Schriften des Instituts für Dokumentologie und Editorik — Band 3

Kodikologie und Paläographie im digitalen Zeitalter 2

Codicology and Palaeography in the Digital Age 2

herausgegeben von | edited by

Franz Fischer, Christiane Fritze, Georg Vogeler

unter Mitarbeit von | in collaboration with

Bernhard Assmann, Malte Rehbein, Patrick Sahle

2010
BoD, Norderstedt

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über **http://dnb.d-nb.de/** abrufbar.

Online-Fassung
Herstellung und Verlag der Druckfassung: Books on Demand GmbH, Norderstedt 2010
ISBN: 978-3-8423-5032-8
Einbandgestaltung: Johanna Puhl, basierend auf dem Entwurf von Katharina Weber
Satz: Stefanie Mayer und LaTeX

Semantic Technologies for Manuscript Descriptions — Concepts and Visions

Robert Kummer

Abstract

The contribution at hand relates recent developments in the area of the World Wide Web to codicological research. In the last number of years, an informational extension of the internet has been discussed and extensively researched: the Semantic Web. It has already been applied in many areas, including digital information processing of cultural heritage data. The Semantic Web facilitates the organisation and linking of data across websites, according to a given semantic structure. Software can then process this structural and semantic information to extract further knowledge. In the area of codicological research, many institutions are making efforts to improve the online availability of handwritten codices. If these resources could also employ Semantic Web techniques, considerable research potential could be unleashed. However, data acquisition from less structured data sources will be problematic. In particular, data stemming from unstructured sources needs to be made accessible to Semantic Web tools through information extraction techniques. In the area of museum research, the CIDOC Conceptual Reference Model (CRM) has been widely examined and is being adopted successfully. The CRM translates well to Semantic Web research, and its concentration on contextualization of objects could support approaches in codicological research. Further concepts for the creation and management of bibliographic coherences and structured vocabularies related to the CRM will be considered in this chapter. Finally, a user scenario showing all processing steps in their context will be elaborated on.

Zusammenfassung

Der Beitrag bezieht neue Entwicklungen im World Wide Web auf die kodikologische Forschung. Seit einiger Zeit wird eine informationelle Erweiterung des Internet diskutiert und in vielen Bereichen, auch der digitalen Informationsverarbeitung des kulturellen Erbes, ausführlich erforscht und getestet: das Semantic Web. Das Konzept beinhaltet, dass Daten auf der Ebene ihrer Bedeutung miteinander verknüpft werden, damit Computer diese verarbeiten und weitere Informationen daraus gewinnen können. Im Bereich der Kodikologie gibt es schon seit einigen Jahren Bemühungen, handschriftliche Kodizes online verfügbar zu machen. Wenn auch diese den Schritt

Kodikologie und Paläographie im Digitalen Zeitalter 2 – Codicology and Palaeography in the Digital Age 2. Hrsg. Franz Fischer, Christiane Fritze, Georg Vogeler, unter Mitarbeit von Bernhard Assmann, Malte Rehbein, Patrick Sahle. Schriften des Instituts für Dokumentologie und Editorik 3. Norderstedt: Books on Demand, 2010. 133–154.

in das Semantic Web vollziehen würden, könnten daraus nicht unerhebliche Forschungspotenziale abgeleitet werden. Die Datengewinnung aus wenig strukturierten Datenquellen ist dabei nicht unproblematisch. Insbesondere Daten aus unstrukturierten Quellen müssen zunächst mittels Verfahren der Informationsextraktion einer weiteren Verarbeitung im Sinne des Semantic Web zugänglich gemacht werden. Im Umfeld der Museumsforschung wird das CIDOC Conceptual Reference Model (CRM) ausführlich diskutiert und bereits gewinnbringend eingesetzt. Das CRM lässt sich gut auf die Forschung des Semantic Web beziehen und seine Konzentration auf Kontextualisierung von Objektzusammenhängen könnte der kodikologischen Forschung entgegenkommen. Weitere Konzepte und Standards im Umfeld des CRM zur Erstellung und Verwaltung bibliographischer Zusammenhänge und strukturierter Vokabulare werden in die Überlegungen einbezogen. Abgerundet wird die Betrachtung durch ein Benutzungsszenario, an dem verschiedene Verarbeitungsschritte in ihren Zusammenhang gestellt werden.

1. Semantic Codicology

How can the methods and tools of the Semantic Web be applied to the domain of codicology? Many handwritten codices have already been published online, mainly for viewing. Catalogs and common information retrieval techniques (e.g. full-text searching) enable discovery of information. But could additional research potential be unlocked by also making this information available according to the concepts of the Semantic Web? Could we ask and approach other questions by processing this information with the tools that have been developed in this area?

For the study and description of a specific codex, knowledge from several disciplines needs to be considered such as, for example, philology. In addition, statistical techniques have been employed to elaborate stemmata for single texts. Geographic and chronological dissemination of scripts and decorations have been considered as significant features. With regard to the individual codex, manual and technical aspects of production require study, for instance, queries regarding material (papyrus, parchment or paper), binding of folios and quires, ink and writing utensils, book decorations and provenance. Codicology simultaneously treats its research objects as material artifacts and as abstract documents. Thus, an analogy between codicology and archaeology can be drawn to a certain extent. For example, during his studies of the Rothschild collection, Delaissé showed a strong commitment to what he called "the archaeology of the book" (Delaissé, Marrow, and de Wit; Maniaci).

In order to assess the research potential of the Semantic Web for the domain of codicology, this standard should be evaluated with a focus on how methods of codicology translate into methods of Semantic Web research; explicit contextual modeling of information could be the key method that is common to both. In particular, focusing on contextual coherences of objects, as the Semantic Web does, could support the methods of codicology.

The following passages provide an overview of Semantic Web concepts and tools. Semantic Web research itself needs to be considered as part of research in information integration and artificial intelligence. Findings in these research areas will not be exhaustively presented but rather mentioned when appropriate. Concepts and tools that have evolved as part of Semantic Web research will be introduced by an example that relates to codicology. However, no suggestions for concrete applications will be made. User scenarios have been considered helpful both for envisioning future software applications and implementing existing ideas (Alexander). An exhaustive user scenario would certainly help to understand where the ideas of the Semantic Web could support codicology in the future. Hopefully, this contribution will help to create such a user scenario for “Codicology and the Semantic Web”.

2. Semantic Web Research

Usually, information on a specific research topic is scattered among several cultural heritage information systems. In many cases, information can only be processed according to user-needs if it has been integrated. Integrated information systems can process data in a more complete fashion and usually provide better results. Additionally, they offer one consistent way of dealing with the data instead of users having to learn many user metaphors. Thus, if data stemming from different information systems is to be processed in a uniform way, it needs to be harmonized in terms of syntax and semantics. For many operations, the integrated information must reside in the main memory of a single computer to be processed efficiently and according to the needs of users.

Berners-Lee, Hendler, and Lasilla conceptualized a so-called “vision piece” that describes an infrastructure to provide greater capabilities. The authors argue that the available data on the World Wide Web has been designed for humans to read and process. They point out the importance of particular pieces of software, so-called intelligent software agents. These software artifacts are reminiscent of rational agents described by Russell and Norvig. An agent, in this sense, is designed to aid humans in information processing by acting rationally to collect, process and share data. To enable this, data currently published as part of the World Wide Web needs to be represented in certain ways. This holds true for web pages but also for databases that are considered to be part of the “deep web”.[1] Consequently, research in the area of the Semantic Web has developed from several branches of information technology: building

[1] Some web sites are generated dynamically each time a user requests a page. This part of the web is difficult to index by search engines like Google. Usually, the data is managed in some kind of proprietary structure

models, computing with knowledge and exchanging information (Hitzler, Krötzsch, and Rudolph).

In order to make information accessible for automatic processing, it has to be formalised. In the scope of the Semantic Web several concepts have been proposed. XML seems to be well established in the digitisation community, and is often used for encoding information about a codex and its contents. The Text Encoding Initiative (TEI) provides a set of XML tags for this task in chapter ten of the TEI guidelines. The Semantic Web community has proposed a recommendation that can be expressed in XML. The Resource Description Framework (RDF) can be used to express so-called triples that are simple statements which take the ordered form: 'subject', 'predicate', 'object'. RDF is a data model that allows us to make "statements" about subjects. The World Wide Web Consortium provides an excellent introductory text on RDF (Manola and Miller). A statement like " 'De natura rerum' is written by Beda Venerabilis" can be easily expressed as the RDF triple: " 'De natura rerum' " (subject): "is written by" (predicate): "Beda Venerabilis" (object). Subject, predicate and object may each be identified by a Universal Resource Identifier (URI). A URI is a simple string of characters that is used to identify a thing. The most commonly used form of a URI is an URL (Uniform Resource Locator), which are used daily to direct a web browser to a web page like "<http://example.org/>". It is common practice to use URIs that have the form of URLs in the Semantic Web community to refer to a thing.

Another interesting aspect of the Semantic Web is that its community actively researches techniques from artificial intelligence. Many Semantic Web tools make use of so-called "inference engines" to deduct new knowledge from databases. Thereby, structured and sophisticated queries can rely on a larger amount of information than originally available. Furthermore, so-called "ontologies" comprise "taxonomies" and rules that can be deployed to process data according to the intended meaning.[2]

XML-based data, RDF, URIs and ontologies provide the tools for manipulating data as information, and even as knowledge, but can also support information sharing. With the help of ontologies different communities can agree on the meaning of certain concepts. By coordinating definitions that different communities have developed, a shared understanding of concepts can be achieved. This allows data stemming from different information systems to be processed according to the intended and agreed meaning.

that makes it difficult to share and process outside the boundaries of the system; clearly an issue that the Semantic Web tries to deal with.

[2] Although it has its origin in philosophy, in computer science the term "ontology" refers to a formal representation of knowledge about a certain domain. The notion of an ontology will be elaborated on in chapter 4.

3. Extracting and Modeling Information

The previous section has introduced a suite of concepts that should help to put the main ideas of the Semantic Web into practice. The central concept to encode, share and process information is the triple. So-called “Triplestores” are computer programs that control the creation, use and maintenance of data that has the form of triples. Unlike traditional relational databases, Triplestores are purpose-built for dealing with data encoded according to RDF. However, relational databases may be used to make data persistent. The RDF data model builds on the notion of a graph.[3] The process of filling these stores with data will be described in this section.

It becomes apparent that the subject and object of a triple need to be atomic units of discourse that can be identified by a URI (e.g. “<http://example.org/cod/123>”: “<http://example.org/writtenIn>”: “<http://example.org/scriptorium/321>”). We want to be able to refer to exactly one concept or thing to make a statement about it. But information sources need to deliver information of high quality and granularity to establish these triples. Therefore, in some cases, information needs to be extracted rather than mapped to each data source if it does not provide enough structure.

Many information systems that deal with cultural heritage material use information retrieval methods to provide searching capabilities. In fact, traditional information retrieval tries to deal with material that is not very well-structured, such as full-text. In contrast, information extraction aims at extracting useful structured information from, for example, a text document (Konchady). In the field of Semantic Web research it would be desirable to extract triples from semi-structured sources as well as from highly structured sources (like formal descriptions of manuscripts in a catalogue). In Section 6, a user scenario will be developed that relies on highly structured data and would not be possible with traditional information retrieval. Often, information retrieval starts with identifying named entities such as people, places and institutions in documents (Cardie), not unlike a traditional printed index.

Where do we find data in the field of codicology? Many institutions have decided to publish digital information about codices according to certain standards of description. Listing 1 shows how information about a manuscript can be encoded using TEI. It is neither completely structured nor completely unstructured. Some information is highly structured, for example, the reference to material in line 12. Other information is less structured, like the information about the history of the manuscript in line 18. Inside this element some information is structured, like the reference to the archbishop

[3] A graph is a concept from mathematics that is structured as sets of ordered pairs. Each pair consists of two edges that are connected by arcs. This is reminiscent of a triple where subject and object are the edges that are connected by the predicate (arc). If one edge connects to many other edges, a whole network of knowledge can emerge from simple triple statements.

of Cologne, but other information lacks precision, like the reference to a Cistercian convent.

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Biblia Sacra</title>
    </titleStmt>
    <sourceDesc>
      <msDesc xml:id="kn28—0001" xml:lang="de">
        <physDesc>
          <objectDesc form="codex">
            <supportDesc material="perg">
              <support>
                <material>Parchment</material>
              </support>
            </supportDesc>
          </objectDesc>
        </physDesc>
        <history>
          <origin>Liber sancti petri a pio patre herimanno datus (<date>9/10c</date>, <
                locus> f. 1r</locus> — <persName>Herimann Abp. of Koeln <date>890—923</
                date></persName>);</origin> <provenance><persName> Rutgheri </persName>. (
                <date>9— or 10c?</date>, <locus>f. 1r</locus>); <quote>Hic liber est
                sancti petri in colonia concessus conventui de prato sancte marie per
                manum domini alberti subdecani, quem idem conventus reddet sine
                contradictione, cum <sic>repitittus</sic> fuerit a capitulo sancti petri,
                sicut continetur in litteris , quibus se predictus sanctimonialium
                conventus obligavit. Et in eo sunt multa folia truncata. <date>Anno MCCXLI
                 </date>.</quote> (<locus>f. 1r</locus>, notice dated <date>1241 </date>,
                that this book was lent by cathedral to the convent of the Prata S. Mariae
                , also called Benden; this Cistercian convent for women was founded <date>
                 1207 </date> in the area of Bruehl [about 10 km south of Cologne]; [...]<
                /provenance>
        </history>
      </msDesc>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```

Listing 1. Manuscript description as part of a TEI encoded document.

What can we do about semi-structured text? Highly structured data usually can be extracted very well. The tag “`<material>`” indicates that the contained value will denote a certain material. A triple like “kn28-0001”, “consists of”, “Pergament” can easily be constructed if the meaning of the attribute “`xml:id`” is known in this context. However, the tag “`<history>[...] this book was lent by cathedral to the convent of the Prata S. Mariae, [...]</history>`” will be harder to extract unless a well maintained list of cathedrals and convents supports the information extraction tool.

The result of the extraction process should be a triple like “kn28-0001”, “was lent to”, “Prata S. Mariae”.[4]

Structured queries that rely on the semantics of information are only possible if the data model is also highly structured. Therefore, the migration of information to the Semantic Web cannot be limited to adopting its concepts but needs to aim at making information explicit that was implicit before. Listing 2 shows how some of the TEI information has been encoded according to Semantic Web concepts. In this case the information on the manuscript has been saved as a text file encoded in Turtle.[5]

```
@prefix rdf: <http://www.w3.org/1999/02/22—rdf—syntax—ns#>.
@prefix crm: <http://erlangen—crm.org/100302/>.
@base <http://ceec.uni—koeln.de/>.

:kn28—0001
        rdf:type crm:E22_Man—Made_Object;
        crm:P1_is_identified_by [
                rdf:value "Koln, Dombibliothek, Codex 1."@de
                ].
        crm:P128_carries :kn28—0001doc;
        crm:P45_consists_of :parchment.

:kn28—0001doc
        rdf:type crm:31_Document;
        crm:P1_is_identified_by [
                rdf:value "Biblia Sacra"@de
                ];
        crm:P1_is_identified_by [
                rdf:value "Vulgata Bible"@en
                ].

:parchment
        rdf:type crm:E57_Material:
        crm:P1_is_identified_by [
                rdf:value "Pergament"@de
                ];
        crm:P1_is_identified_by [
                rdf:value "parchment"@en
                ].
```

Listing 2. Manuscript information modelled in Turtle.

[4] This triple is a shortcut of a more complex set of triples that include the actor who surrendered the custody and the actor that the custody was surrendered to. A special interest group has been formed to reasearch the relation of markup like TEI to ontologies (Eide and Ore).

[5] Turtle (Terse RDF Triple Language) is a serialization format for RDF. In this context, serialization means to dump triple data to a file for persistence or transmission. Turtle is a popular systax for RDF because it is more human-readable than XML. However, according to the recommendation, RDF should be serialized as RDF/XML (the XML syntax for expressing RDF).

Lines 1 to 3 define different namespaces that can be reused throughout the document to guarantee that keywords are unique although they have been collected from different information sources.[6] The rest of the document can be read as an aggregation of simple subject, predicate and object statements. The information in the figure already adheres to the CIDOC CRM that will be described in the following section. Lines 5 and 6 express that there is an entity "`:kn28-0001`" which is an instance of the class "`E22_Man-Made_Object`". Line 11 adds the information that "`:kn28-0001`" consists of parchment. The expression "`:kn28-0001`", of course, denotes the physical codex that is part of the collection of the "Diözesan- und Dombibliothek Köln."[7] This information is highly structured and additional semantic information has been made explicit, thus satisfying the precondition for complex query processing.

4. The Role of Ontologies

In the field of Semantic Web research the Resource Description Framework (RDF) has been proposed as an approach to conceptually model data of a certain domain. An example of how data can be encoded according to RDF has been presented in listing 2. However, RDF does not make any recommendations as to how a certain domain could be structured or which terminology should be used. Like a traditional relational database it does not make any statements about the meaning of data, and many of the semantics have to be modeled as part of the application logic of a computer program. Ontologies have been proposed as a much richer approach to model the semantics of information. The CIDOC CRM mentioned before has been explicitly modeled as an ontology and its inventors introduced it as an "ontological approach" (Dörr).

Information technology took the word "ontology" from philosophy in an analogy but redefined the term to fit its needs. In fact, ontologies have been considered as being a "silver-bullet" for information integration (Fensel). Basically, ontologies have been introduced to support communication processes in larger groups. They have been developed to help organisations find a common language and understanding of important domain concepts. In comparison to flat glossaries or terminology lists, ontologies comprise a complex thesaurus-like structure, additional rules and

[6] Because it would be cumbersome to write the full URI of each part of the triple, so-called namespaces can be defined. A namespace definition binds a part of the full URI to a qualified name that can be used throughout the document. The base namespace is applied to all names that omit the qualified name in front of the colon connecting the qualified name with its suffix. For example, "`:kn28-0001`" translates to "<http://ceec.uni-koeln.de/kn28-0001>" and "`rdf:value`" to "<http://www.w3.org/1999/02/22-rdf-syntax-ns#value>".

[7] The digital facsimiles of the "Diözesan- und Dombibliothek Köln" have been published as "Codices Electronici Ecclesiae Coloniensis" (Thaller and Finger). "kn28" denotes the identification code of the "Diözesan- und Dombibliothek".

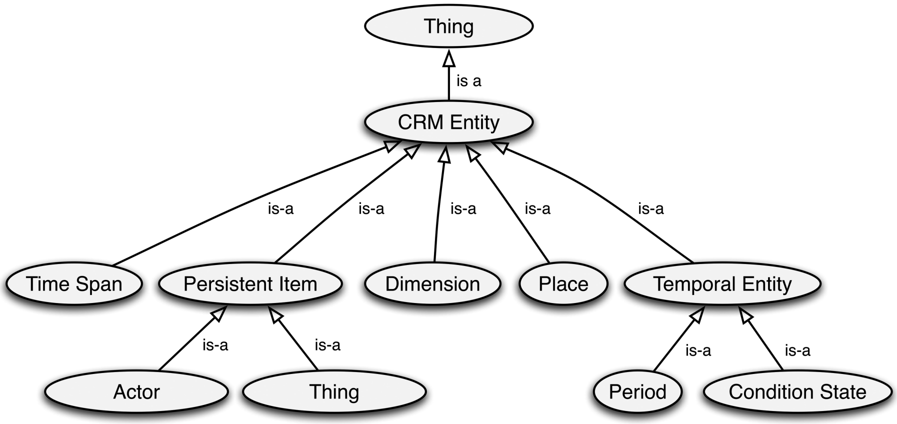

Figure 1. The Class hierarchy of the CIDOC CRM.

restrictions.[8] Thus, they not only define certain notions but can also encode complex interrelations. Although there is no canonical definition of what an ontology is in information technology, Gruber was the first to formalise the topic. Ontologies can be constructed with the Web Ontology Language (OWL).[9]

In 2006, the CIDOC Conceptual Reference Model was accepted as official standard ISO 21127:2006 (Crofts et al.). It provides a taxonomy for expressing information about material objects in the cultural-heritage area. Like any ontology, it can be used both to support communication processes in larger communities that strive for sharing of information and to implement software systems that integrate information from different information systems. A hierarchy of classes defines concepts that are commonly referred to in museum documentation practice. And so-called properties form relations between these conceptual classes. Up to now, the CRM has been used in several integration projects.[10]

Figure 1 shows a part of the class hierarchy provided by the CIDOC CRM. The visualization has been generated from an OWL implementation of the “Erlangen CRM”

[8] For example, a rule that states the uncle relationship in a fictional family ontology can have the form `[rule1: (?f pre:father ?a) (?u pre:brother ?f) -> (?u pre:uncle ?a)]` (the rule is written in the syntax of the Jena Semantic Web framework; the example is taken from <http://jena.sourceforge.net/inference/#rules>). Rules are evaluated and processed by inference engines to create new facts (triples). An example of a restriction is that a human always has, at most, two arms.

[9] More information about OWL can be found at Smith, Welty, and McGuiness.

[10] Two examples would be SCULPTEUR (Giorgini) and BRICKS.

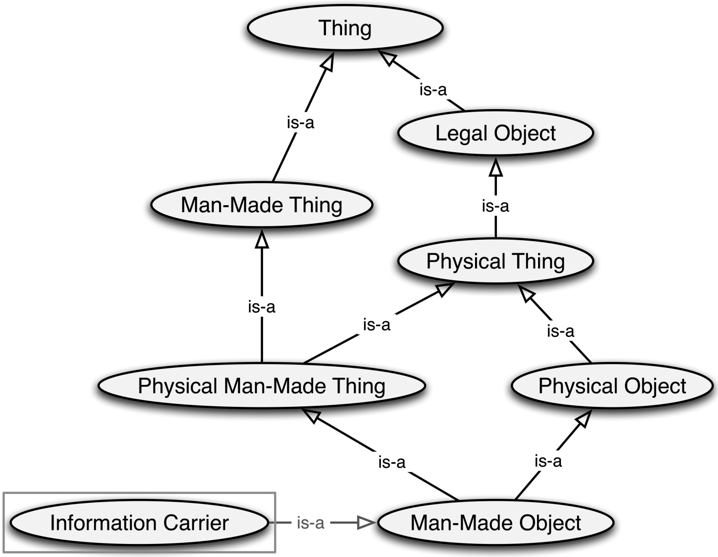

Figure 2. An information carrier in the CRM hierarchy.

(Schiemann et al.). For modeling data in the field of codicology, the CRM provides both classes for modeling a codex (for example, as a physical man-made object or information carrier) and the contained work (for example, as a document). But also classes for describing additional contextual information such as the condition of a codex, people involved in its creation and the history of ownership. The following section will introduce two relevant classes: "`E84_Information_Carrier`" and "`E31_Document`".

In order to describe a codex as a material thing, for example, one might use the CRM class "`E84.Information Carrier`". From the official CIDOC CRM documentation: "This class comprises all instances of E22 Man-Made Object that are explicitly designed to act as persistent physical carriers for instances of E73 Information Object. This allows a relationship to be asserted between an E19 Physical Object and its immaterial information contents" (Crofts et al. 67). Figure 2 shows the class as part of the inheritance hierarchy of the CRM. It is important to keep in mind that each class inherits all the features of its super class.

The contained textual material considered as a conceptual object can be modeled as "`E31.Document`". The official documentation defines that this class "comprises identifiable immaterial items, which make propositions about reality. These propositions may be expressed in text, graphics, images, audiograms, videograms or by other similar means"

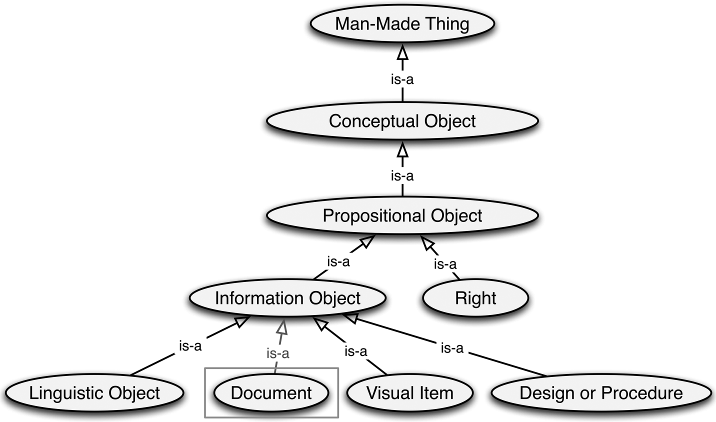

Figure 3. A document in the CRM hierarchy.

(Crofts et al. 48). Figure 3 shows how a document according to the CRM is also a man-made thing but without material character. However, it would be beyond the scope of this contribution to discuss if a document in this sense can describe the features of a certain hand writing or whether another class like “`E36.Visual_Item`” would be a better fit for this.

The structure of the CIDOC CRM relies heavily on the notion of events. Dörr and Kritsotaki argue that modeling events in metadata is helpful for dealing with cultural heritage information. For example, the notion of an event can be helpful for expressing uncertain information. The class “`E13.Attribute_Assignment`”, which is a sub-class of “`E7.Activity`”, has been provided to emphasize how a statement about something came about. Opinions of different authors can be distinguished by using “`E13.Attribute_Assignment`” for each researcher’s assertion about a codex. Additionally, the history of ownership of a codex can be modeled by using events. Although developed in a museum context, classes like “`E10.Transfer_of_Custody`” and “`E8.Acquisition`” (also both sub-classes of “`E7.Activity`”) suggest that the CRM provides structures that can be adapted to the needs of research in the field of codicology.

But how do codices relate to the contained works in the world of CIDOC CRM? The class hierarchies shown in figures 1, 2 and 3 do not display the properties mentioned above which are needed in order to relate instances of these classes to each other. Figure 4 highlights another perspective. Instead of the class hierarchy, the relations between

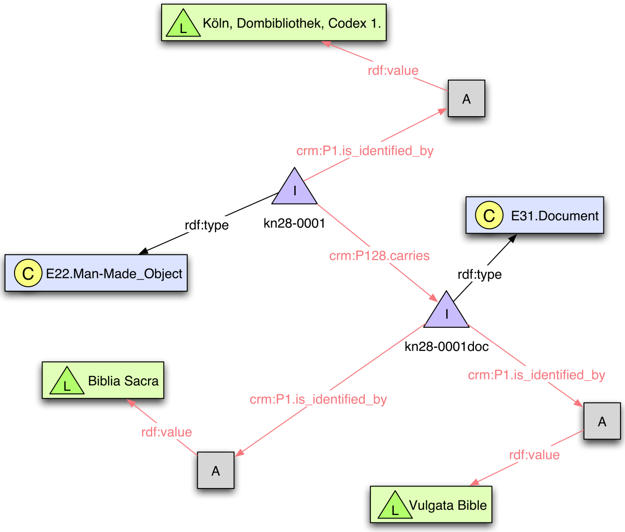

Figure 4. Manuscript information graph visualization.

the individuals are presented. Please note that this visualization has been automatically generated from the Turtle code in listing 2.[11]

Two further developments worth mentioning that deal with structured vocabularies and bibliographies are SKOS and FRBRoo. Up to now, the discussion has focused on integrating different data models and schemas. However, different groups tend to refer to the same thing by different names. Just think of an international research environment where codex materials are referred to using national languages (e.g. "Papier", "paper" and "páipear"). Lines 22 to 29 in listing 2 demonstrates how two different names have been assigned to the URI "`:parchment`". Here the URI denotes the material itself and the different names have been associated by using the CRM property "`crm:P1_is_identified_by`". One way to approach the terminology problem is to provide structured controlled vocabularies by using SKOS (the Simple Knowledge Organization System). It is a family of formal languages designed for any type of structured controlled vocabulary (Miles and Bechhofer). While CIDOC CRM is a formalisation of how cultural heritage content can be encoded, SKOS is a formalisation of how structured terminologies can be encoded. Figure 5 illustrates how appellations of different materials can be expressed according to SKOS. It shows that the material which

[11] RDF Gravity has been used to generate the visualization (Goyal and Westenthaler). For better readability the figure has been reworked by using a charting tool.

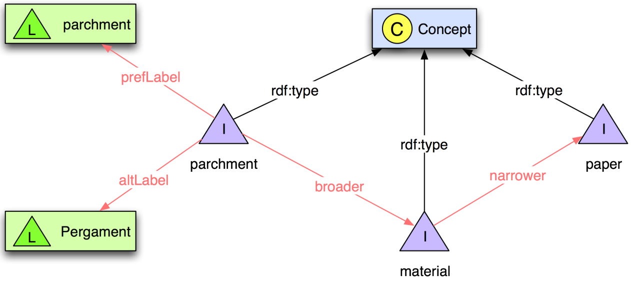

Figure 5. A graphical visualization of a vocabulary as SKOS.

the URI "`:parchment`" refers to has been associated with the SKOS class "`Concept`".[12] The hierarchical links "`:broader`" and "`:narrower`" indicate that "`:material`" is more general than "`:parchment`" and "`:paper`". Of course, SKOS data is not valuable in itself but needs to be made available for information systems so that they can make use of the structured vocabularies.

Another standard with a strong connection to Semantic Web research has been proposed in the field of library science. The Functional Requirements for Bibliographic Records (FRBR) form a conceptual model developed by the International Federation of Library Associations Institutions (Tillett). FRBR distinguishes the notions *Work*, *Expression*, *Manifestation* and *Item* (IFLA Study Group on FRBR). According to the definition document, a *Work* is an intellectual creation (for example "Moby Dick") and the *Expression* is the realization of this creation in its distinct form (e.g. German translation of "Moby Dick"). A *Manifestation* is "the physical embodiment of an expression of a work. As an entity, manifestation represents all the physical objects that bear the same characteristics, in respect to both intellectual content and physical form" (e.g. a certain edition of the German translation). Finally, an *Item* is "a single exemplar of a manifestation. The entity defined as item is a concrete entity" (a certain copy of a certain edition). A medieval codex would be defined as an *Item* in the terminology of FRBR. And since for each manifestation there is just one item, the distinction between *Manifestation* and *Item* does not seem to be pertinent to practical research in the field of codicology. Although FRBR is powerful at modeling relations between these four layers, it does not come with the means to express the history of development for an old manuscript.

[12] The SKOS reference document defines the class "`Concept`": "A SKOS concept can be viewed as an idea or notion; a unit of thought. However, what constitutes a unit of thought is subjective, and this definition is meant to be suggestive, rather than restrictive" (Bechhofer and Miles).

FRBR has been harmonised with the CIDOC CRM. Therefore, it has been expressed as a formal ontology that links to the classes of the CRM. The harmonisation project strives for better related representations of bibliographic and museum information, and to facilitate the "integration, mediation and interchange" of information (Dörr and Le Boef). In June 2009 the latest version of the standard was released (Aalberg et al.).

5. Putting it Together

We have looked at how different cultural heritage information systems can use information extraction (for unstructured and semi-structured texts) and mapping (for databases or the "deep web") to get a grip on relevant entities. We have discussed how these entities and their relations can be modeled as RDF triples in a way that conforms to a standard like the CIDOC CRM. Adhering to this standard enables software to process data according to the intended meaning. It is helpful for further data fusion tasks and complex querying of information to integrate the information acquired from different sources in one physical place. This enables comprehensive mining, indexing and querying.

Different architectures have been developed to tie together complex distributed systems ranging from distributed databases over middleware software to Service-Oriented Architectures with Web Services.[13] In the cultural heritage domain it has become very common to publish information using the OAI Protocol for Metadata Harvesting (OAI-PMH), which relies on the HTTP protocol. The OAI-PMH has been suggested by the Open Archives Initiative for publishing and collecting metadata. Data providers, such as archives, publish their metadata as XML and service providers harvest that data in order to offer further services (Lagoze et al.). Using this protocol it would be possible to publish both the TEI documents and additional semantic data in RDF. Recent suggestions in Semantic Web research tend to avoid cumbersome approaches in favor of light-weight infrastructures. In 2006, Berners-Lee articulated his thoughts on a concept that he called "Linked Data". The core of this concept is not only to put data on the web in a certain manner but also to link related data. Since then, the notion of Linked Data refers to a set of best practices for publishing and connecting structured data on the World Wide Web. While the concept of Linked Data requires service providers to systematically crawl linked data to acquire information, OAI-PMH offers guided and streamed pulling of data.

[13] A Service-Oriented Architecture provides loosely coupled software components that provide services. These services can be combined to solve certain tasks. A Web Service provides an application programming interface that can be called via the HTTP protocol. HTTP (Hypertext Transfer Protocol) is a standard that is used to transmit data over a network (in particular for the Internet). Therefore it is ubiquitously available.

Once the data has been integrated, tools are needed for further processing. Semantic Web Frameworks like Jena (Carroll et al.) support software developers in creating Semantic Web applications. They provide an integrated set of tools that facilitate the design and operation of a knowledge base that can deal with triples. Most Semantic Web frameworks provide a so-called SPARQL endpoint to query the knowledge base. SPARQL is an acronym that stands for “SPARQL Protocol and RDF Query Language” (Prud’hommeaux and Seaborne). An endpoint is an entry point for a service that can be called over a network. SPARQL allows the formulation of queries that are highly structured. It does have some similarities to SQL.[14]. But while SQL is commonly used to query or manipulate data in a relational model, SPARQL can be used to formulate complex graph-like query structures to query RDF data.[15] SPARQL queries can be transmitted over the HTTP protocol. Larger projects are picking up the idea of the Semantic Web and developing advanced applications. Information extraction projects like DBpedia mine the World Wide Web for structured data and make it accessible for the Semantic Web (Auer et al.). Another example is the SIMILE project (Mazzocchi, Garland, and Lee). It is more end-user-oriented than DBpedia and develops tools that examine the possibility of semantically processing digital assets.

6. User scenario

The Semantic Web has been introduced and codicological material referenced, only the coherent user scenario remains to be described. Therefore, a simple use case for further elaboration will be presented here. It will develop around a simple question and highlight the implications of how Semantic Web technology will be affected. In the above-mentioned vision piece a user scenario has been developed that should motivate future research and funding in the area, but scenarios can also be used in the microcosm of system development. They help to reflect on functional requirements that a single piece of software needs to fulfill in order to meet a user’s needs. User scenarios facilitate communication between software designers, programmers and end-users by providing a shared example.

Imagine a researcher working on medieval codices. To push forward his current research project, he is interested in how texts spread in certain institutions in a specific region.[16] He has a good friend in the IT department of the university who enthusiastically reported on a new strain of research, the Semantic Web. According to this concept, digital information will be managed in a way that supports machine-

[14] SQL is ISO standard ISO/IEC 9075:2008 and stands for Structured Query Language.

[15] Up to now there is no W3C recommendation for the manipulation capabilities of SPARQL. However, most Semantic Web toolkits provide capabilities for data manipulation outside SPARQL and extensions for SPARQL are being developed.

[16] I want to thank Almut Breitenbach and Patrick Sahle for their support in creating this user scenario.

processing. The researcher wonders if this new technology could meet a requirement he has formulated as follows: "For the geographic area of northern Germany, show all codices that contain texts of Classic Latin authors and that have been written in the 13th century. Draw the results as circles on a map and use different colors for monasteries and nunneries." Many requirements need to be fulfilled to enable a system to process such a question.

Certainly, the data that is needed to compile the results resides on scattered information systems, preferably encoded as structured manuscript descriptions. As a first step, this data needs to flow from one information system to another. Catalogue data from different information systems has been published and is exposed via OAI-PMH as TEI. Imagine an information system that strives to support the researcher. It will request information from several data providers and gather it for further processing. This approach requires little effort for data providers. Other architectures could demand that one or more of the following pre-processing steps be performed by data providers before the data is published. As a first step, information extraction needs to be performed on the acquired data by the central information system. The system aims to extract named entities and to assign the right unique identifier (i.e. URI) to each entity. This step is rather important because without canonical names across the participating information systems all following steps will fail.

Once entities and the relations that exist between them are represented as URIs, they can be stored as triples in some serialisation of RDF. To be available for processing, triples are held in main memory according to a suitable data structure. For exchanging information between different information systems, this data needs to be serialised in a file. An example for a serialisation has been given in listing 2. After ingesting the triples in a triplestore, the data will be available for further processing.

The extracted entities alone are of very limited use unless they are aligned with additional background knowledge. This knowledge will be provided by specialised knowledge bases as triples. It comprises, for example, the geographic region, the monastery and religious order mentioned in the manuscript description. Without this background, none of which is contained in the metadata of the codex alone, the query of the researcher cannot be answered. But after adding the supplementary knowledge to the triplestore, additional facts are available that can be considered for query processing. For example, the three triples "codex123", "carries", "document123", "document 123", "has author", "Cicero" (both extracted from codex information) and "Cicero", "has genre", "Classical Latin work" (added from background knowledge base) can be combined to reason that the codex contains a text of an author that has been attributed "Classical Latin work". Additional facts can be derived by applying rules. And plausibility checks can be conducted to disclose contradictory information that may emerge by considering additional knowledge.

We assume that the information about the author of a text and its place of creation could be extracted from the codex metadata. Additionally, a group of theoretical researchers recorded their findings by putting results in a specialized information system (for example that texts of a certain author usually can be ascribed to a specific genre). Another system contributes the geographic coordinates for a certain geographic region. If the information system that the researcher is using has access to all the above systems, they can now formulate queries that could not have been formulated before. Listing 3 shows a selected aspect of the aforementioned query in a formalised way. Its formalisation is little more than preliminary but seems to be sufficient to discuss the process of formalisation.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX crm: <http://erlangen-crm.org/100302/>
PREFIX cod: <http://codicology.org/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
BASE <http://ceec.uni-koeln.de/>

SELECT ?codex, ?gender, ?geo
WHERE {
   ?codex rdf:type crm:E22_Man-Made_Object .
   ?codex crm:P108I.was_produced_by ?codexProduction .
   ?codexProduction crm:P7.took_place_at ?monastery .
   ?monastery crm:P2.has_type ?gender .
   ?gender skos:broader cod:gender .
   [...]
}
```

Listing 3. A SPARQL query that formalizes a research problem.

As in listing 2, the query starts by defining a couple of namespaces including the base namespace to ensure that the used names are unique. The rest of the query is based on the syntax of SPARQL that serves to query triplestores. Then, the three variables “`?codex`”, “`?gender`’ and “`?geo`” are defined to carry the results. The result will hold identifiers of codices together with the information about the geographic unit the monastery belongs to and whether it is a nunnery or monastery. The following part of the query re-uses these variables and defines additional ones that are only used temporarily. The temporary variables like “`?monastery`” are needed to build the “factual bridges” from one piece of information to another. The property “`crm:P2.has_type`” connects the “`?monastery`” with its “`?gender`”. Multiple types can be connected with a “`crm:E22.Man-Made_Object`” and therefore the property “`skos:broader`” has been used to restrict the value of the assigned type to be a specialisation of “`cod:gender`”, which could be either male or female.

For real world use a graphical user interface needs to be implemented. The query demonstrated in listing 3 could not be formulated by a researcher in the field of codicology or palaeography without undertaking significant additional training. On the foundation of the mentioned Semantic Web framework, additional software layers need to mediate between the end-user and the bits and bytes. Many questions arise when thinking about such a system like questions of helpful interface design for easy interaction and formulation of very complex queries. Research in artificial intelligence strives for processing queries in natural language. The Companions Project, for example, explores software that is reminiscent of the intelligent agents mentioned before (Benyon and Mival).[17]

The information that has been acquired by evaluating the example query has the form of a table with the columns holding an identifier for the codex, the information if it is a monastery or nunnery and its geographic region. Although this information is helpful for the researcher it would be useful to display the results as a map. With the advent of Web 2.0, mashups have become quite popular.[18] The fictional researcher could use a similar service to display a map of the region of his interest. The web application would draw a circle for each monastery on the map with different colors for monasteries and nunneries. By querying a geographic database, the geographic identifier can be resolved to coordinates that are needed for the drawing task.

It is obvious that only a few of the infrastructural elements which would facilitate the user scenario are available today. Scientists in the field of codicology would need to make their factual knowledge available as a domain-specific knowledge base. Additionally, dealing with consistent and canonical URIs is anything but easy. Information extraction usually can only extract entities that it can look up in some kind of authority file or that it has been trained to find by machine learning techniques. Other entities can be identified but not resolved to a canonical name, especially in the case of unstructured text. Another problem is the scalability of current Semantic Web Triplestores. Unlike relational databases they are still not well understood and cannot deal with massive amounts of data. However, the Semantic Web community has recognised this problem and is working on scalable solutions. One example is the OWLIM Semantic Repository (Ontotext AD) that scales to several billion triples. The W3C maintains a Wiki that lists Triplestores, sorted by their scalability.

[17] The aforementioned SIMILE project is experimenting with different user interfaces that provide facetted browsing, timelines and maps. Another example would be PhiloSpace, a piece of software that has been developed within the COST framework. It can be used to establish semantic relations among entities of the philosophical domain.

[18] Among other things, the concept of Web 2.0 means that users can interact with the web site and actively contribute to it. Mashups are one example for a web page that combines data and functionality from other resources to create a custom service.

7. Concluding Remarks

This contribution aims to introduce key concepts and tools of Semantic Web research. It does not claim to articulate future directions of research but tries to provide criteria and background information for researchers which may help with their decision processes. Semantic Web technology certainly offers new perspectives on how data will be published, shared and processed in the future. However, the concepts of Semantic Web research have also been criticised. The idea that was formulated in 2001 has not yet been realised (Shadbolt, Berners-Lee, and Hall). Consequently, doubts remain about the practical feasibility of the concept because it is resource consuming to create knowledge bases and to add comprehensive structure to data. With billions of facts that have been published as triples on-line there will be scalability issues. Projects like the Large Knowledge Collider are exploring reasoning with incomplete knowledge due to limited resources (Fensel). Large ontologies tend to be cumbersome and difficult to understand, RDF with its explicit mode of expression becomes verbose and bulky. It has also been argued that the Semantic Web is not semantic. Gärdenfors for example doubts the practical feasibility of the Semantic Web because of its focus on formal syllogisms that stem from formal logic and research in artificial intelligence. These cover only a (dispensable) fraction of semantic operations that scientists want to have performed on their data. And although URIs are proposed as a unique way of identifying things, no data provider is forced to use canonical URIs. The database community can look back on a long research tradition in information integration that could (and already does) contribute valuable input (Leser and Naumann).

However, the vision of the Semantic Web has promoted a plethora of research projects in different domains (some of them mentioned in this contribution). Because of the data model that can represent data with rich and varying structure, it seems to be well suited for the humanities. Since RDF relies on the notion of a graph as its data model, it facilitates the construction of semantic networks of huge complexity and high flexibility. Cultural heritage information models often “suffer” from relying on inflexible structures that do not explicitly model the intended meaning of information objects. Again, this could limit the opportunities for helpful applications. Additionally, RDF handles missing data very well, the concept relies on the “open world assumption”.[19]

Thus, the Semantic Web seems to be both a blessing and a curse for information integration and processing in cultural heritage. It envisions new and interesting approaches that could be very useful for humanities information science. But research projects cannot just adopt the concepts and hope for the best. These projects should be

[19] The “open world assumption” is used in knowledge representation because nobody can comprehensively model the knowledge of a certain domain. By that, one has to conclude that a software system needs to deal with incomplete knowledge and that the kinds of inference which a piece of software can perform are limited to those statements which are available.

prepared to actively engage in Semantic Web research and adequate resources should be allocated (fortunately, a very lively field at the moment). Its data model seems to be well suited to encode codicological data but its mechanisms for manipulating that knowledge seem to be restricted to formal syllogisms. Provided that the means to deal with uncertain and contradictory information are developed, the Semantic Web could foster research in areas that heavily rely on qualitative data, such as codicology. So far, many projects in the field of "humanities information science" focus on encoding information to make it available to a greater audience for searching and browsing. Manipulation of data as the primary method to generate significant insights seems to be restricted to problems that are clearly quantifiable or that can be dealt with by statistical analysis. Certainly, the Semantic Web will not provide for all the information needs of a researcher, but it could begin to play out its strength in well defined and carefully bounded applications.

Bibliography

Aalberg, Trond et al. *FRBR - Object-Oriented Definition and Mapping to FRBRer.* International Working Group on FRBR and CIDOC CRM Harmonisation, 1.0 ed., 2009.

Alexander, Ian and Neil Maiden. *Scenarios, Stories, Use Cases: Through the Systems Development Life-Cycle.* John Wiley & Sons, 2004.

Auer, Sören. et al. "DBpedia: A Nucleus for a Web of Open Data." *Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007).* eds. Karl Aberer et al., vol. 4825, chap. 52. Berlin, Heidelberg: Springer, 2007. 722–735.

Benyon, David and Oli Mival. "Introducing the Companions Project: Intelligent, Persistent, Personalised Interfaces to the internet." *BCS-HCI '07: Proceedings of the 21st British HCI Group Annual Conference on HCI 2008.* Swinton, UK: British Computer Society, 2007, 193–194.

Berners-Lee, Tim. "Linked Data - Design Issues." World Wide Web Consortium 2006-2009. <http://www.w3.org/DesignIssues/LinkedData.html>.

Berners-Lee, Tim, James Hendler, and Ora Lassila. "The Semantic Web." *Scientific American* 284 (2001).5: 34–43.

BRICKS. "BRICKS Project. Building resources for Integrated Cultural Knowledge Services." Bricks Project 2004-2007. <http://www.brickscommunity.org>.

Cardie, Claire. "Empirical Methods in Information Extraction." *AI Magazine* 18 (1997).4: 65–80.

Carroll, Jeremy J. et al. "Jena: Implementing the Semantic Web Recommendations." *WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters.* New York, NY, USA: ACM, 2004, 74–83.

COST. "Working Group 2: Software". COST Action 32 2009. <http://www.cost-a32.eu/wg-2.html>.

Crofts, Nick. et al. "Definition of the CIDOC Conceptual Reference Model." ICOM/CIDOC 2003-2005. <http://www.cidoc-crm.org/docs/cidoc_crm_version_4.2.pdf>.

Delaissé, Leon M. J., James H. Marrow, and John de Wit. "Illuminated Manuscripts: The James A. de Rothschild Collection at Waddesdon Monor." Fribourg: Office du Livre [et al.], 1977.

Dörr, Martin and Athina Kritsotaki. "Documenting Events in Metadata." *The 7th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST (2006).* eds. M. Ioannides et al. 2008.

Dörr, Martin. "The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata." *AI Mag* 24 (2003).3: 75–92.

Dörr, Martin and Patrick Le Boeuf. "FRBRoo Introduction." ICOM/CIDOC 2009.
<http://www.cidoc-crm.org/frbr_inro.html>.

Eide, Øyvind and Christian-Emil Ore. "SIG:Ontologies." Text Encoding Initiative 2004-2010.
<http://wiki.tei-c.org/index.php/SIG:Ontologies>.

Fensel, Dieter. "Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce." Springer-Verlag New York, Inc., 2003.

Fensel, Dieter et al. "Towards LarKC: A Platform for Web-Scale Reasoning." *ICSC.* 2008, 524–529.

Gärdenfors, Peter. "How to Make the Semantic Web More Semantic." *Formal Ontology in Information Systems: proceedings of the third international conference (FOIS-2004).* eds. Achille C. Varzi and Laure Vieu, vol. 114 of *Frontiers in Artificial Intelligence and Applications.* IOS Press, 2004, 17–34.

Giorgini, Fabrizio. "SCULPTEUR (Semantic and content-based multimedia exploitation for European benefit)." Sculpteur 2002-2005. <http://www.sculpteurweb.org>.

Goyal, Sunil and Rupert Westenthaler. "RDF Gravity (RDF Graph Visualization Tool)." Salzburg: Salzburg Research Forschungsgesellschaft 2004.
<http://semweb.salzburgresearch.at/apps/rdf-gravity>.

Gruber, Thomas R. "A Translation Approach to Portable Ontology Specifications." *Knowl. Acquis.* 5 (1993).2: 199–220.

Hitzler, Pascal, Markus Krötzsch, and Sebastian Rudolph. *Foundations of Semantic Web Technologies.* London: Chapman & Hall/CRC, 2009.

IFLA Study Group on the Functional Requirements for Bibliographic Records. *Functional Requirements for Bibliographic Records - Final Report.* UBCIM Publications - New Series Vol 19. K . G. Saur München, 2008.

Konchady, Manu. *Text Mining Application Programming.* Boston, Mass.: Charles River Media, 2006.

Lagoze, Carl et al. "Open Archives Initiative Protocol for Metadata Harvesting." Open Archives Initiative 2002-2008. <http://www.openarchives.org/OAI/openarchivesprotocol.html>.

Leser, Ulf and Felix Naumann. *Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen.* Heidelberg: dpunkt-Verl., 2007, 1. ed.

Maniaci, Marilena. *Archeologia del manoscritto: metodi, problemi, bibliografia recente.* I libri di Viella. Roma: Viella, 2002, 1. ed.

Manola, Frank and Eric Miller. "RDF Primer." World Wide Web Consortium 2004.
<http://www.w3.org/TR/rdf-primer>.

Mazzocchi, Stefano, Stephen Garland, and Ryan Lee. "SIMILE: Practical Metadata for the Semantic Web." O'Reilly 2005. <http://www.xml.com/pub/a/2005/01/26/simile.html>.

Miles, Alistair and Sean Bechhofer. “SKOS Simple Knowledge Organization System Reference.” World Wide Web Consortium 2009. <http://www.w3.org/TR/skos-reference>.

Norvig, Peter and Stuart Russell. *Artificial Intelligence: A Modern Approach.* Prentice Hall International, 2003, 2. ed.

Ontotext AD. “OWLIM Semantic Repository.” Ontotext 2010. <http://www.ontotext.com/owlim>.

Prud’hommeaux, Eric and Andy Seaborne. “SPARQL Query Language for RDF.” World Wide Web Consortium 2008. <http://www.w3.org/TR/rdf-sparql-query>.

Schiemann, Bernhard et al. “Short Documentation of the CIDOC CRM (4.2.4) Implementation in OWL-DL.” Erlangen: Friedrich-Alexander-Universität Erlangen 2008. <http://erlangen-crm.org/docs/documentation_crm_owl-dl_4.2.4.pdf>.

Shadbolt, Nigel, Tim Berners-Lee, and Wendy Hall “The Semantic Web Revisited.”

Intelligent Systems, IEEE [see also IEEE Intelligent Systems and Their Applications] 21 (2006).3: 96–101.

Smith, Michael K., Chris Welty and Deborah L. McGuinness. “OWL Web Ontology Language Guide.” World Wide Web Consortium 2004. <http://www.w3.org/TR/owl-guide>.

TEI Consortium. “Text Encoding Initiative.” Text Encoding Initiative 2010. <http://www.tei-c.org/index.xml>.

Thaller, Manfred and Heinz Finger. “Codices Electronici Ecclesiae Coloniensis (CEEC).” Köln, Universität zu Köln 2002. <http://www.ceec.uni-koeln.de>.

Tillett, Barbara B. “What is FRBR?: A Conceptual Model for the Bibliographic Universe.” Washington (DC): Library of Congress 2004. <http://www.loc.gov/cds/downloads/FRBR.PDF>.

W3C. “LargeTripleStores - ESW Wiki.” World Wide Web Consortium 2010. <http://esw.w3.org/LargeTripleStores>.