MORGAN & CLAYPOOL PUBLISHERS

Publishing and Using Cultural Heritage Linked Data on the Semantic Web

Eero Hyvönen

Synthesis Lectures on The Semantic Web: Theory and Technology
James Hendler and Ying Ding, *Series Editors*

Publishing and Using Cultural Heritage Linked Data on the Semantic Web

Synthesis Lectures on Semantic Web: Theory and Technology

Editors
James Hendler, ***Rensselaer Polytechnic Institute***
Ying Ding, ***Indiana University***

Synthesis Lectures on the Semantic Web: Theory and Application is edited by James Hendler of Rensselaer Polytechnic Institute. Whether you call it the Semantic Web, Linked Data, or Web 3.0, a new generation of Web technologies is offering major advances in the evolution of the World Wide Web. As the first generation of this technology transitions out of the laboratory, new research is exploring how the growing Web of Data will change our world. While topics such as ontology-building and logics remain vital, new areas such as the use of semantics in Web search, the linking and use of open data on the Web, and future applications that will be supported by these technologies are becoming important research areas in their own right. Whether they be scientists, engineers or practitioners, Web users increasingly need to understand not just the new technologies of the Semantic Web, but to understand the principles by which those technologies work, and the best practices for assembling systems that integrate the different languages, resources, and functionalities that will be important in keeping the Web the rapidly expanding, and constantly changing, information space that has changed our lives.
Topics to be included:

- Semantic Web Principles from linked-data to ontology design
- Key Semantic Web technologies and algorithms
- Semantic Search and language technologies
- The Emerging "Web of Data" and its use in industry, government and university applications
- Trust, Social networking and collaboration technologies for the Semantic Web
- The economics of Semantic Web application adoption and use
- Publishing and Science on the Semantic Web
- Semantic Web in health care and life sciences

Publishing and Using Cultural Heritage Linked Data on the Semantic Web
Eero Hyvönen
2012

VIVO: A Semantic Approach to Scholarly Networking and Discovery
Katy Börner, Mike Conlon, Jon Corson-Rikert, and Ying Ding
2012

Linked Data: Evolving the Web into a Global Data Space
Tom Heath and Christian Bizer
2011

Publishing and Using Cultural Heritage Linked Data on the Semantic Web

Eero Hyvönen

www.morganclaypool.com

ISBN: 9781608459971 paperback
ISBN: 9781608459988 ebook

DOI 10.2200/S00452ED1V01Y201210WBE003

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON SEMANTIC WEB: THEORY AND TECHNOLOGY

Lecture #3
Series Editors: James Hendler, *Rensselaer Polytechnic Institute*
Ying Ding, *Indiana University*
Synthesis Lectures on Semantic Web: Theory and Technology
ISSN pending.

Publishing and Using Cultural Heritage Linked Data on the Semantic Web

Eero Hyvönen
Aalto University

SYNTHESIS LECTURES ON SEMANTIC WEB: THEORY AND TECHNOLOGY
#3

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

Cultural Heritage (CH) data is syntactically and semantically heterogeneous, multilingual, semantically rich, and highly interlinked. It is produced in a distributed, open fashion by museums, libraries, archives, and media organizations, as well as individual persons. Managing publication of such richness and variety of content on the Web, and at the same time supporting distributed, interoperable content creation processes, poses challenges where traditional publication approaches need to be re-thought. Application of the principles and technologies of Linked Data and the Semantic Web is a new, promising approach to address these problems. This development is leading to the creation of large national and international CH portals, such as Europeana, to large open data repositories, such as the Linked Open Data Cloud, and massive publications of linked library data in the U.S., Europe, and Asia. Cultural Heritage has become one of the most successful application domains of Linked Data and Semantic Web technologies.

This book gives an overview on why, when, and how Linked (Open) Data and Semantic Web technologies can be employed in practice in publishing CH collections and other content on the Web. The text first motivates and presents a general semantic portal model and publishing framework as a solution approach to distributed semantic content creation, based on an ontology infrastructure. On the Semantic Web, such an infrastructure includes shared metadata models, ontologies, and logical reasoning, and is supported by shared ontology and other Web services alleviating the use of the new technology and linked data in legacy cataloging systems. The goal of all this is to provide layman users and researchers with new, more intelligent and usable Web applications that can be utilized by other Web applications, too, via well-defined Application Programming Interfaces (API). At the same time, it is possible to provide publishing organizations with more cost-efficient solutions for content creation and publication.

This book is targeted to computer scientists, museum curators, librarians, archivists, and other CH professionals interested in Linked Data and CH applications on the Semantic Web. The text is focused on practice and applications, making it suitable to students, researchers, and practitioners developing Web services and applications of CH, as well as to CH managers willing to understand the technical issues and challenges involved in linked data publication.

KEYWORDS

Semantic Web, linked data, cultural heritage, portal, metadata, ontologies, logic rules, information retrieval, semantic search, recommender system

Contents

Preface . . . . . **xi**

Acknowledgments . . . . . **xiii**

1 Cultural Heritage on the Semantic Web **1**

1.1 Characterizing Cultural Heritage 1
1.2 Information Portals for Cultural Heritage 2
1.3 Challenges of Cultural Heritage Data 4
1.4 Promises of the Semantic Web 5
1.5 Outline of the Book 9
1.6 Bibliographical and Historical Notes 9

2 Portal Model for Collaborative CH Publishing **13**

2.1 Global Access for Local Linked Content 13
2.1.1 Federated Search 13
2.1.2 Data Warehousing 14
2.2 Collaborative Publishing of Linked Data 14
2.3 Benefits for End-users 17
2.4 Benefits for Publishers 17
2.5 New Challenges 18
2.6 Components of a Semantic Portal System 18
2.7 Bibliographical and Historical Notes 19

3 Requirements for Publishing Linked Data **21**

3.1 Five-star Model for Linked Data 21
3.1.1 Publishing Structured Data 21
3.1.2 Open Licensing 24
3.1.3 Open Formats 25
3.1.4 Requirements for Identifiers 25
3.1.5 Linking Data Internally and Externally 29
3.2 Requirements for Interfaces and APIs 31

3.2.1 Linked Data Browsing 31
3.2.2 SPARQL Endpoint 31
3.2.3 Download Facility 32
3.2.4 Human Interfaces 32
3.3 Bibliographical and Historical Notes 33

4 Metadata Schemas 35

4.1 Metadata Types 35
4.2 Web Schemas 37
4.2.1 Dublin Core 37
4.2.2 VRA Core Categories 38
4.3 Cataloging Schemas 39
4.3.1 Categories for the Description of Works of Art (CDWA) 40
4.3.2 SPECTRUM 40
4.3.3 Metadata Formats in Libraries 41
4.3.4 Metadata Formats in Archives 41
4.4 Conceptual Harmonization Schemas 42
4.4.1 Approaches to Semantic Interoperability 42
4.4.2 Europeana Semantic Elements (ESE) 43
4.4.3 Europeana Data Model (EDM) 43
4.4.4 CIDOC Conceptual Reference Model (CRM) 44
4.4.5 Functional Requirements for Bibliographic Records (FRBR) 46
4.4.6 Functional Requirements for Authority Data (FRAD) 47
4.4.7 Functional Requirements for Subject Authority Data (FRSAD) 48
4.4.8 FRBRoo 49
4.5 Harvesting Schemas: LIDO 49
4.6 Harvesting and Searching Protocols 50
4.6.1 Searching with Z39.50, SRU/SRW, and OpenSearch 51
4.6.2 Harvesting with OAI-PMH 52
4.6.3 SPARQL Endpoint for Linked Data 52
4.7 Discussion: Object, Event, and Process Models 53
4.8 Bibliographical and Historical Notes 55

5 Domain Vocabularies and Ontologies 57

5.1 Approaches to Ontologies 57
5.1.1 Philosophy 57
5.1.2 Lexicography and Linguistics 58

5.1.3 Terminology 60
5.1.4 Information and Library Science 60
5.1.5 Computer Science 62
5.2 Semantic Web Ontology Languages 63
5.2.1 RDF Schema 63
5.2.2 Simple Knowledge Organization System SKOS 63
5.2.3 Web Ontology Language OWL 64
5.3 Ontology Types 65
5.3.1 Classifications, Thesauri, and Ontologies 65
5.3.2 Ontology Types by Major Domains 67
5.4 Actor Ontologies 68
5.5 Place Ontologies 71
5.6 Time Ontologies 73
5.6.1 Linear Time 74
5.6.2 Cyclic Time 75
5.7 Event Ontologies 75
5.8 Nomenclatures 76
5.9 Bibliographical and Historical Notes 76

6 Logic Rules for Cultural Heritage 79

6.1 The Idea of Logic 79
6.2 Logical Interpretation of RDF(S) and OWL 80
6.3 Rules for Reasoning 81
6.3.1 Horn Logic vs. Description Logics 82
6.3.2 Closed World Assumption 83
6.3.3 Unique Name Assumption 84
6.4 Use Cases for Rules in Cultural Heritage 84
6.5 Bibliographical and Historical Notes 86

7 Cultural Content Creation 87

7.1 Vocabulary and Ontology Creation 87
7.1.1 Conceptual Levels of Ontology Creation 87
7.1.2 Transforming Legacy Thesauri into Ontologies 88
7.1.3 Terminology Creation 93
7.1.4 Ontology Alignment 93
7.1.5 Ontology Evolution 94

7.2 Transforming Local Content into RDF 96
7.2.1 Transformation Process 97
7.2.2 Transforming Relational Databases into RDF 98
7.3 Content Aggregation and Integration 101
7.4 Quality of Linked Data 102
7.4.1 Data Quality of Primary Sources 102
7.4.2 Metadata Quality 103
7.4.3 Quality of Linked Data Services 104
7.5 Bibliographical and Historical Notes 104

8 Semantic Services for Human and Machine Users 107

8.1 Classical Information Retrieval 107
8.2 Semantic Concept-based Search 109
8.2.1 Handling synonyms 109
8.2.2 Homonyms and Semantic Disambiguation 109
8.2.3 Query and Document Expansion 110
8.3 Semantic Autocompletion 111
8.4 Faceted Semantic Search and Browsing 111
8.5 Semantic Browsing and Recommending 112
8.6 Relational Search 114
8.7 Visualization and Mash-ups 115
8.7.1 Visualizing Dataset Clouds 115
8.7.2 Visualizing Ontologies 115
8.7.3 Visualizing Metadata 116
8.7.4 Visualizing Search Results 117
8.8 Personalization and Context Awareness 117
8.9 Cross-portal Re-use of Content 118
8.10 Bibliographical and Historical Notes 119

9 Conclusions 121

Bibliography 123

Author's Biography 139

Index 141

Preface

Publishing Cultural Heritage (CH) collections and other content on the Web has become one of the most successful application domains of Semantic Web and Linked Data technologies. After a period of technical research and prototype development, boosted by the W3C Semantic Web Activity kick-off in 2001 and the Linked (Open) Data movement later on, major national and international CH institutions and collaboration networks have now started to publish their data using Linked Data principles and Semantic Web technologies.

This work is highly interdisciplinary, involving domain expertise of museum curators, librarians, archivists, and researchers of cultural heritage, as well as technical expertise of computer scientists and Web designers. Applying a new technology in the rapidly evolving Web environment is challenging not only for non-technical personnel in CH institutions, but also for computer scientists themselves.

This book aims at fostering the application of Linked Data and Semantic Web technologies in the CH domain by providing an overview of this fascinating application domain of semantic computing. My own work in this field started in 2001 after the W3C Semantic Web Activity launch by establishing the Semantic Computing Research Group (SeCo) focusing on this field. We first developed a semantic photograph search and recommender system for a university museum, followed by semantic portal prototypes for publishing heterogeneous collections of different kinds, including artifacts in cultural history museums, historical events, folklore, maps, fiction literature, and natural history museum data. This book reflects experiences gained during this work.

From the very beginning in 2002, after developing our first ontologies and transforming the first collection databases into RDF, it became clear that the possibility of reusing existing data, metadata models, and ontologies, and linking it all together in an interoperable way, will be a central benefit of Semantic Web applications. W3C recommendations, such as RDF(S), SKOS, SPARQL, and OWL are the corner stones for facilitating cross-domain, domain-independent interoperability, but this is not enough. We also need domain-dependent metadata-models and domain ontologies based on the generic semantic principles, as well as domain specific datasets. From a practical viewpoint, we also need ontology services so that the shared resources can be published and used in legacy and other application systems in a cost-efficient way. In short, a Semantic Web *content infrastructure* needs to be built in a similar vein as railroad, telephone, and other communication networks were created during earlier technological breakthroughs.

Creating a Semantic Web infrastructure, as well as content for it, requires collaboration between content providers. Co-operation is needed not only for sharing data through joint portals such as Europeana, but also for developing shared metadata models and ontologies used in representing the contents in an interoperable way. Publishing CH content is becoming a game of cross-domain

networking where the traditional boundaries of memory organizations based on content types are breaking down. From a user's viewpoint, the focus is on data, knowledge, and experience, be it based on a book in a library, an artifact in a cultural history museum, a story in an archive, a painting in an art gallery, a photograph taken by a fellow citizen, or a piece of music on a record.

During these years my faith in Semantic Web and Linked Data has become strong even if there are great challenges ahead, too. This is a truly promising way for providing richer content to users through more intelligent and usable interfaces, and at the same time for facilitating memory organization with better tools for collaborative, open content publishing on the Semantic Web.

Eero Hyvönen
October 2012

Acknowledgments

Thanks to the series editors Jim Hendler and Ying Ding for the invitation to write this book, and to Mike Morgan for making the publication possible.

The book's contents are based on collaboration with various students, researchers, and visitors in the Semantic Computing Research Group (SeCo) at the Aalto University and University of Helsinki in different times including (in alphabetical order) Matias Frosterus, Harri Hämäläinen, Tomi Kauppinen, Suvi Kettula, Heini Kuittinen, Jussi Kurki, Nina Laurenne, Aleksi Lindblad, Thea Lindquist, Glauco Mantegari, Eetu Mäkelä, Panu Paakkarinen, Tuomas Palonen, Sini Pessala, Tuukka Ruotsalo, Samppa Saarela, Katri Seppälä, Osma Suominen, Jouni Tuominen, Juha Törnroos, Mika Wahlroos, Mark van Assem, and Kim Viljanen.

Ying Ding, Stefan Gradmann, Patrick Leboeuf, Glauco Mantegari, Katri Seppälä, and Regine Stein made fruitful comments to earlier versions of this manuscript. Special thanks to Jouni Tuominen for several comments, suggestions, and help in proofreading the text. C.L. Tondo's help was invaluable in finalizing the text and layout.

Fruitful collaboration with several museums, libraries, archives, and media organizations in Finland is acknowledged, including (in alphabetical order) Agricola.fi network of historians, Antikvaria Museum Group, Espoo City Museum, Finnish Agriculture Museum, Finnish Broadcasting Company YLE, Finnish Literature Society, Finnish Museum of Photography, Finnish Museum Association, Finnish National Gallery, Finnish Public Libraries (Libraries.fi), Helsinki City Library, Helsinki University Library, Helsinki University Museum, Lahti City Museum, National Board of Antiquities, National Library of Finland, and Suomenlinna Sea Fortress.

The National Funding Agency for Technology and Innovation (Tekes)[1] and consortia of tens of public organizations and companies have supported several research projects of SeCo related to CH, such as Intelligent Catalogs (2002–2004), FinnONTO[2] (2003–2012), Semantic Ubiquitous Services (2009–2012)[3], and Linked Data Finland[4] (2012–). The Finnish Cultural Foundation[5] has supported our research on the CultureSampo system, too.

Thanks to SmartMuseum EU project[6] for funding and collaboration, to European Institute of Technology (EIT) Project EventMAP, as well as to the Network for Digital Methods in the Arts and Humanities (NeDiMAH) (European Science Foundation). Joint work with the University of Colorado regarding war history and linked data is acknowledged. Thanks to collaborations with the

[1] http://www.tekes.fi/en/
[2] http://www.seco.tkk.fi/projects/finnonto/
[3] http://www.seco.tkk.fi/projects/subi/
[4] http://www.seco.tkk.fi/projects/ldf/
[5] http://www.skr.fi/
[6] http://www.smartmuseum.eu/

Continuous Access to Cultural Heritage (CATCH) initiative and colleagues at the VU University and other universities in the Netherlands.

Eero Hyvönen
October 2012

CHAPTER 1

Cultural Heritage on the Semantic Web

Cultural Heritage (CH) refers to the legacy of physical objects, environment, traditions, and knowledge of a society that are inherited from the past, maintained and developed further in the present, and preserved (conserved) for the benefit of future generations[1]. This chapter first characterizes the notion of CH and identifies specific challenges encountered when publishing CH contents, especially collection data, on the Web. After this, Semantic Web and Linked Data technologies are introduced as a novel, promising approach to address the problems. The chapter ends with an overview of the book content.

1.1 CHARACTERIZING CULTURAL HERITAGE

CH can divided into three subareas.

1. **Tangible cultural heritage** consists of concrete cultural objects, such as artifacts, works of art, buildings, and books.

2. **Intangible cultural heritage** includes phenomena such as traditions, language, handicraft skills, folklore, and knowledge.

3. **Natural cultural heritage** consists of culturally significant landscapes, biodiversity, and geodiversity.

The key players in preserving CH are *memory organizations* that include libraries, archives, and museums of different kinds specializing in particular areas of CH, such as art museums, archaeological museums, botanical museums and gardens, cultural history museums, medical collections, science museums, theater history museums, geological and mineralogical museums, and zoology museums. Also media organizations often preserve CH materials, especially more recent ones. There are also lots of CH materials maintained by cultural associations of various kinds and individual persons. Tangible CH objects are stored with attached metadata, intangible heritage is documented using textual descriptions, photographs, interviews, and videos, and there are natural history and other museums specializing in storing traces and knowledge of natural history, geology, and environment.

[1] In this book, the ambiguous term "culture" is used to refer to the "the ideas, customs, skills, arts, etc. of a people or group, that are transferred, communicated, or passed along, as in or to succeeding generations" (Webster's New World Dictionary).

The Web has become an increasingly important medium for publishing CH contents of different kinds. For example, libraries and archives are online with their collections, museums show their collections through collection browsers, and documentation of intangible heritage is available as audio and video recordings and as interactive hypertext applications, even as games. There are large national and multi-national CH portal projects active in harvesting and publishing content from different sources via centralized services.

For the layman end-user, such systems provide a single access point to massive heterogeneous collections and an authoritative source of information. In contrast to traditional physical exhibitions, Web services are open all the time, can be accessed without physical presence at an exhibition, the number of exhibits on the Web is not limited by the physical space available, and the exhibits can be linked and accessed flexibly using different strategies, not only the one used in the physical exhibition. Of course, the Web cannot replace the physical experience of visiting a museum or an exhibition in reality but provides a complementary alternative for accessing collection data virtually at any time and from any place.

For researchers in the humanities, availability of CH data in massive amounts in digital machine processable form has opened up a new research paradigm called Digital Humanities.

1.2 INFORMATION PORTALS FOR CULTURAL HERITAGE

There are several kind of CH publications on the Web. First, there is a large variety of well-curated systems that have been hand-crafted for a specific purpose with a focused closed theme, dataset, and interfaces. Such systems are often implemented using tools such as Adobe Flash with a beautiful game-like appearance. For example, the Lewis and Clark Expedition (1803–1806) is documented on the Web in great detail by several applications. The portal in Figure 1.1[2] provides the end-user with several thematic perspectives to the journey by selecting the buttons on the left, such as "overview," "American nation," "geography," "journal excerpts," "natural history," and "technology used." Such systems may also be available on CD/DVD as stand-alone applications.

On the other end of the spectrum, there are collection search services and browsers providing access to large open collection databases whose content is not thematically focused, and curated access paths and interfaces may be missing. In return, large collection databases originating possibly from several institutions can be accessed. For example, a variety of Australian CH collections can be accessed using the Collections Australia Network system[3]. Similar federated portals for searching and browsing collections can be found in many countries and internationally. A flagship application here is Europeana[4], based on millions of collection objects originating from memory organizations all over Europe. For example, in Figure 1.2 the user has typed in the keyword "chair" in the search field of Europeana and the system has found various chairs in participating collections. The search can be refined further by selecting additional filters on the facets on the left, such as "media type,"

[2] http://lewis-clark.org/
[3] http://www.collectionsaustralia.net/
[4] http://www.europeana.eu/

Figure 1.1: A portal exhibiting versatile content related to the Lewis and Clark expedition (1803–1806) in the U.S. from different perspectives. (Fort Mandan Foundation, North Dakota)

"language," "date," "country," and whether content contributed by users should be included or not. Another portal example, harvesting library data, is WorldCat[5] that contains metadata (without the primary sources) of about 1.5 billion books, DVDs, CDs, and articles in the participating libraries. The World Digital Library[6] is yet another international portal, operated by UNESCO and the United States Library of Congress, that makes available, free of charge, significant multilingual primary materials, such as manuscripts, maps, rare books, musical scores, recordings, films, prints, photographs, and architectural drawings.

In this book, the main focus is on *information portal systems* of the latter kind: CH portals based on large heterogeneous collection datasets are considered, where organizing the contents by hand into a focused thematic application with application-specific visualizations and interfaces is not usually feasible. Such shared publication portals facilitate exchange of knowledge for CH researchers, librarians, and archivists. For the contributing memory organizations, such systems are

[5] http://www.worldcat.org/
[6] http://www.wdl.org/

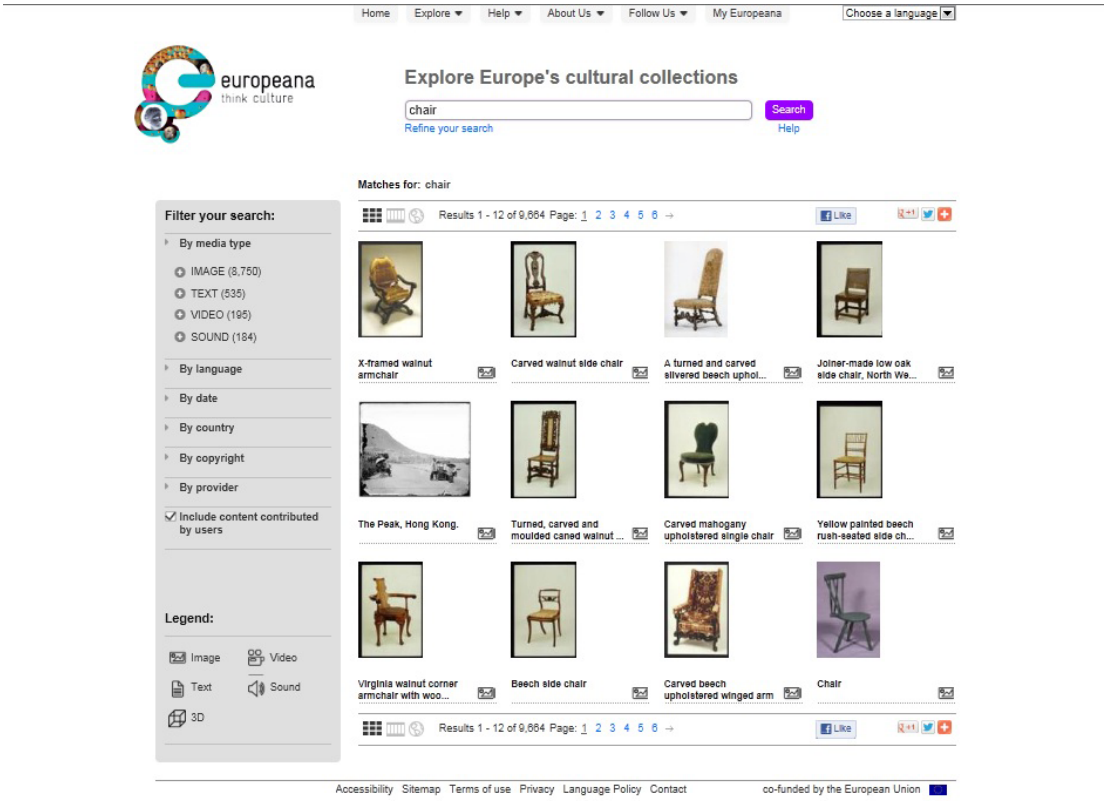

Figure 1.2: Faceted search in Europeana portal exhibiting chairs from different European collections.

an opportunity to reach out to wider audiences on the Web with new ways of interaction, and to collaborate with other organizations. From a societal perspective, publishing CH on the Web stimulates cultural tourism, creative economy, and enhances friendly relationships and unity between parties and nations involved in such initiatives.

1.3 CHALLENGES OF CULTURAL HERITAGE DATA

CH collection data has many specific characteristic features, such as the following.

- **Multi-format**. The contents are presented in various forms, such as text documents, images, audio tracks, videos, collection items, and learning objects.
- **Multi-topical**. The contents concern various topics, such as art, history, artifacts, and traditions.

- **Multi-lingual**. The content is available in different languages.
- **Multi-cultural**. The content is related and interpreted in terms of different cultures, such as religions or national traditions in the West and East.
- **Multi-targeted**. The contents are often targeted to both laymen and experts, young and old.

As a result, a fundamental problem area in dealing with CH data is to make the content mutually *interoperable*, so that it can be searched, linked, and presented in a harmonized way across the boundaries of the datasets and data silos. The problem occurs on a syntactic level, e.g., when harmonizing different character sets, data formats, notations, and collection records used in different collections. Even more importantly, there is the problem of *semantic interoperability*: different metadata formats may be interpreted differently, data is encoded at different levels of precision, vocabularies and gazetteers used in describing the content are different, and so on. The Semantic Web standards[7] and best practices, especially those advocated by the World Wide Web Consortium (W3C)[8], provide a shared basis on which interoperable Web systems can be built in a well-defined manner. The new technologies are of course no panacea for all problems but rather a tool set by which the hard issues can be tackled arguably more effectively than before.

A major reason for interoperability problems in CH content publishing is the *multi-organizational* nature in which CH content is collected, maintained, and published. The content is provided by different museums, libraries, and archives with their own established standards and best practices, by media organizations, cultural associations, and individual citizens in a Web 2.0 fashion. The success of the WWW is very much due to its simple distributed many-to-many publishing paradigm that has few restrictions and shared standards, with the HTML mark-up language combined with the HTTP protocol and the idea of URL addressing as core technologies. However, things get more complicated on the Semantic Web, where content is not published only for human users in HTML form but also as data for machines to use. An additional standard base is needed for the Web of Data. In application domains such as CH more coordinated collaboration is needed between CH publishers and the technical WWW developer community than before.

1.4 PROMISES OF THE SEMANTIC WEB

Semantic Web technologies[9] [34] (SW) are a promising new approach for addressing the problems of publishing CH content on the Web. The term "semantic" here refers to *Semantics*, a discipline studying relations between *signifiers*, such as words, phrases, signs, and symbols, and what they stand for, i.e., *denotata*. In Computer Science semantics refers to the formal meaning and interpretation (declarative or procedural) that has been given to syntactic structures, such as programming languages or symbolic data structures.

[7] Called "recommendations" by the W3C.
[8] http://www.w3.org/
[9] http://www.w3.org/standards/semanticweb/

The Semantic Web can be seen as a new layer of *metadata* being build inside the Web. According to the traditional definition, metadata is data about data. For example, a metadata record of a book (data) may tell its title, author, subject, and publishing year. However, the term "metadata" is used more widely in the Semantic Web context as a synonym for machine processable or interpretable data. The key idea is that syntactic metadata structures make Web content "understandable" to the machines, based on shared semantic specifications founded on formal logic. This makes it possible to create more interoperable and intelligent Web services. A computer that cannot interpret the data it is dealing with is like a telephone just passing information, and cannot be very helpful in more complicated information processing tasks dealing with the meanings of the contents.

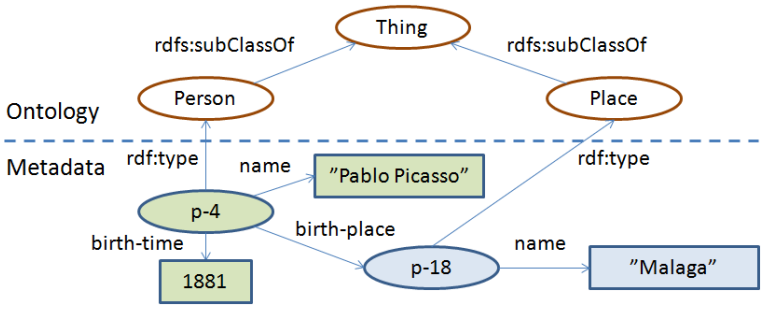

Figure 1.3: The data model of RDF is a directed labeled graph.

The methodology for representing metadata and ontological concepts[10] on the Web is based on a simple data model: a directed labeled graph, i.e., a *semantic net*. For example, Figure 1.3 depicts an RDF graph telling on a metadata level that the identity *p-4* is an individual of the class *Person* (denoted by the arc *rdf:type*) with name "Pablo Picasso" born in 1881 at an instance *p-18* of the class *Place* whose name is "Malaga." In the RDF graph, classes such as places and persons are represented as subclasses (arc *rdfs:subClassOf*) of the class *Thing* on an ontology level, while the individuals of the classes are considered metadata. Both metadata and ontologies are represented uniformly in the same graph. In the figure, identities that may have properties, i.e., may have out-going arcs, are depicted as ovals while literal terminal atomic values without further properties (here strings and numbers) as rectangular boxes.

The figure illustrates that actually there are several levels of descriptions needed on the Semantic Web.

1. **Real world**. On the bottom, there is the real world, i.e., the domain of discourse, such as persons, artifacts, and places.

[10]The notion of "concept" is a complex philosophical notion referring to a general idea or something conceived in the mind. On the Semantic Web, the term "concept" is used for any entity on the Web or outside of it with an identity specified by a URI.

2. **Data level**. Then there is the data level, since real world items have to be represented as data of some kind in a computer. For example, images and documents are data as well as a URI reference to a person.

3. **Metadata level**. After data, there is metadata about the data, e.g., records in a collection database about images, persons, or artifacts.

4. **Ontology level**. Next, ontology level defines the generic classes and properties used in describing a domain, i.e., the vocabularies in terms of which the metadata is represented. The metadata schema used in cataloging and controlled vocabularies of subject headings are part of this level. For example, in Figure 1.3 persons are described in terms of their name, birth time, and birth place, and instantiated from the classes defined on the ontology level. The same ontologies can be used for representing collection metadata of a similar domain area in different memory organizations (e.g., books in libraries).

5. **Metaontology level**. Finally, there are the general cross-domain modeling principles of ontologies that are domain-independent. For example, the notions of subclass-of relation and class are generic and not restricted to a particular domain. Such generic principles are specified by the Semantic Web standards, such as RDF(S) and OWL, and facilitate cross-domain interoperability of contents.

On a global WWW scale, the Semantic Web forms a *Giant Global Graph* (GGG) of connected data resources. The GGG can be used and browsed in ways analogous to the WWW, but while the WWW links associated Web pages with each other for human use, the GGG links associated underlying concepts and data resources together. For example, the GGG may tell that ducks are birds, and that Donald is an instance of a duck (and therefore a bird) while the related WWW pages may constitute a comics book about Donald Duck.

A key idea of linked data is that the different parts of the GGG can come from different data sources. For example, in Figure 1.3 metadata about persons, such as Pablo Picasso, may come from an authority database, information about places, such as Malaga, may be provided by a land survey organization, and the class ontology can be based on an existing keyword thesaurus in use in a library. Different data sources are illustrated in the figure by different colors/densities.

Based on harmonized RDF-based representations of data, more "intelligent" Web applications can be built and with less effort. From a technical application perspective, Semantic Web technologies have many promising features:

- *More accurate content descriptions*. The technology is based on globally unique Universal Resource Identifiers (URI), which makes it possible to refer to meanings more accurately than using literal expressions. For example, person and place names can be disambiguated: there are lots of "John Smiths" around, "Paris" can be found in France, Texas, and in many other places, and the names can have different transliterations in different language systems. In libraries, the notion of, e.g., Shakespeare's play "Hamlet" can refer to the abstract story, its manifestation

as a text or a video of the play, different translations of it, variants of the story, editions of these, and finally individual books or DVDs on the library shelves. Modeling such semantic distinctions can be done using novel "ontology-based" CH standards to be presented in this book.

- *Interoperability*. Semantic Web technologies provide a novel approach to creating interoperable linked data.

- *Simple data model for aggregation*. Two (interoperable) RDF graphs can be joined together technically in a trivial way by simply making the union of them (i.e., the corresponding triple sets).

- *Data aggregation by linked data*. By combining data sources in an interoperable way, data from one source can be enriched with additional linked data from another source. A notable international initiative toward this goal is Linked Data[11] [53], where open datasets such as Wikipedia/DBpedia[12] and Freebase[13] for common knowledge, GeoNames[14] for millions of place names, or Gutenberg project[15] for over 40,000 free ebooks are described in terms of Semantic Web standards and interlinked with each other.

- *Semantic Web services*. Semantic linked data is published not only as passive datasets, but as operational services than can be utilized by legacy and other CH applications via open and generic Application Programming Interfaces (API). By utilizing shared ready-made services, application programmers can re-use work done by others, and save their own programming effort and resources. This idea can be paralleled with Google and Yahoo! Maps that provide map services on a global basis to applications via easy-to-use APIs for mash-up development.

Publishing CH on the Web is not only a technical challenge; issues of trustworthiness of content, copyrights, and licensing are also of concern. Much of CH content is protected by copyright, and there are also other reasons why organizations cannot publish their data openly, e.g., issues of personal privacy. However, based on the ideas of Linked Open Data, the WWW world is clearly taking steps toward publishing open data and free of charge when feasible. The idea is that CH content should be maximally shared. It is also usually produced by public funding and in this sense already paid by the public. Free open data also fosters interoperability and creates a basis on which commercial applications can be built more easily. Trust and copyright issues are important, e.g., in Web 2.0 spirited social cultural portals, where end-users create, tag, and publish content of their own and the others'.

[11] http://linkeddata.org/
[12] http:/www.dbpedia.org/
[13] http://www.freebase.com/
[14] http://www.geonames.org/
[15] http://www.gutenberg.org/

1.5 OUTLINE OF THE BOOK

This book is an introduction to publishing CH contents on the Semantic Web as Linked Data. The idea is to provide a kind of cook book on how to create semantic portals of CH, where heterogeneous content is produced by a multitude of distributed organizations, and is harvested, harmonized, validated, and published as a service for human and machine users.

The text starts (Chapter 2) with presenting a motivating "business model" for this prototypical semantic portal scenario that can be considered a kind of standard model for publishing CH on the Semantic Web. In Chapter 3 requirements for publishing Linked Data are considered. The Semantic Web is based on the "layer cake model" of W3C that adds new standards above the XML[16] standard family, the *lingua franca* of the Web.

- *Metadata level*. The RDF data model[17] is the basis of the Semantic Web and Linked Data, and is used for representing metadata as well as other forms of content on the Web of Data. Metadata models for CH data are considered in Chapter 4.

- *Ontology level*. The RDF Schema and the Web Ontology Language OWL[18] are used for representing ontologies that describe vocabularies and concepts concerning the real world and our conception of it. Domain vocabularies and ontologies for CH are in focus in Chapter 5.

- *Logic level*. Logic rules, to be discussed in Chapter 6, can be used for deriving new facts and knowledge based on the metadata and ontologies. This can be used, e.g., to minimize cataloging work, make searching and browsing more effective, and to find serendipitous semantic links between CH objects.

After presenting technical foundations and models, issues related to annotating and harvesting CH content for a portal are presented in Chapter 7. Chapter 8 discusses intelligent services based on semantic linked data. The book is finally concluded in Chapter 9.

1.6 BIBLIOGRAPHICAL AND HISTORICAL NOTES

The idea of the World Wide Web (WWW) was proposed first in 1989 by Tim Berners-Lee, and more formally with Robert Cailliau in 1990. History of the early WWW is documented in the book "Weaving the Web" [14]. Already in the early days of the WWW the idea of a "Semantic Web," i.e., a web of machine interpretable data, has been around. However, the first generation of the WWW was targeted to humans, and was based on three simple technologies for mediating Web pages between human users: HTML, HTTP, and URLs.

From a scientific viewpoint, the Semantic Web is based on results of Artificial Intelligence, where semantic networks and logic-based knowledge representation have been studied from the

[16] http://www.w3.org/XML/
[17] http://www.w3.org/RDF/
[18] http://www.w3.org/2004/OWL/

late 50's; see, e.g., [126] for a thorough overview of this field. The first Semantic Web standard in use, Resource Description Framework (RDF), was published by W3C already in 1999, only a year after the XML recommendation. As another approach for the Semantic Web, Topic Maps [114] has been developed and published as the ISO standard ISO/IEC 13250:2003[19]. This standard is intended for the representation and interchange of knowledge, with an emphasis on the findability of information. The system originated from the idea of creating semantic indexes for publications. However, Semantic Web development really got off using the W3C standard stack after the publication of the seminal article "The Semantic Web" [15] in *Scientific American,* and the launch of the Semantic Web Activity at W3C.

The semantic technology did not penetrate the market as quickly as many other Web developments, say XML. A reason for this is complexity of some standards and their foundations in logic not so familiar in mainstream computing. In around 2005, the ideas on Linked Data and Web of Data started to gain momentum as a simple approach to the Semantic Web focusing on publishing large existing datasets, and using only simple RDF and lightweight ontologies. Combined with idea of Open Data, the idea of the Semantic Web has been adopted especially by the public sector [158], and several national initiatives have been started in the U.K.[20], U.S[21], and in smaller countries, such as Finland [67].

A thorough overview of Linked Data and Web of Data is presented in [53]. Semantic Web and linked data standards and technology, with pointers to related research and applications, can be accessed at W3C Web pages[22], and at the home pages of the Linked Data community[23]. The W3C Linked Library Data Incubator Group has evaluated the current state of library data management, outlined the potential benefits of publishing library data as Linked Data, and formulated next-step recommendations for library standards bodies, data and systems designers, librarians and archivists, and library leadership in a final report[24]. Another report "Linked Data for Libraries, Museums, and Archives: Survey and Workshop Report" with related goals was published at the same date, based on a workshop at the Stanford University[25]. Major international Semantic Web conferences include the International Semantic Web Conference (ISWC) and Extended Semantic Web Conference (ESWC). The World Wide Web conference (WWW) is the main yearly event for general Web research with a W3C focus.

A wide variety of Web applications in the museum domain have been presented in the proceedings of the Museums and the Web conference series since 1997, with papers available online[26]. The International Federation of Library Associations and Institutions (IFLA)[27] organizes a large annual World Library and Information Congress for libraries, and the International Council on

[19] http://www.isotopicmaps.org/
[20] http://data.gov.uk/
[21] http://www.data.gov/
[22] http://www.w3.org/standards/semanticweb/
[23] http://linkeddata.org/
[24] http://www.w3.org/2005/Incubator/lld/
[25] http://www.clir.org/pubs/abstract/reports/pub152
[26] http://www.archimuse.com/conferences/mw.html
[27] http://www.ifla/

Archives (ICA)[28] has a similar annual congress series, International Conference of the Round Table on Archives (CITRA) for archivists.

The intersection of computing and the disciplines of the humanities are studied in the field of *Digital Humanities*, also called *Humanities Computing*. [105] The general goal here is to develop and apply computational methods in humanities research. Since 1990, the digital humanities community has been organizing the Digital Humanities conference series[29]. A major journal in the field is the *Digital Humanities Quarterly*[30].

[28] http://www.ica.org/
[29] http://digitalhumanities.org/conference
[30] http://www.digitalhumanities.org/dhq/