Archaeology and the Semantic Web—Prospects and Challenges

Robert Kummer[1]

[1]Forschungsarchiv für Antike Plastik am Archäologischen Institut,
Universität zu Köln. Germany.

Abstract

Historians do not want to find database records; they want to understand historical contexts. Consequently, they need to draw sources on a specific subject from the repositories of different cultural heritage institutions to get their work done. For that, interoperability between different and heterogeneous databases needs to be established. This paper deals with experiences and challenges that have been encountered in the course of ongoing data integration efforts. The collaborating parties are the Perseus Project at Tufts University and Arachne, the central object database of the German Archaeological Institute. Together they gauge Semantic Web concepts by integrating data from multiple databases with the aim of establishing interoperable environments for archaeological research. The main challenges experienced during the first project phase were manifold and will be addressed and discussed.

Keywords: *CIDOC CRM, interoperability, semantic web, co-reference resolution*

1 Overview

The paper will report on work that began in 2006 as a still ongoing collaboration of Arachne,[1] the central object database of the German Archaeological Institute[2] and the Perseus Digital Library[3] at Tufts University, Boston. It is currently being advanced by collaboration with the project CLAROS[4] hosted at the Beazley Archive in Oxford. After reflecting on what interoperability could mean in an environment dealing with data from archaeology and classics, the paper will discuss the components that are needed to establish interoperability. A basic workflow has been set up that starts with extracting data from heterogeneous databases and that ends with making this data accessible on-line. While implementing this workflow, developments in the area of Semantic Web research have been exploited and adapted if they turned out to be useful for the project. Applying concepts of Semantic Web research turned out to be both a blessing and a curse since some developments are still work in progress. A lot of effort went into integration of database schemata aiming at establishing not only syntactic but also semantic interoperability on a global scale. Many cases of Semantic Web research are based on referring to bits of information by Uniform Resource Identifiers (URIs). Since databases use different terminology and different national languages as well, it is extremely difficult to generate common global URIs from available data. Consequently, large projects work on reducing the number of identifiers that exist.[5] Future research should focus on applying techniques to Semantic Web data that stem from data and text mining as well as logical and statistical reasoning. The smaller and more focused world of archaeological research could provide a proper environment to study the related problem of semi-automatic co-reference resolution in cultural heritage.

2 Interoperability

When Douglas C. Engelbart defined his position on "augmenting human intellect," he was talking about "increasing the capability of a man to approach a complex situation, to gain comprehension to suit his particular need, and to derive solutions to problems.[6] Solutions to problems usually require complex and complementary mental operations comprising understanding, logical induction and deduction but also creativity and intuition. Scientists need the information at hand that is crucial to address problems and to discover and formulate new ones. Software systems should be designed in ways that support these cognitive and social operations to support historians and archaeologists in addressing their problems and finding

[1] Reinhard Förtsch, "ARACHNE—Datenbank und kulturelle Archive des Forschungsarchivs für Antike Plastik Köln und des Deutschen Archäologischen Instituts," 2007, http://arachne.uni-koeln.de/drupal/node/3.

[2] DAI, "Deutsches Archäologisches Institut," August 2007, www.dainst.org.

[3] Gregory Crane, "Perseus Digital Library," www.perseus.tufts.edu/hopper/.

[4] Beazley Archive, "CLAROS—Classical Art Research Centre Online Service—The University of Oxford," www.clarosnet.org/index.htm.

[5] "Welcome to OKKAM—Enabling the Web of Entities. The OKKAM Project Workspace," www.okkam.org/.

[6] Douglas Engelbart, "Augmenting Human Intellect: A Conceptual Framework," Air Force Office of Scientific Research 3233, October 1962.

approaches to solve them. One step towards this objective is to establish environments that facilitate exchange and sharing of information among multiple cultural heritage information systems.

But what makes the whole become more than the sum of its parts? And what does interoperability mean in the context of cultural heritage in general and archaeological research in particular? A piece of data becomes information if it is being associated with contextual information. The more information is added and linked, the more powerful a research resource will become. One initiative that strives for tight semantic linking of information on the web is called "Linked Data". In 2006, Tim Berners-Lee presented his thoughts on this concept: "The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data."[1] Semantic Web? Related? In which sense? According to the above quotation, linked data is about connecting related information in new and sometimes unanticipated ways.

We believe in the need for putting archaeological data on the web in a manner that allows for tight semantic linking to clear the way for new ways of archaeological research. Many initiatives striving to establish interoperability among cultural heritage information systems are pursuing related research agendas. Since data is controlled by applications, information remains spread all over the world in a fragmented manner. Therefore, some scientific questions cannot be addressed today. The following passages will reflect on how interoperability could be established among the aforementioned project partners and report on respective implementation efforts that have been undertaken.

3 EXAMPLE SCENARIOS

Imagine a library where the books can talk to each other. This is an idea that has been attributed to Marvin Minsky, MIT professor and artificial intelligence researcher. The phrase "books talking to each other" could be interpreted as a metaphor for describing how information that resides in different documents can become more useful if processed in some synergetic way. Of course, the idea abstracts the human being that needs to interpret and make use of the added value stemming from this "discourse". But how can this metaphor be transferred to current cultural heritage information systems? The following two examples should demonstrate different ways of establishing interoperability of such systems to make books talk to each other and to objects. Figure 1 shows a screenshot of the Perseus reading environment displaying Caesar's *Gallic War* in the middle column. The right column shows places that are referenced within this visible passage of text. Associated data about places has been drawn from a different document, an authority file. More significantly, references to other works like commentaries and grammars provide additional related information below that. To make this possible, information stemming from different texts has been encoded and linked in a way that can be processed by machines. These links also enable user interfaces to be built, where users can navigate repositories of related information, resulting in a richer reading experience.

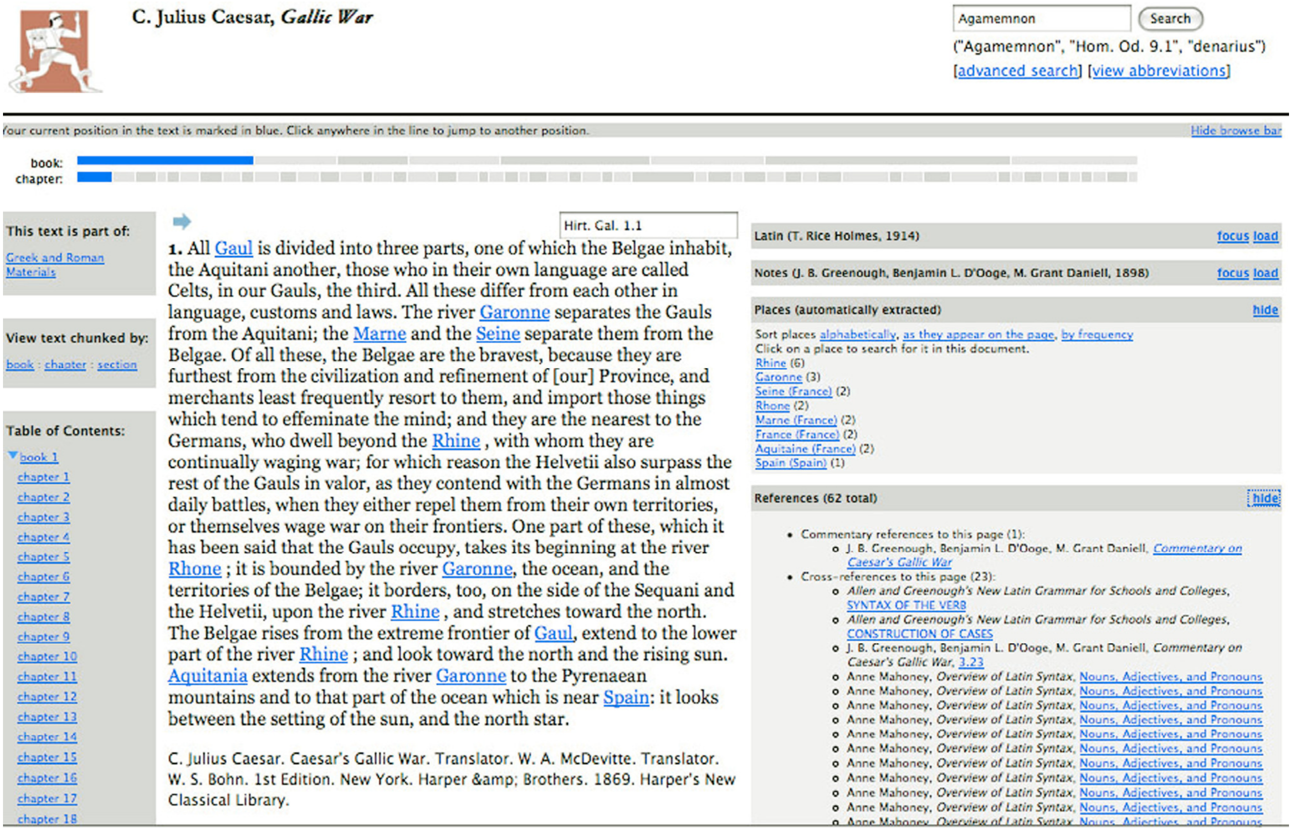
Recently, Arachne introduced a new way to navigate its content that goes beyond traditional browsing of databases. While looking into the archaeological information provided by Arachne, a user might end up seeing one particular archaeological object on his screen. Then, the user can switch to a tool that is called the "context browser". It visualizes links to additional information being related to the object that is currently on the screen. Context is a very fundamental concept in archaeology. Not only does it include the find spot but also further environmental circumstances accompanying the find. That said, Arachne does provide information about the context of a find and defines additional contexts like affiliation to a specific collection or to a historical monument. This tool is meant to help users to recreate different contexts that a material object is embedded in. Figure 2 shows how a node that represents a database record is situated in the middle of two concentric circles. While the inner circle shows database records stemming from different contexts that are immediate neighbors, the outer circle shows indirect neighbors. As in the Perseus example, this visualization has been made possible by explicitly linking objects that are related. By using this tool historians can navigate a set of data that comes with rich structure differently from traditional flat browsing paradigms.

Both systems, Arachne and Perseus, established a certain level of interoperability within the boundaries of their software architectures. Considering the arguments that have been mentioned before, the level of interoperability should be expanded beyond the borders of each system. From a scientific perspective, although Perseus is mainly text-based, its textual material is indispensable for archaeological research and therefore both collections should be linked. Someone doing research on the life of the emperor Augustus, for example, needs the means to bring together information about this specific person from both sources.

4 CHALLENGES

In the course of implementing different components that should form an environment to support interoperability between Perseus and Arachne, we ran into three main obstacles. First, databases are usually crafted according to a specific research problem and therefore differ in their structure. Second, each cultural heritage information system uses a different set of terminology in their respective national language.

[1] Tim Berners-Lee, "Linked Data—Design Issues," www.w3. org/DesignIssues/LinkedData.html.



C. Julius Caesar, *Gallic War*

Agamemnon Search

("Agamemnon", "Hom. Od. 9.1", "denarius")

[advanced search] [view abbreviations]

'our current position in the text is marked in blue. Click anywhere in the line to jump to another position. Hide browse bar

book:
chapter:

This text is part of:

Greek and Roman Materials

View text chunked by:

book : chapter : section

Table of Contents:

- book 1
  - chapter 1
  - chapter 2
  - chapter 3
  - chapter 4
  - chapter 5
  - chapter 6
  - chapter 7
  - chapter 8
  - chapter 9
  - chapter 10
  - chapter 11
  - chapter 12
  - chapter 13
  - chapter 14
  - chapter 15
  - chapter 16
  - chapter 17
  - chapter 18

Hirt. Gal. 1.1

1. All Gaul is divided into three parts, one of which the Belgae inhabit, the Aquitani another, those who in their own language are called Celts, in our Gauls, the third. All these differ from each other in language, customs and laws. The river Garonne separates the Gauls from the Aquitani; the Marne and the Seine separate them from the Belgae. Of all these, the Belgae are the bravest, because they are furthest from the civilization and refinement of [our] Province, and merchants least frequently resort to them, and import those things which tend to effeminate the mind; and they are the nearest to the Germans, who dwell beyond the Rhine , with whom they are continually waging war; for which reason the Helvetii also surpass the rest of the Gauls in valor, as they contend with the Germans in almost daily battles, when they either repel them from their own territories, or themselves wage war on their frontiers. One part of these, which it has been said that the Gauls occupy, takes its beginning at the river Rhone ; it is bounded by the river Garonne, the ocean, and the territories of the Belgae; it borders, too, on the side of the Sequani and the Helvetii, upon the river Rhine , and stretches toward the north. The Belgae rises from the extreme frontier of Gaul, extend to the lower part of the river Rhine ; and look toward the north and the rising sun. Aquitania extends from the river Garonne to the Pyrenaean mountains and to that part of the ocean which is near Spain: it looks between the setting of the sun, and the north star.

C. Julius Caesar. Caesar's Gallic War. Translator. W. A. McDevitte. Translator. W. S. Bohn. 1st Edition. New York. Harper &amp; Brothers. 1869. Harper's New Classical Library.

Latin (T. Rice Holmes, 1914) focus load

Notes (J. B. Greenough, Benjamin L. D'Ooge, M. Grant Daniell, 1898) focus load

Places (automatically extracted) hide

Sort places alphabetically, as they appear on the page, by frequency
Click on a place to search for it in this document.
Rhine (6)
Garonne (3)
Seine (France) (2)
Rhone (2)
Marne (France) (2)
France (France) (2)
Aquitaine (France) (2)
Spain (Spain) (1)

References (62 total) hide

- Commentary references to this page (1):
  - J. B. Greenough, Benjamin L. D'Ooge, M. Grant Daniell, *Commentary on Caesar's Gallic War*
- Cross-references to this page (23):
  - *Allen and Greenough's New Latin Grammar for Schools and Colleges*, SYNTAX OF THE VERB
  - *Allen and Greenough's New Latin Grammar for Schools and Colleges*, CONSTRUCTION OF CASES
  - J. B. Greenough, Benjamin L. D'Ooge, M. Grant Daniell, *Commentary on Caesar's Gallic War*, 3.23
  - Anne Mahoney, *Overview of Latin Syntax*, Nouns, Adjectives, and Pronouns
  - Anne Mahoney, *Overview of Latin Syntax*, Nouns, Adjectives, and Pronouns
  - Anne Mahoney, *Overview of Latin Syntax*, Nouns, Adjectives, and Pronouns
  - Anne Mahoney, *Overview of Latin Syntax*, Nouns, Adjectives, and Pronouns
  - Anne Mahoney, *Overview of Latin Syntax*, Nouns, Adjectives, and Pronouns
  - Anne Mahoney, *Overview of Latin Syntax*, Nouns, Adjectives, and Pronouns
  - Anne Mahoney, *Overview of Latin Syntax*, Nouns, Adjectives, and Pronouns
  - Anne Mahoney, *Overview of Latin Syntax*, Nouns, Adjectives, and Pronouns
  - Anne Mahoney, *Overview of Latin Syntax*, Nouns, Adjectives, and Pronouns
  - Anne Mahoney, *Overview of Latin Syntax*, Nouns, Adjectives, and Pronouns
  - Anne Mahoney, *Overview of Latin Syntax*, Nouns, Adjectives, and Pronouns

Figure 1. *The Perseus reading environment.*

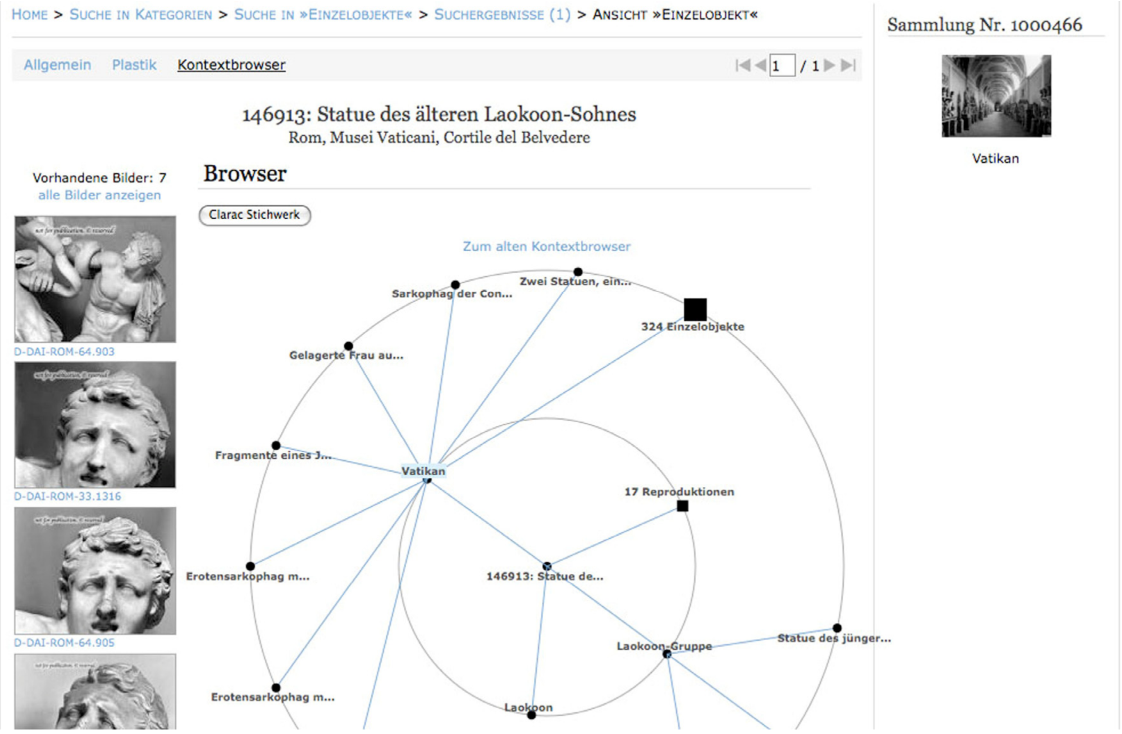


Figure 2. *The Arachne context browser.*

Third, gathering data from many different data providers in a way that supports the vision of the Semantic Web leads to scalability issues that need to be addressed. The CIDOC CRM that will be introduced in more detail later deals with how data of a specific domain is structured.[1] Although it provides basic facilities to deal with controlled vocabularies, it explicitly does not deal with database content. By now it can be said that the necessary infrastructure to overcome the mentioned problems is not in place. However, we found emerging tools and standards that are discussed in the context of the Semantic Web idea to be helpful in establishing an interoperability environment. Even if not everything can be implemented according to the proposed high standards, the discussion does point in the right direction and is necessary in order to consider the chances and risks of an interoperability infrastructure.

5 SEMANTIC WEB

"The Semantic Web is a web of data ..., where on the original Web mainly concentrated on the interchange of documents."[2] These thoughts of Tim Berners-Lee had a huge impact on how people observe the shortcomings of today's Internet and on proposals for new and visionary extensions.[3] However, the idea behind the Semantic Web has been discussed controversially, with opinions ranging from very pessimistic to extremely optimistic. Probably the truth will be found somewhere in between these extreme positions. Different aspects of research related to the idea of the Semantic Web turned out to be helpful for discussing how several cultural heritage resources could interact in the future. Since the term "Semantic Web" denotes many different things to different people, the following remarks will focus on narrowing the scope for the presented project and on discussing the participating components in respect to the needs of the project.

One of the most fundamental concepts in this area is the notion of a Uniform Resource Identifier. If someone needs to talk about something, one relies on some mechanism to clearly refer to what one wants to make a statement about. Each modern information system provides some mechanism to unambiguously identify a certain chunk of data. A relational database, for example, resolves this in the scope of one application. By heavily relying on URIs, a system is currently being promoted that introduces a way to refer to things unambiguously on a global level and puts different communities in a position to talk about similar things by using similar identifiers.

The Semantic Web builds upon a framework that can be serialized as XML and that should facilitate the annotation of data with additional explicit meaning. This framework, called the Resource Description Framework (RDF), is organized as statements, each one forming a triple consisting of subject, predicate and object.[4] This approach to digital representation of knowledge is powerful enough to express data that has been extracted from relational databases. Imagine, for example, a database that contains a table "sculpture" with a field "material." And for the sculpture record with the unique identifier 4711 this field has the value "marble." At least three triples can be derived from this information (I will stick to a non-technical notation here that is inspired by CIDOC CRM): "4711," "is a," "man-made object," "4711," "has type," "sculpture" and "4711," "consists of," "marble." Consequently, tools are available that map the internal structure of a relational database to RDF.[5]

Another concept has been elaborated in the scope of Semantic Web research that improves the expressiveness of RDF—the notion of ontologies. In information science the term ontology describes a formal representation of concepts within a domain and the relationships between these concepts. Since this representation is machine-readable, it can be used to deduct new knowledge from data that has been encoded accordingly. The Web Ontology Language (OWL), which has been published as a recommendation by the W3C, is a compilation of languages for representing such ontologies.[6] All these developments aim at establishing means to clearly define the meaning of structured vocabularies in use. That puts software developers in a position to craft applications that can process the data according to its intended meaning. The CIDOC CRM can be expressed with the means of OWL.

Data that has been richly annotated and encoded as RDF is not useful by itself. There needs to be an mechanism that is able to process the data in a way that is useful and meaningful to human beings. Research in the field of artificial intelligence coined the term "intelligent agent".[7] Such an intelligent agent is a piece of software

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

[2] "W3C Semantic Web Activity," www.w3.org/2001/sw/.

[3] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," www.scientificamerican.com/article.cfm?id=the-semantic–web.

[4] Dan Brickley and R.V. Guha, *RDF Vocabulary Description Language 1.0: RDF Schema*, 2004, www.w3.org/TR/rdf-schema/.

[5] Chris Bizer, "D2RQ—Treating Non-RDF Databases as Virtual RDF Graphs," www4.wiwiss.fu-berlin.de/bizer/d2rq/.

[6] D. L. McGuinness and F. van Harmelen, *OWL Web Ontology Language Overview*, 2004, www.w3.org/TR/owl-features/.

[7] Peter Norvig and Stuart Russell, *Artificial Intelligence: A Modern Approach*, 2nd ed. (Prentice Hall International, 2003).

that bears human-like qualities in respect to cognition and that can process Semantic Web data while interacting with human beings in a beneficial way. The qualities of an agent can include perception, learning, inductive/deductive inference and many more. A user interface that will be described below can be perceived as an agent that interacts with Semantic Web data and at the same time exposes it to users in an easy and understandable way through hiding unnecessary complexity.

3 THE CIDOC CRM

Reference models convey a basic idea of how objects in a certain system are arranged and interrelated. These fundamental paradigms are technology-agnostic but can—for example in software technology—be used to derive standards. The CIDOC Conceptual Reference Model is such an abstract model that has been crafted to support the exchange of information on objects of museum quality. It has been accepted as official standard ISO 21127:2006 and comprises definitions arranged as a structured vocabulary that were developed over a period of more than ten years by the CIDOC Documentation Standards Group. To establish interoperability, a certain community needs at least some agreement on how shared information should be structured and what the meaning of the elements that form the structure is. This enables software developers to craft software that can deal with data that has been annotated in a certain way. Figure 3 (next page) provides a birds-eye perspective on how the CRM has been built.[1]

Technically speaking, the CIDOC CRM is a hierarchy of 84 classes defining concepts that are commonly referred to in museum documentation practice. Each class describes a set of objects that share common features. 141 so-called properties define semantic relations between these conceptual classes. Thus, the CRM builds a foundation for semantic interoperability in the cultural heritage area. By adopting these concepts of formal semantics, the CIDOC CRM is well prepared to play a role in the development of the Semantic Web.

Since the CRM is a reference model, it does not specify the peculiarities of an implementation. However, multiple alternatives to providing standards and formats have been described that encode and submit information about museum objects according to the CRM. The Erlangen CRM, for example, is an interpretation of the CRM as an ontology formulated in OWL.[2] Therefore, it has a very strong affinity to the concepts that are discussed as forming the Semantic Web. CDWA Lite[3] is more harvesting oriented, and Museumdat[4] bridges the gap between the harvesting-oriented format CDWA Lite and the CIDOC CRM. Nevertheless, it is possible to transmit data that has been encoded as RDF via the harvesting format OAI-PMH[5] once the boundaries of a document are defined.

We decided to use the CIDOC CRM as a guiding model to capture the structure of our data for integration because it appears to have a strong affinity to Semantic Web research. Additionally, with the Erlangen CRM, a Semantic Web enabled implementation is in place. Although the CRM originates from the museum community, it can also be applied to related domains like archaeology. One noteworthy feature of the CRM should be mentioned. It relies heavily on the concept of events that are considered as indispensable for digitally representing data about museum objects.[6] Although this approach to capturing metadata is somewhat orthogonal to common documentation practice, it does help to preserve the context of an archaeological object.

7 INTEROPERABILITY WORKFLOW

A workflow needs to be established to integrate data being hosted in different cultural heritage information systems. The whole procedure starts with extracting data from each system, guides over dealing with different kinds of heterogeneity, and ends with finding an adequate paradigm to make the integrated information available for end-users. The following section will describe each step in more detail.

First, both data models were exported to an intermediate representation for further processing. Neither Arachne nor Perseus relies on internal data models that can be instantly expressed in terms of the Semantic Web. Exporting the data to an intermediate XML-format turned out to be helpful because it allows for pre-processing and data cleaning; beyond that, it has proven to be extremely scalable. Chunks of 1000 database records per XML file were made ready for further processing this way.

[1] The figure follows M. Dörr, "The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata," *AI Magazine* 24 (3) (2003): 75–92.

[2] Bernhard Schiemann et al., "Erlangen CRM / OWL," http://erlangen-crm.org/current-version (accessed Feb. 1, 2010).

[3] J. Paul Getty Trust and ARTstor, "CDWA Lite: Specification for an XML Schema for Contributing Records via the OAI Harvesting Protocol," July 17, 2006. www.getty.edu/research/conducting_research/standards/cdw a/cdwalite.html.

[4] Regine Stein et al., "Museumdat–Harvesting Format for Providing Core Data from Museum Holdings," October 26, 2007. www.museumdat.org.

[5] M. L. N. Herbert, "Resource Harvesting within the OAI-PMH Framework," *D-Lib Magazine* 10 (12) (2004), www.dlib.org/dlib/december04/vandesompel/12vandesompel.html.

[6] M. Ioannides et al., "Documenting Events in Metadata," http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.79.6127.

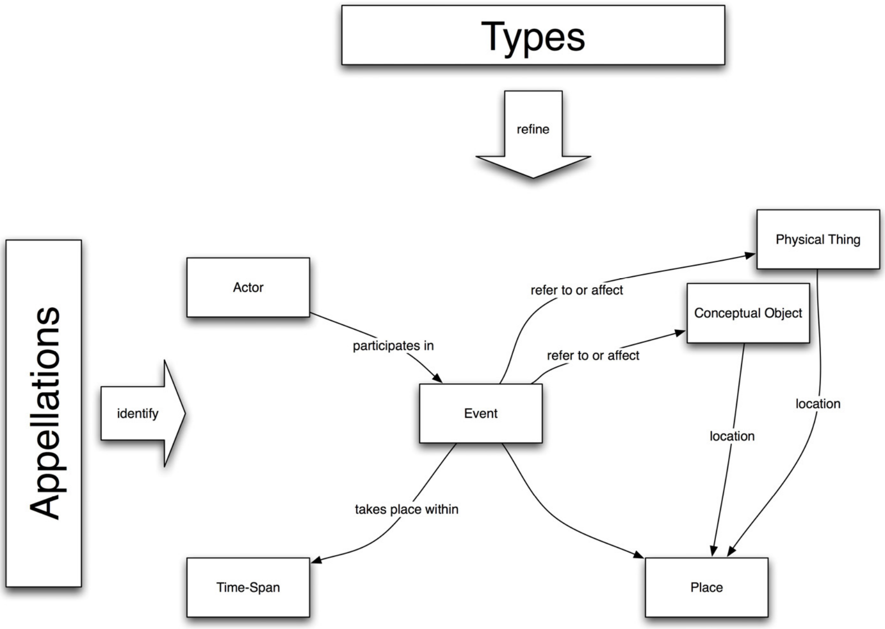


Figure 3. *The general structure of the CIDOC CRM.*

Thereafter, an XSLT style-sheet was used to apply mapping rules to the exported data set. The result of this process is a representation of the extracted data as RDF/XML conforming to the Erlangen OWL ontology. This step also involved assigning unique global identifiers (URI) to each material object and concept. Crafting the XSLT style-sheet turned out to be very labor intensive, because the CRM suggests a rather different approach of representing data than state-of-the art archaeological databases. Some parallel structures need, for example, to be split up and presented as hierarchies. Additionally, current databases only have implicit information about events that have to be made explicit for encoding according to the CRM. Because of these difficulties, it is unlikely that large parts of cultural heritage databases will be mapped to any shared conceptualization, and the amount of completeness should be driven by the overarching interoperability aim. The resulting RDF/XML code was then ingested into a so-called triplestore for further processing and querying. As of today, most triplestores do not perform well enough to establish live-user querying for large data sets. Therefore, an intermediate step has been established that extracts certain information from this store to provide faster querying and facetted browsing. The workflow presented does not contain any provision for duplicate record detection or co-reference resolution. Some records have therefore been assigned different global URIs, although they are referring to the same thing (archaeological object) in the world.

Working through the presented workflow helped with gaining experience in applying techniques and tools related to Semantic Web ideas in general and to RDF/XML in particular. It turned out that cultural heritage data can be expressed in terms of these concepts but there are still issues. The overall workflow needs more automation by adding means to publish, harvest, link, index, and present the information stemming from different resources. However, the main challenge will be to enhance the process with components that facilitate cross-language co-reference resolution. Without this component, data that has been mapped to the CRM still cannot be processed together even if it has been integrated syntactically and semantically. Thus, the problem of distributed terminology will be of dominant interest in future interoperability architectures.

8 BROWSER IMPLEMENTATION

The primary goal of our information integration efforts has been to establish simple means for resource discovery across our systems. Enabling easy and rapid resource discovery seems to be a common sense minimal approach to establishing interoperability among cultural heritage information systems. After having mapped a basic set of database fields to RDF that conforms to CIDOC CRM, we started experimenting with the MIT Longwell[1] Semantic Web browser in 2007 and early 2008, but ran into serious scalability issues. Therefore, we decided to build a new user interface from scratch that communicates with the data back-end in a way that prevents the system from stalling. While relying on Semantic Web concepts, the browser component presents the integrated data with reduced complexity. The new system uses SPARQL[2] queries to extract certain bits of information from RDF data for indexing. An indexer component that is based

[1] *Longwell-SIMILE*, http://simile.mit.edu/wiki/Longwell.

[2] "SPARQL Query Language for RDF," www.w3.org/TR/rdf-sparql-query/.

on Apache Lucene then prepares the data for searching and facetted browsing.[1]

Figure 4 shows a screenshot of the archaeological object browser prototype. Each material object is represented as a thumbnail along with a short description. On the right side of the user interface, tools for information discovery are provided, comprising full-text searching and facetted browsing. Accessing large collections via a paradigm that draws information from facetted classification techniques has become quite popular. We believe that combining this approach with additional full-text querying provides a simple and intuitive way to find information in vast amounts of data. However, the browser does not solve the problems that stem from multilingual data sets. Archaeologists searching for an English term will not find records from a German database. The next logical step would be to equip the back-end with tools that facilitate the merger of archaeological objects stemming from different databases even if they come with differing terminology or terms from different national languages.

Due to aspects of efficiency, not all database fields could be mapped to a structure that conforms to the CRM. Therefore, the presented browser can only show a choice of metadata that has been recorded for one single material object. Additionally, the respective metadata have been taken out of their original digital context; this leads to additional loss of information. Therefore, we decided to provide links back to the contributing information systems so that users can navigate to the respective system and have a look at the object in its original digital context. If data for one single object has been harvested from different CH databases, the system should provide links to both contributing systems.
Again, this raises the question of how digital information on a single object that has been gathered from multiple databases can be merged even if different terminology and multiple national languages are used.

9 Lessons Learned

We found that the Semantic Web does provide a decent set of concepts to think about building blocks for information integration. These include the extensive usage of Uniform Resource Identifiers (URIs) and the ability to express metadata as XML that made processing via XSL Transformations[2] easy. Furthermore, RDF in connection with the Erlangen CRM proved to be flexible enough to express a variety of project-specific data models. Emerging ideas like "Linked Data" aim at finding ways of expressing data that comes with rich structure, as in archaeological research. However, the underlying concepts and standards demand very high levels of semantic precision and data quality. Since URIs are meant to refer to exactly one thing, database fields with internal structure turned out to be problematic. Editorial efforts or automatic data cleaning processes need to leverage the data quality to meet the needed standards. Because of the costs associated with this procedure, many cultural heritage databases will not be instantly in a position to deliver this kind of data in large amounts. Therefore, projects striving to establish a shared environment of any kind should spend time and effort to define the scope of their particular project. This will enable all project partners to estimate the effort for their contribution to the project.

Computers rely heavily on the paradigm of serial symbol processing, and many ideas stemming from discussions in the field of Semantic Web research aim at explicitly formulating information that has been implicit before. Beyond that, applying techniques like reasoning and data merging will create even more information. This bloats the amount of information that needs to be encoded, and from a technical perspective the number of symbols that need to be processed is growing exponentially. That said, it is no surprise that it is very easy to overload most semantic stores with certain queries, resulting in a response time that is not tolerable. At the same time, it is difficult to estimate the performance of semantic stores. The way they are implemented does influence their performance very much. But it seems plausible that performance becomes worse if the considered graph structure becomes larger and more complex. This is analogous to simple JOIN operations of relational databases, which usually slow down query performance. One promising approach could be to use the semantic store for any kind of background operation and to insert a separate indexing layer for direct and rapid communication with end-users.

Because of its complexity, the CRM takes some time to become acquainted with. Institutions striving for an interoperable environment should examine the pros and cons. This complexity results from the challenge of establishing interoperability itself. Archaeological objects and those of museum quality live in a complex and highly contextualized environment that is reflected by the property- and event-centric architecture of the CRM. This degree of complexity may make sense for technical aspects of information integration and those with regard to content, but it is not easy to convey this complexity to the end-user. Therefore, the overall complexity needs to be reduced on the user side so that non-experts can understand and use the resulting system.

The way that data is encoded and manipulated according to Semantic Web standards and technologies enables new ways of interacting with information. The underlying paradigm of information encoding is a graph structure where information is stored as nodes and links. This is the provision for navigating a large amount of

[1] *Welcome to Solr*, http://lucene.apache.org/solr/.

[2] "XSL Transformations (XSLT)," www.w3.org/TR/xslt.

information that is linked with regard to content; exploiting the underlying graph structure could make browsing possible in new ways. However, it has been argued that the CRM deals with the structure of the museum domain and not primarily with the terminology and vocabularies used in databases. The following passage will focus on this specific problem.

10 Co-reference Resolution

We have found that the collections of Arachne and Perseus do overlap to a certain degree.[1] Art and archaeological objects being curated in large and prominent museum collections tend to be present in more than one (international) database. And even if the collections do not overlap with respect to material finds or objects in general, they most probably will with respect to concepts and terminology. In this regard, it would be beneficial to reduce the number of names (identifiers) that refer to one conceptual or material thing in the world, resulting in a need for data analysis and fusion. Thus, beyond the challenge of integrating heterogeneous data models and establishing a certain data quality lies the problem of merging heterogeneous data records. This problem is usually referred to as record linkage and has been described as a difficult and resource consuming challenge.[2] Nevertheless, this step is indispensable if related information needs to be drawn together according to the principles of Linked Data that have been described above.

The task of record linkage results in linked data, i.e. data that is marked as belonging together in a certain way, for example, by assigning a common identifier. Most concepts of Semantic Web research will not function properly without consistent identifiers. Reasoning, for example, will only work if all facts that are known about a certain topic are considered. This is clearly not possible if many names are used for the same thing. In historical research, record linkage was already used in the 1980s to study data from census records and parish registers to perform longitudinal studies. Several approaches to the problem of record linkage have been developed since then, ranging from rule-based approaches to probabilistic methods like Naive-Bayes algorithms.

It is at least doubtful that enough training data will be available in an international environment to apply data mining tasks. But libraries traditionally put immense efforts into crafting authority files that could help with resolving co-references and disambiguating entities. Moreover, users could provide additional input in some kind of crowd-sourcing process. This way, co-reference resolution could be performed in a supervised fashion, where users provide input by correcting decisions or proposals that have been made by algorithms. Training data generated thereby can be used to enhance the reliability of the above-mentioned methods. The following section will provide a real-world example and reflect on possibilities for approaching the challenges.

11 Example

Figure 5 shows two database records, one originating from Perseus and the other from Arachne. Interestingly, both refer to the same object in the world, found in Ariccia near Rome in Italy. It is now on display at the Museum of Fine Arts in Boston. The figure shows a simplified metadata record of each object after being mapped to the CRM. The way of representing the metadata makes sure that everything that is known about a specific object is associated with a global object identifier. One could imagine several approaches to enable machines to identify the fact that the two identifiers “Perseus:Boston 99.334” and “Arachne:2913” refer to the same entity. Different kinds of metadata need different approaches to detect similarities.

One not-so-obvious approach to detect similarities could be the application of image recognition software to images that are associated with a specific material object. But more and more information about places and their names has been made available online. Projects like Geonames,[3] the Getty Thesaurus of Geographic Names,[4] and the Alexandria Digital Library Project[5] are collections that have geo-referenced materials online.

Geonames, for example, provides a web service to publish its data according to the principles of the Linked Data initiative. This makes it easy not only to embed the service in some kind of co-reference resolution infrastructure but also to use identifiers that have been provided thereby. Other developments, like the LC Name Authority Service (www.oclc.org/research/ researchworks /authority/default.htm), provide means to check names against the Library of Congress Author-

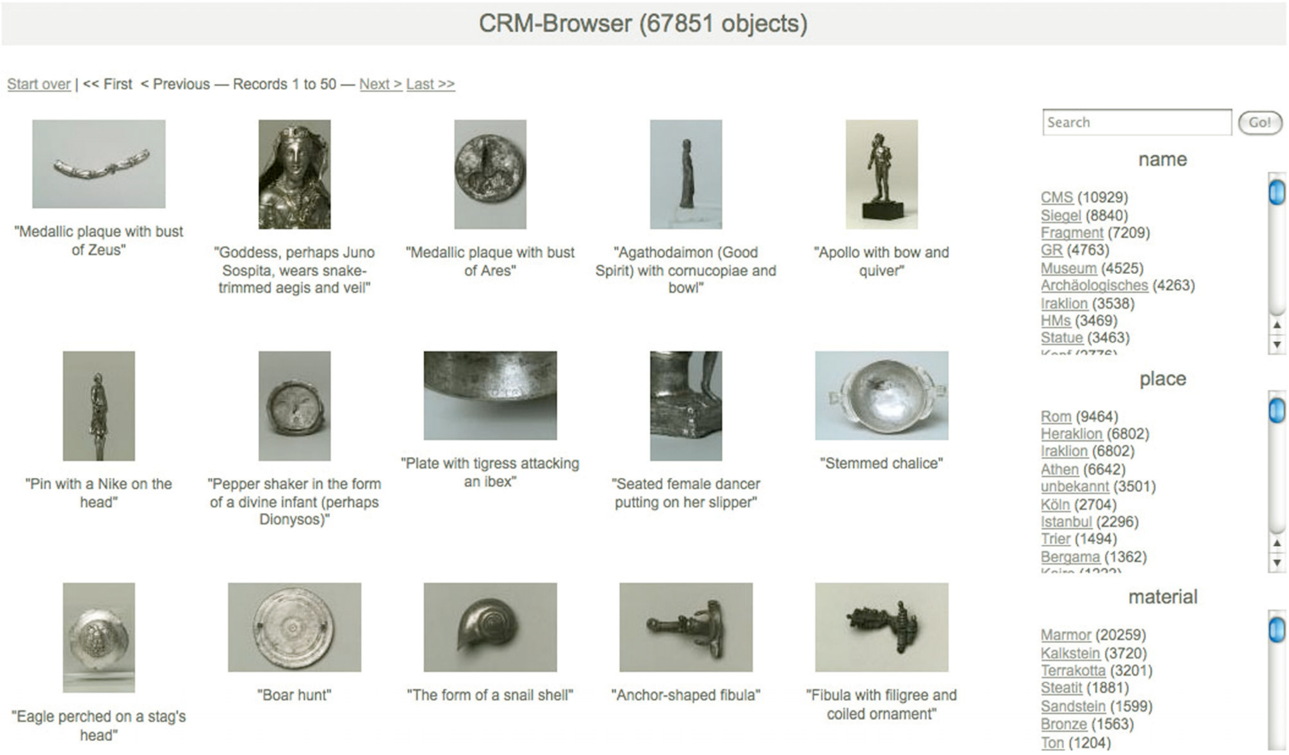
[1] Alison Babeu et al., “Named Entity Identification and Cyberinfrastructure,” *Research and Advanced Technology for Digital Libraries* (2007): 259–270. http://dx.doi.org/10.1007/978-3-540-74851-9_22.

[2] H. Zhao and S. Ram, “Entity Identification for Heterogeneous Database Integration: A Multiple Classifier System Approach and Empirical Evaluation,” *Information Systems* 30 (2) (2005): 119–132.

[3] “GeoNames,” www.geonames.org/.

[4] Getty Research Institute, *The Getty Thesaurus of Geographic Names Online*, 2007, www.getty.edu/research/conducting_research/ vocabularies/tgn/index.html.

[5] “Alexandria Digital Library Project,” www.alexandria.ucsb.edu/.



CRM-Browser (67851 objects)

Start over | << First < Previous — Records 1 to 50 — Next > Last >>

"Medallic plaque with bust of Zeus"

"Goddess, perhaps Juno Sospita, wears snake-trimmed aegis and veil"

"Medallic plaque with bust of Ares"

"Agathodaimon (Good Spirit) with cornucopiae and bowl"

"Apollo with bow and quiver"

"Pin with a Nike on the head"

"Pepper shaker in the form of a divine infant (perhaps Dionysos)"

"Plate with tigress attacking an ibex"

"Seated female dancer putting on her slipper"

"Stemmed chalice"

"Eagle perched on a stag's head"

"Boar hunt"

"The form of a snail shell"

"Anchor-shaped fibula"

"Fibula with filigree and coiled ornament"

name

CMS (10929)
Siegel (8840)
Fragment (7209)
GR (4763)
Museum (4525)
Archäologisches (4263)
Iraklion (3538)
HMs (3469)
Statue (3463)

place

Rom (9464)
Heraklion (6802)
Iraklion (6802)
Athen (6642)
unbekannt (3501)
Köln (2704)
Istanbul (2296)
Trier (1494)
Bergama (1362)

material

Marmor (20259)
Kalkstein (3720)
Terrakotta (3201)
Steatit (1881)
Sandstein (1599)
Bronze (1563)
Ton (1204)

Figure 4. *The archaeological object browser prototype.*

ity File. Other authorities exist that associate preferred names of specific persons with variants in spelling and language.

Many archaeological databases record the provenance of the information that has been entered by its curators. If two digital surrogates of a material object are referred to by a similar bibliographic reference, they are most probably connected in terms of content. Additionally, both metadata records provide information about the dimensions of an object. The example at hand provides two measurements for the height of the statue that could provide additional hints on similarity if processed correctly. A software component that can distinguish and convert different measures and dates could be helpful.

To put it in a nutshell, there are at least three areas that could be exploited in order to see if two digital representations of an archaeological object are referring to the same thing. First, methods that stem from data mining research; these comprise similarity measures as well as techniques that rely on machine learning. Second, external references published via services like the aforementioned Geonames. If these resources are joined with data mining techniques, a powerful tool for resolving co-references could be the result. Third, attributes that can be compared by logical induction, such as measurements, should be included in the resulting co-reference resolution framework. The more similarities can be observed for a set of data, the more likely they will refer to the same thing in the world and therefore should be equipped with the same identifier.

This summary emphasizes that systems performing co-reference resolution or record linkage face many problems. Different approaches for building a framework that performs co-reference resolution have been presented above. Some matching techniques focus on archaeological content itself by computing similarities of instances or looking at how data is structured. Other matching strategies draw additional information from resources that reside outside of the content to be linked. But computing similarities of strings will most certainly fail in an international environment, and external resources like domain-specific multilingual thesauri are still not in place. In any case, scientists are doing co-reference resolution as part of their day-to-day work. A good start would be to build a system that makes it easy for them to resolve co-references online. The resulting data then could be used to enable systems to make recommendations for improvement on possible further co-references to be resolved.

Exploiting external resources, considering features that have been drawn from data itself, and interacting with professional users form the building blocks of resolving co-references. Thus, the following remarks will focus on these areas.

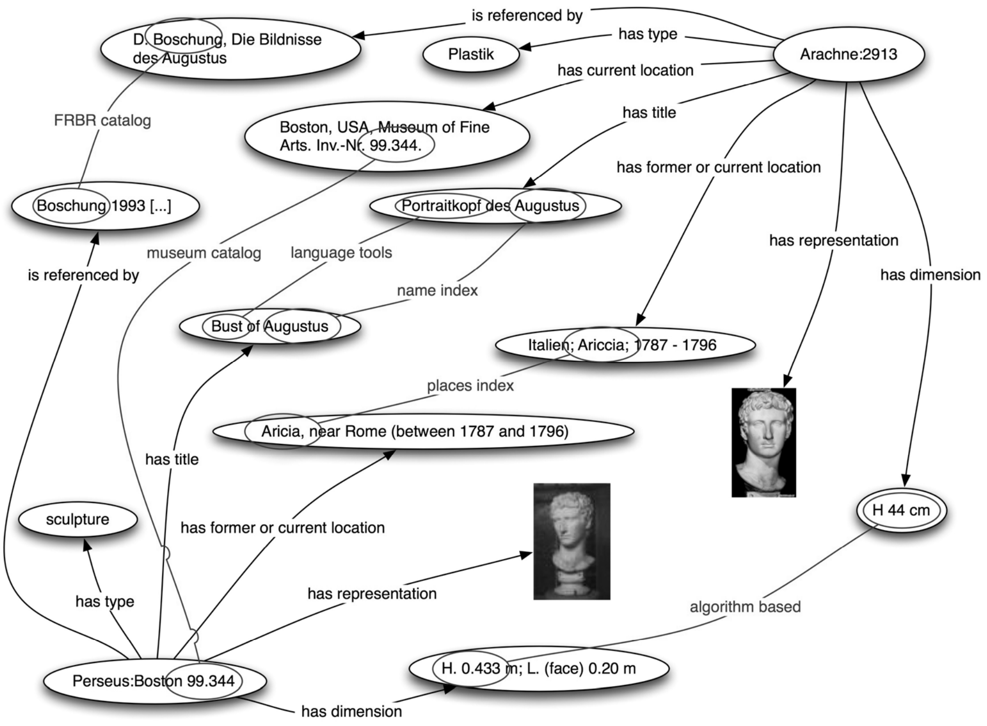


Figure 5. *The problem of co-reference resolution.*

12 AUTHORITY CONTROL

Traditionally, libraries invest immense amounts of human resources in establishing and curating authority files. But in digital collections with massive amounts of digital-born data these standards of curation cannot be achieved manually. However, we are expecting that smaller communities will establish and publish vocabularies to enhance information sharing within the scope of their projects. Sieglerschmidt, for example, argues for the evaluation and use of structured controlled vocabularies for information exchange in the cultural heritage domain.[1] He focuses on how knowledge organization in an international environment can be supported by such vocabularies. Consequently, more and more initiatives concentrate on developing ways to encode and publish structured vocabularies.

In the area of Semantic Web research one development stands out. While the Web Ontology Language (OWL) has been built to express complex conceptual structures, SKOS intends to provide a more straightforward approach to publishing multilingual structured vocabularies. SKOS builds upon the foundation of RDF that makes seamless processing of vocabulary data and database content possible.[2] Consequently, initiatives such as "museumsvokabular" publish their vocabularies as XML, HTML and SKOS.[3] Binding and Tudhope describe a service that adds behavior to these published thesauri, including search, browsing, and semantic expansion across structured vocabularies.[4] First efforts to integrate this service to establish semantic interoperability have been made.[5]

13 DATA MINING

Data mining plays a major role in trying to get information out of a lot of data. Research in data mining focuses on solving problems by analyzing data that has

[1] Jörn Sieglerschmidt, "Knowledge Organization and Multilingual Vocabularies. Vortrag auf der Jahrestagung 'Managing the Global Diversity of Cultural Information' des Comité International pour la Documentation (CIDOC), Vienna, 20–22 August 2007," http://opus.bsz-bw.de/swop/volltexte/2008/280/.

[2] "SKOS Simple Knowledge Organization System Primer," www.w3.org/TR/skos-primer/.

[3] M. Hagedorn-Saupe et al., "museumsvokabular.de," http://museum.zib.de/museumsvokabular/.

[4] Ceri Binding and Douglas Tudhope, "Using Terminology Web Services for the Archaeological Domain," *Research and Advanced Technology for Digital Libraries* (Berlin: Springer, 2008) 392–393, http://dx.doi.org/10.1007/978-3-540-87599-4_42.

[5] Ceri Binding, Keith May, and Douglas Tudhope, "Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM," *Research and Advanced Technology for Digital Libraries* (Berlin: Springer, 2008) 280–290, http://dx.doi.org/10.1007/978-3-540-87599-4_30.

been drawn from existing information systems.[1] Most techniques deal with finding meaningful patterns in the data that can be exploited to generate some (economic) advantage. Machine learning algorithms, for example, are the foundation of recommender systems that are gaining popularity in e-commerce applications. Could these techniques also be applied to cultural heritage data that has been expressed in terms of Semantic Web concepts? It seems to be obvious that string distance metrics could help with co-reference resolution. Cohen et al. compiled a comparison of string distance metrics that can be applied to arbitrary strings.[2] Some of these measures work well with names, but in an international environment the overall performance will most probably be poor.

By contrast, machine learning is based on training a piece of software with example instances. The software then deducts certain structural relationships that are internally represented as a learned model. By evaluating these relationships, the system tries to assess new and unknown chunks of knowledge. A useful outcome of this data analysis process is knowledge in the form of rules that can be used to classify instances, or that enlightens structural correlations that have not been obvious before. Bayesian models have been used to perform co-reference resolution in text corpora,[3] but also in more general contexts.[4] It would certainly be interesting to evaluate the application of these techniques to more structured data in the world of OWL and RDF. Additionally, another area of research called "link mining" is exploring richly structured data sets.[5] In this field, predictive link mining could be fruitful for linking data sets that are richly structured.

Euzenat and Shvaiko present a classification of matching approaches that aim at a holistic matching framework.[6] This classification includes element-level techniques and structure-level techniques. Some techniques only use data that can be found inside the examined set of data, while others draw additional information from external resources like thesauri. Of extraordinary importance in the field of humanities is the concept of authorship, a person or entity that makes the final decision that two entities refer to the same thing. The author could be a human being or an intelligent software agent that makes a decision on the basis of acquired data. Therefore, a matching framework should make decisions transparent and rely on user input and validation. If two digital surrogates have been identified as referring to the same entity in the world, a new co-reference has been established. This alignment should be encoded, stored, and published so that other information systems can benefit.

14 Towards a Co-reference Resolution Framework

So far, different approaches to co-reference resolution have been presented. An effective framework, which is able to resolve co-references on information that has been encoded in RDF/XML according to the CRM, has been motivated. Such a framework would support establishing, encoding, sharing, and exploiting co-references. Data that has been encoded as RDF can be used as input for data mining processes resulting in additional information that can be added. Such information, for example, can be co-reference information.

A framework needs to be established that not only implements a thorough set of the above-mentioned techniques but that is also embedded in an infrastructure that allows for the management of co-reference information. Eide drafted a possible architecture of such a system.[7] This co-reference system comprises an application server that relies on a database for storing co-references. Additionally, the established information can be exposed to a particular community in different ways. Users can view and edit co-referenced information via web interfaces. A web service provides means for information systems to integrate this service as part of their software infrastructure. And most importantly, the system can also contribute its knowledge to a "network of identity," so that other institutions can benefit from information that has been established at one place.

Data Mining and the Semantic Web are two approaches that are aiming at related objectives. Both concepts describe techniques that process data in some kind of synergistic way. Data mining is biased towards deriving new and useful information from data that is controlled by a different system (by using neural networks, decision trees and so on). Semantic Web efforts focus

[1] Ian H. Witten and Eibe Frank, *Data Mining. Practical Machine Learning Tools and Techniques*, second ed. (Morgan Kaufmann, 2005).

[2] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg, "A Comparison of String Distance Metrics for Name-Matching Tasks," in *IIWeb* (2003) 73–78.

[3] Aria Haghighi and Dan Klein, "Unsupervised Coreference Resolution in a Nonparametric Bayesian Model," *Proceedings of the 45th annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, June 27–30, 2007*, 848–855.

[4] William E. Winkler, "Overview of Record Linkage and Current Research Directions," *Bureau of the Census*, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.79.1519.2006.

[5] Lise Getoor and Christopher P. Diehl, "Link Mining," *ACM SIGKDD Explorations Newsletter* 7 (2005): 3–12.

[6] Jérôme Euzenat and Pavel Shvaiko, *Ontology Matching* (Berlin: Springer-Verlag GmbH, 2007).

[7] Øyvind Eide, "The Unit for Digital Documentation (EDD) System for Storing Coref Information," September 2008, http://cidoc.mediahost.org/eddSystemCoref.pdf.

on explicitly encoding information that does not need to be processed with complex algorithms but with straightforward calculations. Both approaches should team up to facilitate recognition and resolution of co-references.

15 CONCLUSION

This paper reported on information integration efforts that have been undertaken as a collaboration of Arachne and Perseus. Concepts that are currently being discussed in the context of Semantic Web research turned out to be helpful for establishing a shared research environment for archaeologists. The main challenges and approaches to problems have been presented. Pressing problems for future research have been identified, namely linking digital surrogates that refer to the same entity in the world. While this issue needs to be addressed to make Semantic Web efforts work as a whole, the archaeological domain provides an adequate research environment with pressing needs in this respect. A basic example has been used to demonstrate this issue and to discuss first thoughts on resolving co-references for cultural heritage data, in this case data stemming from archaeological databases. Multiple suggestions for approaching the problem of data fusion have been made, but the overall framework that has been elaborated is still missing.

ACKNOWLEDGEMENTS

This paper was made possible by the generous financial support of the German Archaeological Institute. I would also like to thank Prof. G. Crane, Tufts University, and Prof. R. Förtsch and Prof. M. Thaller, University of Cologne for sponsoring me and keeping me grounded. This work is part of a Ph.D. project that is currently being funded by the Cusanuswerk, Bischöfliche Studienförderung.

BIBLIOGRAPHY

"About LC Name Authority File [OCLC-ResearchWorks]." www.oclc.org/research/researchworks/authority/default.htm.

"Alexandria Digital Library Project." www.alexandria.ucsb.edu/.

Babeu, Alison, David Bamman, Gregory Crane, Robert Kummer, and Gabriel Weaver. "Named Entity Identification and Cyberinfrastructure." *Research and Advanced Technology for Digital Libraries* (2007): 259–270. http://dx.doi.org/10.1007/978-3-540-74851-9_22.

Beazley Archive. "CLAROS—Classical Art Research Centre Online Services—The University of Oxford." www.clarosnet.org/index.htm.

Berners-Lee, T., J. Hendler, and O. Lassila. "The Semantic Web." *Scientific American* (May 2001). www.scientificamerican.com/article.cfm?id=the-semantic-web.

Berners-Lee, T. "Linked Data-Design Issues." www.w3.org/DesignIssues/LinkedData.html.

Binding, Ceri, Keith May, and Douglas Tudhope. "Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM." *Research and Advanced Technology for Digital Libraries* (2008): 280–290, http://dx.doi.org/10.1007/978-3-540-87599-4_30.

Binding, Ceri, and Douglas Tudhope. "Using Terminology Web Services for the Archaeological Domain," in *Research and Advanced Technology for Digital Libraries* (2008): 392–393. http://dx.doi.org/10.1007/978-3-540-87599-4_42.

Bizer, Chris. "D2RQ-Treating Non-RDF Databases as Virtual RDF Graphs. www4.wiwiss.fu-berlin.de/bizer/d2rq/.

Brickley, Dan, and R. V. Guha. *RDF Vocabulary Description Language 1.0: RDF Schema*, 2004. www.w3.org/TR/rdf-schema/.

Cohen, William W., Pradeep Ravikumar, and Stephen E. Fienberg. "A Comparison of String Distance Metrics for Name-Matching Tasks." *IIWeb* (2003): 73–78.

Crane, Gregory. "Perseus Digital Library." www.perseus.tufts.edu/hopper/.

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

Dörr, Martin, and Dolores Iorizzo. "The Dream of a Global Knowledge Network—A New Approach." *Journal on Computing and Cultural Heritage* 1 (2008): 1–23.

Eide, Øyvind. "The Unit for Digital Documentation (EDD) System for Storing Coref Information." September 2008. http://cidoc.mediahost.org/eddSystemCoref.pdf.

Engelbart, Douglas. *Augmenting Human Intellect: A Conceptual Framework.* Air Force Office of Scientific Research 3233 (October 1962).

Euzenat, Jérôme, and Pavel Shvaiko. *Ontology Matching*. Berlin: Springer-Verlag GmbH, 2007.

Förtsch, Reinhard. *ARACHNE-Datenbank und kulturelle Archive des Forschungsarchivs für Antike Plastik Köln und des Deutschen Archäologischen Instituts*, 2007. http://arachne.uni-koeln.de/drupal/node/3.

"GeoNames." www.geonames.org/.

Getoor, Lise, and Christopher P. Diehl. "Link Mining." *ACM SIGKDD Explorations Newsletter* 7 (2) (2005): 3–12.

Getty Insitute. *The Getty Thesaurus of Geographic Names Online*, 2007. www.getty.edu/research/conducting_research/standards/cdw a/cdwalite.html.

Hagedorn-Saupe, Monika, Carlos Saro, Axel Ermert, and Lütger Landwehr. "museumsvokabular.de." http://museum.zib.de/museumsvokabular.

Haghighi, Aria, and Dan Klein. "Unsupervised Coreference Resolution in a Nonparametric Bayesian Model," in *Proceedings of the 45th Annual Meeting of the Association for Computer Linguistics, Prague, Czech Republic, June 23–30, 2007* (2007) 848–855.

Herbert, M. L. N. "Resource Harvesting within the OAI-PMH Framework." *D-Lib Magazine* 10 (12) (2004). http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html.

Ioannides, M., D. Arnold, F. Niccolucci, and K. Mania. "Documenting Events in Metadata." http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.79.6127.

J. Paul Getty Trust, and ARTstor. "CDWA Lite: Specification for an XML Schema for Contributing Records via the OAI Harvesting Protocol," July 17, 2006.

Longwell-SIMILE. http://simile.mit.edu/wiki/Longwell.

McGuinness, D. L., and F. van Harmelen. *OWL Web Ontology Language Overview*, 2004. www.w3.org/TR/owl-features/.

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

Schiemann, Bernhard, Martin Oischinger, Günther Görz, and Georg Hohmann. "Erlangen CRM / OWL." http://erlangen-crm.org/current-version (accessed Feb. 1, 2010).

Sieglerschmidt, Jörn. "Knowledge Organization and Multilingual Vocabularies." Vortrag auf der Jahrestagung "Managing the Global Diversity of Cultural Information." Comité International pour la Documentation (CIDOC) Vienna, 20–22 August, 2007. http://opus.bsz-bw.de/swop/volltexte/2008/280/.

"SKOS Simple Knowledge Organization System Primer." www.w3.org/TR/skos-primer/.

"SPARQL Query Language for RDF." www.w3.org/TR/rdf-sparql-query/.

Stein, Regine, Axel Ermert, Jürgen Gottschewski, Monika Hagedorn-Saupe, Regine Heuchert, Hans-Jürgen Hansen, and Angela Kailus. *museumdat—Harvesting Format for Providing Core Data from Museum Holdings*, October 26, 2007.

"W3C Semantic Web Activity." www.w3.org/2001/sw/.

"Welcome to OKKAM —Enabling the Web of Entities—The OKKAM Project Workspace." www.okkam.org/.

Welcome to Solr. http://lucene.apache.org/solr/.

Winkler, William E. "Overview of Record Linkage and Current Research Directions." *Bureau of the Census* (2006). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.79.1519.

Witten, Ian H., and Eibe Frank. *Data Mining. Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann, 2005.

"XSL Transformations (XSLT)." www.w3.org/TR/xslt.

Zhao, H., and S. Ram. "Entity Identification for Heterogeneous Database Integration: A Multiple Classifier System Approach and Empirical Evaluation." *Information Systems* 30 (2) (2005): 119–132.