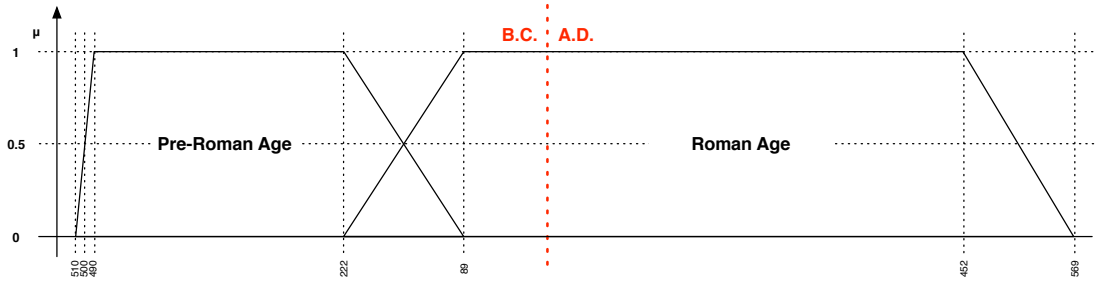


Figure 4.13: *A fuzzy temporal interval "Pre-Roman Age" and its intersection with another fuzzy temporal interval "Roman Age".*

The results of calculation of the values of the functions *overlap* and *overlappedBy* between Q and A are:

$$overlaps = \frac{|A \cap Q|}{|A|} = 0.0974\ldots \approx 0.1 \tag{4.7}$$

[1] See Zimmermann (1996, p.16) for a discussion about cardinalities of fuzzy sets.

[2] The criteria used for this example are a personal choice related to the case study chronology discussed in chapter 6.

$$overlappedBy = \frac{|A \cap Q|}{|Q|} = 0.0505\ldots \approx 0.05 \tag{4.8}$$

These calculations show that "Roman Age" overlaps around 10 percent of "Pre-roman age", while "Pre-Roman Age" is overlapped around 5 percent by "Roman age". The values obtained can be used for measuring the relevance of the annotation period A with respect to the query period Q, and these measurements in turn can be used for defining information retrieval mechanisms based on imprecise temporal intervals. However, a potential problem is that in cases where two intervals do not overlap but are still close, the results would be unsatisfactory. For this reason a **closeness function** *c* between Q and A is introduced as a defuzzified value of their distance *D*. *D* in turn is a fuzzy temporal interval and is calculated using a fuzzy subtraction operation $\ominus$ (Dubois and Prade, 1988, p. 50) between Q and A, which can be defined using the Left-Right notation (LR) as follows:

$$\begin{aligned} D = Q \ominus A \quad &= (Q_{begin} - A_{end}, \\ &\quad Q_{end} - A_{begin}, \\ &\quad \alpha_Q + \beta_A, \\ &\quad \beta_Q + \alpha_A)_{LR} \end{aligned} \tag{4.9}$$

The fuzzy extensions α and β for Q and A can then be calculated as shown in 4.10, and then used in order to get the fuzzy temporal interval *D* expressing the distance as shown in 4.11:

$$\begin{array}{lll} \alpha_Q = Q_{begin} - Q_{fuzzybegin} & = 89\ \ B.C. - 222\ \ B.C. & = 133 \\ \beta_Q = Q_{fuzzyend} - Q_{end} & = 569\ \ A.D. - 452\ \ A.D. & = 117 \\ \alpha_A = A_{begin} - A_{fuzzybegin} & = 510\ \ B.C. - 490\ \ B.C. & = 20 \\ \beta_A = A_{fuzzyend} - A_{end} & = 89\ \ B.C. - 222\ \ B.C. & = 133 \end{array} \tag{4.10}$$

$$\begin{aligned} D = Q \ominus A \quad &= (89\ \ B.C. - 222\ \ B.C., \\ &\quad 569\ \ A.D. - 510\ \ B.C., \\ &\quad 133 + 133, \\ &\quad 117 + 20)_{LR} \\ &= (133, 1079, 266, 137)_{LR} \end{aligned} \tag{4.11}$$

In order to measure the closeness of the two intervals, i.e. obtain the value for the closeness function c, three main steps are required:

1. Distance D is defuzzified to a crisp value d_{df} e.g. by calculating the Center of Area (COA) (Zimmermann, 1996, pp. 212–214), and taking the absolute value of it. By using COA we aim to take into account the fuzzy parts of the temporal intervals. For example, the Mean of Maxima (MOM) (Zimmermann, 1996) only considers the area where $\mu = 1$ of the fuzzy set in defuzzification. In our example we get d_{df} = 503.11 for the defuzzified value of *D*.

2. The defuzzified distance —d_{df}— is normalized to d_n, meaning its value is between [0,1]. The normalization is made by dividing —d_{df}— with the maximum d_{max} of all defuzzified distances between all examined temporal interval pairs (in our case d_{max} = 1111). After this step, the closer d_n is to 0, the closer the two intervals are considered to each other.

3. The closeness measure c is calculated as $c = 1 - d_n$. As a result c gets values close to 1 (i.e. better values[1]) when d_n is close to 0 (i.e. when intervals are close to each other).

The full equation for calculating the closeness value from the defuzzified distance is:

$$c = closeness = 1 - \frac{|d_{df}|}{|d_{max}|} \tag{4.12}$$

which in our case results in

$$c = 1 - \frac{503.11}{1111.11} \approx 0.55 \tag{4.13}$$

Finally, the three measures that have been calculated, i.e. *closeness c*, *overlaps* o_t, and *overlappedBy* o_b can be combined in a relevance measure *r* whose values will be in range [0,1]:

$$r = \frac{w_c * c + w_{ot} * o_t + w_{ob} * o_b}{w_c + w_{ot} + w_{ob}} \tag{4.14}$$

w_c, w_{ot}, w_{ob} are the weights that can be introduced in the relevance measure. Therefore, the combined relevance measure is a weighted average of the three individual measures.

[1]Values for overlap and overlappedBy are similarly considered better when close to 1.

5

MANTIC: The Archaeology of Milan on the Semantic Web

This chapter introduces a case study for the research proposals that have been discussed in the previous chapter; the case study is related to the development of a semantic portal (MANTIC, sections 5.1 and 5.2) for the archaeology of the city of Milan in Italy. In particular, the characteristics of relevant and heterogeneous data sources taking part in MANTIC are described in section 5.3, while section 5.4 provides the discussion of the mappings of their metadata schemata to the CIDOC CRM.
Section 5.5 provides an in-depth discussion of the issues connected to the usage of the CIDOC CRM in real application scenarios that have been highlighted in section 4.2.2 and evaluated in MANTIC mappings, as well as the other relevant aspects that have emerged from our experimentation.

5.1 The *Milano Antica* Project

The city of Milan's important ancient history dates back to the 5th century B.C., when Insubrian Celts settled in the area for the first time. From that period on the city extended progressively, especially under Roman domination. During the Tetrarchy period, between 286 and 310 A.D., Milan became the seat of the Maximinian imperial residence, and it acquired great importance in the Roman Empire. Emperor Constantine emanated his edict granting freedom of worship to Christians in the city in 313 A.D.; while, from 374 to 397 A.D. Bishop Ambrose ruled the spiritual and political life of the city giving it a Christian physiognomy. However, during the 5th century A.D. the ancient city started its fall: the imperial seat was transferred to Ravenna (402 A.D.) after the invasion of the Visigoths, and shortly thereafter the city was sacked by Attila (452 A.D.) and finally destroyed by the Goth Uraia in 538-539 A.D.

This millenary history has a rich archaeological heritage (see Caporusso et al., 2007). Research activities in the field are a continuous and challenging activity, because urban works in the center of the city often reveal significant and fragmentary portions of the complex archaeological stratification of ancient Milan. Consequently, the activities concerning the documentation of these archaeological elements and their interpretation, as well as the communication and sharing of the research results to different audiences (from the specialist to the occasional tourist) are complex tasks, which necessarily involve the joint efforts of a number of different Institutions, such

as the local Archaeological Museum, the Regional Offices on cultural heritage, and the Universities.

Figure 5.1: *Excavations of the Santa Tecla basilica in the central area of the city in 1961 (Caporusso et al., 2007, fig. 245, p. 211).*

The "Milano Antica" (Ancient Milan) project is the development of a previous project funded by Regione Lombardia which concerned the study and experimentation of innovative archaeometrical techniques for the chronological attribution and certication of materials coming from the excavations in the city. The leading partner of that project is the University of Milano-Bicocca Centre for Dating (CUDaM[1]), which works in collaboration with the Archaeological Museum of Milan. The scope of that project has been recently extended to include a series of activities concerning the development of innovative ICT systems supporting the digital representation of the heritage, and the design of strategies and methods for multi-level access to information in digital media environments, have been defined.

The principal partner of this new initiative at the University of Milano-Bicocca is the Cultural Resources Management Research Centre (CResM[2]) which has been re-

[1] http://cudam.mater.unimib.it/
[2] http://www.cresm.unimib.it/

cently set up in order to aggregate research competencies on cultural heritage that are already present in different Research Centres and Departments at the same University. Among these are the Complex Systems and Articial Intelligence Research Centre (CSAI[1]) and the Innovative Technologies for Interaction and Services Laboratory (ITIS[2]) which are involved in research on IT models and technologies supporting cultural heritage activities since a few years.

"Milano Antica" is part of an Agreement Program among the Regione Lombardia - General Directorate for Cultures, Identities and Autonomies, the University of Milano-Bicocca (CResM), the University of Milano, the University of Pavia, the Politecnico University of Milano, the National Research Council (CNR) and the National ACLI Agency for Professional Education (ENAIP). This program aims to define a new regional pole for the valorization of cultural heritage (Regione Lombardia, 2007) which among other activities foresees the development of specific actions related to applied research, innovation and technology transfer.

5.2 MANTIC: a Semantic Portal for *Milano Antica*

The "Milano Antica" project is complex and ambitious, and it embraces a long term perspective where a number of complementary activities will be part of a composite research and development process. On the other hand, when evaluating the overall scenario, two basic but still fundamental aspects have been identified:

- Relevant data and information are produced, managed and owned by different institutions (as well as by non-institutional subjects) that use heterogeneous models and systems.
- There is a lack of appealing and highly interactive applications for the fruition of these data and information. Therefore, new and engaging applications must be developed in order to involve users in the discovery of the archaeological heritage of the city.

These aspects led to the evaluation of an approach that is substantially based on the new Web technologies, and the design and development of a semantic portal ("MANTIC") is being proposed.

The definition of the preliminary requirements, which has been made in collaboration with the project partners, took three main objectives into consideration:

1. To support an effective and "archaeologically aware" visit of the heritage that is still preserved in the urban area, through the deployment of systems making digital information easily accessible and richly contextualized.

[1] `http://www.csai.disco.unimib.it/`
[2] `http://siti-server01.siti.disco.unimib.it/itislab/`

2. To stimulate users of the systems in discovering (rather than searching for) relevant information according to different profiles and skills. A central concern should be the design and the development of highly dynamic interaction experiences mediated by the new technologies.

3. To define new modalities for a timely and extensive communication and sharing of the research results beyond traditional paper publishing. This way, specialized research should benefit from the quasi-immediate availability of new resources and data, and the public awareness of the richness of the archaeological heritage of the city should be dramatically improved.

Moreover, a set of basic dimensions according to which data should be represented, displayed and retrieved in MANTIC have been defined, according to four main classes of properties:

- spatial properties: the location, extent and spatial relationships each structure had and currently has within the context of the city
- temporal properties: the phases that characterized each structure during time, including eventual changes in its functional use and its post-abandonment transformations
- qualitative properties: describing, for example, differences in functions, materials, styles, etc.
- causal properties: the relationships among the activities carried out on and about the cultural heritage

Finally, it has been suggested that information collected in the portal should embrace both a traditional perspective concerning the "archeological time" and the past life of the heritage, and a contemporary perspective connected to the activities of study, preservation and valorization. In fact, the co-occurrence of this information makes it possible not only to have a comprehensive historical view of the heritage, but also to appreciate the activity of different actors on it (be they single persons or Institutions) which often receives minor attention.

The "Milano Antica" project, and specifically the MANTIC initiative, therefore represent an ideal context and case study for the application, evaluation and discussion of the research elements and proposals that have been introduced in the previous chapter. In fact, our PhD research contributes to this scenario by taking into consideration aspects of fundamental relevance for MANTIC and in particular, the integration of existing and heterogeneous data sources and the evaluation of innovative forms of information retrieval according to the temporal dimension. The first aspect is the specific object of this chapter, while the second is discussed is chapter 6. In addition, the experimentation of a prototype system, and especially of dynamic Web interfaces supporting the navigation in the semantic repository, will be described in chapter 7.

5.3 Selection of Relevant Data Sources

A number of different Institutions (notably the Regional Offices for cultural heritage, the Archaeological Museum of Milan, the Regional Directorate for cultural heritage and Universities) produce high quality and scientific data for different purposes, from the creation of general inventories of cultural heritage to specific databases concerning e.g. the archaeometrical analyses on the preserved ancient structures. On the other hand, a relevant body of contributions comes from non-institutional subjects, such as media organizations, associations and individual citizens in the form of Web sites [1] or more frequently in the form of user generated content. This latter is created and published mostly using Web 2.0 applications: a growing quantity of specific and highly heterogeneous content is available, for example in the form of Wikipedia pages, Flickr image sets (fig. 5.2), Wikimapia polygons (fig. 5.3), and discussions on Facebook groups[2].

Naturally, the quality and trustworthiness of these contents vary a lot in reason of the different subjects producing them, and the specific objectives of content creation. Therefore, it is fundamental, at least for the initial development of the MANTIC portal, to select the most relevant and scientifically trustable data sources. Our selection took into consideration those coming from the most important Institutions dealing with the archaeological heritage of Milan. The reason for this choice is strictly connected to the characteristics of the data, and in particular their availability, their scientific quality, and the heterogeneity of their scopes, granularities and formats. In fact, these characteristics make the case study particularly stimulating with respect to the problems of information integration through the CIDOC CRM domain ontology. Moreover, each data source takes into consideration different categories of archaeological items, and the integration of this information for the first time will allow to connect different archives on the archaeology of Milan and to create an integrated corpus of semantically related data on which to design and test transversal forms of access and retrieval.

The data sources that have been selected for the project are:

- the SIRBeC database by Regione Lombardia, containing data about archaeological artifacts
- the IDRA database by the Regional Directorate on Cultural and landscape Resources, containing data about archaeological excavations
- the MANTIC database, which was created by using additional digital material provided by the Archaeological Museum of Milan, concerning the archaeological sites, the structures, and the monuments of the ancient city

[1] For example: `http://www.serenoeditore.com/milano/archeo/`

[2] `http://www.facebook.com/pages/Milano-Italy/Museo-Archeologico-di-Milano/100194754615/`

Figure 5.2: *Archaeological excavations in Milan: photos posted by a Flickr user (*`http://www.flickr.com/photos/7581350@N02/`*).*

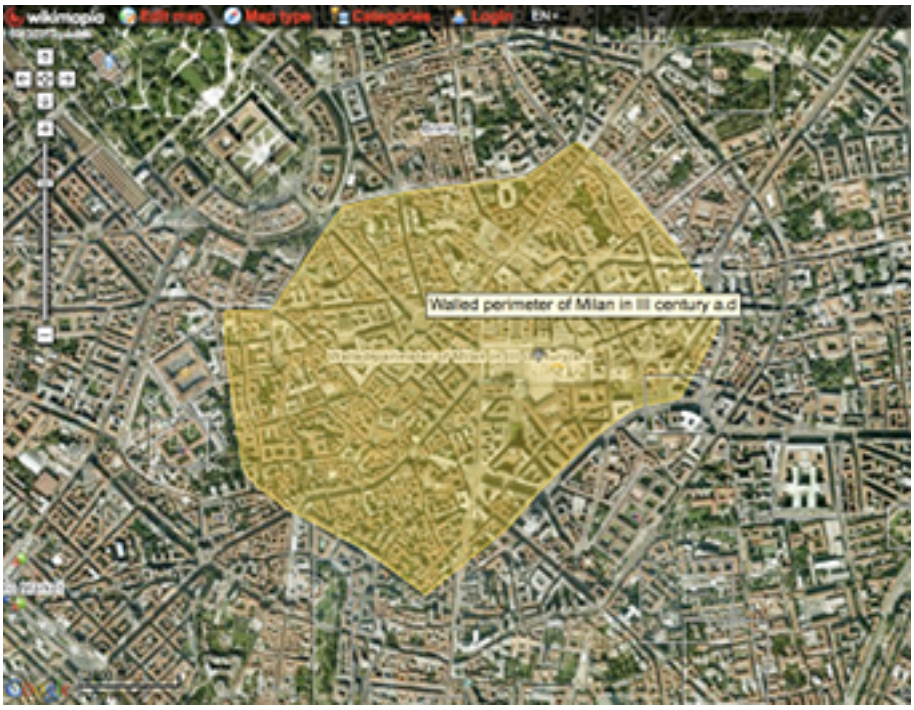


Figure 5.3: *A polygon related to archaeological features of Milan on Wikimapia (*`http://wikimapia.org`*).*

5.3.1 SIRBeC

The Regione Lombardia is an extremely active promoter of cultural initiatives, which range from cultural events to libraries, museums and heritage access, and specialized training[1]. Within these activities the cataloging and valorization of the local cultural heritage[2] represents a case of excellence in the Italian scenario and can boast a consolidated experience in the use of ICT methods and tools. SIRBeC[3] (the Regional Information System on cultural heritage) is the main technological platform created by the Regione Lombardia for cataloging cultural heritage: since 1992 it has collected continuously updated information about the items that are present in the territory or are preserved in museums, collections and other cultural institutions.

A list of the categories that fall under this activity clearly shows how extensive the cataloging is:

- architectures: monumental complexes, public and religious buildings, rural historical buildings, aristocratic homes, private buildings, industrial archaeology buildings
- art objects: paintings, drawings, sculptures, religious and liturgical items, furniture, textiles
- photographs: photographic collections of artistic, historical, or documental interest
- prints and engravings: prints and engraving matrices of historical-artistic relevance
- archaeological artifacts: ceramics, coins, jewellery, epigraphs, glyptic, mosaics, glass, weapons
- etno-antrophological items: work tools, domestic and personal objects, popular art, toys
- scientific and technological heritage: tools, machines and all the items that are relevant for the history of science and technology

Since 1998 the system has become compatible with the national cataloging standards defined by the Central Institute for Cataloging and Documentation[4] (ICCD), which is the body within the Italian Ministry for cultural heritage and Activities (MiBAC) promoting the creation of a unified national catalog of cultural heritage, and defining the cataloging standards and tools in agreement with the Italian Regions.

[1] See: `http://www.lombardiacultura.it/`
[2] `http://www.lombardiabeniculturali.it/`
[3] `http://www.lombardiabeniculturali.it/sirbec/`
[4] `http://www.iccd.beniculturali.it/`

Every year, the Regione Lombardia supports cataloging projects using the SIRBeC softwares. Between 1992 and 2005 nearly 350 projects have been undertaken, with the collaboration of nearly 200 partners, which include more than 100 museums, as well as public and personal collections. As of November 2008 (Regione Lombardia, 2008, p. 4) nearly 600.000 items were already archived. The control over the quality of the cataloging process and the analysis of the possible improvements of the system represent a key activity, which is the object of specific and extensive reports (Degiarde, 2007).

Each item is described by specialists in the single disciplines, taking into consideration information such as typology, materials and techniques, denomination, author, current location, and period. Information is then further detailed on the basis of the specific characteristic of the category the object belongs to, and it is enriched by one or more images depicting it. The cataloging process is very articulated; it is composed of different stages where intermediate and preliminary data are progressively refined towards the creation of final high quality datasets. Figure 5.4 provides a simplified schema of the process: first the sets of items to be cataloged and the groups of experts ("operative unit") in charge of it are identified, and preliminary datasets are produced; then, the datasets are examined in order to verify their validity with reference to eventual missing information and the adherence to the vocabularies and norms defined by the cataloging committee. This process is iterative and may be performed at different stages and times, also depending on priorities and funding possibilities. Once the dataset has been verified and normalized, it is approved as an official cataloging source, and very rich and detailed cards are made available.

Catalogers use SIRBeC software running on local machines; the produced files, both in their preliminary or final form, are progressively uploaded into a central database management system, which is hosted on the Regione Lombardia servers.

Over the years initiatives concerning the online publishing of the SIRBeC datasets have been promoted (e.g. Gagliardi, 2003); in March 2006, the "Lombardia Beni Culturali" website[1] was launched in order to allow access to simplified versions of the cataloging carts, concerning approximately 65.000 items. A successive agreement was defined in 2008 in order to develop an integrated Web portal for the cultural heritage of the Lombardia Region (Regione Lombardia, 2008), linking together the existing Web site and other online archives concerning the history of the Region[2], its libraries [3] and the Archive of Etnography and Social History[4] that collects oral documents, text transcripts, musical transcripts, audio documents and photographs. Even if the new Web site is well designed with respect to current Web standards, such as CSS and XHTML, it relies on a proprietary data schema, and it does not provide any means for easily

[1] http://www.lombardia.beniculturali.it/

[2] http://www.lombardiastorica.it/

[3] http://www.biblioteche.regione.lombardia.it/OPACRL/cat/SF/

[4] http://www.aess.regione.lombardia.it/index.htm

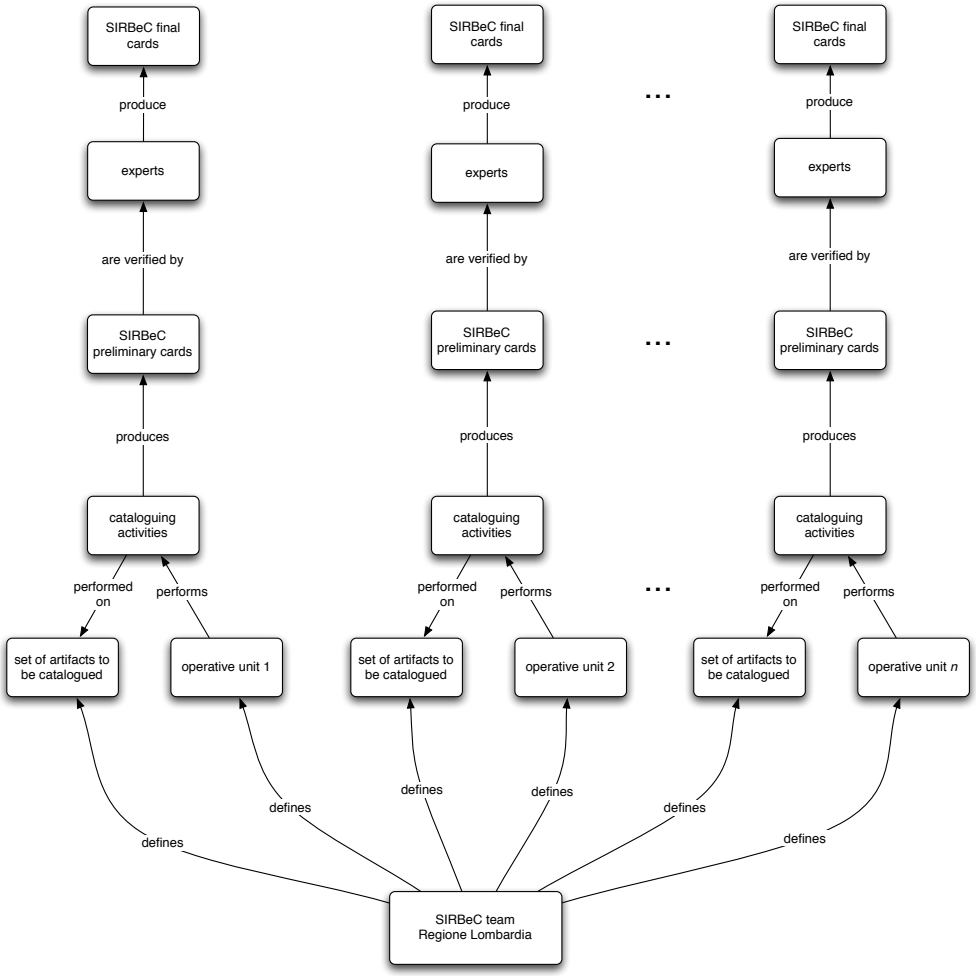


Figure 5.4: *A general schema of the SIRBeC cataloging process.*

accessing information by external applications (e.g. through Web services), therefore making it very difficult to integrate data in external applications.

SIRBeC represents an extremely important and authoritative data source for the MANTIC project. Thanks to the support of the SIRBeC team, an evaluation of the quantity and nature of archaeological data has been made; the database currently contains data about approximately 400 artifacts related to the period of the ancient city; these artifacts were discovered occasionally during public works or in the context of regular archaeological research projects. The system provides a means to collect data concerning structures, monuments, and archaeological areas; however this activity has not yet been done; specific cataloging campaigns are planned for the near future.

An agreement was made with the data owners (which include the Regione Lom-

bardia, as well as the the Archaeological Museum of Milan, and the University of Milano-Cattolica) in order to a obtain copy of relevant data for the case study. SIRBeC data have been exported as a dump in MS Access format, together with the photographs that are attached to each single artifact card.

5.3.2 IDRA

IDRA[1] is the Information Database on Regional Archaeological-Artistic-Architectural heritage which connects several databases related to Lombardia's heritage, which were created within the activities of the Ministry offices in the regional territory. The system aims to make information about the cultural heritage accessible in a unified and coordinated format, both for the general public and for cultural heritage professionals.

IDRA collects information about different categories of archaeological items, from sites to artifacts. Information can be visualized and queried through traditional interfaces based on drop down lists and single detailed cards, as well as through interactive maps. In addition to this, a sub-project is dedicated to the archaeological map of Milan, which filters IDRA data and allows the users to visualize and search them within a WebGIS system.

The archaeological data come from bibliographical and archival sources, in particular from the archives of the Superintendency of Archaeological Heritage of Lombardia, which is moreover the promoter of the archaeological map project.

Due to the complexity of the IDRA archive, a limited set of relevant data coming from the system has been semi-automatically obtained from the Web site, in order to provide a complementary source, with specific reference to information about archaeological excavations, and it has been stored in a MS Excel spreadsheet. These data concern excavations made in the city, together with information about the chronological attribution of the excavated sites and different kinds of other cataloging parameters.

Formal agreements with the institutions that maintain the IDRA system are currently being made in order to evaluate the characteristics of the databases in depth, and to proceed with a more extensive integration of the available data. Here the mapping of the spreadsheet data is discussed, which gives interesting contributions to the scenario that characterizes the MANTIC project.

5.3.3 MANTIC

During a preliminary experience concerning the experimentation of different Web navigation and presentation methods (such as faceted browsing, and interactive maps), a database was created containing synthetic information about the archaeological sites, structures and monuments of the city. This database represents an interesting comple-

[1] http://www.lombardia.beniculturali.it/Page/t01/view_html?idp=96

mentary source to SIRBeC and IDRA; therefore it has been considered as a relevant source to be integrated with the others.

More specifically, the database is available as a PostgreSQL application and collects the information and data contained in the most recent and comprehensive publication concerning the archaeology of Milan that has been produced by the Archaeological Museum (Caporusso et al., 2007). Most of the materials, including maps and photographs, have been made available by the Museum's staff, thus making the population of the database easier and error-free.

MANTIC collects information about:

- archaeological sites
- single structures, such as tombs or Roman houses (*domus*)
- major monuments, such as the Roman amphitheatre, the Roman circus, and the ancient churches

Basic data concerning these items are available, such as their chronological attribution, and their denomination. In the case of major monuments and complexes, additional information is available, such as a brief description, and the current preservation state. In addition to this, different kinds of spatial information have been acquired, both in terms of indications of modern addresses and in terms of georeferenced points that were created from graphic files concerning the positioning of the archaeological items on the modern cartography.

5.4 Conceptual Mapping to the CIDOC CRM

The mapping activity, which has been carried out according to the workflow introduced in section 4.2.2.1, brought to the identification of top-level events concerning:

- the production, use, modification and destruction of items (be they artifacts, features, stratigraphic units, buildings, etc.)
- the activities of archaeological research in the field (e.g. the excavation)
- the discovery of items
- the documentation and cataloging processes, which also include the interpretations given by different scholars during their research and study activities
- the definition of legal constraints, e.g. for the assessment of the legal property of items or the safeguard of archaeological sites

The following sections describe the details of the mapping activity. The extensive documentation of the mapping templates according to the criteria introduced in section 4.2.2.1 is provided in appendix A.

5.4.1 SIRBeC Mapping

SIRBeC implements the MiBAC "Reperto Archeologico" (RA) card[1] for cataloging archaeological artifacts. Minor custom extensions have been created by the SIRBeC team for the specific necessities and peculiarities of the local cataloging activity; however, they will not be discussed here but can be evaluated by consulting the official documentation in Lombardia Informatica s.p.a (1999). The RA card is extremely rich and detailed: it contains more than 250 fields which are selectively used on the basis of the specific characteristics of each single artifact, which can vary from a simple pottery sherd to more complex items, such as statues or inscriptions. According to the objectives and scopes of the MANTIC project, approximately 40 fields were selected. Table 5.1 lists the selection and presents it according to the hierarchical structure and the scope notes contained in Lombardia Informatica s.p.a (1999).

Table 5.1: *List of the metadata that have been selected from the SIRBeC-RA card schema.*

Identifiers	
UOP	Operative unit
NSK	Card number
Current localization	
Specific localization	
LDCM	Name of the collection the object belongs to
Discovery	
I.G.M. coordinates	
LGII	Institute
LGIT	Name of the tablet
LGIN	Sheet number
LGIQ	Quadrant code
Cadastral parcels	
LGCC	Name of the Municipality
LGCM	Code of the sheet
LGCA	Date of the sheet
LGCR	Codes of the parcels
SPR	Details on the discovery of the object
Excavation data	
DSCF	Author of the archaeological excavation (legal body)
DSCD	Date of the archaeological excavation
Object	
Object	

continued on the next page

[1] http://www.iccd.beniculturali.it/Catalogazione/standard-catalografici/normative/scheda-ra/

OGTD	Typology of the object
OGTT	Sub-typology of the object
CLS	Class and production attribution
Chronology	
General chronology	
DTZG	Chronology (century)
DTZS	Chronology (portion of century)
Technical data	
MTC	Material and technique
Measures	
MISU	Measurement unit
MISA	Height
MISL	Width
MISP	Depth
MISD	Diameter
MISN	Length
MISS	Thickness
MISG	Weight
Analytical data	
Description	
DESO	Indications on the object
DESS	Indications on the subject
Inscriptions	
ISRL	Language of the inscription
ISRI	Transcript of the inscription
Sources and reference documents	
Linked images	
IMCX	Name of the image file
Region images	
IMRT	Typology of the image
Compilation	
Compilation	
CMPD	Date of compilation
CMPN	Name of the compiler
FUR	Name of the officer in charge

The analysis of the metadata schema and the actual values contained in SIRBeC, allowed to group data according to three main activities: the production of an artifact, its discovery and its classification and cataloging. These activities, together with the

related sub-activities, are represented in figure 5.5, and the detialed sub-graphs are shown and discussed in the following sections.

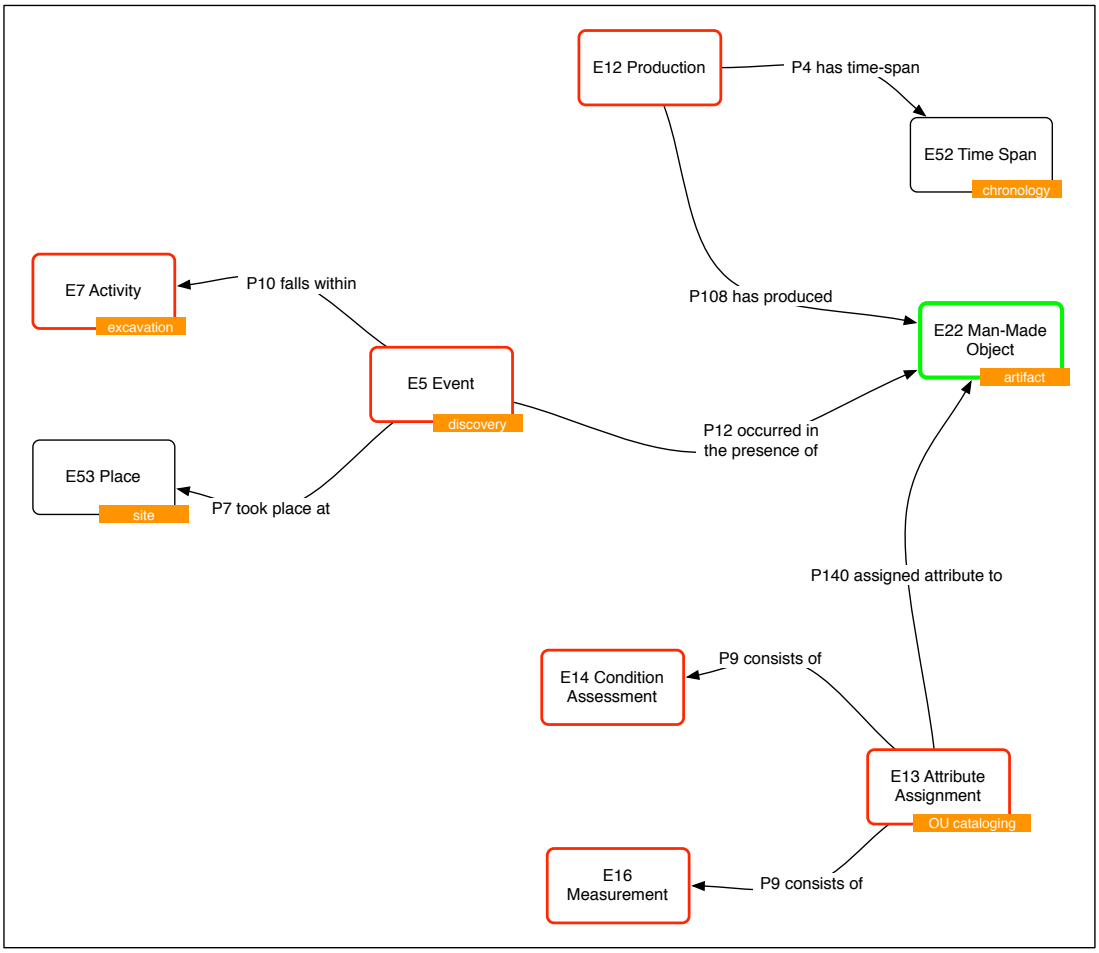


Figure 5.5: *A graph representing the main activities identified in SIRBeC.*

5.4.1.1 Data About the Production of Artifacts

The CRM offers different classes and properties for the precise representation of the activities carried out on a physical object. For the representation of the production of an archaeological artifact the `E11 Modification` and the `E12 Production` classes have to be evaluated. The CRM definition of `E11 Modification` states that "*This class comprises all instances of E7 Activity that create, alter or change E24 Physical Man-Made Thing. Since the distinction between modification and production is not always clear, modification is regarded as the more generally applicable concept. This implies that some items may be consumed or destroyed in a Modification, and that others may be produced as a result*

of it" (Crofts et al., 2009, p. 7). Instead `E12 Production`, which is a sublaclass of `E11 Modification` "'*...comprises activities that are designed to, and succeed in, creating one or more new items. It specializes the notion of modification into production. The decision as to whether or not an object is regarded as new is context sensitive.*" (Crofts et al., 2009, p. 7). In the case of SIRBeC, as well as in most archaeological documentation, the chronological attribution of an object usually refers to the production of the object itself, rather than its use (which in exceptional cases can even continue in contemporary times); therefore, the `E12 Production` has been used here.

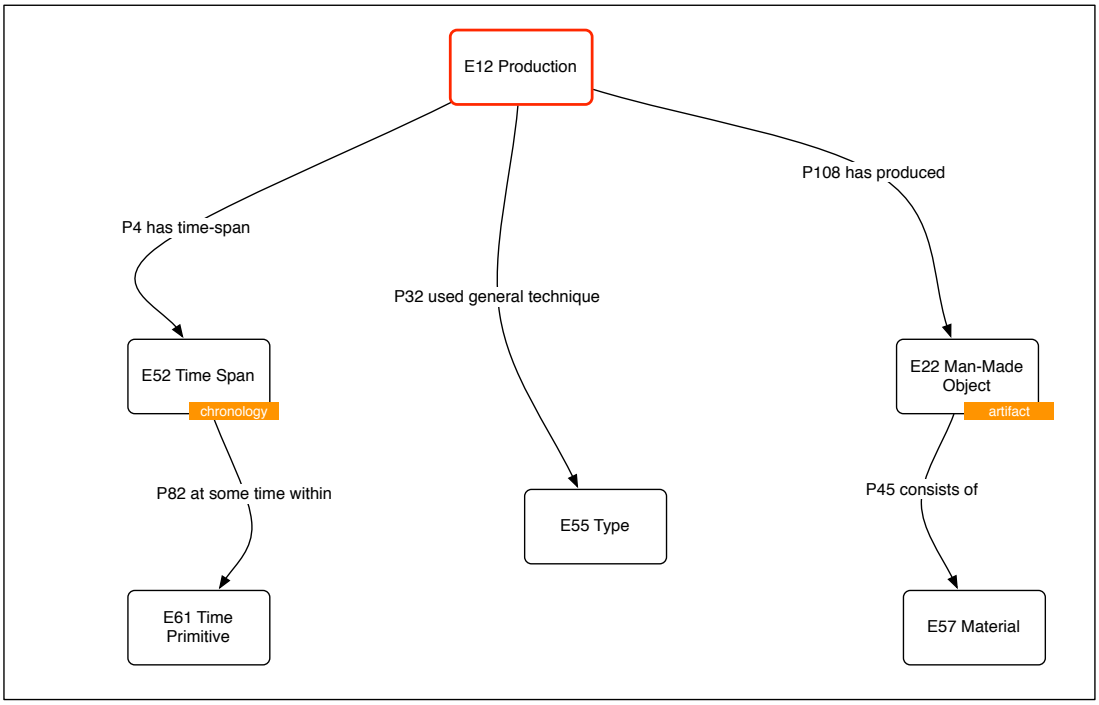


Figure 5.6: *The production of artifacts in SIRBeC and the corresponding graph in CIDOC CRM.*

The chronological attribution of an artifact (with respect to its production) is represented in SIRBeC by two fields, which together define the time span. DTZG contains the indication of the centuries (e.g. I B.C. – I A.D.), while DTZS refines this general indication with portions of the centuries (e.g. beginning/fisrt half). The combination of the two provides the complete chronological information (e.g. beginning I B.C. – first half I A.D.). This choice creates some problems with respect to the CRM, in terms of the direct mapping of these fields to its classes and properties; in reality this difficulty stresses the fact that a bad design choice was made for the SIRBeC schema. Consequently, methods would be required for managing this case, and the template table provides an unified mapping of the twos to a CRM chain. The time span is

linked to the time primitive representing the actual temporal interval through the `P82 at some time within` property, which allows to define uncertain temporal extents within which the time span falls, and specifically the interval with the maximum possible extent.

On the contrary, the MTC field groups together two different kinds of data, concerning respectively the technique used in the production and the material the artifact is made of. Methods would also be required for this situation, because data contained in MTC refer to different classes and properties of the CRM, and specifically `P32 used general technique` and `E55 Type`, and `P45 consist of` and `E57 Material`, which moreover relate to different classes, and respectively `E12 Production` and `E22 Man-Made Object` (see fig. 5.6).

An alternative representation choice can be to treat these data as text annotations attached to the artifact, by using the `P3 has note` property and the `E62 String` class. However, this means loosing significant data, and dramatically reducing the possibility to cross-link and retrieve information about the artifacts on the basis, e.g. of the material they are made of.

5.4.1.2 Data About the Discovery of Artifacts

The discovery of an artifact can happen by chance or during regular research activities, such as an archaeological excavation. In SIRBeC, the circumstances during which the discovery took place are described in free text form in the SPR field. This makes it very difficult or even impossible to automatically distinguish between the different cases, and the SPR has to be mapped as a note using the `P3 has note` property and the `E62 String` class.

However, the fact that an artifact was discovered during an archaeological excavation may be deduced by the presence of data in the DSCF and the DSCD field, which contain respectively the appellation of the body who conducted the excavation, and the date when it took place. Therefore, in all cases where at least one of these data are present, an instance of `E7 Activity` representing the excavation needs to be created, and the discovery event can be linked to it through the `P10 falls within` property.

The discovery of the artifact, be it occasional or in the context of an archaeological excavation, happened at a given spatial location (which can be considered as a "site"), which in SIRBeC is defined by two different kinds of geographical coordinates. In particular LGII, LGIT, LGIN, and LGIQ refer to the I.G.M. (Military Geographical Institute (Istituto Geografico Militare) system, while LGCC, LGCM, LGCA, and LGCR refer to the cadastral system. In both systems one field is related to the other in a hierarchy that provides increasingly accurate spatial information; for example, the name of the tablet (LGIT) is followed by the number of the sheet within the tablet (LGIN), which is in turn followed by the code of the quadrant within the sheet (LGIQ). The mapping template table groups the sets of data related to the two different systems in order to provide a more concise view of them.

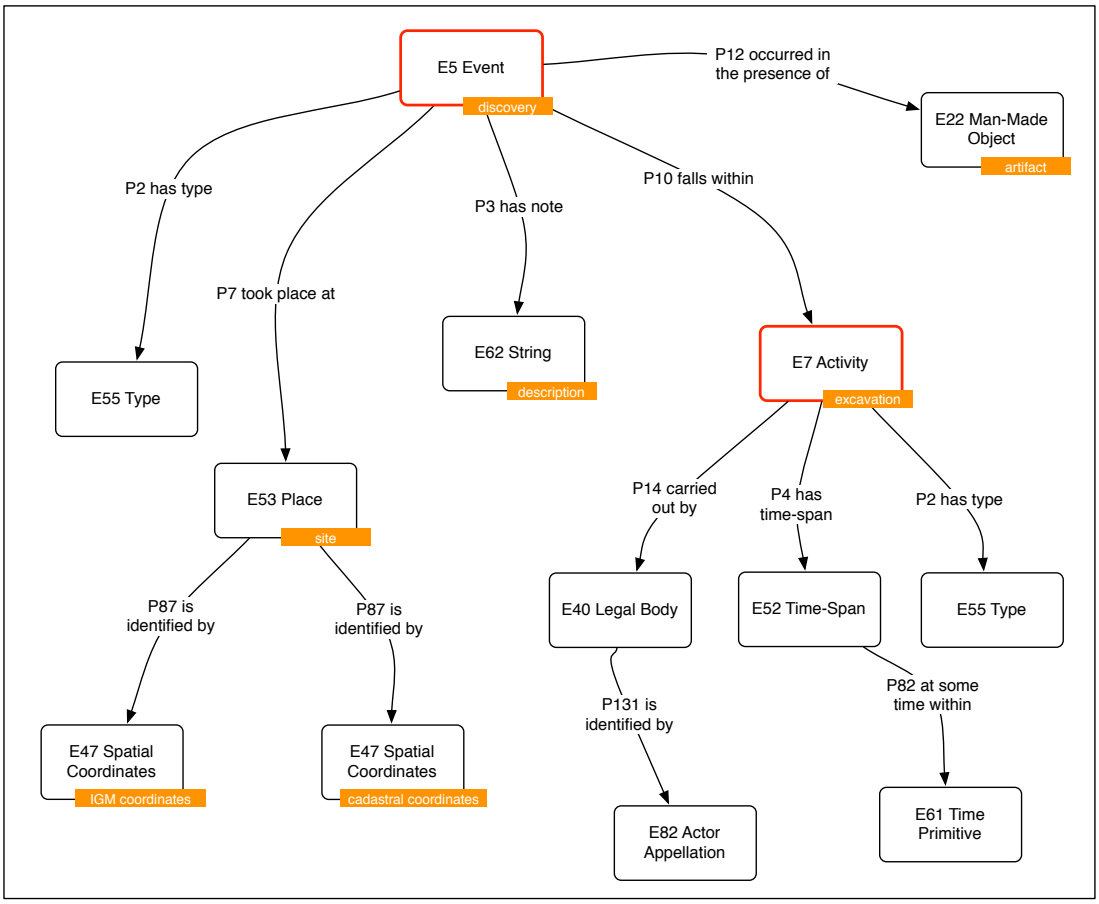


Figure 5.7: *The discovery of artifacts in SIRBeC and the corresponding graph in CIDOC CRM.*

Finally, it is interesting to stress how the discovery event relates to the artifact through the `P12 occurred in the presence of` property. This is coherent with the event-centric perspective of the CRM, even if it seems counterintuitive, especially with reference to the traditional methods of archaeological documentation. On the other hand, the introduction of an additional "discovered" property would mean defining a "Discovery" sub-class as a specialization of the `E5 Event` class, which in turn would require analogous choices for other specific events (e.g. the archaeological excavation) and raise issues of compliancy of the extensions with the original model.

5.4.1.3 Data About the Cataloging Process

It is no surprise that most part of the data contained in SIRBeC are related to the cataloging process, since the system has been created for this specific purpose. As

was introduced in section 5.3.1, the cataloging process is very articulated, and it can be divided into a number of activities carried out at different times and eventually by different actors. On the other hand, the event-centric perspective of the CRM offers a number of possibilities to model the process with different levels of details: several classes and properties exist for this, from the more general (e.g. `E7 Activity`) to the more specific (e.g. `E13 Attribute Assignment`, `E17 Type Assignment`, `E14 Condition Assessment`). The choice of the most suitable classes and properties strongly depends on the available documentation: "*Which kind of such assignments and statements need to be documented explicitly in structures of a schema rather than free text, depends on if this information should be accessible by structured queries*" (Crofts et al., 2009, p. 8).

With reference to the schema represented in figure 5.4, one choice can be, for example, to define two main kinds of related activities: one for the cataloging process as a whole, i.e. the process promoted by Regione Lombardia, and the other for the single activities carried out by each operative unit. These can be modeled using respectively the `E7 Activity` class, and the `E13 Attribute Assignment` class, as is shown in figure 5.8.

However, information about the Regione Lombardia process is not directly contained in the data schema, and would need to be manually added; for this reason, the representation takes into consideration only the `E13 Attribute Assignment` activities related to the Operative Units. These, in turn, consist in the definition of an identifier for the object, and in the assignment of a type and a class to it; moreover two sub-activities are related to the assessment of the object's condition state, and to various measurements (height, depth, length, weight, etc.). This representation allows to better explicit the cataloging process as it actually is, and provides means for identifying the observations and attributions on the object each operative unit makes.

In fact, the specific characteristics are associated to the object through the activity of a specific cataloging group. This way the representation of different interpretations made by different groups (and possibly in different contexts and projects) is possible, and methods can be designed for the comparison of e.g. contrasting hypotheses. The CIDOC CRM strongly encourages this approach; the scope note of `E13 Attribute Assignement`, for instance says that "*This class comprises the actions of making assertions about properties of an object or any relation between two items or concepts. This class allows the documentation of how the respective assignment came about, and whose opinion it was. All the attributes or properties assigned in such an action can also be seen as directly attached to the respective item or concept, possibly as a collection of contradictory values.*" (Crofts et al., 2009, p. 8).

On the other hand, the mapping template tables represented in appendix A clearly show how this approach makes the complexity and quantity of the mapping chains grow considerably, mostly because most of the information concerning the cataloging process is only implicitly represented in the SIRBeC schema.

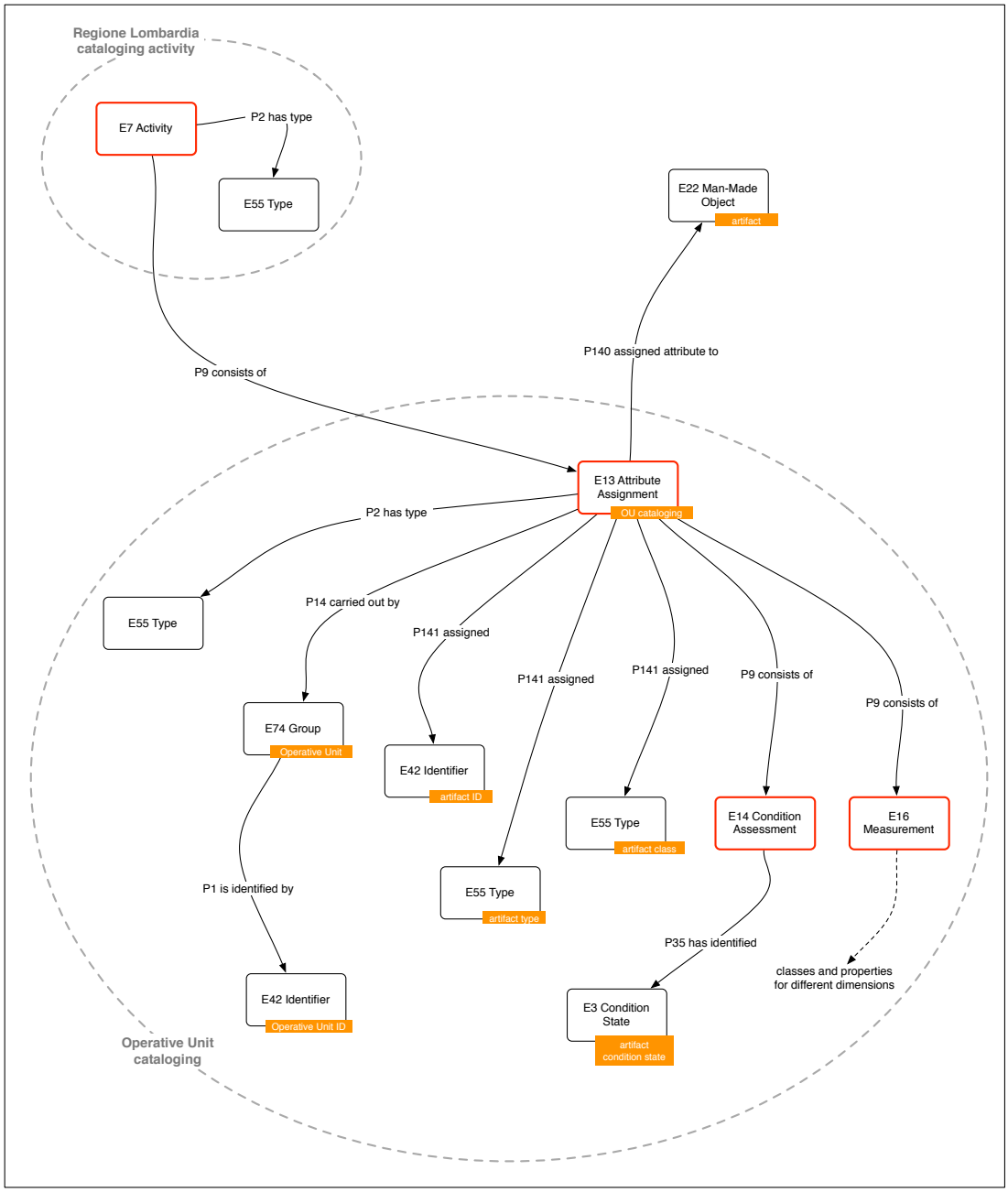


Figure 5.8: *A simplified view of the cataloging process in SIRBeC and the corresponding graph in CIDOC CRM.*

In addition to this, there are situations where apparently simple data (such as those related to the measurements) involve the use of a number of the CRM classes and properties. For example, weight measurement is expressed in terms of a measurement unit and a numeric value in the MISG field, e.g. "kg 22.7". The resulting mapping chain involves the use of the `E54 Dimension` class, which is linked to the actual numeric value through the `P90 has value property` to the measurement unit through the `P91 has unit` property and to its characterization as a weight measurement through the `P2 has type` property. Moreover, several issues are related to the SIRBeC schema: the measurement unit for height, width, depth, diameter, length and thickness is represented in the MISU field (thus making it impossible, for example, to define different measurement units for each measurement), while the measurement unit for weight is contained in the MISG field, together with the actual measurement value (thus combining two very different data types into a single representation). In these cases, as with the data concerning chronology, techniques and materials that were discussed above, methods will be needed for the actual mapping to CRM triples chains.

Finally, there are some cases where the actual semantics of one field may differ from the asserted semantics described in the documentation. For example, the OGTD field usually contains information about the condition state of the object, while this information would be normally included in the STCC field.

The discussion of the SIRBeC mapping templates, clearly shows that the expressiveness of the model helps in precisely identifying the structure of the documentation and in making implicit information explicit and accessible. On the other hand, it makes the representation grow considerably, thus deeply influencing the choices that can be made when e.g. defining efficient methods for the retrieval of relevant information from a Web-based system. These issues will be discussed in more detail in section 5.5.

5.4.2 IDRA Mapping

Sample data from the IDRA system have been semi-automatically obtained from the system's Web site (see section 5.3.2). These data refer to a very limited yet relevant subset of the data schema; the stylesheet with the IDRA data is far less articulated than the relational structure of the SIRBeC system, and the mapping activity has been more simple and straightforward. Table 5.2 lists the selected fields of the IDRA system, which are presented according to the hierarchical structure that can be identified in the IDRA Web site.

Table 5.2: *List of the selected fields from the IDRA database.*

Codes	
Identificativo del bene	Item ID
Object	
General excavation data	
Tipo di evidenza	General typology of the archaeological items
Descrizione degli scavi	Excavation description
Autori degli scavi	Authors of excavation
Date degli savi	Excavation dates
Localization	
Geographical and administrative localization	
Indirizzo	Address
Cadastral Localization	
Cadastral Localization	
Riferimento Catastale	Cadastral Reference
Chronology	
General Chronology	
Secolo	Century
Legal condition and constraints	
Preservation acts	
Estremi provvedimento	Preservation act details
Cataloging	
Cataloging	
Data	Date
Nome	Name
Funzionario responsabile	Officer in Charge

Data have been grouped according to three main activities: the archaeological excavation, the cataloging process, and the definition of legal constraints on the archaeological site. These activities, together with the related sub-activities, are represented in figure 5.9, and the detailed sub-graphs are shown and discussed in the following sections.

5.4.2.1 Data About the Archaeological Excavation Activity

The IDRA Web site contains relevant information about the archaeological excavations that have been carried out in the city. For the purposes of this research data concerning

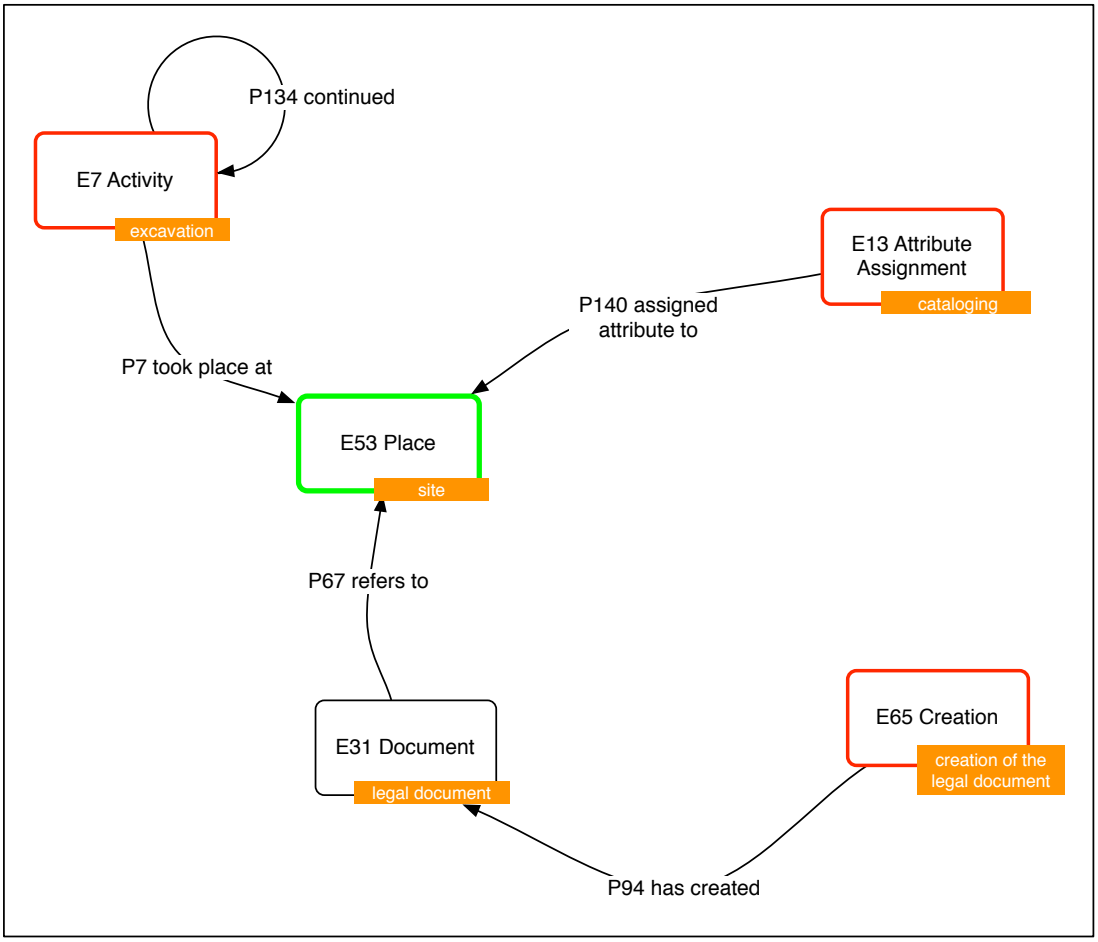


Figure 5.9: *A graph representing the main activities identified in IDRA.*

the identification of the archaeological site, the legal bodies that promoted and participated in the excavation, and the time spans during which the excavation took place have been selected.

A number of archaeological excavations can be carried out in succession on the same site. This kind of information is implicit in the sequence of the excavation dates, and it has been made explicit through the property `P134 continued`. This choice allows to represent a chain of related research activities in a meaningful temporal sequence, thus enriching the representation, and the possibilities of e.g. the retrieval of relevant information concerning the study processes.

Archaeological sites, which have been represented through the `E53 Place` class, can be identified by means of their current address, which is contained in the "Indirizzo" field. However, there are some cases where other indications are present in

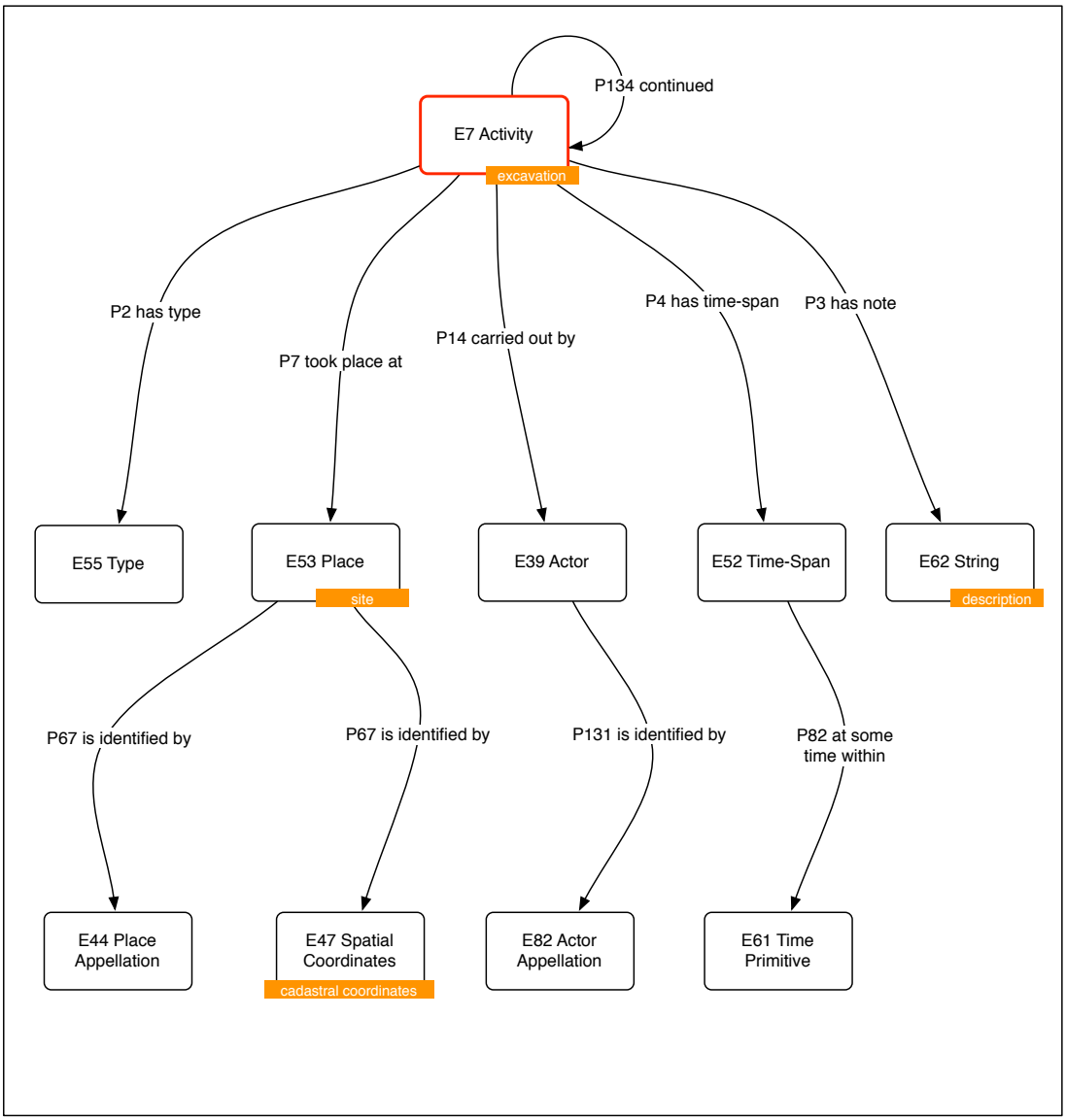


Figure 5.10: *The archaeological excavation activity in IDRA and the corresponding graph in CIDOC CRM.*

addition to the address. for this reason the more generic `E44 Place Appellation` class was used instead of `E45 Address`.

5.4.2.2 Data About the Cataloging Activity

Data about the cataloging activity are the most numerous, because of the nature and the purposes the IDRA system was created for. The representation of the cataloging activity is simpler than that of SIRBeC, but it makes use of similar classes and properties. This confirms the generality of the model, which allows coherent and sound representations of similar concepts coming from different sources.

The chronology of an excavated archaeological site, which is represented through a `E4 Period` that took place on a given place (site), has been presented here together with the cataloging activity. The choice is motivated by the fact that, in order to provide a sufficiently simple representation based on the available data, no specific activity was defined at this stage for the past history of an archaeological site. Successive and more extensive evaluations of the IDRA database would probably suggest the definition of more articulated representations, which at the moment would not provide specific advantages.

5.4.2.3 Data About the Definition of Legal Constraints on Archaeological Sites

The definition of legal constraints on the archaeological sites is one of the principal activities that are performed by the Institutions working on the preservation of cultural heritage. Therefore, the specific representation of this activity was deemed necessary. However a few details are provided in the data source, concerning mainly the legal documents which represent the end result of the legal constraint definition. In absence of further information about the entire process (such as the actors who promoted and participated in it, or the time spans during which this process was performed), the `E65 Creation` class was used to represent the document creation.

5.4.3 MANTIC Mapping

Since the MANTIC database was created from scratch for a very initial phase of the project (see section 5.3.3), and it is maintained by our research group, it shows the advantage of being open to schema modifications in order to ease the mapping to the CRM. On the other hand, it collects data coming from the activities of the Archaeological Museum of Milan (which have been made available in digital formats, and are presented in Caporusso et al. (2007), which are organized along a series of descriptive dimensions and concepts.

In particular, the available data concern: the archaeological sites (i.e. all the items that are defined as “area of findings”, “work area”, or “necropolis”) and the structures (i.e. all the items defined as “monument”, “domus”, or “city gate”).

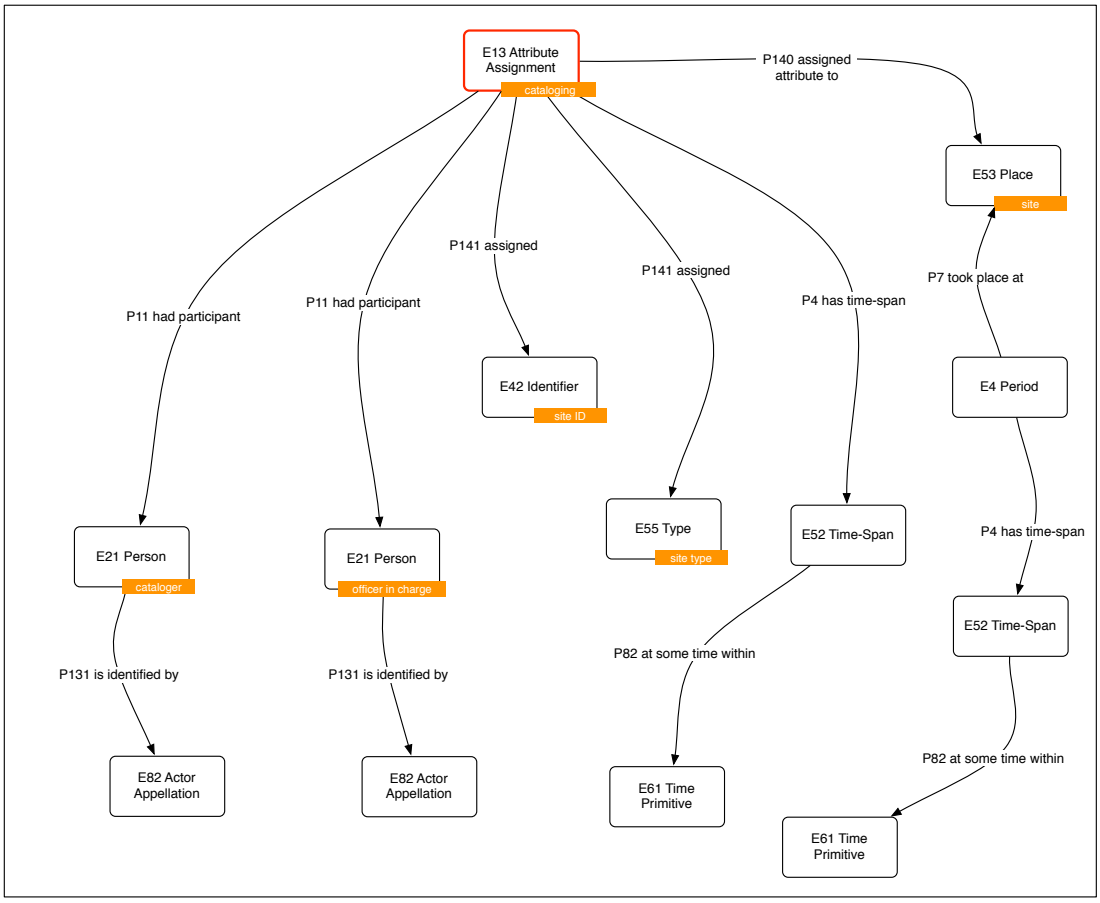


Figure 5.11: *The cataloging activity in IDRA and the corresponding graph in CIDOC CRM.*

With respect to these data, three main activities have been identified: the building of structures, the cataloging of archaeological sites, and the cataloging of structures. These activities, together with the related sub-activities, are represented in figure 5.13, and the detailed sub-graphs are shown and discussed in the following sections.

5.4.3.1 Data About the Building of the Structures

The place where the structures have been built is always defined by spatial coordinates, in terms of latitude and longitude, which have been acquired through georeferencing processes on the digital maps available; in the case of domus, the place is also identified in terms of an address.

The chronology of the production is always present as an indication of a period (e.g. "the Christian city"); in the case of monuments, more detailed chronologies are available in the form of time spans related to their building. Therefore, a double

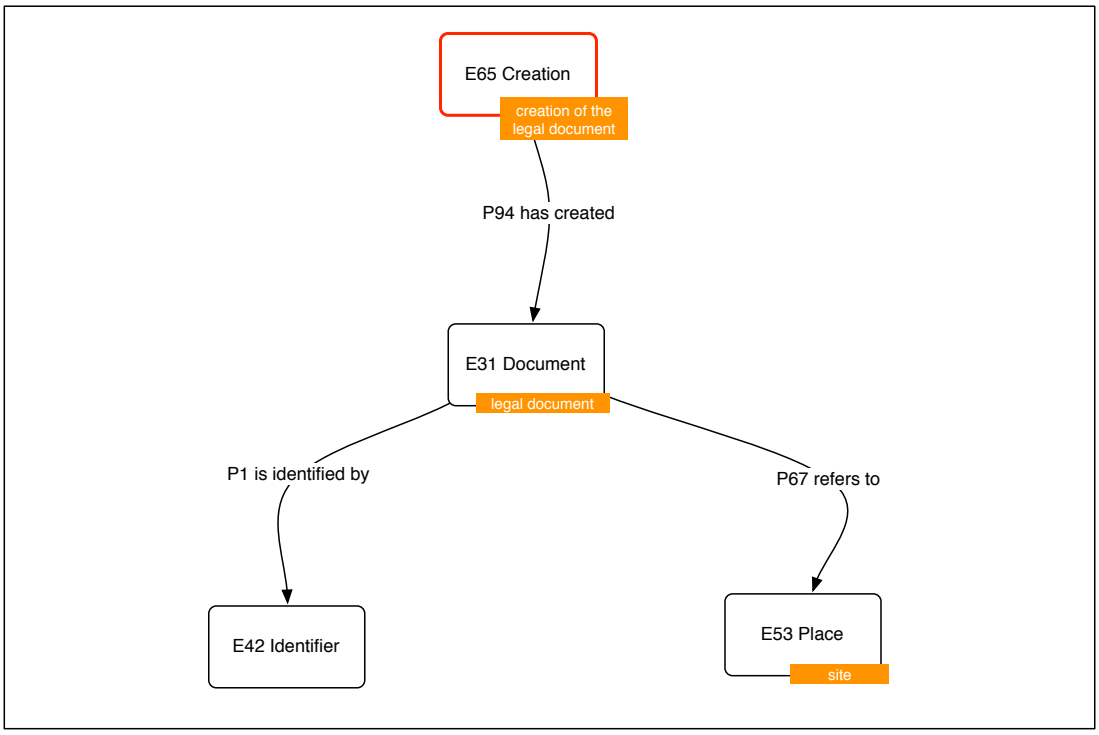


Figure 5.12: *The definition of legal constraints on an archaeological site in IDRA and the corresponding graph in CIDOC CRM.*

chronological representation for the building activity has been provided.

5.4.3.2 Data About the Cataloging of Sites and Structures

Data present in MANTIC are related to two different cataloging activities, which are respectively related to the archaeological sites, and to the structures. Coherently with the SIRBeC and the IDRA mapping, this activity is identified by `E13 Attribute Assignment`; moreover, in the case of structures, a `E14 Condition Assessment` activity is introduced.

The archaeological site is always defined by latitude-longitude coordinates, and can be additionally defined by an address. Since there are cases where the spatial indications differ from the address, the `E44 Place Appellation` has been used. The chronology of an archaeological site has been modeled in analogy with the IDRA case by adding a `E4 Period` to the place.

Structures are identified by an `E41 Appellation`, which is either the single name of the monument (e.g. "Roman Amphitheater") or a combination of the structure's typology and the address where it is located (e.g. "domus via Necchi").

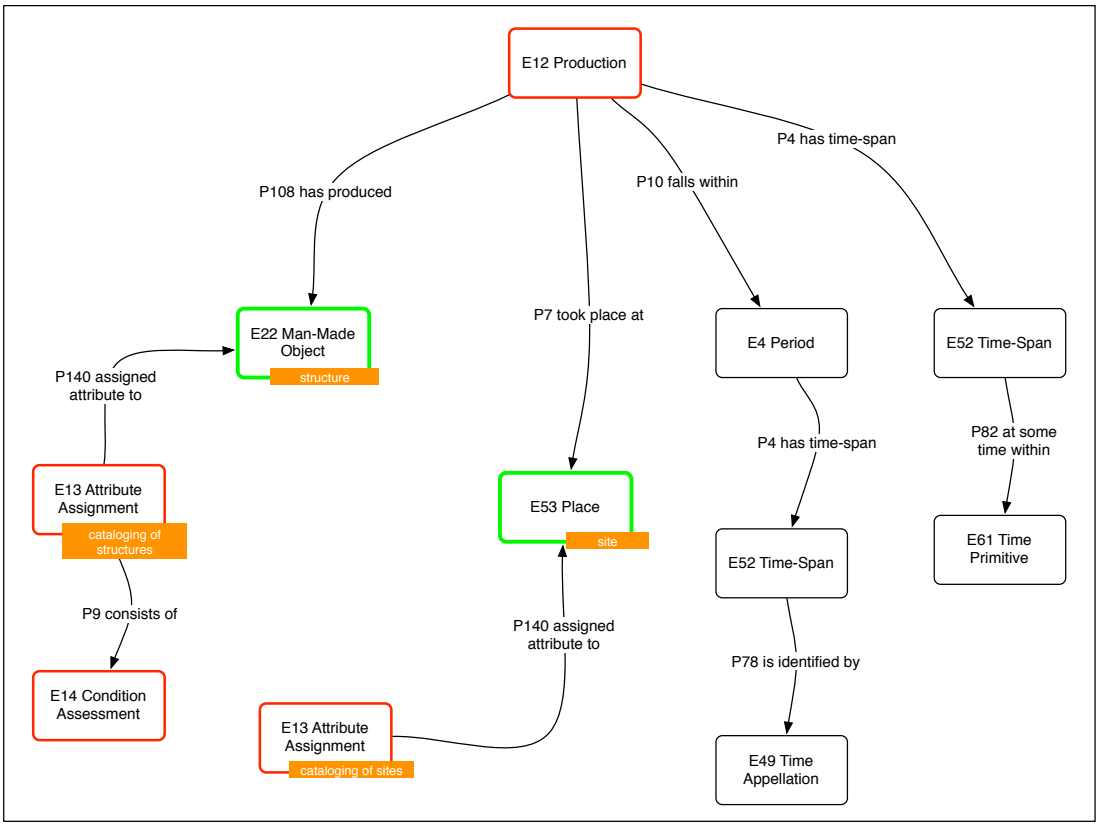


Figure 5.13: *A graph representing the main activities identified in MANTIC.*

5.5 Discussion

The mapping activity gave the opportunity to evaluate the CIDOC CRM in detail, both for what concerns its structure and peculiarities and its actual usage as a component of a Semantic Web system for cultural heritage. This section discusses the main elements and issues that arose from this activity, with reference to the specific results as well as the outcomes and contributions of similar documented experiences. The section is organized around a set of crucial key points, whose discussion aims to provide a framework for the discussion of futher improvements of the MANTIC project, and to make a contribution to the general scenario of Semantic Web applications for the integration of and access to cultural heritage information.

5.5.1 General Considerations

The CIDOC CRM documentation (Crofts et al., 2009) strongly supported the evaluation of the structure and principles of the model, and the available case studies allowed

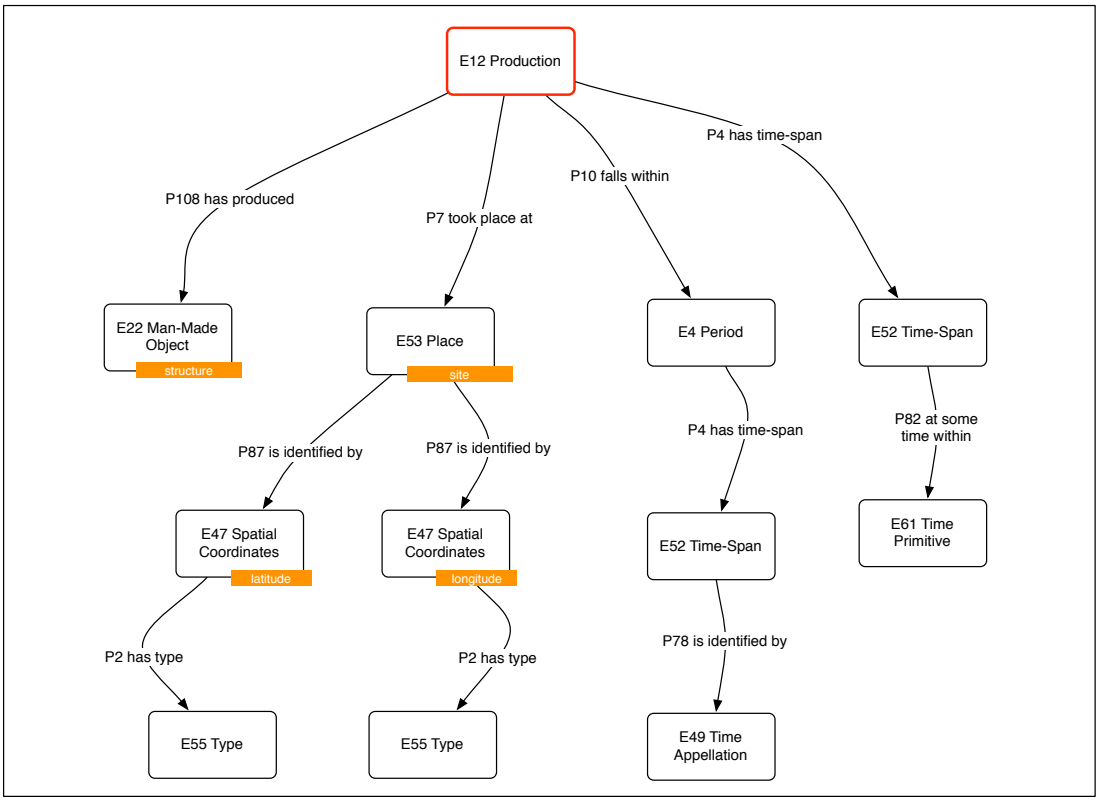


Figure 5.14: *The production of structures in MANTIC and the corresponding graph in CIDOC CRM.*

to better understand its actual use in real application scenarios. The documentation provides a well organized description of the CRM's classes and properties, through the use of textual scope notes that illustrate their semantics, along with examples and best practices. Moreover, particular attention is dedicated to the documentation of the evolution of the model through the release of successive versions. These consists in better specifications of the existing concepts, or even in the removal or addition of new classes and properties; amendments from version 3.2 on are accurately documented in an appendix. In general for an effective understanding of the documentation, only a few basic notions about the principles of object-oriented modeling and the cardinalities and constraints of properties are required, thus making the model appraisable by a number of different professional users, notably in the fields of cultural heritage, computer science, and related disciplines.

However, since the CIDOC CRM is a a core ontology, with specific focus on cultural heritage documentation, the role of domain experts with respect to the definition of mappings from heterogeneous documentation systems to the CRM becomes

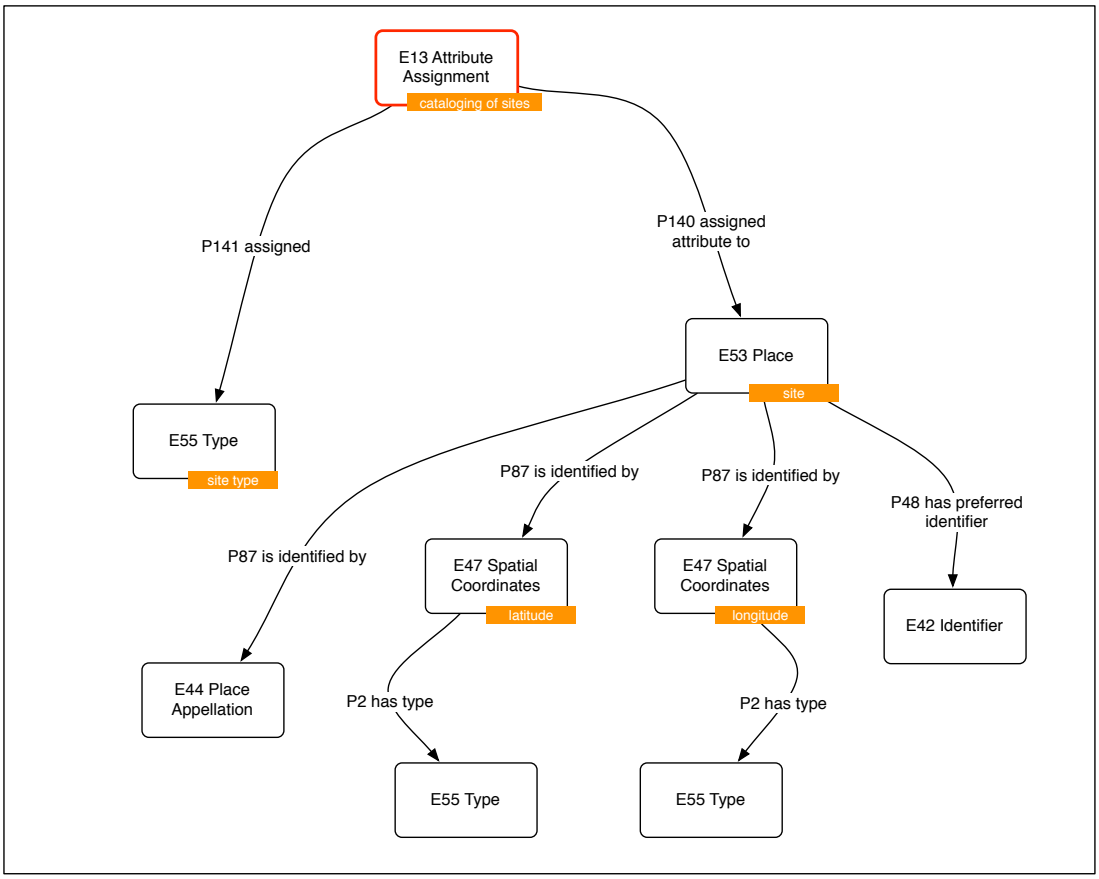


Figure 5.15: *The cataloging of sites in MANTIC and the corresponding graph in CIDOC CRM.*

crucial. Not only domain experts may approach the specificity of the CRM cultural heritage-related concepts with more ease than others, but their familiarity with the traditional documentation systems allows them to provide more accurate and sound mappings. This apparently trivial consideration is in reality of primary relevance: the importance of domain expert work is acknowledged in most literature concerning cultural heritage information integration projects through the use of the CIDOC CRM (e.g. Nussbaumer and Bernhard, 2007; Binding et al., 2008); the experience in the context of the MANTIC project fully supports and confirms this. The evaluation and understanding of legacy metadata schemata has been carried out by using the available documentation and by discussing them with the system's maintainers, while the definition of mappings to the CRM has been conducted with a more centralized approach in order to maintain direct and close control over the process. During these activities, different levels of knowledge (up to the very domain specific) about the characteris-

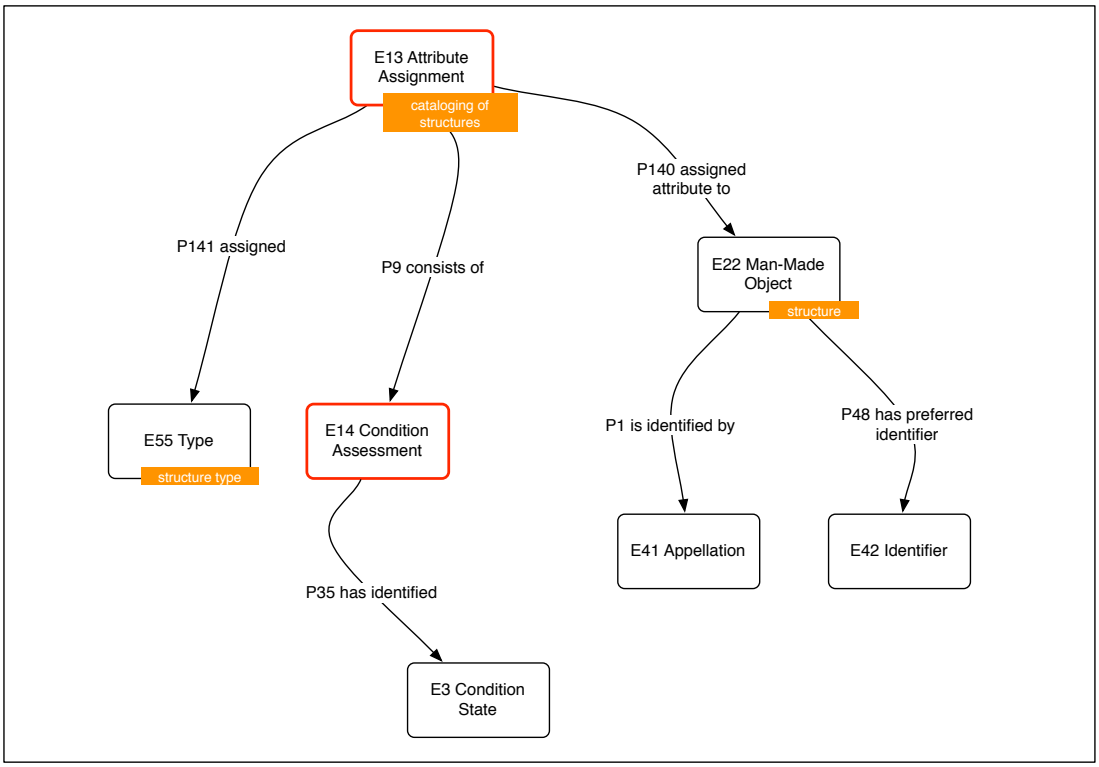


Figure 5.16: *The cataloging of structures in MANTIC and the corresponding graph in CIDOC CRM.*

tics of archaeological concepts and data and the documentation and cataloging criteria have been fundamental, and they have profoundly conditioned the mapping choices, especially in those cases where alternative representations were possible.

On the other hand, the understanding of the CIDOC model took more time than the initial estimates. This is related mainly to the fact that even if the model is remarkably compact with respect to the domain it takes into consideration, it is still complex, and there has been no previous personal experience in using it. Moreover, this characteristic sometimes goes hand in hand with certain levels of ambiguity in the semantics of the CRM concepts and relations, at least as they are described in the model specifications. As Goerz et al. (2008, pp. 2–3) points out, this can be considered an intrinsic drawback of textual documentation: "*As with any natural language text, there will always be a range of interpretation as opposed to a formal, mathematical specification with clear (and unambiguous) definitions and a well-defined semantics [...] A critical investigation of the scope notes in the CRM document shows that the description of intricate semantic problems in common language is not only error-prone, but also in danger of vagueness and a certain degree of ambiguity — despite all efforts in the precision of the argumentation.*". The work

done for the implementation of an OWL-DL version of the model, which required the clear definition of its semantics with respect to the characteristics of a language based on description logics, examines several of these aspects. More about general issues related to the correct understanding of the CRM is mentioned and discussed in almost every contribution dealing with the use of the model in real application scenarios, which provide useful case studies and lessons learned for new projects.

In addition to these considerations, it can be observed that the generality of the model, while showing the great potential of capturing the semantics of a wide range of cultural heritage documentation, is at the same time characterized by a major drawback, which consists in the abstractness of its concepts that makes them too ambiguous for any human user, as has been pointed out by (Nussbaumer and Bernhard, 2007, p. 19). This characteristic slowed the initial understanding of the CRM, and at a very basic level is shown by the necessity to introduce additional labels to the mapping sub-graphs in order to discuss them with reference to domain specific concepts. Other projects using the CRM faced this issue by implementing the model with custom-named and context-specific classes; the COINS project[1] for instance, introduces the `owl:equivalentClass` construct for setting correspondences between its CRM-compliant domain ontology and more general CIDOC CRM concepts:

```
<owl:Class rdf:about=#Coin_material>
<owl:equivalentClass
rdf:resource=&cidoc;E57.Material/>
</owl:Class>
```

This approach may complicate the modeling process and the management of the domain ontology when e.g. modifications of the model are defined and new versions become available (even if the issue concerning the compatibility of old mappings with new versions is more general and transversal). On the other hand, introducing domain specific named classes may show several advantages when e.g. formulating queries that take into account specific and more clearly identifiable concepts rather than abstract and general ones. Due to the experimental nature of MANTIC, and the limited set of schemata that were taken into consideration, this approach was not adopted, but it will be evaliuated at more advanced stages of the project.

For the same reasons, no custom extensions were developed, even if the CIDOC CRM SIG encourages this approach for the needs of more specialized communities and applications (Crofts et al., 2009, p. i), and sound and articulated extensions have been created and successfully tested, such as the CRM-EH for the archaeological research processes in the context of the STAR project. In fact, the STAR experience reports of not having had problems with the CRM abstractness, but it also points out that this can be considered a positive consequence of the extension approach, because

[1] `http://www.coins-project.eu/coins_ontology.owl`

"*[...] a more detailed model does afford more meaningful mappings from highly specific data elements than the (non-extended) CRM.*" (Binding et al., 2008, p. 289).

Given these general characteristics of the model, and their related issues, the mapping process became quite a complex and time-consuming task. Similar experiences (even if they were carried out on considerably larger scales) show the same situation, and they stress how several iterations of the mapping process are usually required. In some cases the ambiguity of the CRM specifications brings to the formulation of different mappings by different domain experts that were moreover familiar with the CIDOC CRM and the institution-specific schemata (Nussbaumer and Bernhard, 2007, p. 19). In other cases, the situation seems to be the opposite, and a counterposition between the CRM concepts and the less well defined concepts which archaeologists use every day is recognized (Cripps et al., 2004, p. 26). In the context of the MANTIC project, at least three iterations of the mapping process have been necessary, and particularly problematic cases required even more effort.

A tentative conclusion to date is that both the issues introduced above and the initial non-familiarity with the model are the main factors that influenced the progress of the mapping process. In addition to this, the event-centric nature of the CRM as opposed to the object-centric nature of the data sources, and the consequent amount of implicit information that needed to be identified, analyzed and explicitly represented further complicated the process. These latter aspects deserve more specific discussion in the next section.

5.5.2 Defining Event-Based Mapping Chains: Issues and Challenges

The discussion of the mapping templates and the related graphs contains different cases showing CRM concepts present in the legacy data schemata as implicit information. The identification of this information and its explicit representation has been a challenging process, but at the same time it has highlighted the value of the model with respect to the "*...analysis of the intellectual structure of cultural documentation in logical terms*" (Crofts et al., 2009, p. i). In fact, the evaluation of the data schemata in light of the CRM allowed to better understand their characteristics and to identify the relevant semantic relations they contain in detail, for example for what concerns the study and cataloging processes they document. At the same time, during this activity several issues arose which, beyond characterizing other similar experiences from a general point of view, influenced the actual process connected to the definition of event-based chains.

Data quality issues have been encountered transversally in all the data sources the project deals with, even where detailed norms concerning data entry criteria exist, such as in the case of SIRBeC. On the other hand, a certain amount of inconsistencies was expected since the very beginning of the activity, since the Institutions maintaining the databases are constantly performing data quality checks and data normalization procedures, which for example, in the case of SIRBeC are still ongoing. Moreover,

data quality issues are a traditional and central concern of almost every project dealing with cultural heritage and the CIDOC CRM and in most cases are intrinsically related to the high heterogeneity and complexity of cultural heritage documentation systems. In order to face them, a number of different procedures for data cleansing (e.g. by using pre-processing methods) have been defined, and their relevance for the creation of sound mappings has been stressed in several contributions (see e.g. Binding et al., 2008, p. 289). Kummer (2007, pp. 48–50), while discussing these issues in the context of the Aracne-Perseus project, provides a clear separation between schema level quality and instance level quality problems, and he introduces a set of real examples and possible solutions for both cases. In the context of the MANTIC project, schema level issues are related mainly to:

- fields containing overlapping information that needs to be merged (e.g. the case of the chronological attribution of artifacts; see section 5.4.1.1)
- fields with heterogeneous information that needs to be split (e.g. the case of materials and techniques; see section 5.4.1.1)

A practical approach for facing these issues is the creation of specific code and applications which are able to recognize and merge or tokenize relevant information; however, in general a more extensive analysis together with the database developers and maintainers would be suitable, and it should make an important contribution to the cataloging process quality control, such as the one reported in Degiarde (2007). On the contrary, problems at the data level are mainly related to inconsistent spelling and to missing information. These issues will certainly benefit in the short period from the ongoing check and normalization activities each Institution is currently undertaking; in the meantime, they have been preliminarily faced with the introduction of semi-automatical data cleansing methods.

The creation of normalized datasets, greatly eased the definition of event-based mapping chains, but still some problems connected to the generality of the CRM concepts had to be faced. In fact, as Nussbaumer and Bernhard (2007, p. 9) points out, "*The CIDOC CRM provides no guidance for the domain or schema experts which metadata attributes of the source schema to map, and which classes and properties should be used in doing so*." and it identifies two groups of possible mapping inconsistencies. The first one is related to the creation of **different chains for equivalent metadata**, i.e. where semantically equivalent database fields coming from two different sources have been mapped differently, and the introduction of an internal vocabulary, which defines concepts for grouping mappings with the same semantics is suggested. This kind of inconsistency can be the result of mappings carried out independently on different sources from different actors, as well as of different choices for different legacy data schemata defined by a single person. The latter is the case of the MANTIC experience, where each data source has been mapped independently from the others; however, iterations of the

mapping process, as well as a more precise understanding of the CRM and the legacy metadata schemata led to the transversal harmonization of the representation.

The second kind of inconsistency is related to the possibility of defining **identical chains for different metadata**. This has been a frequent situation in our case; consider this simple example, which shows the mapping chains on the left and the possible instance values on the right:

E16 Measurement		
P39 measured		
E22 Man-Made Object	⟵	epigraph
P82 observed dimension		
E54 Dimension		
P90 has value		
E60 Number	⟵	40
P91 has unit		
E58 Measurement Unit	⟵	cm
P2 has type		
E55 Type	⟵	**length**
P82 observed dimension		
E54 Dimension		
P90 has value		
E60 Number	⟵	40
P91 has unit		
E58 Measurement Unit	⟵	cm
P2 has type		
E55 Type	⟵	**width**
P82 observed dimension		
...		

These chains map the measurement activity performed on each artifact, and the related values and measurement units. Without the introduction of additional "assertion chains" (Nussbaumer and Bernhard, 2007, p. 11) it is impossible to solve ambiguity in the subgraphs, which in this specific cases comes from having the same representation for different measurements (40 cm). The additional chains (highlighted in the example) provide a typification of the measurement, which is only implicitly represented at the schema level of the data source (in the form of field name), therefore allowing to uniquely identify the subgraphs.

This in turn makes the mapping chains grow further on and combined with the richness of the model and the necessity to explicitly represent implicit information, makes the overall graph grow considerably in dimensions if compared to the compactness (and of course less expressivity) of the original data schemata. A potential

negative side-effect of this phenomenon is represented by the necessity to deal with long chains when querying the graph for the retrieval of even simple information. The example of measurements shown above gives an idea of this, but this consideration can be extended to several cases, including critical ones, such as the retrieval of chronological attribution of artifacts:

E22 Man-Made Object ⟵ cup
P108I was produced by
E12 Production
P4 has time-span
E52 Time-Span
P82 at some time within
E61 Time Primitive ⟵ 1st century A.D.

According to the CRM, the chronological attribution of an artifact (i.e. the date it was created) is associated to `E12 Production`, which in turn is chronologically defined through the `P4 has time-span` property and the related classes and properties. Therefore, the retrieval of chronological information has to take 4 classes and 3 properties into consideration, making e.g. a SPARQL query become very articulated and probably time-consuming, with direct influence on the efficiency of the retrieval mechanisms, while a SQL query on the original database would require a couple of basic statements. Therefore, the CRM shows high expressive power and great accuracy for the representation of cultural heritage information, but the complexity that descends from this expressiveness propagates until the more practical and application-oriented levels, where suitable choices have to be made e.g. for defining efficient methods of information retrieval and presentation.

Moreover, the need to explicitly represent events often goes hand in hand with the lack of actual knowledge about the events themselves. This is similar to what Binding et al. (2008, p. 284) refers to as the definition of "**virtual entities**", when discussing the mapping of their relational databases to an RDF CIDOC-CRM compliant graph structure. In our project the identification of events, besides being in most cases postulated on logical bases (e.g. every artifact is evidently the physical result of a production event), additional knowledge was required. A particularly clear example of this is related to the SIRBeC cataloging process, which is quite articulated (see fig. 5.4), and its characteristics (e.g. the actors taking part in it with different roles; the logical sequences of activities constituting the whole process) can only be partially determined with a simple analysis of the data schema. If one wants to model the cataloging process in its entirety, additional knowledge coming from the SIRBeC documentation and the discussion with the SIRBeC team is required. In this specific case it has brought either to the introduction of "virtual entities" representing sub-activities (e.g. `E14 Condition Assessment` and `E16 Measurement`), or the exclusion of

activities that are not represented even implicitly in the available data source (e.g. the `E7 Activity` concerning the general cataloging activity promoted by the Regione Lombardia; see section 5.3.1).

On the contrary, there have been some cases where the identification of activities has been intentionally skipped in this first experimentation, and a simpler representation has been adopted. For example, the `E4 Period` related to the chronological attribution of an archaeological site in IDRA (see fig. 5.11) may refer to a number of activities that took place in the past on the site, but in absence of detailed information concerning them (and without the introduction of a too general `E11 Modification` to group them), the choice was to present the chronological attribution of the site together with the cataloging activity. This choice may be changed in the future, when the IDRA data schema will be analyzed in more detail and the specific case will be discussed with the IDRA team.

The event-based chains that emerged from the mapping experience are related to two high level groups of events, with a perfect analogy with the results obtained in other projects, notably the experience of the English Heritage Centre for Archaeology (Cripps et al., 2004): the first group is related to the events that happened in the "archaeological time" (e.g. the production of artifacts, the building of structures, etc.) resulting in the formation of the archaeological record; the second concerns the activities performed in modern times by archaeologists and cultural heritage professionals for the exploration, study, cataloging and preservation of archaeological items. The possibility to access and query this integrated information in a unified manner comes from the ability to explore the integrated graph in effective and efficient ways. Long and complex event-based chains should hinder this possibility, but on the other hand, the CIDOC CRM provides methods to simplify the representation that can be used e.g. for creating different versions of the graph which can be navigated selectively on the basis e.g. of specific query requirements. **Shortcut properties** in the CRM allow to substitute fully developed paths with more compact ones; for example measures can be directly linked to the artifact (through a `P43 has dimension` property) without the need to introduce a measurement activity. Concerning shortcuts, the CRM documentation (Crofts et al., 2009, p. xvi) says that : "*An instance of the fully-articulated paths always implies an instance of the shorctuct property. However, the inverse may not be true; an instance of the fully-articulated path cannot always be inferred from an instance of the shortcut property.*. Therefore, the simplification of the representation through the use of shortcuts means some information will be lost. This is particularly true (and it is moreover stressed by the documentation) for the `E13 Attribute Assignment` class, which allows to link every characteristic of e.g. an artifact to the actors that defined it. On the other hand, introducing redundant modeling with different, more or less detailed graphs, should allow to take advantage of shortcuts, while at the same time, preserving a rich representation; the CRM documentation, even if in the specific context of the `E13 Attribute Assignment` class description points out that :"*[...]*

many implementations may have good reasons to model either the action or the short cut, and the relation between both alternatives can be captured by simple rules." (Crofts et al., 2009, p. 8). This consideration can be extended to all cases where the introduction of shortcuts together with fully developed paths allows to retrieve more efficiently relevant information, while preserving the possibility of retrieving specific details by navigating through more articulated graphs. Figure 5.17 shows an example from the SIRBeC cataloging activity mapping, where shortcut properties are highlighted in green.

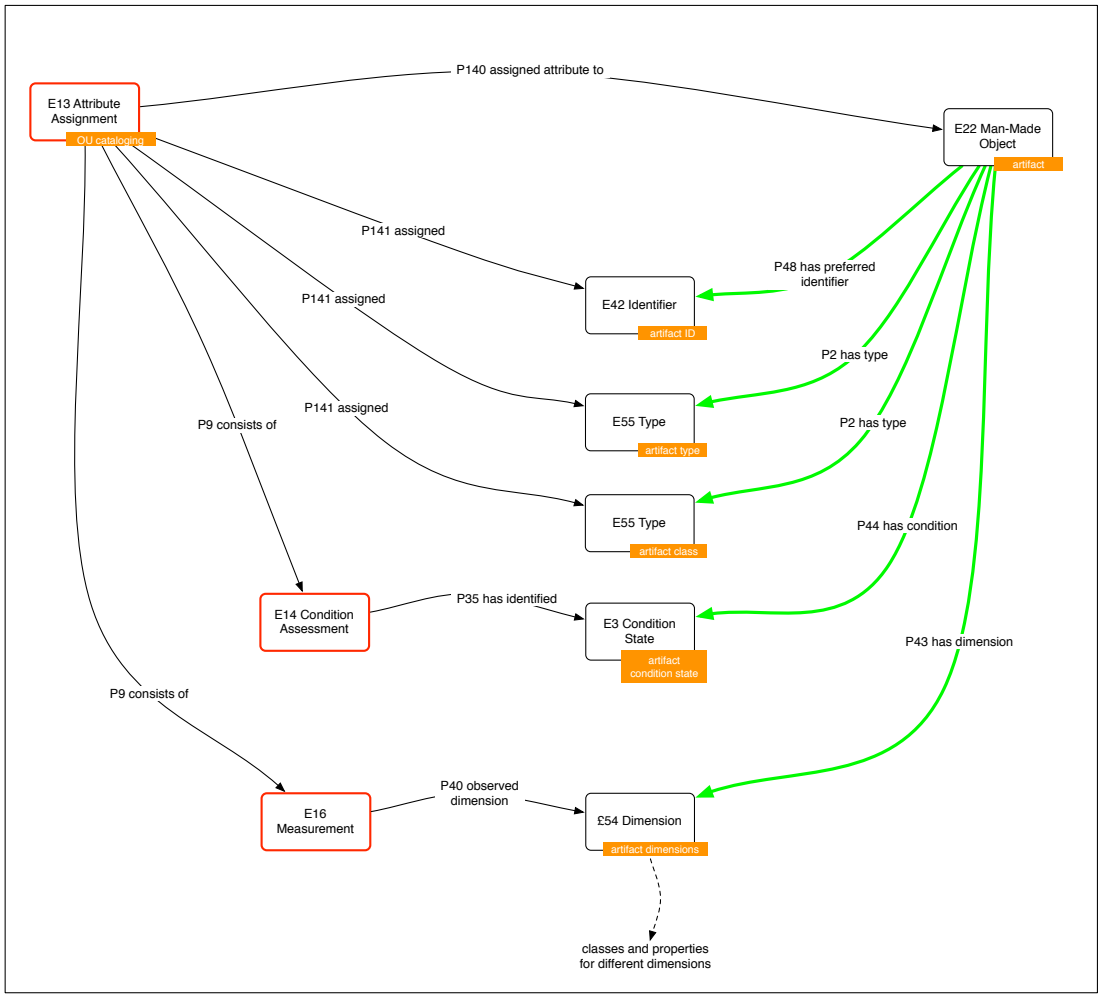


Figure 5.17: *Redundant modeling using additional properties and shortcuts.*

In this example, if basic information is required about an artifact (such as its type, dimensions, and preservation state), shortcut properties can be exploited for the query, while specific details about e.g. the type attribution can be retrieved either by following the full path or by exploiting the alternative path enabled by the shortcut. This

redundant modeling approach is currently under evaluation in our project, and it will possibly be part of an improved mapping activity.

5.5.3 Representing Actual Instance Data Values

Being a conceptual model, the CRM "*. . . is not optimized to implementation-specific storage and processing aspects*." (Crofts et al., 2009, p. i), and has little intrinsic provision for the representation of actual data instance values (Binding et al., 2008, p. 284). In particular, the `E59 Primitive Value`, and its sub-classes (i.e. `E60 Number`, `E61 Time Primitive`, and `E62 String`) while used for the representation of primitive values used as documentation elements, are not further elaborated upon within the model, as it is declared in the class scope notes. However, in real world applications, the representation of actual instance data values becomes crucial, and suitable strategies have to be defined for it, depending on the actual languages and technologies with which the model is implemented and used. In the context of information integration projects dealing with Semantic Web languages, different approaches can be found.

Binding et al. (2008, pp. 284–285) make use of `rdf:value` relationships as an additional property to model instance data for entities wherever appropriate, such as:

```
crmeh:EHE0022.rrad.context.contextno.110575
rdf:value ``98000E 56879N''
```

Here `rdf:value` is used in order to represent coordinate information in terms of easting and northing values. On the contrary, other proposals rely on the use of `xsd:type`, which offers the possibility to precisely define standard XML schema data types (XSD) that machines can correctly identify and process. However, this approach at the same requires the extension of the implementation of the CRM with the introduction of `owl:DatatypeProperties` linking the `E59 Primitive value` subclasses to XSD types . For instance, the Erlangen OWL-DL serialization of the model (Goerz et al., 2008) introduces 4 of these properties, whose respective domains and ranges are shown in table 5.3.

Table 5.3: *The Erlangen CRM implementation of datatype properties.*

Property	Domain	Range
`has_FloatNumber`	`E60 Number`	`xsd:float`
`has_IntNumber`	`E60 Number`	`xsd:int`
`has_PrimitiveString`	`E60 Number` or `E62 String`	`xsd:string`
`has_PrimitiveTime`	`E61 Time Primitive`	`xsd:dateTime`

Being based on the Elangen CRM, our project makes use of the data type properties of the implementation and the respective XSD types, while for `E55 Type` the `rdfs:label` property has been used (vocabularies are more extensively treated in section 5.5.4). One example is:

E16 Measurement
  P39 measured
    E22 Man-Made Object
  P82 observed dimension
    E54 Dimension
      P90 has value
        E60 Number
        has_floatNumber
        xsd:float ⟵ "40.0"
      P91 has unit
        E58 Measurement Unit
        rdfs:label ⟵ "cm"
      P2 has type
        E55 Type
        rdfs:label ⟵ "length"

5.5.4 Managing Vocabularies

The definition of vocabularies, thesauri and authority lists is a traditional concern in cultural heritage documentation; in archaeology, moreover, it constitutes a fundamental activity which is mainly related to the creation of typological lists. Information systems dealing with cultural heritage and archaeological data usually provide different structures and tool for the representation, management and use of lists of terms, which often are also the privileged access key for querying the datasets and retrieving relevant information. Consequently, the management of vocabularies and thesauri is of crucial importance in information integration contexts; at the same time it requires that structured lists exist and most of all, that the concepts and terms in these lists be shared or be semantically integrated. In our project vocabularies cover a wide spectrum of different cases. SIRBeC is the data source that provides the most extensive and complex set of vocabularies (in the form of open and closed lists), mostly because they have been defined by ICCD in the context of the activities for the creation of a national and unified catalog of cultural heritage resources. The ICCD vocabularies that are relevant for our context are those related to the definition of the artifacts' type and classes, and to the materials and the techniques. They are available in the ICCD

Web site in the form of text lists in pdf format[1], where narrower/broader relationships between different terms are represented by delimiting the terms with a slash, such as:

argilla/ a tornio
argilla/ a tornio/ a barbottina
argilla/ a tornio/ a barbottina/ ingobbiatura
argilla/ a tornio/ a barbottina/ verniciatura
argilla/ a tornio/ a ditate

On the contrary, vocabularies for IDRA and MANTIC can be derived respectively from the analysis of the scraped data, and from the data that MANTIC models, i.e. those provided by the Archaeological Museum of Milan. In both cases their structure is far less complex than that of SIRBeC vocabularies; moreover, vocabularies for the typification of implict information that has been made explicit (e.g. the different types of activities) has been modeled from scratch. This high heterogeneity, and the lack of ready-to-use structured digital representation of the SIRBeC vocabularies suggested us to adopt a simple approach which completely relies on the CRM classes and properties, similarly to what has been done in other projects, such as Arachne-Perseus (Nussbaumer and Bernhard, 2007). In fact, the CRM defines `E55 Type` as a class comprising "*...concepts denoted by terms from thesauri and controlled vocabularies used to characterize and classify instances of the CRM classes* which is conceived as "*the CRM's interface to domain specific ontologies and thesauri.*" (Crofts et al., 2009, p. 23). Moreover, the class provides a property for the composition of hierarchies of terms through narrower/broader linking. Using this approach, several vocabularies of extremely different consistencies have been implemented; they are listed in table 5.4 together with the legacy data source they refer to.

This approach is deemed sufficient for our first experimentation, and allows not to introduce extension subclasses related to the CRM concepts, which is the other possible way suggested by the documentation and should be done when "*...the concept is sufficiently stable and associated with additional explicitly modeled properties specific to it*" (Crofts et al., 2009, p. xvii), which evedently is not our case. From the point of view of the OWL-DL implementation, some issues with the `E55 Type` class have been indicated (Goerz et al., 2008, p. 10) but they do not affect our representation.

At the same time the approach based on the `E55 Type` class and its properties, while providing a certain degree of flexibility, is characterized by several drawbacks

[1]Concerning the RA card, several specifications are available. See: `http://www.iccd.beniculturali.it/Catalogazione/standard-catalografici/strumenti_di_ausilio_e_di_controllo/strumenti-di-ausilio-e-di-controllo/`. With reference to our project, relevant documents are the thesaurus concerning the definition of the object (OGTD - `http://www.iccd.beniculturali.it/Catalogazione/standard-catalografici/strumenti_di_ausilio_e_di_controllo/resolveUid/d58416c1a9a968dde9f719721c5488d1/`, and the vocabulary concerning the object's class and production (CLS - `http://www.iccd.beniculturali.it/Catalogazione/standard-catalografici/strumenti_di_ausilio_e_di_controllo/resolveUid/378187ff14afd45d76bd9281f0322fd0/`

Table 5.4: *List of the implemented vocabularies.*

Vocabulary	Data sources
Types of activities	SIRBeC, IDRA, MANTIC
Types of artifacts	SIRBeC
Classes of artifacts	SIRBEC
Materials	SIRBeC
Techniques	SIRBeC
Types of structures	MANTIC
Types of sites	IDRA, MANTIC
Languages of inscriptions	SIRBeC
Types of documents	SIRBeC, IDRA, MANTIC
Types of measures	SIRBeC
Measurement units	SIRBeC
Spatial reference systems	SIRBeC, IDRA, MANTIC

that will need to be evaluated at more mature stages of the project, and would possibly require the substantial re-design or extension of the current solution. One principal drawback is the impossibility of modeling other kinds of relationships (other than narrow/broader) between concepts and terms that are related to the same real world entities coming from different resources (e.g. the sites' types from IDRA and MANTIC). In fact, even if specific extensions of the `E55 Type` class can be theoretically defined, more effective and sound models exists that can be used for the specific needs of thesauri creation and alignment, such as the Simple Knowledge Organization System (SKOS). The choice of SKOS has been made e.g. in the COINS project, where a compact domain-specific and multilingual thesaurus has been created for the concepts and terms related to ancient coins[1]. Another and much more complex example is that of the STAR project, where SKOS thesauri have been created together with a set of terminology services[2] (Binding et al., 2008, p. 288) that provide a further development of the SKOS API and have been integrated with the Delos Digital Library Management System. Furthermore, high-level thesauri, vocabularies and lists providing general concepts and terms that can be used in the cultural heritage context exist, such as those defined by the Getty Foundation, and notably the Art and Architecture Thesaurus (AAT)[3], the Union List of Artist Names (ULAN)[4] and a the new Cultural

[1]The thesaurus is available as a project deliverable at `http://www.coins-project.eu/downloads/reports/Coins-044450-D5.pdf` and a SKOS file at `http://www.coins-project.eu/coins_thesaurus.owl`

[2]`http://hypermedia.research.glam.ac.uk/kos/terminology_services/`

[3]`http://www.getty.edu/research/conducting_research/vocabularies/aat/`

[4]`http://www.getty.edu/research/conducting_research/vocabularies/ulan/`

Object Name Authority (CONA)[1] which is currently under development and will be released in 2011. These can be used e.g. as a "semantic glue" between different and domain-specific thesauri, thanks to the fact that they are released in different standard formats; SKOS implementations are planned or are already available.

5.5.5 Coreference Linking

The integration of complementary sources through a common model is intrinsically characterized by the problem of identifying possible different resource identifiers (e.g. URIs) that refer to the same real world entities. This problem is a crucial one for the actual creation of global knowledge networks and cyberinfrastructures (see section 3.3), and deeply influences the possibilities of cross- retrieval of relevant data. With reference to our specific scenario, the different subgraphs that have been created from the mapping of every single legacy data source to the CRM model are a clear example of this situation. Figure 5.18 depicts simplified versions of the subgraphs with the indication of the main activities and the related archaeological concepts; links between these subgraphs can be made on the basis of the identification of equivalences between the archaeological sites and the archaeological excavation activities.

Setting these equivalences allows to enrich the possibility of retrieval and the cross-comparison of different cataloging criteria and opinions concerning the same entities. For example, information about an archaeological site can be shown to the user together with information about the structures and the artifacts that have been discovered during the excavations carried out on it. Currently, this kind of query is not possible within the single legacy systems, while in the integrated graph is should be made on the basis of this path:

E53 Place	⟵	MANTIC site = IDRA Site
P7I witnessed		= SIRBeC site
E5 Event		
P12 occurred		
in the presence of		
E22 Man-Made object	⟵	SiRBeC artifact
P7I witnessed		
E12 Production		
P108 has produced		
E22 Man-Made object	⟵	MANTIC structure

Moreover, linking the different subgraphs with the identification of coreferences would allow the enrichment of information on the same entity, with different details

[1] http://www.getty.edu/research/conducting_research/vocabularies/contribute.html#cona

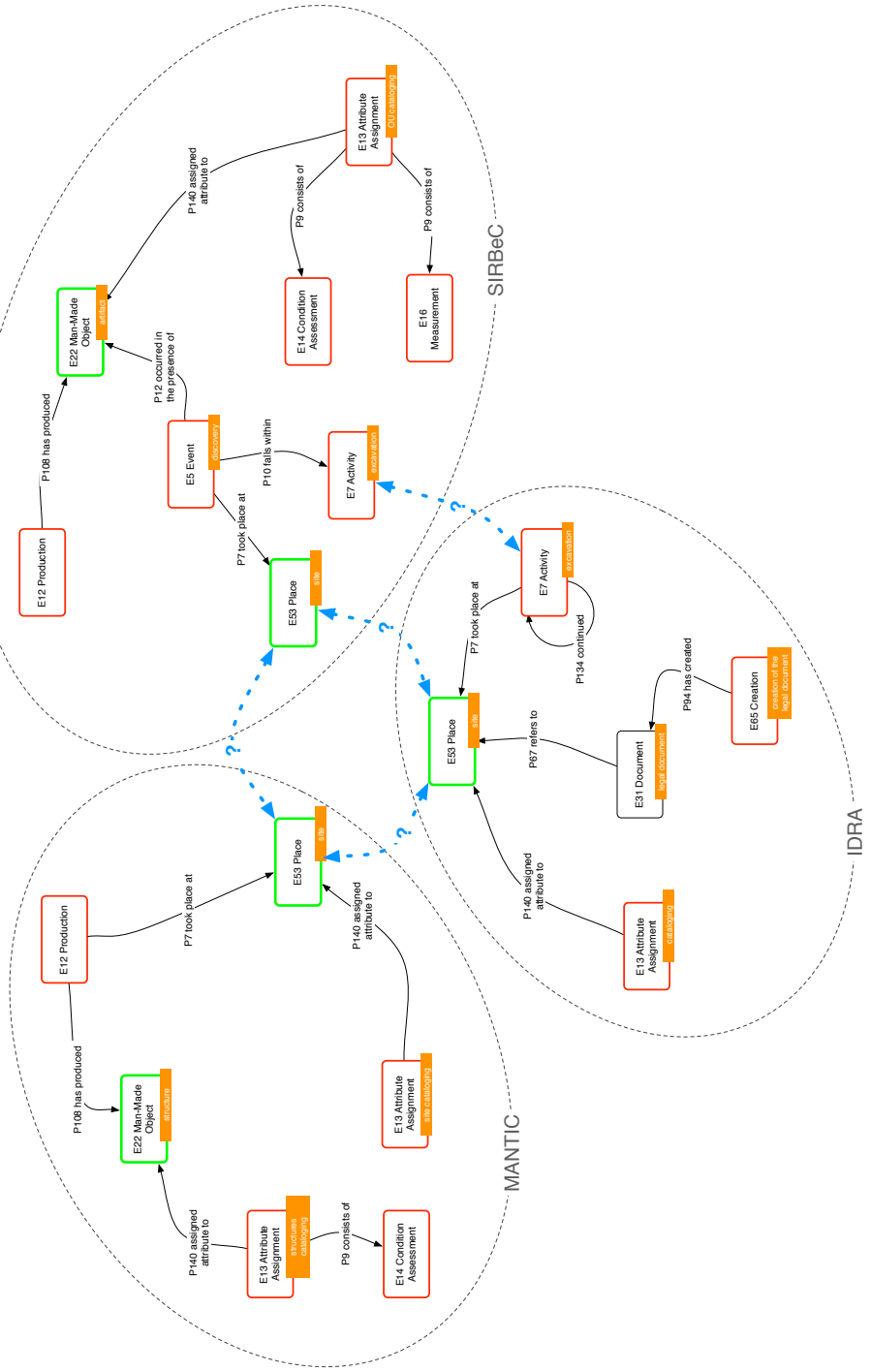


Figure 5.18: *Coreference graph.*

coming from the integrated sources, such as information about the archaeological excavations contained in IDRA and in SIRBeC. Several other possibilities linked to cross-query and retrieval can be defined: they are currently being evaluated in detail. Their effectiveness with reference to the retrieval of relevant information will be tested with different categories of end-users.

On the other hand, no standard theoretical approach or practical solution currently exists to face the issues of coreference linking, such as the automatic identification of equivalent URIs. From a general perspective, these are crucial concerns in the Semantic Web and Web of Data research, which to date has brought to the definition of heterogeneous approaches and proposals. One recent initiative in this direction is the "Okkam" EU-funded project[1], which promotes the vision of a "Web of entities" characterized by the systematic reuse of single and globally unique identifiers for any entity that is named on the Web. Within this project specifications and the general infrastructure that would enable the vision are being developed, and their results are too preliminary (at least for our case) to be taken into consideration. More specific work carried out under the CIDOC CRM framework further discusses the problem in the context of cultural heritage information integration, and provides suggestions related to e.g. semi-automatic coreference detection (Doerr and Iorizzo, 2008). Similar issues are also discussed in (Kummer, 2007; Babeu et al., 2007), which shows the approach and methods for the identification of named entities within the Arachne-Perseus project.

In the context of MANTIC a sufficiently simple and flexible approach should consist in manually correlating different entities related to the same real world things by using additional properties. This process would not be too time-consuming (due to the relatively limited set of data this first experimentation deals with), but at the same time it would provide a proof of concept for future and more articulated work. In particular, the `owl:sameAs` property should be used because it allows to link an individual to another, indicating that two URI references actually refer to the same thing, or the individuals have the same "identity", such as:

```
<rdf:Description rdf:about="#SIRBeC_site_A">
<owl:sameAs rdf:resource="#IDRA_site_B" />
</rdf:Description>
```

The `owl:sameAs` property is often used for defining mappings between different ontologies; a suitable strategy would be to create a simple and autonomous RDF file containing the relevant equivalent individuals linked through this property, which can then be used at query time in order to allow the formulation of query chains, connecting the different available subgraphs.

[1] `http://www.okkam.org/`

6

ChronoMANTIC: Representing and Retrieving Fuzzy Chronologies

This chapter discusses the application and evaluation of the methods for modeling and retrieving fuzzy chronological information introduced in section 4.3 for the case study introduced in the previous chapter. The description of the different kinds of chronologies available and their representation with the proposed fuzzy set-based model represent the starting point of our discussion (sections 6.1 and 6.2).
Thereafter, an evaluation setting, the related criteria, and an interface enabling the actual evaluation through an interactive timeline are introduced (section 6.3). The results of the evaluation (section 6.4), which involved 8 domain experts and 4 average users, have been analyzed with respect to precision and recall, and they provided the elements for the discussion of both the model and the retrieval methods (section 6.5).

6.1 MANTIC Chronologies

The analysis of the data sources and the conceptual mapping to the CIDOC CRM revealed the existence of different temporal annotations which are related to different chronologies. Two main groups can be identified:

1. annotations concerning the chronological attribution of archaeological elements with reference to ancient events, such as the production of an artifact, the building of a structure or the periods during which activities were conducted on an archaeological site

2. annotations concerning the chronology of modern processes of discovery, documentation and analysis of archaeological elements, such as archaeological excavations, cataloging activities, restorations, etc.

In reality, this clear cut distinction is not always valid, because in several cases there is a *continuum* linking the archaeological and the present time: a typical example is that of structures (such as churches) which have never ceased to be modified and used until today. However, the distinction will be maintained here for its intuitive value, and coherently with similar approaches such as the ones introduced in section 5.5.2.

In MANTIC, the intervals and the related chronologies are characterized by different kinds of temporal imprecision; for example:

- the production of an ancient artifact is attributed to an interval "between the end of the 2nd century and the first quarter of the 1st century B.C."
- an archeological site is attributed to the "late republican period"
- the documentation concerning an archaeological excavation has been lost, but from indirect sources we can reconstruct that it was done "before 1939"
- the discovery of an artifact happened during public works in the city centre in the years 1955–1956, without more precise indications

The representation of imprecise temporal intervals according to the model introduced in section 4.3.3 and the evaluation of the associated retrieval possibilities will take into consideration the chronologies that are related to the "archaeological" time, which offers the richest and most composite case. Taking into consideration the general categorization of chronologies introduced in section 4.3, the temporal intervals we deal with are defined by:

- references to absolute chronologies, in terms of calendric elements such as centuries and/or parts of centuries
- references to historical and archaeological periods (e.g. the Roman Age), which are in turn definable in "absolute" terms

6.1.1 References to Absolute Chronology

References to absolute chronology are provided as highly consistent labels assuming one of the following forms[1]:

- a single reference to a century: e.g. "1st century B.C."
- a single reference to a part of a century: e.g. "middle 1st century B.C."
- a single reference to two parts of a century: e.g. "first half – third quarter 1st century B.C."
- a combination of those; e.g.:
  - "1st century B.C. – 4th century A.D."
  - "end 1st century B.C. – 1st century A.D."

[1]Centuries are indicated with Roman numbers in data sources, as shown in the syntax diagram following this list.

- "end 3rd century A.D. – beginning 4th century A.D."
- "first half/third quarter 1st century B.C. – 6th century A.D.

At a more general level, each label in the dataset is the result of the combination of a few atomic elements, expressing temporal information with different degrees of imprecision, from parts of centuries to the era according to the Gregorian calendar. The possible combinations and the values these elements can assume are expressed in the following syntax diagrams:

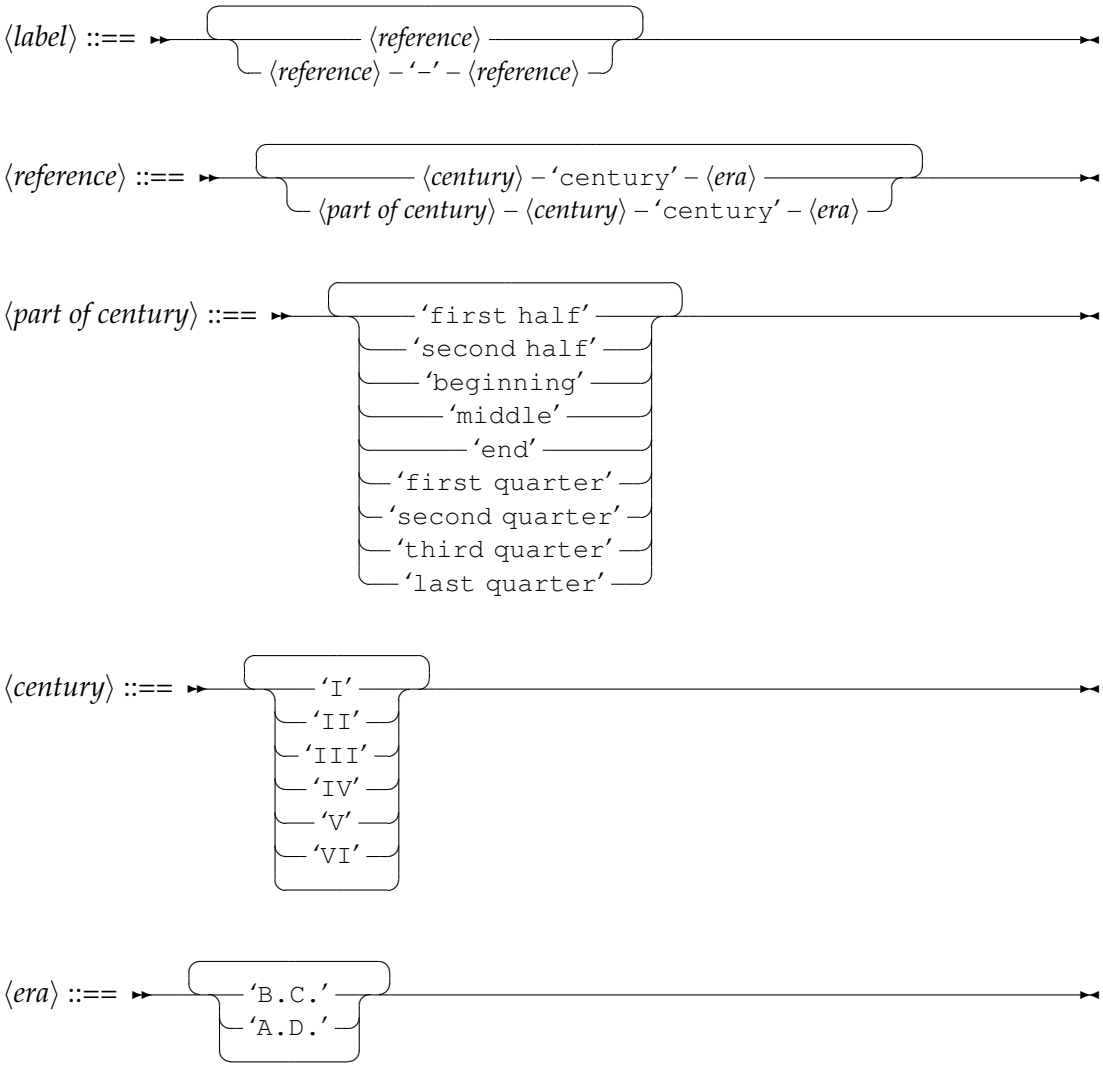


Figure 6.1 represents different possible imprecision levels (X) with reference to the 1st century B.C. and the 1st century A.D (Y).

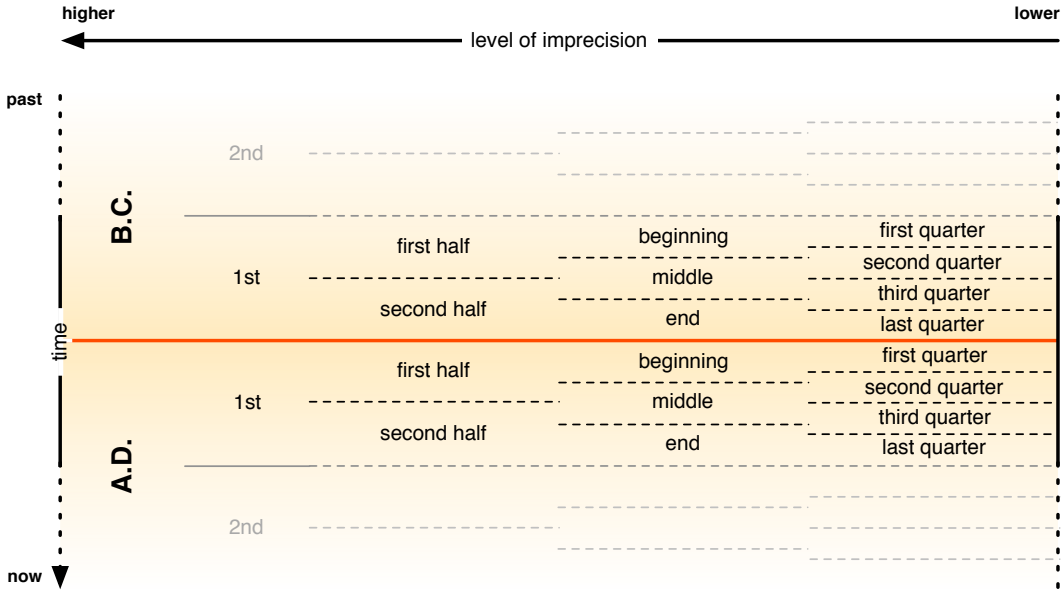


Figure 6.1: *A depiction of the possible references to absolute chronology showing different levels of imprecision.*

6.1.2 References to Historical or Archaeological Periods

References to historical or archaeological periods are also characterized by imprecise temporal intervals, assuming different forms. The calendric dates that are associated to these periods can be categorized according to two main groups:

- Conventional dates that are usually related to major and significant historical events. For example, the end of the period "Milan capital of the Roman Empire" coincides with the transfer of the Imperial seat to Ravenna in 402 A.D.
- More general indications that can cover even a whole century. For example the "Origins" period is in the 5th century B.C.

The definition of the fuzzy temporal boundaries of the relevant periods is complicated by the fact that transitions from a period to another are generally continuous phenomena, and both the specific internal articulation of the transitions, as well as their actual duration may vary a lot from case to case. For example, the shift in the control of Milan from the Celts to the Romans happened quite gradually over more than a century, a period when elements belonging to both cultural spheres were present. A model based on fuzzy sets, such as that introduced in section 4.3 offers the required flexibility for representing these kinds of situations.

6.2 Representing Fuzzy Temporal Intervals

Taking into consideration the different characteristics of the temporal intervals and the chronologies that have been introduced in the previous sections, different criteria and practical approaches to actually represent the 4 dates that are required by our model have been identified. In particular:

- for references to absolute chronologies the dates can be automatically determined from label information, by means of regular expressions and string parsing techniques
- for references to historical and archaeological periods, the dates need to be modeled manually, taking into consideration personal domain knowledge, the analysis of external scholarly resources, and discussion with domain experts

6.2.1 Determining Fuzzy Temporal Intervals from Label Information

Labels containing references to absolute chronologies are the result of a mapping chain based on the `E52 Time-Span` class and the related subclasses and properties. The consistent RDF representation of the labels provides a solid base for determining fuzzy temporal intervals using a combination of regular expressions and string parsing techniques.

The structure of the labels, as has been described in section 6.1.1, allows to split them into atomic elements. The combination of the single elements contributes to the definition of the dates for the fuzzy temporal representation. The criteria that have been used are based on a personal choice, and are represented in figure 6.2 with reference to the 2nd century A.D.

More specifically, a basic choice behind these criteria is the symmetric expansion of conventional dates assuming that the more precise the original chronological indication is, the narrower the expansion should be. For example, if a certain element is chronologically attributed to the 2nd century A.D., usually the year 100 A.D. (or 101 A.D.) is used as the conventional beginning date for the period and the year 200 A.D. is used as the end date. Since the expansion for the definition of centuries has been defined in ten years, the fuzzy interval representing the same chronological attribution according to our model is:

$T_{fuzzybegin}$ = 90 A.D.
T_{begin} = 110 A.D.
$T_{fuzzyend}$ = 190 A.D.
T_{end} = 210 A.D.

The criteria schematized in figure 6.2 are considered a good choice in the context of this initial experimentation; different and possibly more composite choices will be

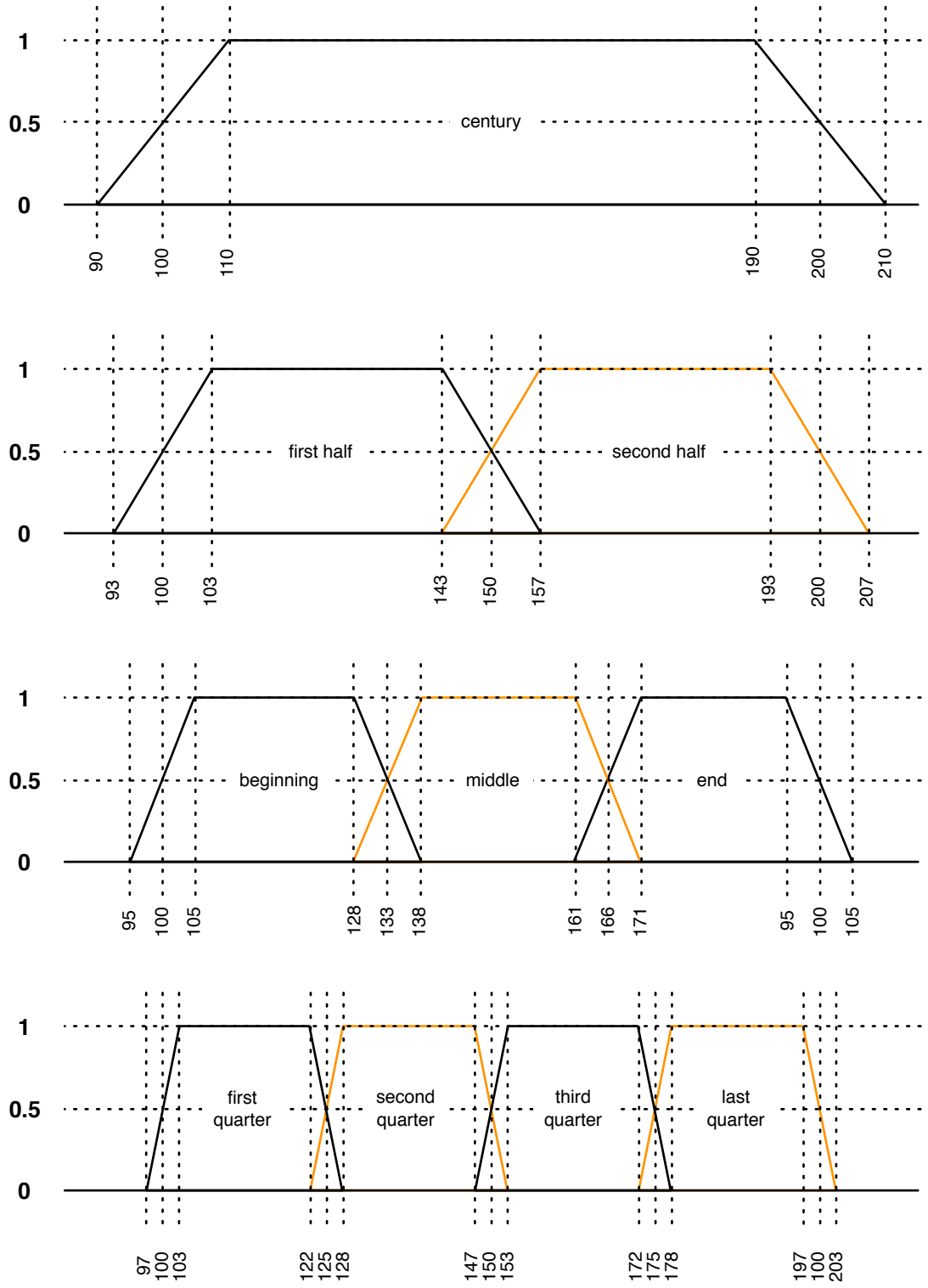


Figure 6.2: *A schema depicting the basic fuzzy intervals that can be represented with reference to the absolute chronology. The example takes into consideration the 2nd century A.D.*

discussed with other domain experts together with the evaluation of the initial results. Intuitively speaking, a possible alternative choice with respect to this example would have to consider:

$T_{fuzzybegin}$ = 100 A.D.
T_{begin} = 120 A.D.
$T_{fuzzyend}$ = 180 A.D.
T_{end} = 200 A.D.

However, this choice would have determined (in our trapezoidal fuzzy set representation) a zero probability for both 100 A.D. and 200 A.D., which is evidently wrong.

The actual implementation of the regular expressions and string parsing techniques determining the fuzzy set representation from label information is part of an *ad hoc* Java program[1] which checks each single labels and determines whether it is formed by a single interval or by two intervals and:

- in case of a single interval the label is processed from right to left, since the further right in the label an atomic element is, the coarser is its granularity in terms of temporal imprecision
- in case of two intervals, both are handled separately and are later combined, using the fuzzified begin of the first interval and the fuzzyfied end of the second interval

Once the labels have been processed, the fuzzy set representations are saved as RDF triples, where each of the temporal properties describes the temporal instance, such as:

```
time-schema:Instance843372351262457
        rdf:type time-schema:Time ;
        rdfs:label "V A.D." ;
        time-schema:earliestStart
                "0390-01-01T00:00:00.0Z" ;
        time-schema:latestStart
                "0410-01-01T00:00:00.0Z" ;
        time-schema:earliestEnd
                  "0490-12-31T23:59:59.999Z" ;
        time-schema:latestEnd
                  "0510-12-31T23:59:59.999Z" .
```

[1]The program, as well as all the other software components that are related to the activities described in this section have been developed by Tomi Kauppinen and Panu Paakkarinen, of the SeCo Research Group at the Helsinki University of Technology

The time-schema structure makes use of `xsd:dateTime`, which is the most diffused and standard data type used for codifying dates according to the XML languages specifications. However, this type is far too precise for the representation of ancient chronologies, since it requires the expression of time up to seconds. For this reason a few conventions have been defined and used consistently throughout the entire dataset for all the elements that are not relevant for our domain, i.e. month, day, hour, minute, second.

6.2.2 Representing Fuzzy Temporal Intervals for Historical and Archaeological Periods

In MANTIC, historical and archaeological periods are defined using the `E4 Period` class and the related classes and properties. The actual dates denoting these periods were not present in the original data sources. Therefore, the periods needed to be temporally located on the calendar by using personal knowledge, the analysis of scholarly sources and the discussion with other domain experts. The result of this activity is a schema 6.3 that have been produced on the basis of the periods mentioned in the data sources, as well as on other periods that have been considered relevant for the research.

It is important to stress that the cultural attributions that are connected to specific periods have not been taken into consideration for this experimentation. In fact, if an artifact is attributed to e.g. the "Paleo-Christian Age", it is evidently considered as the expression of the culture of the first christians, while a contemporary artifact attributed to the "Late Antique" is not. However, the research described here is focused on the temporal and chronological information, and the cultural aspects will be possibly object of future analysis and integration in the system.

The fuzzy set based model we adopted offers the possibility to represent the transition phases between two main periods in a more flexible way. For example, the shift in the control of Milan from the Insubrian Celts to the Romans, and therefore the passage between one period and the other (fig. 6.4) can be modeled using the following conventional dates:

- $T_{fuzzyend}$ (Insubrian period) = $T_{fuzzybegin}$ (Roman period) = 222 B.C. With the battle of *Clastidium* the Romans gained the first significant success on the Insubrian Celts and occupied Milan for the first time.
- T_{end} (Insubrian Period) = T_{begin} (Roman period) = 89 B.C. The *Lex Pompeia* (Law of *Pompeus*) granted *ius Latii* (Latin Law) to the the Trans-Padanian area and the Insubrian mint stopped its activity.

In other cases, a more crisp representation has been possible, such as for the period "Milan capital of the Roman Empire", which has two well defined dates (from 286 A.D. to 402 A.D.).

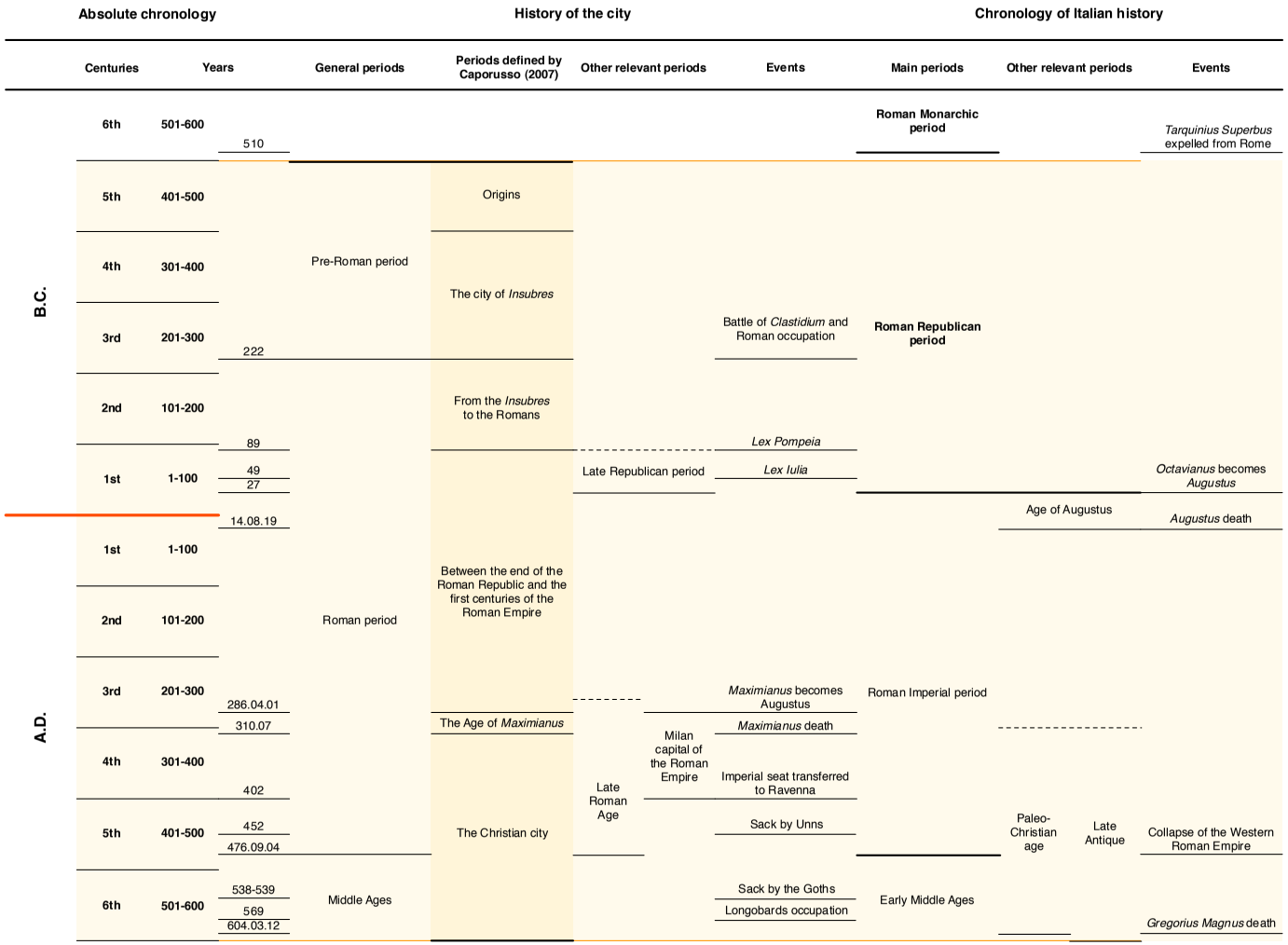


Figure 6.3: *A general chronological schema of the ancient history of Milan.*

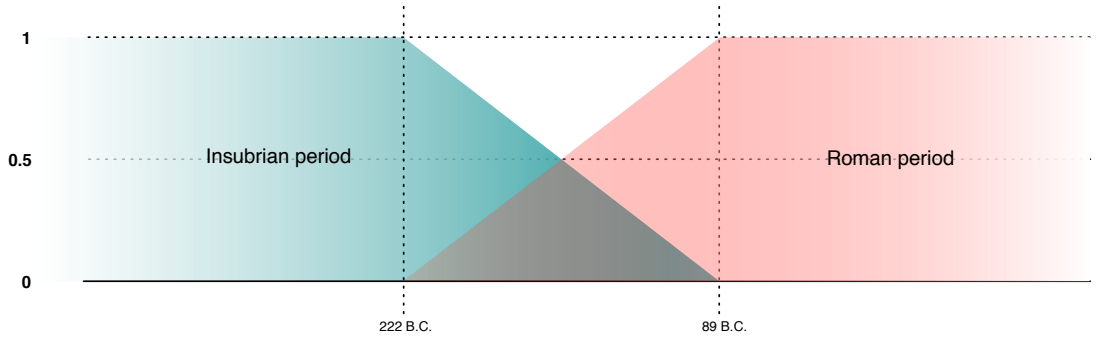


Figure 6.4: *A graphical representation of the transition between the Insubrian and the Roman periods using trapezoidal fuzzy sets.*

6.3 Evaluation Setting and Criteria

The aim of the evaluation has been to measure the correlation between different measures and human opinions[1]. In order to do so, 12 human evaluators were given the task to asses annotation intervals with respect to single query intervals. In particular:

- the annotation intervals refer to the chronological attributions of artifacts. The chronological attribution of structures and monuments has been integrated where available
- the query intervals have been selected according to the schema presented in figure 6.3 and the criteria exposed in the previous section

Therefore, evaluators were asked to asses the relevancy of the chronological attributions of artifacts with respect to a given historical or archaeological period; in more intuitive terms, this corresponds to evaluating the attribution of an artifact to a given historical period.

Twelve query intervals were taken into consideration:

- Six intervals belong to the periodization proposed by Caporusso et al. (2007): the origins, the city of Insubres, from the Insubres to the Romans, between the end of the Roman Republic and the first centuries of the Roman Empire, the age of Maximianus, the christian city. This periodization offers the advantage of being specific to the city, and at the same time, it is easily understandable with little background knowledge.

[1]The definition of the analysis methods and the actual analysis using the methods were carried out by Tomi Kauppinen, Panu Paakkarinen and Heini Kuittinen of the SeCo Research Group at the Helsinki University of Technology.

- Six intervals have been added by the domain expert, for their relevance in the context of the project: the pre-Roman Age, the late Republican period, the Roman Age, Milan capital of the Roman Empire, the late Roman Age, the Late Antique.

The representation of the query periods using trapezoidal fuzzy sets is represented in figure 6.5.

The selected query periods mostly refer to the Roman Age, which represents the biggest and most signicant phase in the ancient history of Milan. They show heterogeneous durations and degrees of temporal uncertainty; therefore, they provide an interesting and varied scenario for the evaluation of the system.

Eight of the evaluators were domain experts, with backgrounds from the fields of history, archaeology and museology; four evaluators were considered as average users. The choice of a mixed group of evaluators represents an interesting choice. On the one hand, expert evaluation offers the possibility to analyze the results provided by the system by comparison to the results expected by professionals, and to fine-tune the retrieval mechanisms accordingly; on the other hand, evaluations by average users provide the possibility to understand how non-professionals perceive the interaction with the system and the relevance of its retrieval capabilities.

6.3.1 An Interface for the Evaluation

In order to efficiently evaluate the suitability of the calculated relevance measures on the annotation and query intervals, a Simile Timeline[1] was created representing the temporal intervals (fig. 6.6). The interface is divided into two bands. The upper band shows the query periods in the form of blue bars, with sky-blue boundaries representing the fuzzy begin and the fuzzy end. Users select one query period by clicking on its bar, and the lower band reconfigures itself in order to display the annotation periods (in red).

When users select an annotation period, the system displays the artifacts (on the right side of the interface) whose chronological attribution coincides with the selected annotation period. The selected query and annotation periods become highlighted in yellow; on the upper-right side, users can then attribute a star rating assessing the evaluation of the relevancy of the annotation period with respect to the query period. Star ratings range from one to ten; evaluators were instructed that all query/annotation pairs with no explicit rating will be treated as having zero stars. The evaluation happens in a highly interactive and visual way; the 12 evaluators evaluated all 12 query intervals with respect to all 66 annotation intervals. The average time employed for evaluating all the possible combinations of the query and annotation periods (nearly 800 ratings) was four hours.

[1] http://www.simile-widgets.org/timeline/

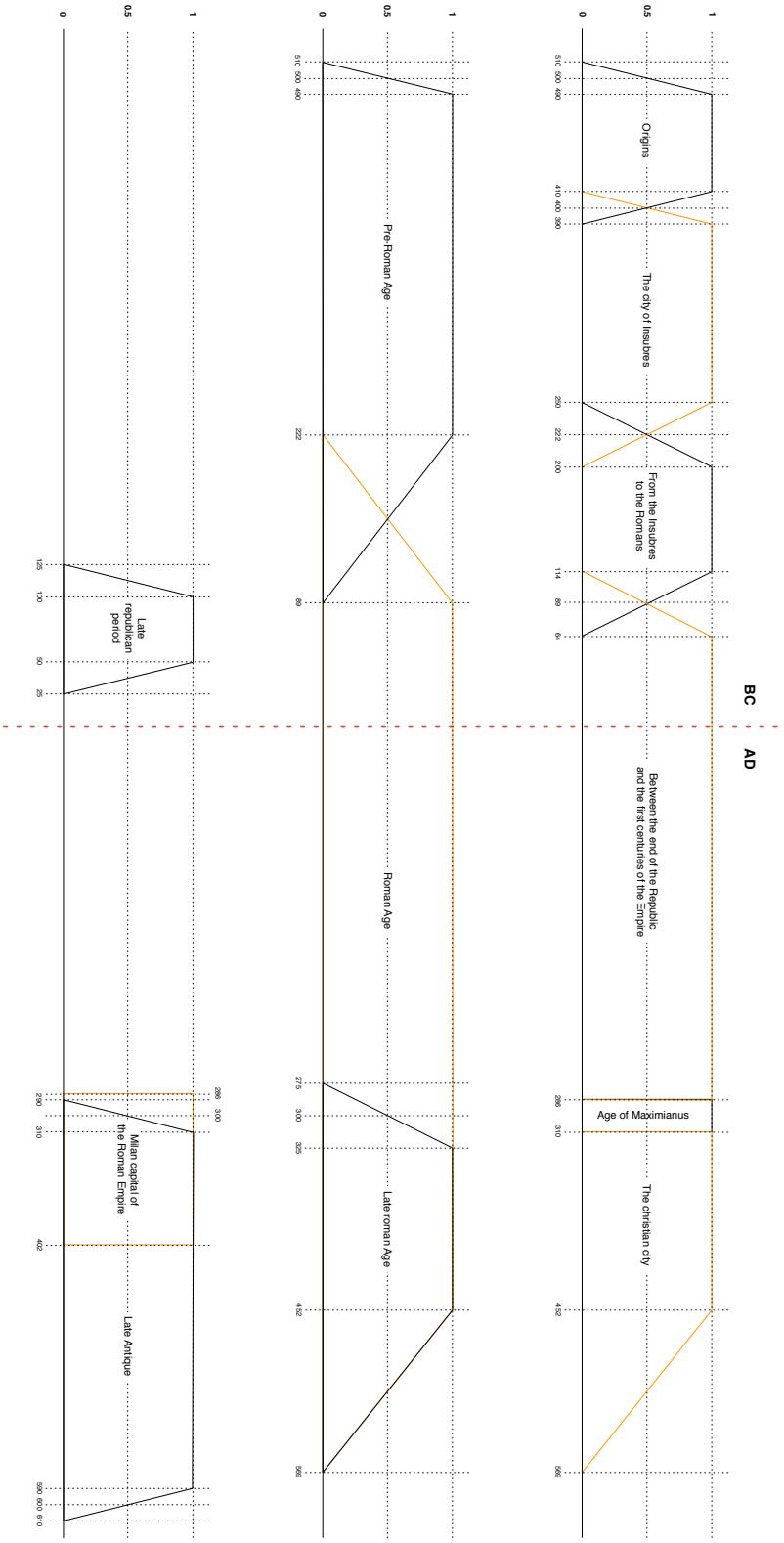


Figure 6.5: *The selected periods and their representation as trapezoidal fuzzy sets.*

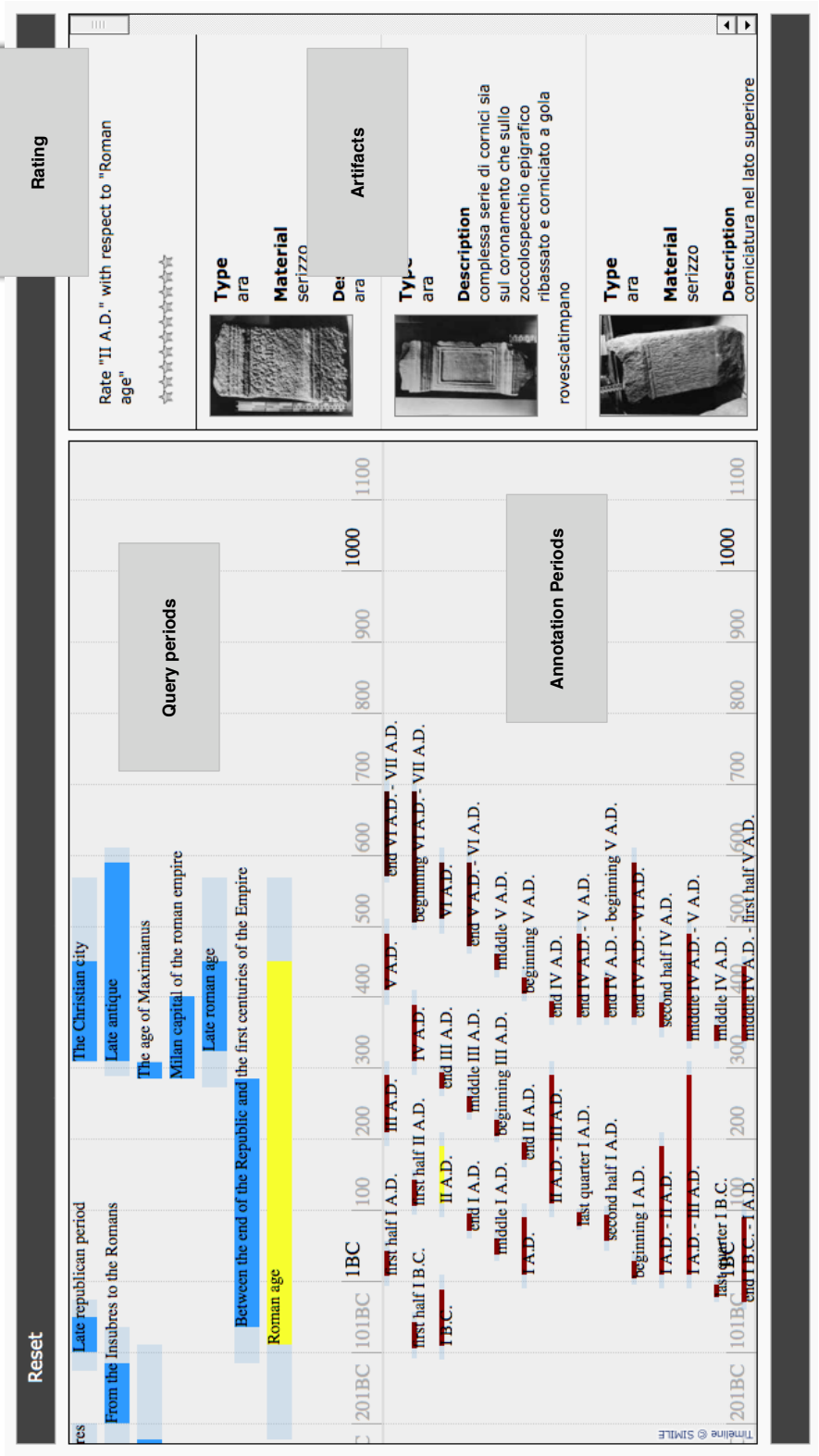


Figure 6.6: *The interface for the evaluation.*

6.4 Results

Weighted kappa (Cohen, 1968) was used in order to analyze the user agreement about relevance levels between query and annotation. The examination of pairwise kappa's revealed that one evaluator among the 12 gave notably different relevance assessments than the others. However, since the other 11 evaluators showed higher agreements, the average of the kappa was 0.83. Therefore, values above 0.8 are considered to reflect perfect agreement; the average of the user opinions is considered as the gold standard for weighting the relevance measures, and for the comparison of the final results.

6.4.1 Defining the Weights for Relevance Measure

In order to define the weights for relevance measures with respect to different possible combinations (e.g. all the three measures or different combinations of two of them) linear regression was used. The values obtained are:

- w_c = 0.13 (closeness)
- w_o = 0.73 (overlaps)

w_{ob} (overlappedBy) has been considered in this setting as 0, since it did not enhance the results as the precision and recall analysis (see section 6.4.2). Therefore overlappedBy was not used as part of the combined measure.

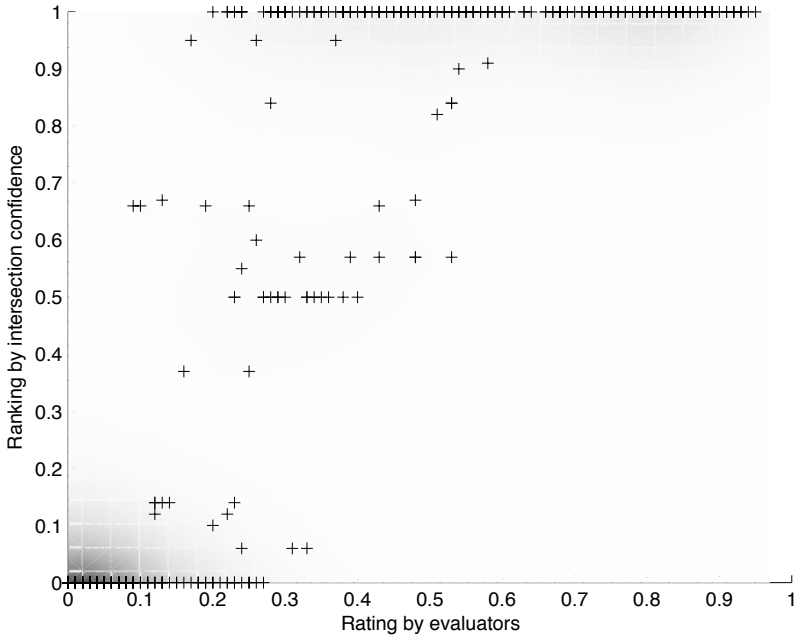
Figure 6.7(a) shows a correlation graph between the evaluators' ratings (X) and the intersection confidence (Y), this latter being the evaluation (in range [0,1]) of the existence of a crisp temporal relationship *intersects* between two fuzzy temporal intervals. The graph can be compared to that of figure 6.7(b), which shows the correlation between the average of the evaluators' ratings (X) and the combination of the closeness and overlap weighted measures (Y). It is evident how the correlation between the combined measure and the evaluators' ratings is remarkable.

6.4.2 Analyzing Precision and Recall

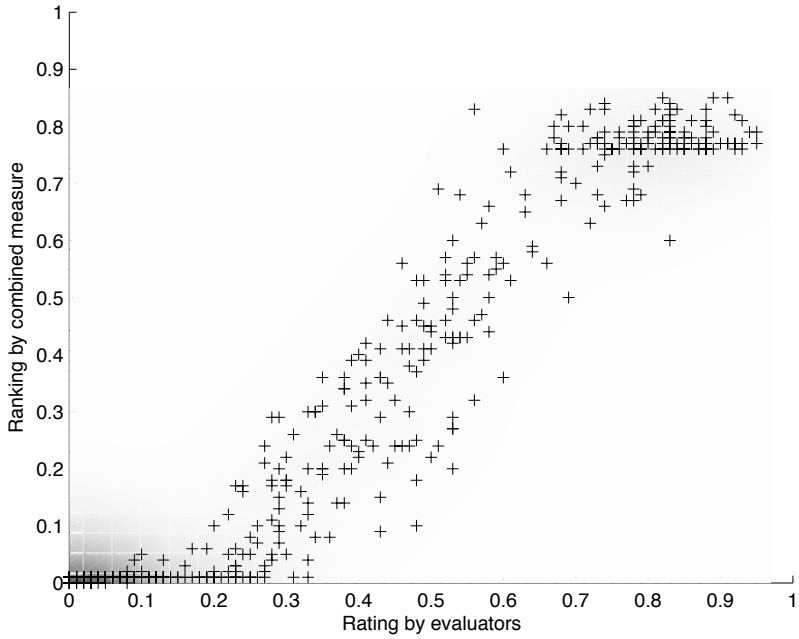
The performance of different measures was evaluated quantitatively using precision figures of 11 recall levels. In particular, the following analyses have been done:

- comparison of the performance of the individual measures (*overlaps*, *overlappedBy* and *closeness*) and the combined measure (using *overlaps* and *closeness*)
- calculation of the performance of the *intersection confidence*

The baseline of the analysis is the binary overlap measure between the crisp intervals, where the value is 1 (relevant) for crisp areas of intervals that overlap, and 0 (non-relevant) for crisp areas that do not overlap. The precision and recall curve shown in



(a)



(b)

Figure 6.7: *Scatter plots of relevance ranking based on (a) intersection confidence and (b) combination of weighted measures (figures by T. Kauppinen).*

figure 6.8(a) shows the results of this analysis, where ratings in range 1–10 stars were considered relevant, and ratings having 0 values were considered non-relevant. The curve shows that the combined measure performs best, the closeness measure is second and *overlaps* and *overlappedBy* are at third place, with *overlaps* giving slightly better results.

In order to improve our evaluation beyond simple relevant vs. non-relevant values, a new analysis taking into consideration the multiple grade relevance assessments (0 to 10) was performed, using generalized precision and recall (Kekäläinen and Järvelin, 2002). The results shown in figure 6.8(b) confirm the best performance for the combined measure, while *overlaps* is second, the intersection confidence is third and *closeness* is fourth.

6.5 Discussion

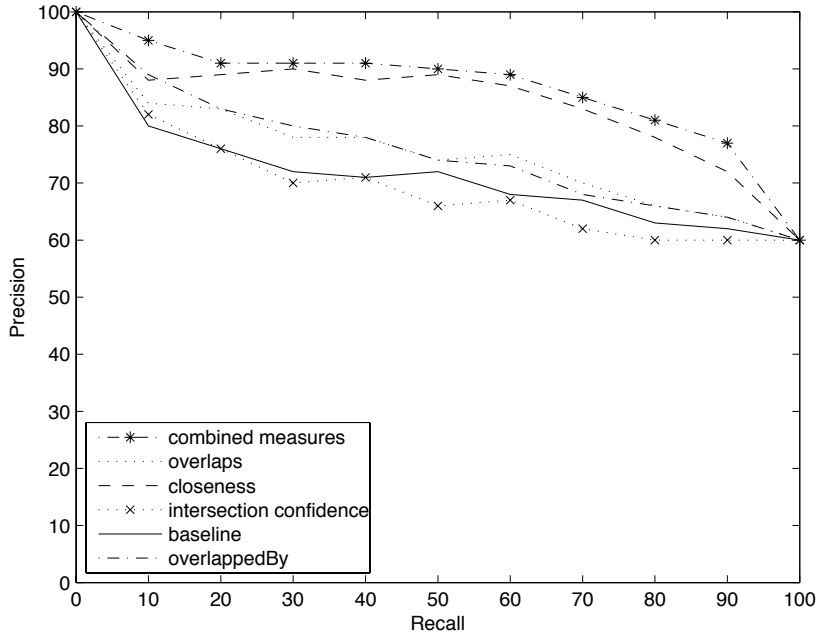
The proposal introduced in section 4.3 and its experimentation in a real application scenario provide an original contribution to the problem of matching query and annotation temporal intervals for cultural heritage information retrieval. Temporal imprecision was modeled using fuzzy sets, and a method was developed in order to obtain a better match between human and machine interpretations of fuzzy temporal intervals in information retrieval. For this we proposed to calculate overlappings and closenesses between annotation and query intervals, and showed how they can be combined together.

The method could be used in e.g. suggesting items from approximately the same period as the reference query period, and also for ranking the relevance of more distant periods of time.

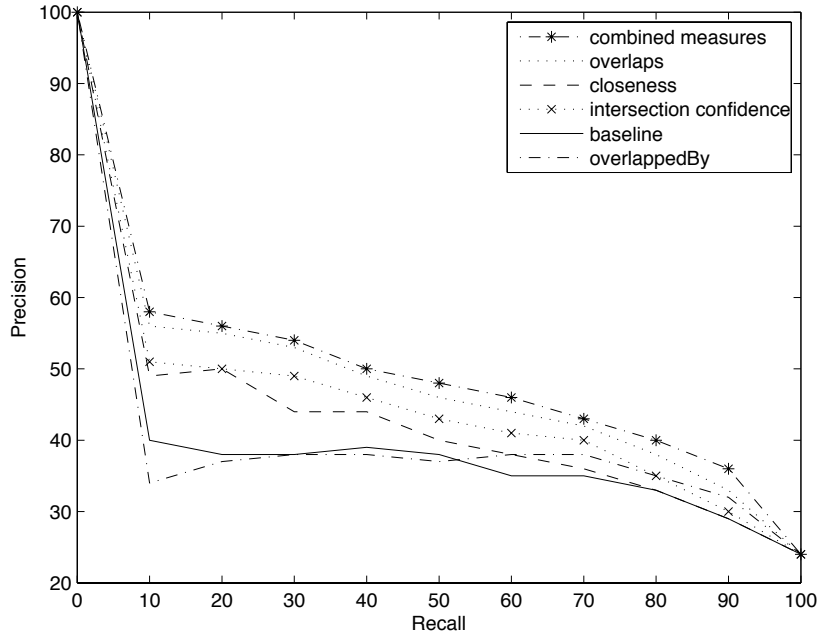
6.5.1 The Model

The definition of a new model for the representation of imprecise temporal intervals was motivated by the necessity to enable effective and efficient information retrieval in a simpler way than using standard models such as the CIDOC CRM and the Erlangen CRM. These model offer a sound and rich approach to temporal representation supporting chronological reasoning, but at the same time they are too complex for the needs of scenarios such as the one we investigated (see section 4.3.3.1). In this regard, our model offers the advantage of providing a much simpler representation, while at the same time, it enables effective and efficient forms of retrieval.

The definition of calendric dates for the representation of fuzzy intervals according to the quadruple $< T_{fuzzybegin}, T_{begin}, T_{fuzzyend}, T_{end} >$ may be seen as an oversimplification of real situations. While this is true for scientific research, where primary evidence is characterized by heterogeneous forms of temporal properties that cannot be reduced to this representation, in the more general situations beyond research and closer to general fruition, the model is considered sufficiently accurate.



(a)



(b)

Figure 6.8: *Precision and recall curves: (a) average recall versus precision for each measure used to rank the results (b) generalized curve of the results (figures by T. Kauppinen).*

For example, the choice of trapezoidal fuzzy sets surely do not take into consideration the internal discontinuity of the temporal characterizations connected to historical and archaeological periods that was introduced in section 4.3. However, also the CIDOC CRM and Erlangen CRM do not specifically mention and face this aspect, which would require a specific research project. Our proposed "fictional" regularity is therefore considered adequate for the context we are dealing with.

More articulated fuzzy set-based models, such as the the one depicted in figure 6.9, should have been used.

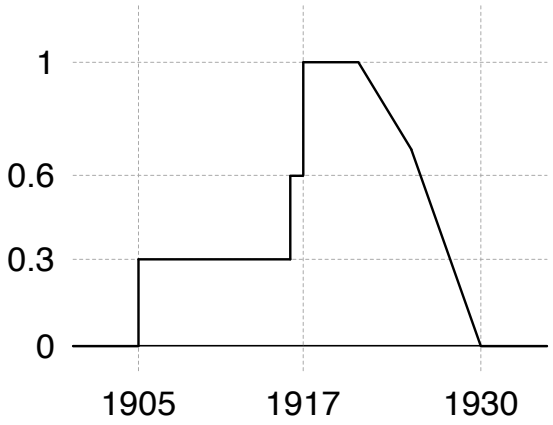


Figure 6.9: *A graphical representation of the fuzzy interval "Russian Revolution happens" (redrawn from Nagypál and Motik, 2003, fig. 1(c), p. 916).*

However, this approach would have required to develop far more complicated reasoning methods, therefore making the advantages of our approach pointless and justifying, on the contrary, the full adoption of the standard models.

On the other hand, the actual implementation of instances of the quadruple required by our model made it necessary to introduce subjective choices, and in particular:

- for references to absolute chronology, the symmetrical expansion of conventional dates and the actual number of years characterizing it, which were based exclusively on personal domain knowledge and experience
- for references to historical and archaeological periods, the review of the scientific literature and the selection of the solutions on which most part of the scholars agree, or the ones that were considered most convincing

The results of the evaluation of our method however seem to positively support the choices, even if the evaluation needs to be extended both for what concerns the dataset and the number of the evaluators.

The integration of our model in the CIDOC CRM structure requires a detailed analysis, which falls is of the scopes of this work. In particular, the correspondence of our concepts and relationships with the CRM's, and the evaluation of the conformity of a possible extension with respect to the model's principles will be possibly conducted once more extensive evaluations of our approach is available.

6.5.2 The Retrieval Method

The evaluation of the method combining calculations on overlappings and closenesses between fuzzy temporal intervals with precision and recall analyses have produced encouraging results.

However, the standard precision and recall analysis and the generalized one provide contrasting indications with respect to the significance of the single measures. In fact, while in the traditional analysis *closeness* performs better than *overlaps*, in the generalized analysis the situation is exactly the opposite. Assuming that the generalized analysis better represents human opinions (since it models relevance with different degrees: 0–10) the result indicates that the more the annotation interval was immersed within the query period, the better rating evaluators gave for the query-annotation pair. On the contrary, the result of the traditional analysis can be explained with the fact that other measures measure the level of overlap in different ways, and they do not notice if intervals are quite close, but do not overlap.

It should be recalled that the research experience described here took into consideration only the chronological information which the references to intervals carried, without addressing the characteristics and value of cultural attributions, which can be very significant for information retrieval, as was previously introduced. Moreover, the fact that the actual representation is based on subjective choices, and that some periods are more open than others to different possible degrees of subjective temporal characterization, inevitably influences the analysis of the results of the retrieval possibilities, even if the results seem to confirm the validity of the personal choices, as was mentioned in the previous section.

The results are moreover restricted to a specific case study, which mainly concerned the Roman Age with a limited internal chronological articulation. Taking into consideration that more extensive experimentation of the model and the method is required in order to further evaluate its effectiveness, the results of the retrieval method based on our combined measure represent an advancement in the area of domain-specific applications for cultural heritage on the Semantic Web.

From a more general perspective, our results extend similar work by Nagypál and Motik (2003), who measured the level of confidence associated to the detection and measurement of a crisp relationship *intersects* between two fuzzy temporal intervals. The approach of Nagypál and Motik (2003) has been evaluated in our case for relevance calculation; however, we demonstrated that the combined measure (overlaps and closeness together) provides better results in terms of precision and recall anal-

ysis. Our experience also extends the work of Visser (2004), providing an evaluation of the proposed calculation of overlap between two fuzzy temporal intervals. Moreover, it introduces the closeness calculation, which was neither considered nor tested in Visser (2004).

7

DemoMANTIC: Design and Development of a Prototype System

The design and development of a prototype MANTIC portal that integrates the proposals, experimentations and outcomes of the research discussed so far are the objects of this chapter. The portal is based on previous experience which made use of a Web mashup approach providing browsing and navigation functionalities on a repository based on a relational database (section 7.1).
Instead, the new prototype system is based on a triplestore managing RDF graphs which has been chosen and set-up after a preliminary analysis and benchmark of prominent existing products (section 7.2). The architecture of the new system, which is introduced in section 7.3, has been compared to the previous version, in particular for what concerns the system performances suggesting the need to define optimization mechanisms (section 7.4).

7.1 Previous Work: The MANTIC 1.0 System

A first prototype Web portal for the MANTIC project has been developed in order to experiment and evaluate different modalities of interactive access to data, according to the new approaches concerning the integration of existing services, and mashups (see chapter 2). The identification of three main use cases (fig. 7.1) served as a guide for the design and deployment of this system.

In particular the faceted paradigm, which has been the object of different and successful experimentation in cultural heritage projects (see e.g. Hyvönen et al., 2004a; van Ossenbruggen et al., 2007), has been investigated. Faceted browsing aims to overcome the problems of traditional query and navigation modalities based on the use of keywords, such as the need to know the specific vocabularies associated to the content, and the target information to be retrieved (see e.g. Hyvönen et al., 2004a).

Using **faceted browsing**, contents can be explored along sets of key dimensions or "facets" (such as style, historical period, material, etc.) and the related vocabularies. This moreover prevents the possibility of performing queries ending with no results, since only the actual categories and terms of the repository are presented. Different facets are combined in a way that the selection of values (i.e. terms) on one of them determines the application of filters on the values of the other facets. Therefore, faceted browsing can be designed with increasing levels of articulation and precision,

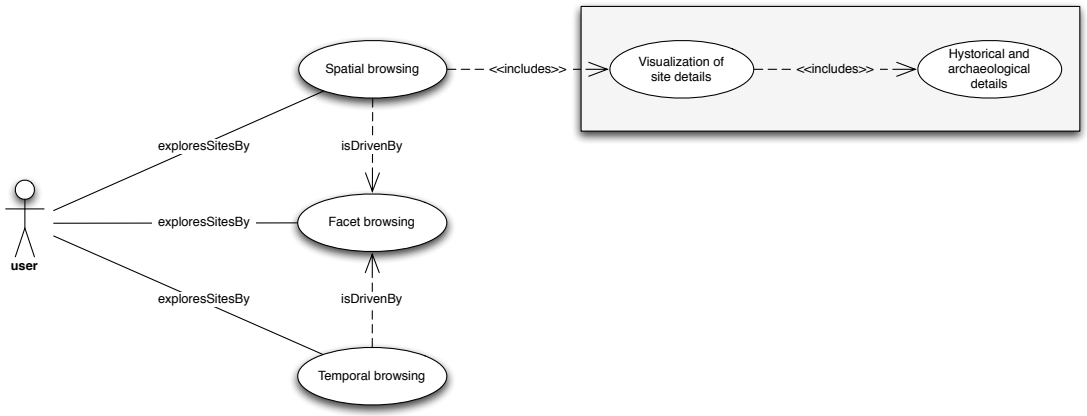


Figure 7.1: *Use cases of the MANTIC 1.0 system.*

allowing intuitive and effective forms of navigation on even complex data structures.

The facet approach has been first demonstrated by the Flamenco system (Yee et al., 2003); specific research effort in the last few years has been devoted to the development of tools for better integration of this approach in existing Web sites (Hildebrand et al., 2006). In particular, work in the context of the SIMILE project[1] has produced an effective faceted browsing service, along with other ready-to-use tools and widgets for managing and publishing structured digital resources (in the RDF and JSON formats).

In the context of our research, a set of basic attributes characterizing archaeological sites, such as their appellation, address, chronological attribution and typology have been considered relevant for query and browsing. However, only the chronological attribution and the typology seem to be perfect for a faceted approach, since they rely upon suitable domain vocabularies. In addition, a third facet related to the appellation of archaeological sites has been implemented, in order to allow a more effective filtering of data.

In addition to faceted browsing, means of visualization and interaction with **the spatial and the temporal dimensions** have been identified as key requirements. The spatial dimension includes the localization, extent and spatial relationships each site had with the others in the urban scale; the temporal dimension includes the historical chronology of the sites. In order to effectively support the navigation through this kind of information, specific interfaces based on WebGIS technologies and interactive timelines have been deployed.

Finally, interfaces for visualizing **detailed information** concerning the archaeological sites (such as textual descriptions, and the specific chronological attributions and geographic location of the site's structures) had to be designed and developed. These interfaces make it possible to also access multimedia materials (images, videos and

[1] `http://simile.mit.edu/`

three-dimensional models), which are increasingly becoming available.

7.1.1 System Architecture

The first version of the MANTIC system has been designed and implemented using a three-tier architecture, which is based on the separation of the data logic, the application logic and the presentation logic (fig. 7.2).

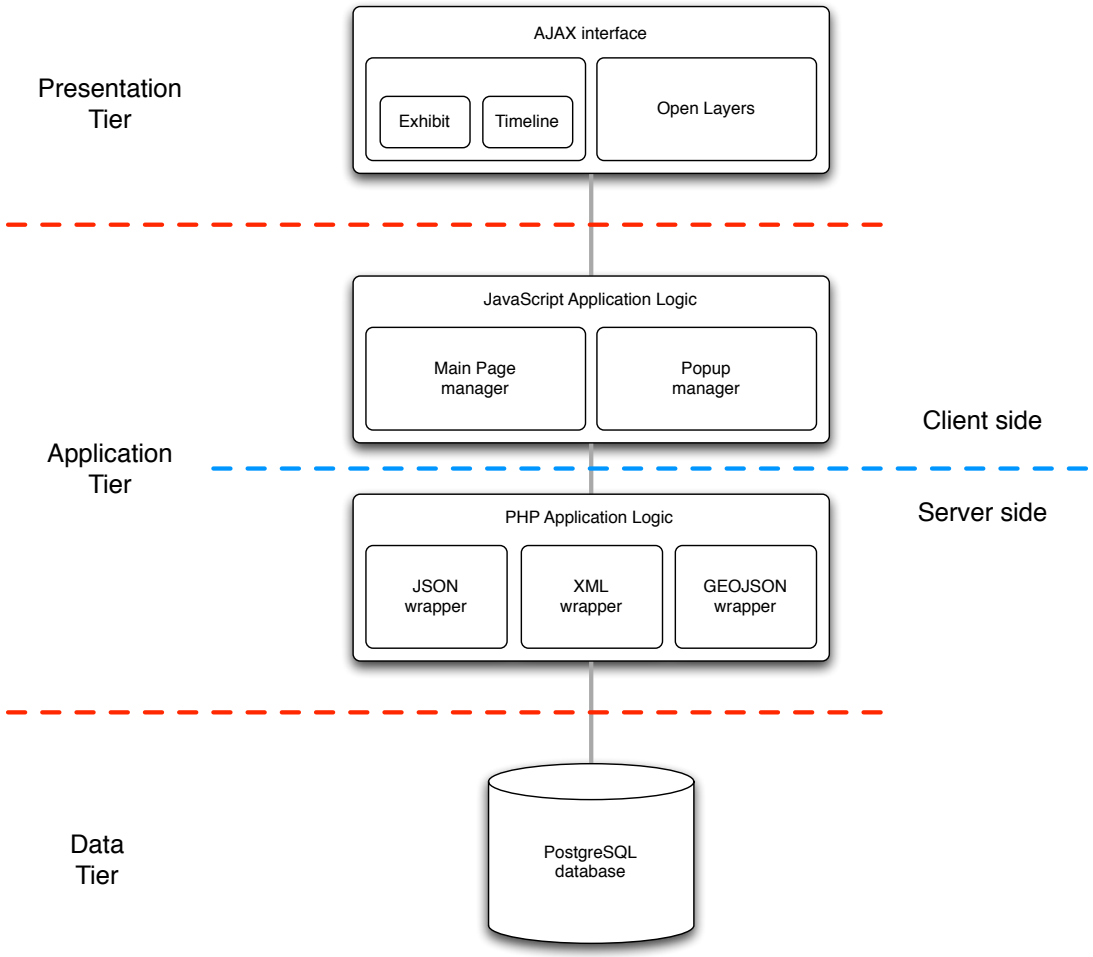


Figure 7.2: *General architecture of the MANTIC 1.0 system.*

Data are stored in the PostgreSQL Object-Relational Database Management System; archaeological sites have been georeferenced manually from maps available in (Caporusso et al., 2007), and their coordinates are managed through the PostGIS extension of the PostgreSQL system.

The application logic is made by server-side PHP scripts that allow to connect

with the database, and to transform data into the specific formats required by the user interface (i.e. JSON, XML, and GeoJSON). On the client side JavaScript code allows to manage user interactions with the interface modules.

The system interface takes the form of a mashup, integrating different content and service sources. In particular, Google Maps[1] (contents) and OpenLayers[2] (services) have been included in order to implement WebGIS capabilities. Faceted browsing relies on the SIMILE Exhibit framework[3], while browsing on temporal elements concerning historical and archaeological chronologies is based on the SIMILE Timeline application[4].

7.1.2 Browsing Functionalities

When accessing the Web site, the interface provides an interactive map with the archaeological sites; the timeline shows the reference chronological interval (500 B.C. to 600 A.D.), while facets do not show any filters.

By selecting a value from a facet, data are filtered: the map and the timeline change their contents according to the selection, and the other facets reconfigure themselves in order to show the values that are relevant to the selected data (fig. 7.3). Navigation through the map is possible thanks to the tools that are present in the upper-left side of the interface, which offer pan and zoom functionalities. Where available, a point or a polygonal representation of the site is visualized, according to different zoom levels. Selecting a point on the map, a bubble showing basic information about the archaeological site appears (fig. 7.4).

Clicking on the "dettagli storici e archeologici" (historical and archaeological details) link in the bubble gives access to the details card (fig. 7.5), which is made of different sections dealing with specific aspects of the archaeological site: a brief textual description, its plan overlaying contemporary aerial photographs, a timeline with chronological attributions obtained with archaeometrical analysis and a thumbnail gallery showing photographs of the site's structures.

7.1.3 Towards a Semantic Web System

The MANTIC 1.0 system represented an important experience and is the result of an intense phase of domain requirement definition, mostly concerning the modalities of browsing and user interaction. The design and the deployment of the system offered the possibility to test different Web mashup tools and services and to evaluate several solutions in detail, ending with the choices described in the previous sections.

[1] http://code.google.com/intl/it-IT/apis/maps/
[2] http://openlayers.org/
[3] http://www.simile-widgets.org/exhibit/
[4] http://www.simile-widgets.org/timeline/

Figure 7.3: *The Mantic 1 interface: selection of a value from the typology facet.*

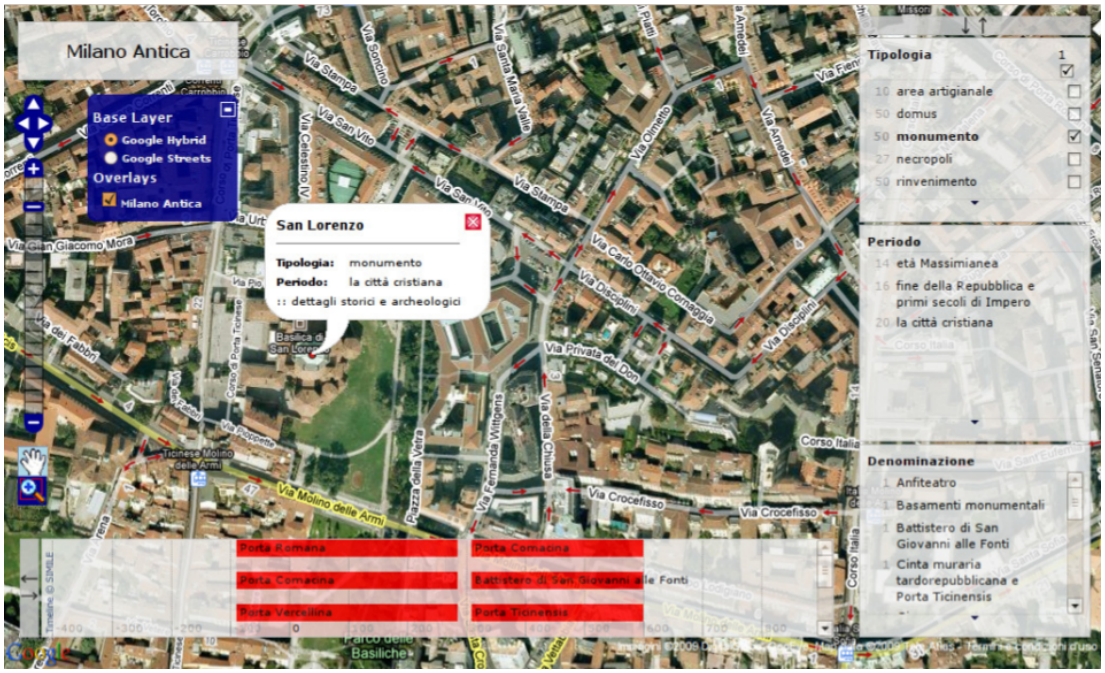


Figure 7.4: *The Mantic 1 interface: selection of a point on the map.*

Figure 7.5: *The Mantic 1 interface: details of a monument.*

Research on the case study presented in chapter 5, and in particular the integration of heterogeneous data sources by mapping to the CIDOC CRM ontology, required the re-engineering of the first prototype, with particular respect to the data-logic, which has to manage RDF data structures.

Therefore, an analysis of the most diffused systems for the storage and management of RDF data structures has been made. The application logic and the presentation logic layers in turn needed to be changed according to the new data logic, and a new architecture has been produced. The following sections summarize the key points of these activities and introduce the basic elements characterizing version 2.0 of the prototype system.

7.2 Semantic Repositories: General Remarks

The architecture of a generic semantic repository can be described using an "onion-model" representation, in which different functional layers cooperate. More specifically, three sub-systems (figure 7.6) can be identified:

- the framework sub-system
- the triplestore sub-system

- the reasoner sub-system

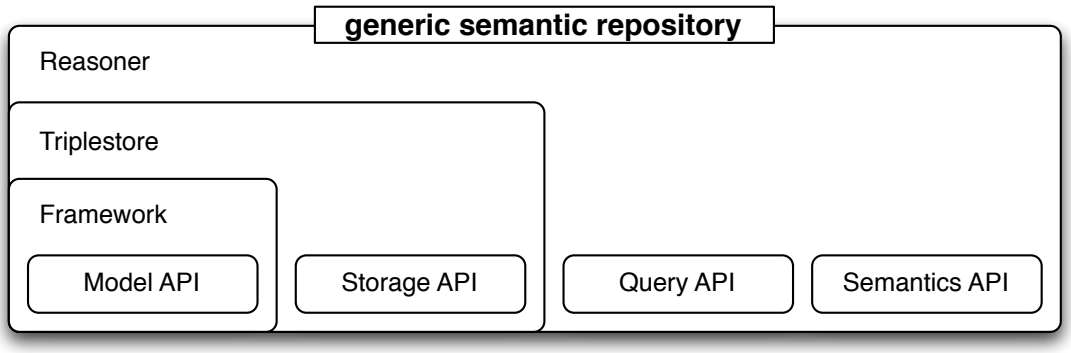


Figure 7.6: *The structure of a generic semantic repository.*

7.2.1 The Framework Sub-System

The framework sub-system represents the core of a semantic repository, and it basically provides two classes of fucntionalities exposed through the "Model API" interface:

- the creation, navigation and manipulation of RDF models in the memory with a support for the representation of resources (S-O), properties (P), triples (S-P-O), and graphs (C-S-P-O)
- the import and export of RDF models in different serialization formats (e.g. RDF/XML, N-Triples, and N3)

7.2.2 The Triplestore Sub-System

The triplestore sub-system makes it possible to provide persistency to RDF models through a native or non-native memorization schema. The "Storage API" is the interface that allows to access this functionality. Native schemata are implemented by the product itself, while non-native ones are managed by third-parties products, which need to be properly configured.

The lightest persistence modality, which is called "in-memory", allows to use the main memory as a storage space for the graph. A dump of the graph for successive use in another session is possible, even if this functionality is not necesseraily offered by semantic repositories. The in-memory modality represents an effective solution as far as the dimension and the complexity of the RDF graph do not exceed a hundred-thousand triples; beyond this threshold, more stable and robust persistence modalities are required in order to preserve scalability.

On the contrary, native persistence makes use of persistent memory (hard disks). In this case, the semantic repository relies on a proprietary memorization schema, which is normally modeled as a quad-store. This solution is extremely efficient, since it makes it possible to exploit proprietary optimization mechanisms that increase the product's scalability up to millions of triples.

Non-native persistence is usually implemented through a schema based on relational databases. The semantic repository makes use of a proprietary relational schema in order to model the database and an interface through which data are managed on the database. Regardless of the specific database management system used, this solution is extremely robust because it relies on the transactional functionalities and the large storage capabilities of relational databases.

Finally, remote persistence is possible in some semantic repositories. It is not a true memorization schema; instead it represents the possibility to deploy the datastore on a server which, in turn, makes an HTTP endpoint available for queries. The server's reply to the query consists of an RDF file that contains the results of the elaboration; this service is usually implemented using a REST protocol.

7.2.3 The Reasoner Sub-System

The reasoner sub-system represents the external layer of a general semantic repository model. It provides two functionalities that are exposed by specific APIs:

- Query of the RDF model (→ Query API): allows to send a query in a suitable language and to obtain a binding of variables corresponding to the selected triples. This functionality is conceptually similar to queries on relational databases. The query language should at least satisfy four properties:
  - full support to the RDF model and absence of syntactical constructs that are specific to a given serialization syntax
  - support to the semantic extension of the RDF model, at least to the RDF Schema level of expressivity
  - support to XML Schema datatypes
  - support to incomplete or contradictory information in the query, since it is not possible to assume that it always carries complete information on each resource
- Semantic extension of the RDF model (→ Semantics API) through a series of primitives for the formalization of domain ontologies at different levels of expressivity.

A semantic repository can implement and expose a proprietary Semantics API, as well as integrate a third-party one. In the latter case an external reasoner is added to the the native one in order to improve the inference capabilities of the product.

7.2.4 Technological Survey

This section introduces a brief overview of some of the semantic repositories that are available today, both in the form of open source and proprietary software. The goal of the survey is to identify the main characteristics of each product , in order to identify a suitable solution for the MANTIC 2.0 prototype.

Figure 7.7 represents a schema of the analyzed solutions according to the semantic repository model introduced in the previous section.

The solutions supporting the Jena model through their Model API can be considered an optimal option, since Jena offers a complete implementation of the RDF model. With the exception of Sesame, all the products show this characteristic. On the contrary, the possibility to import and export RDF models in different serialization formats is always present, and it is generally implemented with read and write operations of RDF/XML and N-Triples serializations.

With reference to the triplestore sub-system, a native memorization schema is considered the best option, since it allows to fully exploit the optimization mechanisms provided by the products, and it offers optimal performances from the point of view of e.g. scalability. Native memorization is present in all the analyzed products, with the exception of Jena. In particular, Mulgara, AllegroGraph and Virtuoso offer better scalability than Jena and Sesame.

The possibility to expose a datastore through a SPARQL endpoint that can be accessed via HTTP, and the REST protocol is a value added and a useful functionality for the creation of Web Services. SPARQL endpoints can be enabled in all the considered products, even if its configuration is not always straightforward. For example, Jena does not support the initialization of an RDF server by default, and the Joseki server needs to be added; on the contrary, Sesame makes it possible to deploy a .war archive in a server container (such as Apache Tomcat) and to easily create a SPARQL endpoint. Mulgara, AllegroGraph and Virtuoso require a preliminary configuration through initialization scripts in order to be loaded with a SPARQL endpoint support.

SPARQL is currently the best option for the Query API of the reasoner sub-system. Even if all the products support this language, only AllegroGraph and Virtuoso support it natively, while the other products support other proprietary query languages. Regardless of the characteristics of these languages (RDQL for Jena, SeRQL for Sesame, and TQL for Mulgara), it needs be stressed that they may introduce proprietary restrictions on the RDF model, which can negatively influence the queries. With respect to the Semantics API, the current optimal solution seems to be the support for OWL, at least at the basic Lite level, together with the possibility of defining personalized rules for the reasoners: Sesame and Virtuoso offer support to inferences on RDFS; Mulgara and Jena offer reasoners based on proprietary syntax; AllegroGraph integrates a Prolog interpreter which eases the definition of logical rules.

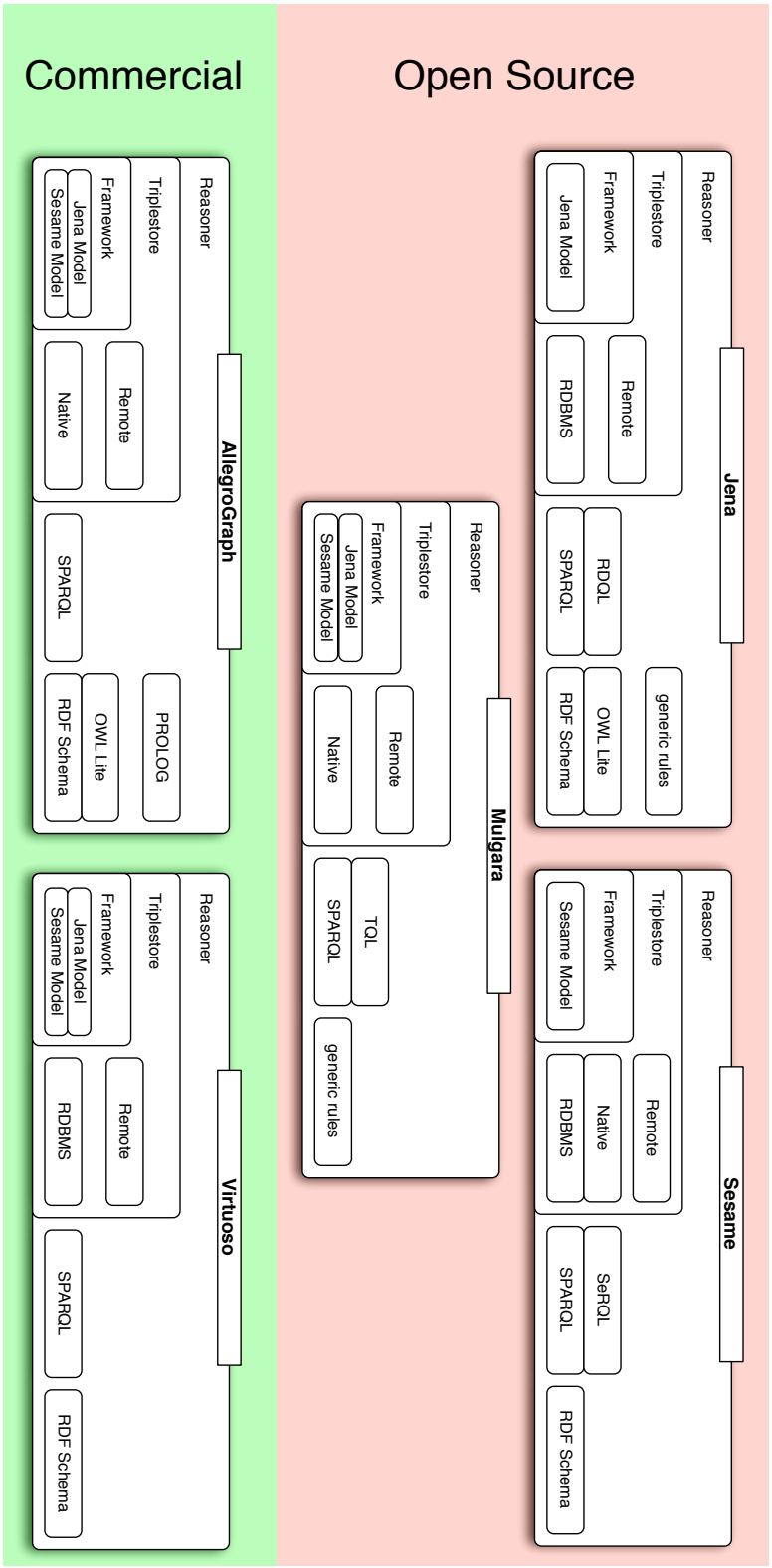


Figure 7.7: *The analyzed technologies and their models.*

7.2.5 Evaluation

In order to better compare the products and choose a suitable solution for the MANTIC 2.0 system, an evaluation has been made on the basis of the basic dimensions of the Model API, the Storage API, the Query API, and the Semantics API. In addition, other parameters have been taken into consideration, such as the presence and quality of the documentation and the availability of management tools.

The evaluation is based on scores that have been subjectively attributed to the aforementioned characteristics, and according to this schema:

- Model API:
  - 1: third-party Model API
  - 2: proprietary Model API
- Storage API:
  - 0: no SPARQL endpoint
  - 1: SPARQL endpoint
  - 2: SPARQL endpoint and native schema
- Query API:
  - 0: no SPARQL
  - 1: non-native SPARQL
  - 2: native SPARQL
- Semantics API:
  - 0: no minimum requirements
  - 1: inference on rules
  - 2: inference on rules and OWL-Lite
- Documentation:
  - 0: incomplete
  - 1: complete, without examples
  - 2: complete, with examples
- Managament tools:
  - 1: command line
  - 2: dedicated Desktop application
  - 3: dedicated Web application

- Distinctive and relevant characteristics of the single product:
  - 1 for each characteristic

Table 7.1 and 7.2 show the results of the evaluation.

Table 7.1: *Evaluation of the basic dimensions of different semantic repositories.*

Product	Model API	Storage API	Query API	Semantics API	Total Score
Jena	2	1	1	2	**6**
Sesame	2	2	1	0	**5**
Mulgara	1	2	1	1	**5**
AllegroGraph	1	2	2	2	**7**
Virtuoso	1	2	2	0	**5**

Table 7.2: *Evaluation of supplementary dimensions of different semantic repositories.*

Product	Documentation	Management Tools	Distinctive Characteristics	Total Score
Jena	2	1	1	**4**
Sesame	2	3	2	**7**
Mulgara	0	1	0	**1**
AllegroGraph	2	3	3	**8**
Virtuoso	1	3	0	**4**

The final evaluation combines the scores defined for the basic and the supplementary dimensions of each product, and it is shown in table 7.3.

According to our evaluation, AllegroGraph represents a complete and suitable solution for the MANTIC 2.0 system. Distinctive characteristics of this product are the availability of APIs for modeling spatial and temporal data, which open interesting perspectives for testing sophisticated forms of spatial and temporal reasoning. Moreover, the APIs for the federation of triplestores allow to efficiently organize the design of the global datastore.

7.3 The MANTIC 2.0 Architecture

The flexibility of the three tier model simplified the re-engineering activity of the data tier (figure 7.8). In particular, the introduction of a semantic repository in the applica-

Table 7.3: *Final evaluation of different semantic repositories.*

Product	Basic Dimensions	Supplementary Dimensions	Total Score
Jena	6	4	**10**
Sesame	5	7	**12**
Mulgara	5	1	**6**
AllegroGraph	7	8	**15**
Virtuoso	5	4	**9**

tion stack made it necessary to modify the application logic on the server side, adding specific components for the translation of data from the RDF model to an interchange format based on the JSON format. This component is made of several Java servlets, managing user interactions with the interface through the translation of requests into SPARQL queries. Queries are then sent to the Query API, and the results are produced in formats that can be directly managed by the user interface (fig. 7.9).

The triplestore is populated using an Extract-Transform-Load (ETL) approach, i.e. data are retrieved from the single data sources, normalized, translated into an RDF structure according to CIDOC model and the mapping templates, and finally loaded into the triplestore. This approach has been preferred to dynamic ones, not only because of its simpler deployment, but also since the rate of update of the legacy data sources is very low (approximately a few times a year), therefore making "live update" of the system currently irrelevant (in a perspective similar to Isaksen et al., 2009). Moreover, specific code has been written in order to face data quality issues at the data level that have been discussed in section 5.5.2.

Data and control flows between the application and the presentation tiers are managed by AJAX-based methods; the modalities through which data are managed by the SIMILE Exhibit and Timeline services have been completely re-designed. More specifically, data structures are currently managed in the main memory, and they are sent to the interface through real-time API calls. This approach, which substantially differs from the one adopted in the MANTIC 1.0 system, simplifies the maintenance of the new prototype and improves the overall protection of confidential data.

A servlet is available for every single component of the interface, therefore making it possible to better control the sending of the requested data during browsing. A typical data and control flow related to the interaction with the interactive map is characterized by the following steps:

1. event: user selects a value from a facet

2. control flow (client): the system catches the user's request and analyzes the new

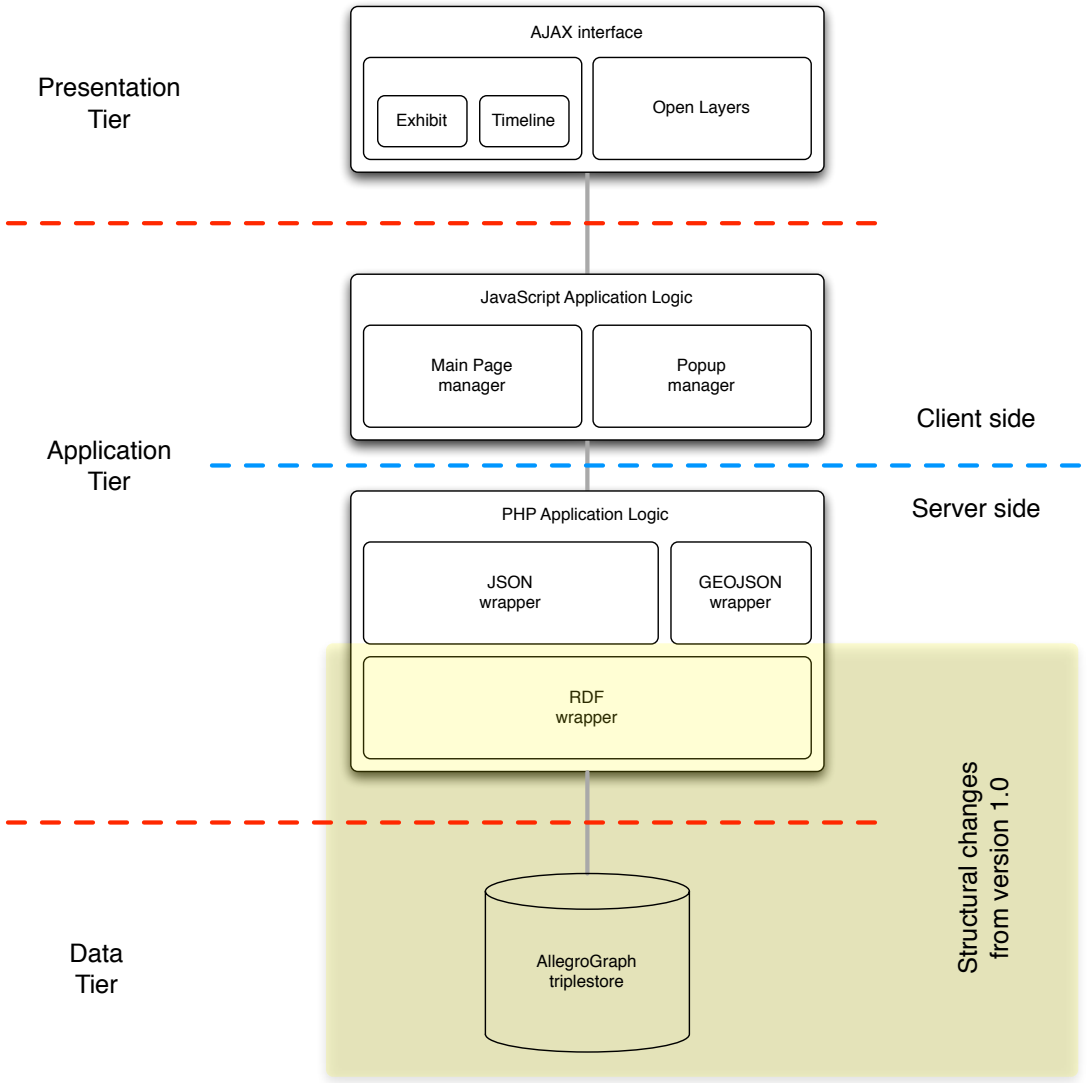


Figure 7.8: *General architecture of the MANTIC 2.0 system.*

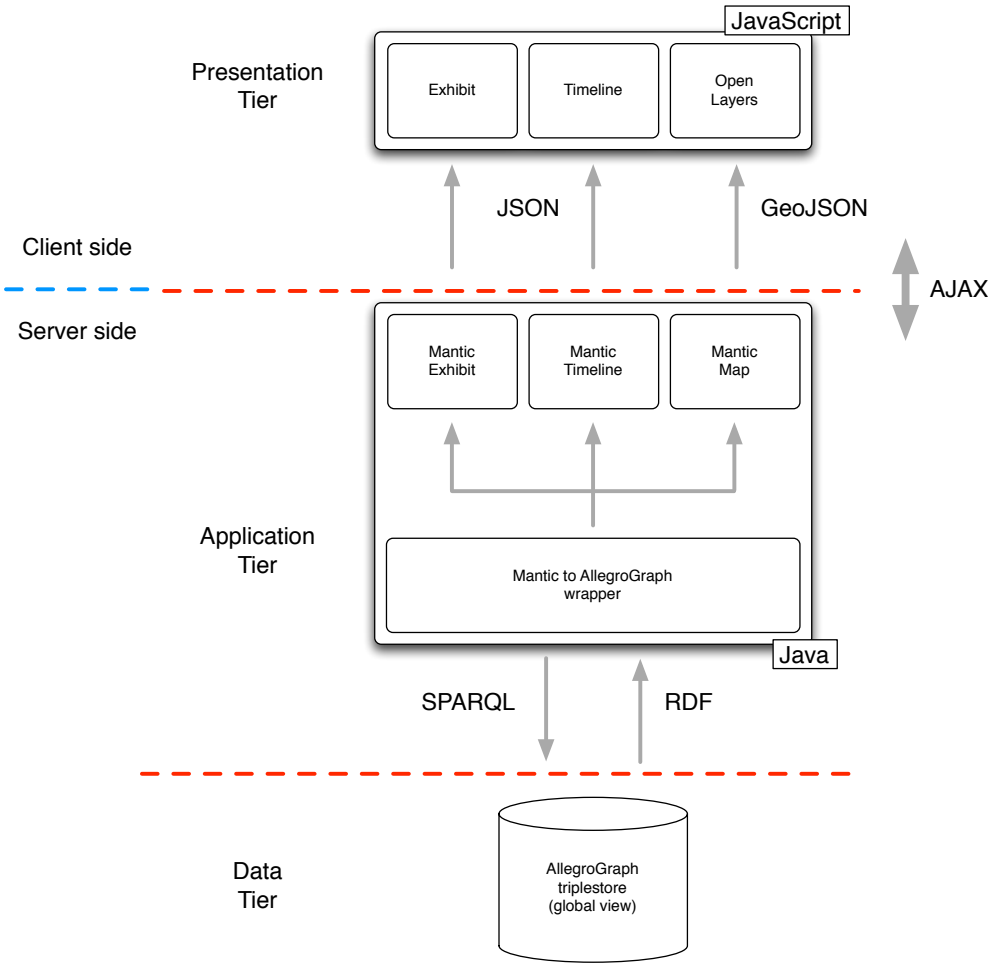


Figure 7.9: *A general schema of the data and control flows in the MANTIC 2.0 system.*

facet configuration

3. control flow (client → server): the system activates a connection through AJAX with the MANTICMap servlet, and sends the configuration as a parameter
4. control flow (server): the MANTICMap servlet, which is invoked through the doGet method, sends a SPARQL query to the reasoner sub-system of the semantic repository, using the Query API through the MANTICAGWrapper library
5. data flow (server): the MANTICMap servlet obtains the results of the query from the semantic repository, and it creates a GeoJSON file on them
6. data flow (server → client): the MANTICMap servlet sends the GeoJSON data structure to the system

7. control flow (client): the system invokes OpenLayers in order to create a new map containing the points corresponding to the results; the map is then updated and the request is completed

7.4 Concluding Remarks

The MANTIC 2.0 system has been designed and implemented aiming to substantially improve the previous version substantially, while at the same time preserving all the elements that have been proved to be satisfactory, such as the end-user browsing functionalities. Most of the work concerned the re-design of the data-tier and the subsequent revision of data access methods, while the visual appearance of the interface underwent minor changes.

The design and deploy of a semantic backend required both conceptual and technical work, and the advantages of the new solutions are gradually paying off in terms of e.g. the flexibility of the system in the integration of additional information. Moreover, the passage to a semantic system makes it finally possible to investigate semantic services (such as those discussed in Hyvönen, 2009) in more detail and to experiment with innovative functionalities for end users.

However, an empirical comparison of the system performances with version 1.0, indicates that the introduction of a semantic backend greatly increased the time required for loading the interface and for interacting with it. Therefore, a benchmark of the performances has been made (fig. 7.10).

The benchmark consisted of two phases:

1. The temporary modification of the servlets' code, in order to introduce timestamp during specific executions.

2. The execution of calls to specific servlets using well-formed URIs, in order to simulate traditional HTTP/GET requests, which are normally received through AJAX methods by the client-side application logic. In particular, three invocations have been executed for each selected servlet, in order to evaluate the execution flow of every single doGet method precisely, and the following evaluations have been made:

   (a) the time required to federate the triplestore and to connect to it, i.e. the interval from the beginning of the doGet method to the availability of the global view

   (b) the time required to execute SPARQL queries through the servlet, i.e the interval from the query request to the MANTICAGWrapper to the restarting of the servlet's execution flow

   (c) the time required to generate the JSON data structure, i.e. the interval required in order to obtain the JSON file to be sent through AJAX to the interface

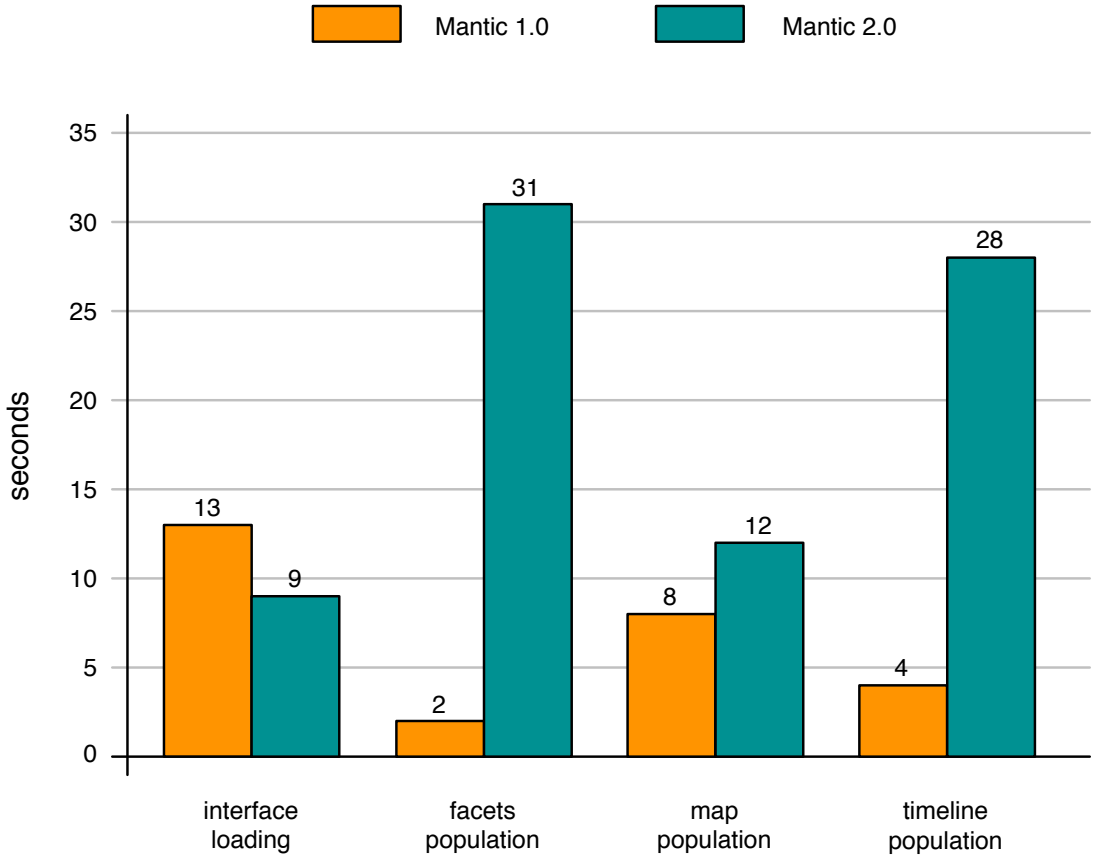


Figure 7.10: *Benchmark of the performances of the MANTIC systems.*

The MANTICExhibit servlet is invoked once during the session; when the application is launched, it retrieves the binding of variables containing data about the typological and chronological classification of sites and structures from the semantic repository, which is requested in order to populate the SIMILE Exhibit facet browser. The benchmark shows that most of the computation time for this servlet is actually used in order to execute the SPARQL query on the typological classification. In this regard, the AllegroGraph SPARQL engine does not seem as efficient as traditional DBMS SQL query engines. Qualitative tests have shown that splitting the query into parts that are then combined through JavaScript code makes it possible to decrease the execution time while retrieving the same results. On the contrary, the chronology classification query shows a reasonable execution time, which does not require such a work-around.

Therefore, it seems that the query optimization algorithms that are typical of RDBMS still do not have an adequate counterpart in the young sector of semantic repositories.

However, the AllegroGraph reasoner subsystem is constantly subjected to improvements that aim to define better strategies of query management[1], and it will likely perform better in the near future.

The MANTICmap servlet is invoked when the system is launched, after the execution of the MANTICExhibit servlet, and every time facets change their state. The task of this servlet is to retrieve the binding of variables containing data about the geolocalization of archaeological entities from the semantic repository, which is required by the OpenLayers interface component. Since this servlet performs a complex query, its long execution time is partially justified; however, user experience suffers from this delay tremendously, therefore requiring consistent optimization.

The MANTICTimeline servlet is invoked when the system is launched, after the execution of the MANTICMap servlet and whenever a facet changes its state. Its task is to retrieve the binding of variables containing data about the chronological context of the archaeological entitites from the semantic repository, which is required by the SIMILE Timeline interface component. The benchmark shows a situation that is similar to the MANTICExhibit case.

A preliminary solution to the performance problems introduced here may consist in the re-engineering of the AJAX methods for the invocation of the servlets. More specifically, control flows should be parallelized in order to improve the interface response times to changes in the facets configuration.

[1] See: `http://www.hpl.hp.com/techreports/2005/HPL-2005-170.html`

8

Concluding Remarks and Future Directions

There is no doubt today that the trajectories along which the Web is evolving and the richness and heterogeneity of the cultural heritage domain make interdisciplinary research combining the two areas an exciting challenge, as well as a fertile ground for original contributions to the study of the Information Society. On the other hand, the frequent skeptical attitude of the communities of cultural heritage professionals towards the promises of the new technological era can only be overcome with specific contributions providing effective results for the cultural heritage domain.

This PhD research aimed to contribute to this scenario from a research perspective where the evaluation of the feasibility of the new approaches (with particular respect to the few standard solutions existing to date) is coupled with more "vertical" contributions concerning specific domain aspects and issues. A transversal and interdisciplinary research path going from the representation to the fruition of data, information and knowledge has been proposed as the high-level framework on which relevant and inter-related key aspects have been identified.

The analysis of both the current trends in the development of the Web, and the existing proposals and projects in cultural heritage and archaeology represented the starting point of our investigation. On the one hand, the Semantic Web/Web of Data scenarios, together with the new approaches in the deployment of Web applications by combining existing services and data, emerged as the technological reference frameworks for the research. On the other hand, the comprehensive review of the most relevant projects that adopted these approaches in the context of cultural heritage and archaeology (with particular respect to the Semantic Web) made it possible to precisely identify the key aspects on which to focus, according to the aforementioned interdisciplinary perspective.

This review demonstrated that even if the more theoretical principles and proposals of the Semantic Web (and the related visions and frameworks) can be fruitfully understood on the basis of the vast literature existing today, the analysis of real applications and systems is of primary relevance, since it dramatically improves the evaluation of the feasibility of the new approaches. For this reason, a key direction and contribution of our research has been the joint discussion of theoretical and methodological aspects, together with the discussion of more practical issues at the application

level. These discussions have greatly benefited from the design of a well-delimited and real case study and the development of a prototype system.

A central aspect this PhD research investigated has been the issue of representation, in terms of the approaches for modeling domain data, information and knowledge on the Semantic Web. Beyond being a fundamental aspect in the context of the Semantic Web, which deeply influences the design of applications and systems, the area of domain knowledge representation is the one in which a notable result has been obtained in the last few years, with the standardization of a core and domain ontology (the CIDOC CRM). Our discussion specifically took into consideration the in-depth analysis, experimentation and discussion of this model, which is based on an event-centric approach that differs from traditional cultural heritage documentation consistently, and therefore required a preliminary introduction. Thereafter, crucial issues that are connected to the use of the CRM in real systems have been identified as the elements motivating new research, in terms of novel experimentation and discussion of the model in relation to the identified case study. Finally, the proposals of a mapping workflow and mapping documentation criteria, as well as the evaluation of the characteristics of the available serializations of the CRM in Semantic Web languages (principally RDFS and OWL) have been set as the methodological framework underlying and guiding this experimentation and discussion.

The results of the research following this approach are particularly valuable for cultural heritage professionals who aim to design and/or take part in the increasing number of projects where the semantic integration of heterogeneous metadata schemata using the CIDOC CRM is required. In fact, there is no doubt that the role of domain experts in defining appropriate mappings of legacy metadata schemata to the model is fundamental (as has been particularly stressed in Nussbaumer and Haslhofer, 2007; Binding et al., 2008), but there are still a few contributions that provide a description and a discussion of this activity in all its complexity, together with full documentation of the mapping templates. Nevertheless, this approach is fundamental in order to make it possible for cultural heritage professionals to evaluate, understand and effectively use the model, and it should constitute the norm rather than the exception in similar experimentations.

Our research moreover represents the first extensive contribution available in the Italian context to date, and it provides a new proposal for mapping a small set of national standard metadata elements for cataloging archaeological artifacts to the CRM.

The discussion of the elements taking part in this scenario, as well as the documentation of the mapping templates have been presented according to the event-centric perspective of the CIDOC CRM. This peculiar characteristic of the model usually complicates the initial understanding of the CRM by domain experts, especially if there is in parallel little familiarity with the principles of object-oriented modeling; in addition the ambiguities that are inevitably connected to the textual scope notes of the model's documentation further increase the learning curve. As a result, domain ex-

perts may feel disoriented and discouraged in approaching the CRM; therefore, a clear case study and a consistent documentation today represent references of primary relevance. In fact, once the principles and the general structure characterizing the CRM are clearly understood, the model is surely an effective resource since it combines great expressivity and a sound formal framework with notable compactness with respect to the vast and complex domain of cultural heritage.

Nevertheless, the research also confirmed the central relevance of the issues that have been identified in the use of the CRM in real application scenarios. The extensive evaluation and discussion of these issues with reference to the case study offer new contributions to the debate on the structure, principles and practical implementations of the model.

In particular. the definition of event-based mapping chains has been the basis for verifying and highlighting situations where alternative choices in mapping metadata schemata to the CRM are possible (different chains for equivalent metadata), as well as situations where additional information need to be modeled in order to prevent the creation of identical chains for different metadata. The introduction of "virtual entities" is moreover required in a number of cases in order to explicitly represent events that are only implicitly present in legacy metadata schemata. All these factors determine the exponential growth of the number of triples constituting the mapping chains, which besides confirming the richness of the model and its effective support in making implicit information explicit, should require optimization techniques at different levels. A solution that is provided directly by the model are the so called "shortcut properties" which can be used in all cases where the loss of information that is inevitably connected to substitute fully developed paths with shortcuts is not critical.

On the other hand, other characteristics such as the generality of the model and the absence of implementation-specific storage and processing criteria (which is coherent with the purely conceptual nature of the CRM) required to enlarge our analysis perspectives towards the existing solutions and serializations. Among these the Erlangen CRM has been identified as the most accurate OWL-DL serialization, since it relies on well advised research work and introduces the required elements, such as data types that are compatible with the CRM classes.

Finally, the problems of managing vocabularies and establishing coreference links between different subgraph resources that relate to identical real world entities have been introduced, but their analysis did not go into details, therefore requiring future and specific work.

On the other hand, a more "vertical" research contribution has been proposed in the context of the analysis of fuzzy temporal information and fuzzy chronologies, and the possibility to design innovative methods exploiting them for the retrieval of relevant cultural heritage information. The choice of Achaeology as the domain on which to concentrate with respect to these aspects proved to be particularly significant, since it made it possible to examine a complex scenario where fuzzy temporal information

assumes a crucial relevance.

In fact, temporal fuzziness is an intrinsic characteristic of archaeological chronologies, since archaeological dating is in the quasi totality of cases an estimate rather than an "absolute" and precise chronological attribution. A number of contributions dealing with archaeological dating techniques and chronology-building investigated the possibilities of improving the precision of temporal data, for example using statistical methods and techniques. Nevertheless, temporal imprecision and fuzziness are inevitable in archaeology, and they necessarily have to be deeply taken into consideration when designing Semantic Web systems.

While there is general agreement on the importance of these aspects, only an extremely small number of specific researches in this area exists. The most relevant effort to date (the CIDOC CRM temporal representation model) is exceptionally accurate and sound, and it offers the possibility to represent both qualitative temporal relationships between temporal intervals (on the basis of Allen's temporal interval algebra; see Allen, 1983) and more quantitative aspects, such as duration. Thanks to this the model supports consistent chronological reasoning, for example by drawing chronological consequences in order to validate existing scientific interpretations or by producing new interpretations. However, in contexts where the principal aim is not to support automatic interpretation, but to improve the retrieval of meaningful information in and beyond the community of experts, the model may be too rich and complex to be efficiently managed and exploited.

Therefore, our proposal made use of a different approach for representing temporal annotations and queries for Semantic Web systems, which is based on the fuzzy-set theory. Our model in particular made use of a quadruple $< T_{fuzzybegin}, T_{begin}, T_{end}, T_{fuzzyend} >$ for the representation of the imprecision of fuzzy temporal intervals with trapezoidal fuzzy sets. Moreover, the definition of methods for temporal reasoning on the modeled fuzzy sets, according to *overlaps*, *overlappedBy* and *closeness* calculations, and the introduction of a combined measure for determining the relevance of an annotation period according to a query period have been introduced.

The experimentation of our approach on data coming from the case study made it possible to define an evaluation setting and criteria for the retrieval methods. This evaluation, which has been made by 8 domain experts and 4 average users, to our knowledge represents the only case existing to date. Moreover, the precision and recall analysis of the evaluation results demonstrated the effectiveness of our approach, and they provide a significant contribution to the general scenario of fuzzy temporal representation and reasoning in archaeology, as well as to its practical implementation and exploitation in real Semantic Web systems.

Since this research work is the result of the activity of a single person, the innovative possibilities that are connected to the adoption of the newest Web technologies in the case study have been necessarily limited to priority aspects, according to the key elements and results synthesized so far. Nevertheless, the idea of applying these tech-

nologies in the context of the archaeological heritage of the city of Milan is completely original, and should show an extremely positive impact on the activities of cataloging and valorization of the rich heritage of the city. In fact, a fragmentary situation exists today, as different Institutions cooperate but are not in the condition to easily integrate the results of their activities easily. The Semantic Web, and in particular the integration of the systems and data of every single institution in a network, preserves the original systems and work protocols and at the same time represents a perspective of sure interest in this context, whose basic building blocks have been set with this PhD work. Moreover, the fruition of integrated repositories through highly interactive Web interfaces may constitute crucial improvement for the valorization and understanding of the ancient past of the city at different levels, from the specialist to the occasional tourist.

In this regard, the development of a prototype system only scratched the surface of the technical aspect involved in the full realization of this scenario. Nevertheless, it contributed to highlight the basic aspects connected to the practical development of a system with the newest Web technologies, which still show significant disadvantages (e.g. in terms of performances) with respect to different but more consolidated solutions, such as those based on relational databases. On the other hand, it has to be acknowledged that our technological evaluation may suffer from a partial and to a certain extent naïve approach, which is justified by the absence of personal previous technological experience in this field.

The optimization of the current system and further improvements of its efficiency represent necessary conditions in order to make users feel encouraged in using it. On the other hand, the re-design of the current prototype interface, integrating new interactive tools and giving access to the totality of the available semantic repository is ongoing according to the a partial revision of the user interaction requirements. This activity has recently started, and therefore it has not been been described in this work; however, significant results concerning the availability of a completely renovated frontend for MANTIC are expected soon. Specific attention is being paid in particular to the possibility of integrating a recommender system capable of suggesting contents that are semantically related to the ones users progressively browse and discover, in the vein of existing experiences such as that of the "Culture Sampo" project. Parallel to the full integration of the retrieval methods on fuzzy temporal information is being evaluated, since it represents a distinctive characteristic of our project, which greatly improves the effective access to relevant data contained in our repository.

Moreover, the design of a section of our portal with a predefined "tour" of the ancient city according to 14 relevant areas is being undertaken. The areas are being modeled with the CIDOC CRM; semantic relationships linking the areas to the elements that are connected to them (e.g. visible structures, archaeological excavations conducted in the area, archaeological artifacts found in the area, etc.) open a new scenario for the experimentation of new modalities in the suggestion of a tour to end

users.

From a more general perspective, the selection and integration of other relevant data sources will constitute a central activity for future development. The familiarity with the CRM that has been achieved during this research will surely ease this process, not only with respect to the conceptual level, but also for what concerns the possible issues that have been analyzed and faced during the work. It is likely to say that the adoption of a documentation approach for the mapping templates that has a more formal structure than the one we adopted in this research (e.g. using one of the existing mapping languages) will become necessary in order to manage the mapping process more precisely. Moreover, methods for transforming actual data contained in the data sources into RDF triples, for example based on XML and XSLT, will be investigated in order to improve our current solution.

"Doing the Semantic Web" (or more in general, exploiting the newest Web technologies) in cultural heritage is still a complex and resource-consuming activity, which cannot be approached easily. An interdisciplinary perspective is at the same time necessary, since domain and technological knowledge is the key combination for desiging and deploying effective solutions and for facing the number of different issues that emerge, as this PhD work demonstrates. However, investment in these directions are today more than necessary in order to improve the effectiveness of cultural heritage activities within the technological frameworks that are changing our everyday life and that are shaping the Information/Knowledge society of the future,

A

Appendix: The Mapping Templates

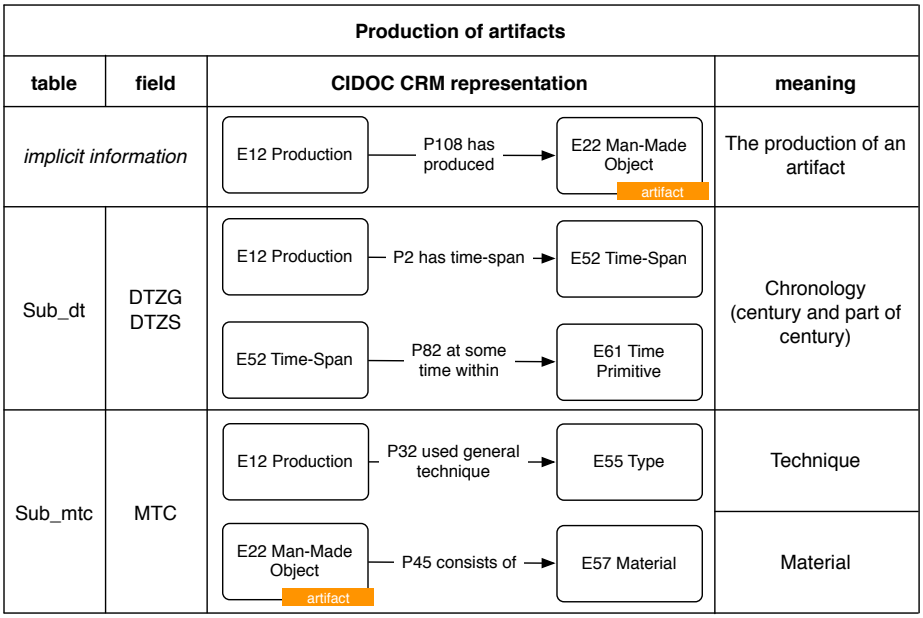


Figure A.1: *SIRBeC to CIDOC CRM mapping template: production of artifacts.*

Discovery of artifacts			
table	**field**	**CIDOC CRM representation**	**meaning**
implicit information		E5 Event (discovery) — P10 falls within → E7 Activity (excavation) E7 Activity (excavation) — P2 has type → E55 Type	The discovery event within the excavation activity
Scheda_ RARL	SPR	E5 Event (discovery) — P3 has note → E62 String	Notes describing the discovery of the artifact
implicit information		E5 Event (discovery) — P7 took place at → E53 Place (site)	The place where the discovery happened
implicit information		E5 Event (discovery) — P12 occurred in the presence of → E22 Man-Made Object	The discovery of the artifact
Scheda_ RARL	LGII LGIT LGIN LGIQ	E53 Place (site) — P87 is identified by → E47 Spatial Coordinates E47 Spatial Coordinates — P2 has type → E55 Type	Spatial coordinates with refernce to the IGM cartography
Scheda_ RARL	LGCC LGCM LGCA LGCO	E53 Place (site) — P87 is identified by → E47 Spatial Coordinates E47 Spatial Coordinates — P2 has type → E55 Type	Spatial coordinates with reference to the cadastral system
implicit information		E7 Activity (excavation) — P14 carried out by → E40 Legal Body	The excavation as an activity by a Legal Body
Scheda_ RARL	DSCF	E40 Legal Body — P131 is identified by → E82 Actor Appellation	Name of the Legal Body responsible for the archaeological excavation
Scheda_ RARL	DSCD	E7 Activity (excavation) — P2 has time-span → E52 Time-Span E52 Time Span — P82 at some time within → E61 Time Primitive	Time span of the excavation activity

Figure A.2: *SIRBeC to CIDOC CRM mapping template: discovery of artifacts.*

Cataloging of artifacts (part I)			
table	**field**	**CIDOC CRM representation**	**meaning**
implicit information		E13 Attribute Assignment (OU cataloging) — P140 assigned attribute to → E22 Man-Made Object (artifact) E13 Attribute Assignment — P2 has type → E55 Type	The OU cataloging activity on the artifact
Scheda_ RARL	UOP	E13 Attribute Assignment (OU cataloging) — P14 carried out by → E74 Group (Operative Unit) E74 Group (Operative Unit) — P1 is identified by → E42 Identifier	Identifier of the Operative Unit
Scheda_ RARL	NSK	E13 Attribute Assignment (OU cataloging) — P141 assigned → E42 Identifier (artifact ID)	Identifier assignements by each Operative Unit
Scheda_ RARL	LDCM	E78 Collection — P1 is identified by → E41 Appellation E78 Collection — P46 is composed of → E22 Man-Made Object (artifact)	Name of the collection the artifact belongs to
Scheda_ RARL	OGTD	E13 Attribute Assignment (OU cataloging) — P141 assigned → E55 Type (artifact type)	Typological attribution
Scheda_ RARL	OGTT	E55 Type (artifact type) — P2 has type → E55 Type (artifact sub-type)	Typological refinement
Scheda_ RARL	CLS	E13 Attribute Assignment — P141 assigned → E55 Type (artifact class)	Class and production
implicit information		E13 Attribute Assignment — P9 consists of → E14 Condition Assessment	Condition assessment as part of the OU cataloging activity
Scheda_ RARL	OGTD	E14 Condition Assessment — P35 has identified → E3 Condition State	Remaining portions of the artifact

Figure A.3: *SIRBeC to CIDOC CRM mapping template: cataloging of artifacts (part I).*

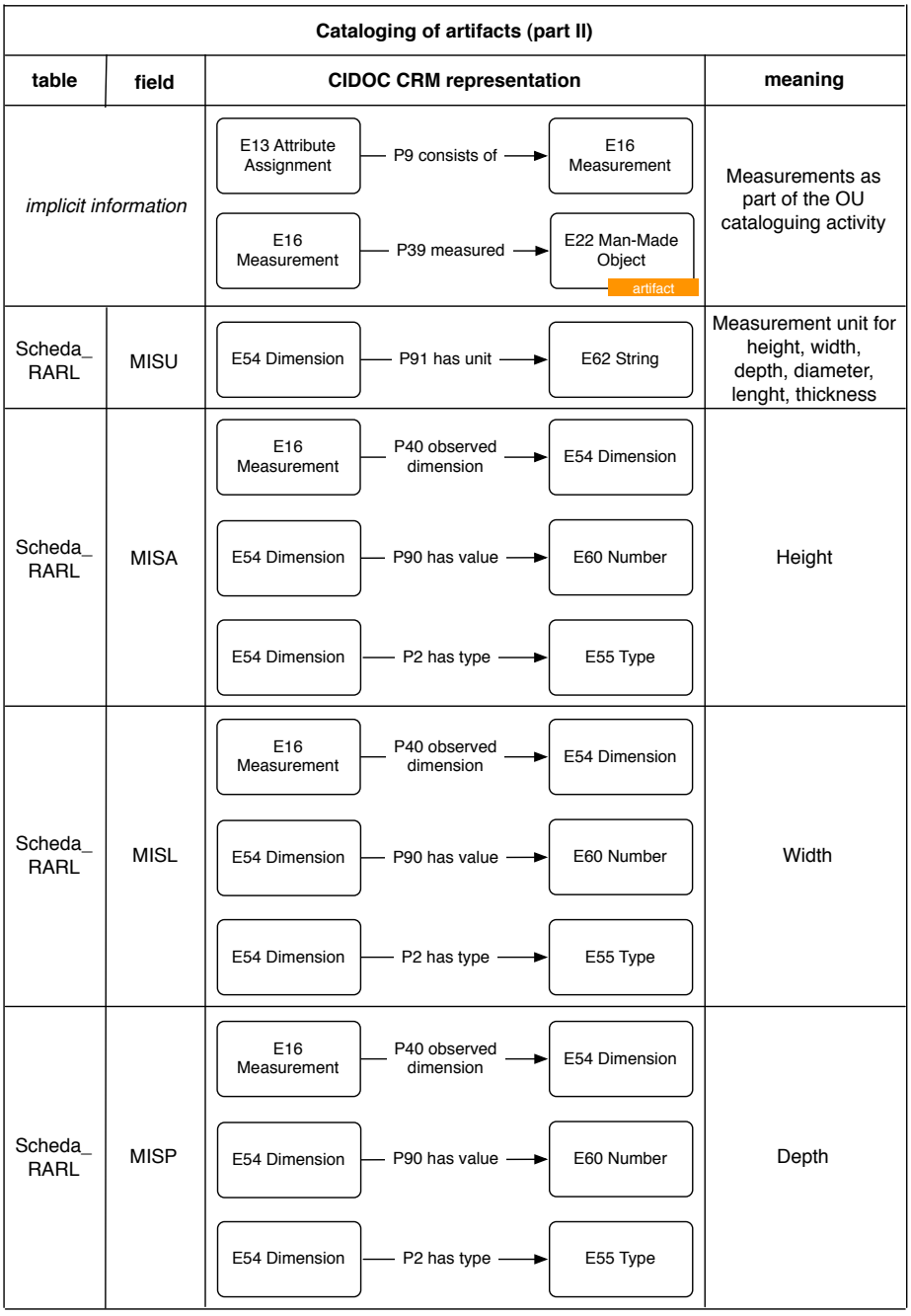


Figure A.4: *SIRBeC to CIDOC CRM mapping template: cataloging of artifacts (part II).*

Cataloging of artifacts (part III)			
table	**field**	**CIDOC CRM representation**	**meaning**
Scheda_ RARL	MISD	E16 Measurement — P40 observed dimension → E54 Dimension E54 Dimension — P90 has value → E60 Number E54 Dimension — P2 has type → E55 Type	Diameter
Scheda_ RARL	MISN	E16 Measurement — P40 observed dimension → E54 Dimension E54 Dimension — P90 has value → E60 Number E54 Dimension — P2 has type → E55 Type	Lenght
Scheda_ RARL	MISS	E16 Measurement — P40 observed dimension → E54 Dimension E54 Dimension — P90 has value → E60 Number E54 Dimension — P2 has type → E55 Type	Thickness
Scheda_ RARL	MISG	E16 Measurement — P40 observed dimension → E54 Dimension E54 Dimension — P90 has value → E60 Number E54 Dimension — P2 has type → E55 Type E54 Dimension — P91 has unit → E62 String	Weight (value and measurement unit)

Figure A.5: *SIRBeC to CIDOC CRM mapping template: cataloging of artifacts (part III).*

Cataloging of artifacts (part IV)

table	field	CIDOC CRM representation	meaning
Scheda_ RARL	DESO	E22 Man-Made Object (artifact) — P3 has note → E62 String	Description of the artifact
Scheda_ RARL	DESS	E22 Man-Made Object (artifact) — P3 has note → E62 String	Description of the subject
implicit information		E22 Man Made Object (artifact) — P46 is composed of → E25 Man-Made Feature (phys. inscription) E25 Man-Made Feature (phys. inscription) — P128 carries → E34 Inscription (inscription)	Inscription on the artifact
sub_isr	ISRL	E34 Inscription (inscription) — P72 has language → E56 Language	Language of the inscription
sub_isr	ISRI	E34 Inscription (inscription) — P3 has note → E62 String	Transcription of the Inscription
sub_imc	IMCX	E31 Document (image) — P70 documents → E13 Attribute Assignment E31 Document (image) — P1 is identified by → E42 Identifier	Identifier of the image file
sub_imr	IMRT	E31 Document (image) — P2 has type → E55 Type	Type of image
sub_cmp	CMPD	E13 Attribute Assignment — P4 has time-span → E52 Time-Span E52 Time-Span — P82 at some time within → E61 Time Primitive	Cataloging date
sub_cmpn	CMPN	E13 Attribute Assignment — P11 had participant → E21 Person (cataloger) E21 Person (cataloger) — P131 is identified by → E82 Actor Appellation	Name of the cataloger
sub_fur	FUR	E13 Attribute Assignment — P11 had participant → E21 Person (officer) E21 Person (officer) — P131 is identified by → E82 Actor Appellation	Name of the Officer in charge

Figure A.6: *SIRBeC to CIDOC CRM mapping template: cataloging of artifacts (part IV).*

Archaeological excavations		
field	**CIDOC CRM representation**	**meaning**
implicit information	E7 Activity (excavation) — P2 has type → E55 Type	The excavation activity
implicit information	E7 Activity (excavation) — P134 continued → E7 Activity (excavation)	Sequences of excavations on the same site
Descrizione degli scavi	E7 Activity — P3 has note → E62 String	General description of the elements of the excavation
Autori dello scavo	E7 Activity (excavation) — P14 carried out by → E39 Actor E39 Actor — P131 is identified by → E82 Actor Appellation	Excavation authors
Data	E7 Activity (excavation) — P4 has time-span → E52 Time-Span E52 Time-Span — P82 at some time within → E61 Time Primitive E7 Activity — P134 continued → E7 Activity	Excavation dates
Indirizzo	E7 Activity (excavation) — P7 took place at → E53 Place (site) E53 Place (site) — P87 is identified by → E44 Place Appellation	Excavation place (name)
Riferimento catastale	E53 Place (site) — P87 is identified by → E47 Spatial Coordinates E47 Spatial Coordinates — P2 has type → E55 Type	Excavation place (cadastral coordinates)

Figure A.7: *IDRA to CIDOC CRM mapping template: the archaeological excavation activity.*

Cataloging of archaeological sites		
field	**CIDOC CRM representation**	**meaning**
implicit information	E13 Attribute Assignment — P140 assigned attribute to → E53 Place (site)	The attribute assignment activity on the site
Identificativo del bene	E13 Attribute Assignment — P141 assigned → E42 Identifier	Site identifier
Tipo di evidenza	E13 Attribute Assignment — P141 assigned → E55 Type	Typology of the site
Secolo	E4 Period — P7 took place at → E53 Place E4 Period — P4 has time-span → E52 Time-Span (site) E52 Time-Span — P82 at some time within → E61 Time Primitive	Chronology of the site
Data compilazione	E13 Attribute Assignment — P4 has time-span → E52 Time-Span E52 Time-Span — P82 at some time within → E61 Time Primitive	Cataloging date
Nome	E13 Attribute Assignment — P11 had participant → E21 Person E21 Person — P131 is identified by → E82 Actor Appellation	Name of the cataloger
Funzionario responsabile	E13 Attribute Assignment — P11 had participant → E21 Person E21 Person — P131 is identified by → E82 Actor Appellation	Name of the officer in charge

Figure A.8: *IDRA to CIDOC CRM mapping template: cataloging of archaeological sites.*

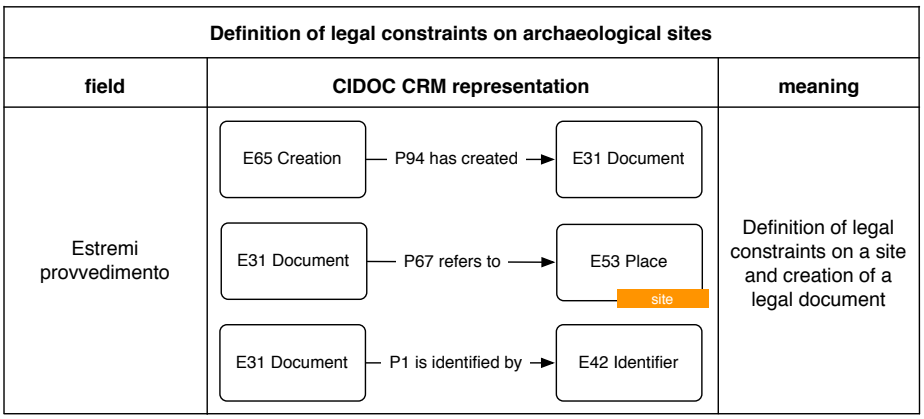


Figure A.9: *IDRA to CIDOC CRM mapping template: the definition of legal constraints on archaeological sites.*

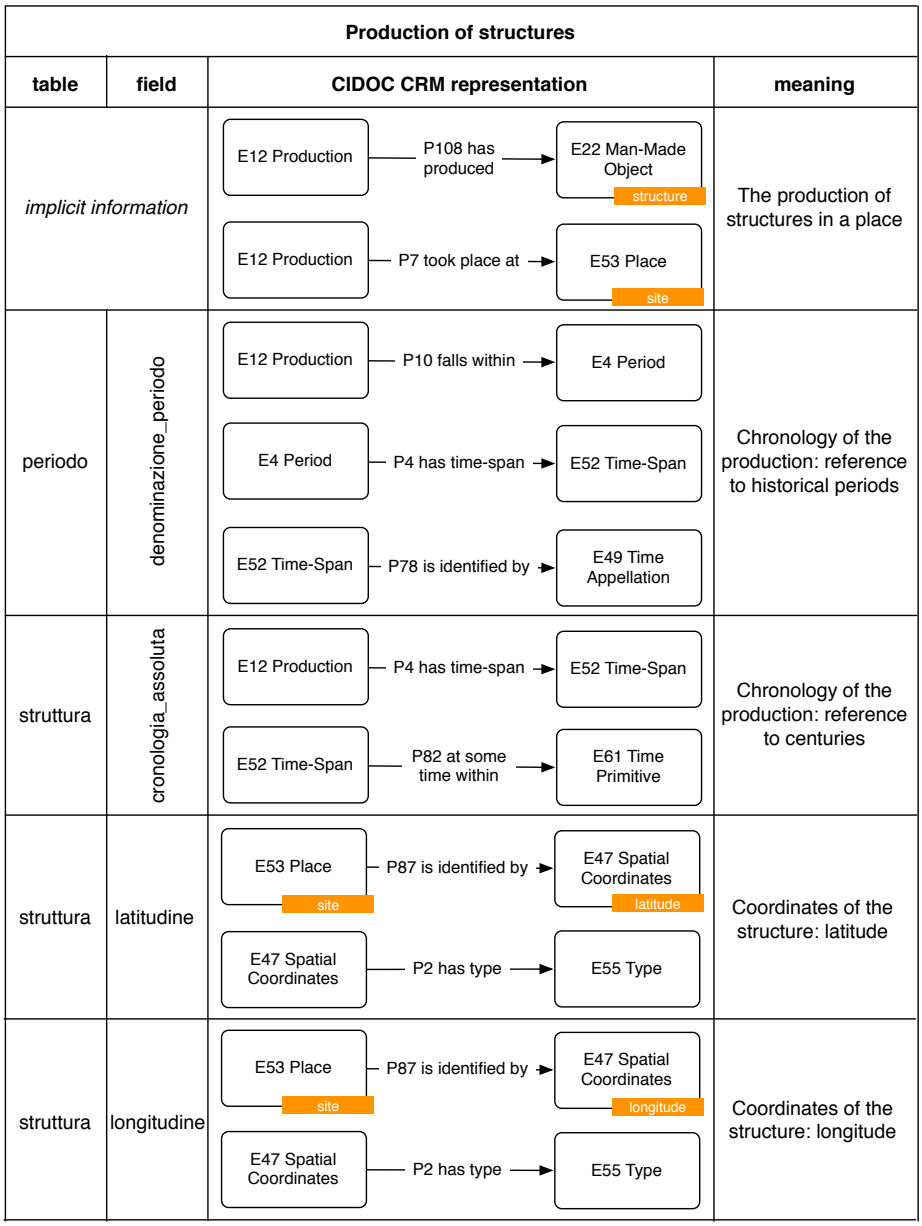


Figure A.10: *MANTIC to CIDOC CRM mapping template: the production of structures.*

Cataloging of archaeological sites			
table	**field**	**CIDOC CRM representation**	**meaning**
implicit information		E13 Attribute Assignment — P140 assigned attribute to → E53 Place (site)	The cataloging activity
tipologia_sito	denominazione_tipologia_sito	E13 Attribute Assignment — P141 assigned → E55 Type	Typology of the site
sito	denominazione_luogo	E53 Place (site) — P87 is identified by → E44 Place Appellation	Appellation of the site
sito	latitudine	E53 Place — P87 is identified by → E47 Spatial Coordinates (latitude) E47 Spatial Coordinates — P2 has type → E55 Type	Coordinates of the site: latitude
sito	longitudine	E53 Place — P87 is identified by → E47 Spatial Coordinates (longitude) E47 Spatial Coordinates — P2 has type → E55 Type	Coordinates of the site: longitude
sito	id_sito	E53 Place (site) — P48 has preferred identifier → E42 Identifier	Id of the site

Figure A.11: *MANTIC to CIDOC CRM mapping template: cataloging of archaeological sites.*

Cataloging of structures			
table	**field**	**CIDOC CRM representation**	**meaning**
implicit information		E13 Attribute Assignment — P140 assigned attribute to → E22 Man Made Object (structure)	The cataloging activity
struttura	denominazione_struttura	E22 Man-Made Object — P1 is identified by → E41 Appellation	Appellation of the structure
struttura	id_struttura	E22 Man-Made Object — P48 has preferred identifier → E42 Identifier	Id of the structure
struttura	cosa_rimane	E13 Attribute Assignment — P9 consists of → E14 Condition Assessment E14 Condition Assessment — P35 has identified → E3 Condition State	Description of the current state of the structures
tipologia_struttura	denominazione_tipologia_struttura	E12 Attribute Assignment — P141 assigned → E55 Type	Typology of the structure

Figure A.12: *MANTIC to CIDOC CRM mapping template: cataloging of structures.*

Bibliography

Accary–Barbier, T. and Calabretto, S. (2008). Building and using temporal knowledge in archaeological documentation. *Journal of Intelligent Information Systems*, 31(2):147–159.

ACLS (2006). Our cultural commonwealth: The report of the ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences. Accessible at: `http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf`.

Addis, M., Hafeez, S., Prideaux, D., Lowe, R., Lewis, P., Martinez, K., and Sinclair, P. (2006). The eCHASE System for cross-border use of European multimedia cultural heritage content in education and publishing. In Ng, K., Badii, A., and Bellini, P., editors, *Proceedings of AXMEDIS 2006, 2nd International Conference on Automated Production of Cross Media Content for Multi-channel Distribution*, pages 27–32, Florence. Firenze University Press.

Addis, M., Lewis, P., and Martinez, K. (2002). ARTISTE image retrieval system puts European galleries in the picture. Cultivate Interactive, 7. Accessible at: `http://www.cultivate-int.org/issue7/artiste`.

Addis, M., Martinez, K., Lewis, P., Stevenson, J., and Giorgini, F. (2005). New Ways to Search, Navigate and Use Multimedia Museum Collections over the Web. In Trant, J. and Bearman, D., editors, *Proceedings of Museums and the Web Conference 2005*, volume 7, Toronto. Archives & Museum Informatics. Online publication. Accessible at: `http://www.archimuse.com/mw2005/papers/addis/addis.html`.

Adida, B., Birbeck, M., McCarron, S., and Pemberton, S. (2008). RDFa in XHTML: Syntax and Processing. A collection of attributes and processing rules for extending XHTML to support RDF. W3C Working Draft, World Wide Web Consortium, Accessible at: `http://www.w3.org/TR/rdfa-syntax`.

Aitken, M. (1990). *Science-Based Dating in Archaeology*. Longman, London.

Allen, J. (1983). Maintaining Knowledge about Temporal Intervals. *Communications of the ACM*, 26(12):832–843.

Anderson, C. (2006). *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion Books.

Ankolekar, A., Krötzch, M., Tran, T., and Vrandecic, D. (2007). The Two Cultures. Mashing up Web 2.0 and the Semantic Web. In Williamson, C., Mary Ellen Zurko, P. P.-S., and Shenoy, P., editors, *Proceedings of the 16th International World Wide Web Conference*, pages 825–833, Banff, Alberta. ACM Press.

Aroyo, L., Hyvönen, E., and van Ossenbruggen, J., editors (2007a). *Proceedings of Workshop 9: Cultural Heritage on the Semantic Web*, Busan, Korea. 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference, Accessible at: `http://www.cs.vu.nl/~laroyo/CH-SW/ISWC-wp9-proceedings.pdf`.

Aroyo, L., Stash, N., Wang, Y., Gorgels, P., and Rutledge, L. (2007b). CHIP Demonstrator: Semantics-Driven Recommendations and Museum Tour Generation. In Abere, K., Choi, K., Fridman Noy, N., Allemang, D., Lee, K., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., and Cudré-Mauroux, P., editors, *Proceedings of the International Semantic Web Conference and the Asian Semantic Web Conference*, volume 4825 of *Lecture Notes in Computer Science*, pages 879–886, Berlin Heidelberg. Springer.

Astudillo, H., Inostroza, P., and López, C. (2008). Contexta/SR: A multi-institutional semantic integration platform. In Trant, J. and Bearman, D., editors, *Proceedings of Museums and the Web Conference*, volume 10, Toronto. Archives & Museum Informatics. Accessible at: `http://www.archimuse.com/mw2008/papers/astudillo/astudillo.html`.

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., Messina, P., Ostriker, J. P., and Wright, M. H. (2003). Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Accessible at: `http://www.nsf.gov/od/oci/reports/atkins.pdf`.

Babeu, A., Bamman, D., Crane, G., Kummer, R., and Weaver, G. (2007). Named Entity Identification and Cyberinfrastructure. In Kovács, L., Fuhr, N., and Meghini, C., editors, *Proceeding of the 11th European Conference on Research and Advanced Technology for Digital Libraries*, volume 4675 of *Lecture Notes in Computer Science*, pages 259–270, Berlin Heidelberg. Springer.

Barceló, J. (2009). El analysis del tiempo en arqueologia: Incertidumbre y estadística. Draft paper.

Barceló, J., Bogdanovic, I., and Piqué, R. (2004). Tele-archaeology. *Archeologia e Calcolatori*, XV:467–481.

Berners-Lee, T. (1998). Semantic Web Road map. Accessible at: `http://www.w3.org/DesignIssues/Semantic.html`.

Berners-Lee, T. (2007). Linked Data. Accessible at: `http://www.w3.org/DesignIssues/LinkedData.html`.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, pages 34–43.

Binding, C., May, K., and Tudhope, D. (2008). Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction Via the CIDOC-CRM. In Christensen-Dalsgaard, B., editor, *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries*, volume 5173 of *Lecture Notes in Computer Science*, pages 280–290, Berlin Heidelberg. Springer.

Bogdanovic, I., Vicente, O., and Barceló, J. (2004). A Theory of Archaeological Knowledge Building by Using Internet: the DIASPORA Project. In Fennema, K. and Kamermans, H., editors, *Proceedings of the 27th Computer Applications and Quantitative Methods in Archaeology Conference*, Leiden. CAA.

Booth, D., Haas, H., and McCabe, F. (2004). Web Services Architecture. W3C Working Group Note, World Wide Web Consortium, Accessible at: `http://www.w3.org/TR/ws-arch/`.

Borgman, C. (1999). What are Digital Libraries? Competing Visions. *Information Processing & Management*, 35:227–243.

Brickley, D. and Guha, R. V. (2004). RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation, World Wide Web Consortium, Accessible at: `http://www.w3.org/TR/rdf-schema`.

Buck, C. and Millard, A., editors (2004). *Tools for Constructing Chronologies: Crossing Disciplinary Boundaries*, volume 177 of *Lecture Notes in Statistics*. Springer, London.

Byrne, K. (2006). Tethering Cultural Data with RDF. In *Proceedings of the 2006 Jena Users Conference*, Accesible at: `jena.hpl.hp.com/juc2006/proceedings/byrne/paper.pdf`.

Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., KKoutrika, G., Meneghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobreva, M., Katifori, V., and Schuldt, H. (2007). *The DELOS Digital Library Reference Model*. DELOS Network of Excellence on Digital Libraries, Accessble at: `http://www.delos.info/index.php?option=com_content&task=view&id=345#reference_model`.

Caporusso, D., Donati, M. T., Masseroli, S., and Tibiletti, T. (2007). *Immagini di Mediolanum. Archeologia e storia di Milano dal V secolo a.C. al V secolo d.C.* Civiche Raccolte Archeologiche e Numismatiche di Milano, Milano.

Christensen, E., Curbera, F., Meredith, G., and Weerawarana, S. (2001). Web Services Description Language (WSDL) 1.1. W3C Note, World Wide Web Consortium, Accessible at: `http://www.w3.org/TR/wsdl20-primer/`.

Cohen, J. (1968). Weighted kappa: Nominal Scale Agreement Provision for Scaled Disagreement on Partial Credit. *Psychological Bulletin*, 70(4):213–220.

Cripps, P., Greenhalg, A., Fellows, D., May, K., and Robinson, D. (2004). Ontological Modelling of the work of the Centre for Archaeology. Technical report, English Heritage - Centre For Archaeology, Accessible at: `http://cidoc.ics.forth.gr/docs/Ontological_Modelling_Project_Report%_%20Sep2004.pdf`.

Crofts, N., Doerr, M., Gill, T., Stead, S., and Stiff, M. (2009). Definition of the CIDOC Conceptual Reference Model Version 5.0.1. Technical report, ICOM/CIDOC CRM Special Interest Group, Accessible at: `http://cidoc.ics.forth.gr/docs/cidoc_crm_version_5.0.1_Nov09.pdf`.

Degiarde, E. (2007). SIRBeC-Sistema Informativo Regionale dei Beni Culturali: metodologie e strumenti per la qualità del processo di catalogazione e la valorizzazione delle informazioni prodotte. Final Report IReR Code: 2006B047, Istituto Regionale di Ricerca della Lombardia (IReR).

Della Valle, E., Celino, I., and Cerizza, D. (2008). *Semantic Web: modellare e condividere per innovare*. Pearson Addison Wesley.

DigiCULT Project (2003). Towards a Semantic Web for Heritage Resources. DigiCULT Thematic Issue 3. Accessible at: `http://www.digicult.info/downloads/ti3_high.pdf`.

Doerr, M. and Iorizzo, D. (2008). The Dream of a Global Knowledge Network — A New Approach. *ACM Journal on Computing and Cultural Heritage*, 1(1):1–23.

Doerr, M. and Kritsotaki, A. (2006). Documenting Events in Metadata. In Ioannides, M., Arnold, D., Niccolucci, F., and Mania, K., editors, *Proceedings of the 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST)*, Accessible at: `http://cidoc.ics.forth.gr/docs/fin-paper.pdf`.

Doerr, M., Plexousakis, D., Kopaka, K., and Bekiari, C. (2004a). Supporting Chronological Reasoning in Archaeology. In *Proceedings of the 32nd Computer Applications and Quantitative Methods in Archaeology Conference*, Accessible at: `http://cidoc.ics.forth.gr/docs/caa2004_supporting_chronological_reasoning.pdf`. In preparation.

Doerr, M., Schaller, K., and Theodoridou, M. (2004b). Integration of complementary archaeological sources. In Niccolucci, F., editor, *Proceedings of the 32nd Computer Applications and Quantitative Methods in Archaeology Conference*. In preparation.

Dubois, D. and Prade, H. (1988). *Possibility Theory: an Approach to Computerized Processing of Uncertainty*. Plenum Press, New York.

Eckkrammer, F., Eckkrammer, T., and Feldbacher, R. (2008). CIDOC CRM in data management and data sharing: Data sharing between different databases. Paper

presented at the 36th Computer Applications and Quantitative Methods in Archaeology.

Felicetti, A. and D'Andrea, A. (2008a). COINS domain ontology. Accessible at: http://www.coins-project.eu/downloads/reports/Coins-044450-D4.pdf.

Felicetti, A. and D'Andrea, A. (2008b). COINS mapping template. Accessible at: http://www.coins-project.eu/downloads/reports/Coins-044450-D6.pdf.

Felicetti, A. and D'Andrea, A. (2008c). COINS Multilingual Thesaurus. Accessible at: http://www.coins-project.eu/downloads/reports/Coins-044450-D5.pdf.

Fernández González, J., Márquez, A. P., and Cerrillo Cuenca, E. (2007). Bases for the creation of ontology in the context of Archaeology. In Figueiredo, A. and Leite Velho, G., editors, *Proceedings of the 33th Computer Applications and Quantitative Methods in Archaeology*, Tomar. CAA Portugal.

Fielding, R. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine.

Forschungsgesellschaft Wiener Stadtarchäologie (2005). Vbi Erat Lvpa. An Internet project communicating antiquity. Accessible at: http://www.ubi-erat-lupa.org/site/PDF_files/Lupa2005.pdf.

Foster, I. (2005). Service-oriented science. *Science*, 308(5723):814–817.

Fox, E., Akscyn, R., Furuta, R., and Leggett, J. (1995). Digital Libraries. *Communications of the ACM*, 38(4):22–28.

Gagliardi, I. (2003). Integration of Different Consultation Models in Cultural Heritage: The SIRBeC Experiment. In *Proceedings of the ICHIM*. Archives & Museum Informatics. Accessible at: http://www.archimuse.com/.

Gardin, J. (1990). *Mathematics and Information Science in Archaeology. A Flexible Framework*, volume 3 of *Studies in Modern Archaeology*, chapter The Structure of Archaeological Theories, pages 7–25. Holos Verlag, Bonn.

Gardin, J. and Roux, V. (2004). The Arkeotek Project: a European Network of Knowledge Bases in the Archaeology of Techniques. *Archeologia e Calcolatori*, 15:25–40.

Garrett, J. J. (2005). Ajax: A New Approach to Web Applications. Accessible at: http://www.adaptivepath.com/ideas/essays/archives/000385.php.

Giustini, D. (2007). Web 3.0 and Medicine. *British Medical Journal*, 335(7633):1273–1274.

Goble, C., Corcho, O., Alper, P., and De Roure, D. (2006). e-Science and the Semantic Web: a Symbiotic Relationship. In Todorovski, L., Lavrač, N., and Jantke, K., editors, *Proceeding of the 9th Discovery Science International Conference*, volume 4265 of *Lecture Notes in Artificial Intelligence*, pages 1–12, Berlin Heidelberg. Springer.

Goerz, G., Schiemann, B., and Oischinger, M. (2008). An Implementation of the CIDOC Conceptual Reference Model (4.2.4) in OWL-DL. In *CIDOC Annual Conference*, Accessible at: `http://www8.informatik.uni-erlangen.de/IMMD8/Services/cidoc-crm/docu/crm_owl_cidoc2008.pdf`.

Gruber, T. (2008). Collective Knowledge Systems: Where the Social Web meets the Semantic Web. *Journal of Web Semantics*, 6(1):4–13.

Gutierrez, C., Hurtado, C., and Vaisman, A. (2007). Introducing Time into RDF. *IEEE Transactions on Knowledge and Data Engineering*, 19(2):207–218.

Harris, E. C. (1979). *Principles of Archaeological Stratigraphy*. Academic Press, London & New York.

Hermon, S. (2007). COINS - an EU Funded Project to "Combat On-Line Illegal Numismatic Sales. In Arnold, D., Niccolucci, F., and Chalmers, A., editors, *Proceedings of the 8th International Symposium on Virtual Reality, Archaeology and Cultural Heritage*.

Hey, T. and Trefethen, A. E. (2005). Cyberinfrastructure for e-Science. *Science*, 308(5723):817–821.

Hildebrand, M., van Ossenbruggen, J., and Hardman, L. (2006). /facet: A Browser for Heterogeneous Semantic Web Repositories. In Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., and Aroyo, L., editors, *Proceedings of the 5th International Semantic Web Conference*, volume 4273, pages 275–285, Berlin Heidelberg. Springer.

Holmen, J. and Ore, C. (2010). Deducing Event Chronology in a Cultural Heritage Documentation System. In *Proceedings of the 38th Computer Applications and Quantitative Methods in Archaeology Conference*, In preparation.

Holmen, J., Ore, C., and Eide, O. (2004). Documenting Two Histories at Once: Digging into Archaeology. In Ausserer, K., Börner, W., Goriany, M., and Karlhuber-Vöckl, L., editors, *Proceedings of the 30th Computer Applications and Quantitative Methods in Archaeology*, volume 1227 of *BAR International Series*, Oxford. Archaeopress.

Hyvönen, E. (2009). *Semantic Portals for Cultural Heritage*, volume Handbook on Ontologies. Springer, Berlin Heidelberg, 2nd edition.

Hyvönen, E., Mäkelä, E., Kauppinen, T., Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Takala, J., Puputti, K., Kuittinen, H., Viljanen, K., Tuominen, J., Palonen, T.,

Frosterus, M., Sinkkilä, R., Paakkarinen, P., Laitio, J., and Nyberg, K. (2009a). CultureSampo: A National Publication System of Cultural Heritage on the Semantic Web 2.0. In Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., and Simperl, E., editors, *Proceedings of the 6th European Semantic Web Conference*, volume 5554/2009 of *Lecture Notes in Computer Science*, pages 851–856, Berlin Heidelberg. Springer.

Hyvönen, E., Mäkelä, E., Kauppinen, T., Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Viljanen, K., Tuominen, J., Palonen, T., Frosterus, M., Sinkkilä, R., Paakkarinen, P., Laitio, J., and Nyberg, K. (2009b). CultureSampo - Finnish Cultural Heritage Collections on the Semantic Web 2.0. In *Proceedings of the 1st International Symposium on Digital humanities for Japanese Arts and Cultures*, In preparation.

Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., and Kettula, S. (2005). MuseumFinland - Finnish Museums on the Semantic Web. *Journal of Web Semantics*, 3(2):224–241.

Hyvönen, E., Saarela, S., and Viljanen, K. (2004a). Application of Ontology Techniques to View-Based Semantic Search and Browsing. In Bussler, C., Davies, J., Fensel, D., and Studer, R., editors, *Proceedings of teh 1st European Semantic Web Symposium*, volume 3053 of *Lecture Notes in Computer Science*, pages 92–106, Berlin Heidelberg. Springer.

Hyvönen, E., Saarela, S., Viljanen, K., Mäkelä, E., Valo, A., Salminen, M., Kettula, S., and Junnila, M. (2004b). A Cultural Community Portal for Publishing Museum Collections on the Semantic Web. In Ding, Y., Fensel, D., Lara, R., Lausen, H., Stollberg, M., and Han, S., editors, *Proceeding of the ECAI Workshop on Application of Semantic Web Technologies to Web Communities*, volume 107 of *CEUR Workshop Proceedings*. CEUR-WS.org. Aceesible at: `http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-107/paper8.pdf`.

Hyvönen, E., Viljanen, K., Tuominen, J., and Seppälä, K. (2008). Building a National Semantic Web Ontology and Ontology Service Infrastructure — the FinnONTO Approach. In Bechhofer, S., Hauswirth, M., Hoffmann, J., and Koubarakis, M., editors, *Proceedings of the 5th European Semantic Web Conference*, volume 5021/2008 of *Lecture Notes in Computer Science*, pages 95–109, Berlin Heidelberg. Springer.

Isaac, A. and Summers, E. (2008). SKOS Simple Knowledge Organization System Primer. W3C Working Draft, World Wide Web Consortium, Accessible at: `http://www.w3.org/TR/skos-primer`.

Isaksen, L. (2008). The TRANSLATION Framework for Archaeological Excavation Data. Accessible at: `http://eprints.soton.ac.uk/63320/`.

Isaksen, L., Martinez, K., Gibbins, N., Earl, G., and Keay, S. (2009). Linking Archaeological Data. In *Proceedings of the 37th Computer Applications and Quantitative Methods in Archaeology*. In preparation.

Johnson, I. (2008). Mapping The Fourth Dimension: a Ten Year Retrospective. *Archeologia e Calcolatori*, 19:31–43.

Kauppinen, T., Mantegari, G., Paakkarinen, P., Kuittinen, H., Hyvönen, E., and Bandini, S. (Submitted for review in October, 2009). Determining Relevance of Imprecise Temporal Intervals for Cultural Heritage Information Retrieval. *International Journal of Human Computer Studies*.

Kauppinen, T., Väätäinen, J., and Hyvönen, E. (2008). Creating and Using Geospatial Ontology Time Series in a Semantic Cultural Heritage Portal. In Bechofer, S., Hauswirth, M., Hoffmann, J., and Koubarakis, M., editors, *Proceedings of the 5th European Semantic Web Conference*, volume 5021 of *LNCS*, pages 110–123, Berlin Heidelberg. Springer.

Kekäläinen, J. and Järvelin, K. (2002). Using Graded Relevance Assessment in IR Evaluation. *Journal of the American Society for Information Service and Technology*, 53(13):1120–1129.

Kinthigh, K. (2006). The Promise and Challenge of Archaeological Data Integration. *American Antiquity*, 71(3):567–578.

Kondylakis, H., Doerr, M., and Plexousakis, D. (2006). Mapping Language for Information Integration. Technical report. Accessible at `http://www.ics.forth.gr/isl/publications/paperlink/Mapping_TR385_December06.pdf`, ICS-FORTH.

Kruk, S. (2009). Semantic Digital Libraries. Tutorial at the 2009 International Conference for Digital Libraries and the Semantic Web. Accessible at: `http://semdl.corrib.org/Tutorial/`.

Kummer, R. (2007). Towards Semantic Interoperability of Cultural Information Systems — Making Ontologies Work. Ma thesis, University of Köln.

Lagoze, C., Krafft, D., Payette, S., and Jesuroga, S. (2005). What is a Digital Library Anymore, Anyway? *D-Lib magazine*, 11(11).

Lombardia Informatica s.p.a (1999). Schema della struttura dati delle schede di catalogo. Technical report, Lombardia Informatica.

Lucas, G. (2005). *The Archaeology of Time*. Routledge, London.

Mäkelä, E., Viljanen, K., Alm, O., Tuominen, J., Valkeapää, O., Kauppinen, T., Kurki, J., Sinkkilä, R., Känsälä, T., Lindroos, R., Suominen, O., Ruotsalo, T., and Hyvönen, E. (2007). Enabling the Semantic Web with Ready-to-Use Web Widgets. In Nixon, L., Cuel, R., and Bergamini, C., editors, *Proceedings of the First Industrial Results of Semantic Technologies Workshop*, volume 293 of *CEUR Workshop Proceedings*, pages 56–69. CEUR-WS.org. Online publication. Accessible at: `http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-293/paper5.pdf`.

Manola, F. and Miller, E. (2004). RDF Primer. W3C Recommendation, World Wide Web Consortium, Accessible at: `http://www.w3.org/TR/rdf-primer/`.

Mantegari, G., Cattani, M., de Marinis, R., and Vizzari, G. (2006). Towards a Web-based environment for Italian Prehistory and Protohistory. In Clark, J. and Hegemeister, E., editors, *Proceedings of the 34th Computer Applications and Quantitative Methods in Archaeology Conference*, pages 374–381, Budapest. Archaeolingua.

Markoff, J. (2006). Entrepreneurs See a Web Guided by Common Sense. Accessible at: `http://www.nytimes.com/2006/11/12/business/12web.html`.

May, K. (2006a). Integrating Cultural and Scientific Heritage: Archaeological Ontological Modeling for the Field and the Lab. In *Proceedings of the CIDOC CRM Special Interest Group Workshop*, Heraklion.

May, K. (2006b). Report on English Heritage Archaeological Application of CRM. In *Proceedings of the CIDOC CRM Special Interest Group Workshop*, Edinburgh.

May, K. (2009). STAR project and SKOS. ATHENA Presentation. Accessible at: `http://www.athenaeurope.org/getFile.php?id=288`.

McAuley, J. and Carswell, J. (2008). Knowledge Management for Disparate Etruscan Cultural Heritage. In Berntzen, L. and Åsa Smedberg, editors, *Proceedings of the 2nd International Conference on the Digital Society*, pages 70–74. IEEE Computer Society.

Merrill, D. (2006). Mashups: The new breed of Web app. Accessible at: `http://www.ibm.com/developerworks/xml/library/x-mashups.html`.

Missikoff, O. (2004). Ontologies as a Reference Framework for the Management of Knowledge in the Archaeological Domain. In Ausserer, K., Börner, W., Goriany, M., and Karlhuber-Vöckl, L., editors, *Proceedings of 30th Computer Applications and Quantitative Methods in Archaeology Conference*, volume 1227 of *BAR International Series*, Oxford. Archaeopress.

Mostern, R. and Johnson, I. (2008). From Named Place to Naming Event: Creating Gazetters for History. *International Journal of Geographical Information Science*, 22(10):1091–1108.

Nagypál, G. and Motik, B. (2003). A Fuzzy Model for Representing Uncertain, Subjective, and Vague Temporal Knowledge in Ontologies. In Meersman, R., Tari, Z., and Smith, D., editors, *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM Confederated International Conferences CoopIS, DOA, and ODBASE 2003*, volume 2008/2003 of *Lecture Notes in Computer Science*, pages 906–923, Berlin Heidelberg. Springer.

Nussbaumer, P. and Bernhard, H. (2007). Putting the CIDOC CRM into Practice – Experiences and Challenges. Technical report, University of Vienna, Accessible at: `http://www.cs.univie.ac.at/upload///550/papers/putting_the_cidoc_crm_into_practice-covered.pdf`.

Nussbaumer, P. and Haslhofer, B. (2007). CIDOC CRM in Action – Experiences and Challenges. In Kovàcs, L., Fuhr, N., and Manghini, C., editors, *Proceedings of the 11th European Conference on Research and Advanced Technology for Digital Libraries*, volume 4675/2007 of *Lecture Notes in Computer Science*, pages 532–533, Berlin Heidelberg. Springer.

OECD (2007). Participative Web and User-Created Content: Web 2.0, Wikis and Social Networking. Technical report, Organization for Economic Co-Operation and Development, Accessible at: `http://213.253.134.43/oecd/pdfs/browseit/9307031E.pdf`.

Oischinger, M., Schiemann, B., and Goerz, G. (2008). Short Documentation of the CIDOC CRM (4.2.4) Implementation in OWL-DL. Accessible at: `http://www8.informatik.uni-erlangen.de/IMMD8/Services/cidoc-crm/docu/documentation_crm_owl-dl_4.2.4.pdf`.

O'Reilly, T. (2005). What is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software. Accessible at: `http://www.oreillynet.com/pub/a/oreilly/tim/`.

Parry, R., Poole, N., and Pratty, J. (2008). Semantic Dissonance: Do We Need (and Do We Understand) the Semantic Web? In *Proceedings of Museums and the Web Conference 2008*, Toronto. Archives & Museum Informatics. Accessible at: `http://www.archimuse.com/mw2008/papers/parry/parry.html`.

Proud'hommeaux, E. and Seaborne, A. (2008). SPARQL Query Language for RDF. W3C Recommendation, World Wide Web Consortium, Accessible at: `http://www.w3.org/TR/rdf-sparql-query/`.

Regione Lombardia (2007). Accordo di programma per la realizzazione del Polo per la Valorizzazione di Beni Culturali in Lombardia. Accessible at: `http://www.lombardiacultura.it/uploads/accordo_di_programma.pdf`.

Regione Lombardia (2008). Portale unico regionale dei beni culturali. Technical report, Regione Lombardia, Online document. Accessible at `www.lombardiabeniculturali.it/docs/PURBeC-progetto-2008.pdf`.

Richards, J. (2006). Archaeology, e-publication and the Semantic Web. *Antiquity*, 80(310):970–979.

Ross, S. (2003). Position Paper. in Towards a Semantic Web for Digital Resources. DigiCULT Thematic Issue 3, pp. 7–11. Accessible at: `http://www.digicult.info/downloads/ti3_high.pdf`.

Schreiber, G., Amin, A., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Hollink, L., Huang, Z., van Kersen, J., de Niet, M., Omelayenko, B., van Ossenbruggen, J., Siebes, R., Taekema, J., Wielemaker, J., and Wielinga, B. (2006). MultimediaN e-Culture demonstrator. In Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., and Aroyo, L., editors, *Proceedings of the 5th International Semantic Web Conference*, volume 4273 of *Lecture Notes in Computer Science*, pages 951–958, Berlin Heidelberg. Springer.

Shadbolt, N., Hall, W., and Berners-Lee, T. (2006). The Semantic Web Revisited. *IEEE Intelligent Systems*, 21(3):96–101.

Signore, O., Missikoff, O., and Moscati, P. (2005). La gestione della conoscenza in Archeologia: modelli, linguaggi e strumenti di modellazione concettuale dall'XML al Semantic Web. *Archeologia e Calcolatori*, 16:291–319.

Smith, M., Welty, C., and McGuinness, D. (2004). OWL Web Ontology Language Guide. W3C Recommendation, World Wide Web Consortium, Accessible at: `http://www.w3.org/TR/owl-guide`.

Snow, D., Gahegan, M., Gilles, C., Hirth, K., Milner, G., Mitra, P., and Wang, J. (2006). Cybertools and Archaeology. *Science*, 311:958–959.

Soergel, D. (2009). Digital Libraries and the Semantic Web. Keynote presentation at the 2009 International Conference for Digital Libraries and the Semantic Web. Accessible at: `http://icsd-conference.org/sites/default/files/SoergelICSD2009Keynote.ppt`.

Studer, R., Benjamins, V., and Fensel, D. (1998). Knowledge Engineering: Principles and Methods. *EEE Transactions on Data and Knowledge Engineering*, 25(1 and 2):161–197.

van Ossenbruggen, J., Amin, A., Hardman, L., Hildebrand, M., van Assem, M., Omelayenko, B., Schreiber, G., Tordai, A., de Boer, V., Wielinga, B., Wielemaker, J., de Niet, M., Taekema, J., van Orsouw, M., and Teesing, A. (2007). Searching and Annotating Virtual Heritage Collections with Semantic-Web Techniques. In Trant, J.

and Bearman, D., editors, *Proceedings of Museums and the Web Conference*, Toronto. Archives & Museum Informatics. Accessible at: `http://www.archimuse.com/mw2007/papers/ossenbruggen/ossenbruggen.html`.

Vila, L. (1994). A Survey on Temporal Reasoning in Artificial Intelligence. *AI Communications*, 7(1):4–28.

Visser, U. (2004). *Intelligent Information Integration for the Semantic Web*. Springer, New York.

Yee, K., Swearingen, K., Li, K., and Hearst, M. (2003). Faceted Metadata for Image Search and Browsing. In *Proceedings of the SIGCHI Conference on human Factors in Computing Systems*, pages 401–408, New York. ACM Press.

Yu, J., Benatallah, B., Casati, F., and Daniel, F. (2008). Understanding Mashup Development. *IEEE Internet Computing*, 12(5):44–52.

Zadeh, L. (1965). Fuzzy Sets. *Information and Control*, 8(3):338–353.

Zhang, C.-X., Cao, C.-G., Gu, F., and Si, J.-X. (2004). Domain-specific formal ontology of archaeology and its application in knowledge acquisition and analysis. *Journal of Computer Science and Technology*, 19(3):290–301.

Zimmermann, H. (1996). *Fuzzy Set Theory and its Applications*. Kluwer, 3rd edition.