Annotation of Heterogeneous Database Content for the Semantic Web

Eero Hyvönen, Mirva Salminen, and Miikka Junnila

University of Helsinki, Department of Computer Science
Helsinki Institute for Information Technology (HIIT)
{firstname.lastname}@cs.helsinki.fi
http://www.cs.helsinki.fi/group/seco/

Abstract. This paper discusses the problem of annotating semantically interlinked data that is distributed in heterogeneous databases. The proposed solution is a semi-automatic process that enables annotation of database contents with shared ontologies with little adaptation and human intervention. A technical solution to the problem based on semantic web technologies is proposed and its demonstrational implementation is discussed. The process has been applied in creating the content for the semantic portal MUSEUMFINLAND, a deployed Semantic Web application.

1 Introduction

A crucial question for the breakthrough of the Semantic Web approach is how easily the needed metadata can be created. Annotating data by hand is laborious, resource-consuming, and usually economically infeasible with larger datasets. Automation of the annotation process is therefore needed. This task is the more severe the more heterogeneous the data is. This paper addresses the problem of annotating heterogeneous and distributed data with a set of shared domain ontologies (within a single application domain). The problem is approached through a real-life case study by describing the annotation process developed for the MUSEUMFINLAND[1] [6, 8, 10] semantic portal. This application publishes cultural collection data from several heterogeneous museum databases in Finland.

We developed the annotation process for MUSEUMFINLAND in order to enable publication of museum collection item data on the Semantic Web. The goal of the annotation process is to transform the heterogeneous local databases into a global, syntactically and semantically interoperable knowledge base in RDF(S) format. The knowledge base is then stored into a common repository. This knowledge base conforms to a set of global domain ontologies, and the services provided by MUSEUMFINLAND to the end-users, i.e., view-based semantic search and browsing [5], are based on it[2].

The users of the process are museum personnel who want to bring their collections into the Semantic Web. Though the process is originally designed for the use of museums, the same approach can be applied to other heterogeneous database contents that

[1] http://museosuomi.cs.helsinki.fi

[2] The (meta)data, ontologies and programs used in the process described in this paper are available as open source at http://www.cs.helsinki.fi/group/seco/museums/dist/.

need to be annotated with shared domain ontologies. We will discuss the issues that affect the applicability and the workload of the process, and give examples based on the MUSEUMFINLAND case.

The annotation process was designed to meet two requirements: First, new museum collections need to be imported into the MUSEUMFINLAND portal as easily as possible and with as little manual work and technical expertise as possible. Second, the museums should not be forced to change their cataloging conventions for creating collection item descriptions. For example, two museums may use different terms for the same thing. The system should be able to accept the different terms as far as the terms are consistently used and their local meanings — with respect to the global reference ontologies — are provided.

Figure 1 depicts the whole annotation process that consists of three major parts:

1. **Syntactic Homogenization.** Since the data in museum databases is syntactically heterogeneous, the first step involves reaching syntactic interoperability by representing the database contents in a common syntax. A way of defining the common syntax is to specify an XML schema that all the different content providers can agree on. This task is simplified by the fact that the heterogeneous databases have a homogeneous domain: they contain cultural metadata about artifacts and historical sites, which means that the data items have similar features. For instance, all museum artifacts have features such as object type, material, place of usage, etc. This data can be exported from the different databases into a syntactically uniform XML form [12] (arrow on the left in figure 1).
2. **Terminology Creation.** To define the meaning of the terms and linguistic patterns used in the XML representation (and in the databases), we need to connect them to the global ontological concepts shared by the portal content providers. The mapping from literal values to concepts is called a *terminology*. In MUSEUMFINLAND, the terminology is created with the help of a tool called Terminator (lower arrow in figure 1).

   A problem in terminology creation is that the museums and catalogers use different vocabularies and describe their collection contents in differing manners. From a practical viewpoint, such local variance should be tolerated and should not impose terminological restrictions on other museums. In order to make MUSEUMFINLAND flexible with respect to variance in terminologies used at different museums, the terminology has been separated from the domain ontologies. In our approach, the museums can share globally agreed term definitions but also override them with their own local term definitions without any need to change the shared domain ontologies or global term definitions.
3. **Annotation Creation.** During the annotation creation process the XML data containing the museum item descriptions is enriched with references to the ontological definitions. This process is based on the terminologies and makes the heterogeneous collection data semantically interoperable with respect to the set of underlying domain ontologies. In MUSEUMFINLAND, a tool called Annomobile has been created to automate the annotation process (arrow on the right in figure 1).

In the following, these three parts of the process are discussed in more detail.

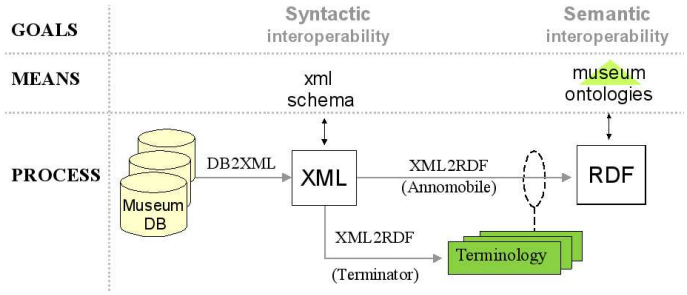

Fig. 1. The content creation process in MUSEUMFINLAND.

2 Syntactic Homogenization

The museum databases are both distributed and heterogeneous, i.e. the databases are situated in physically different places, the used database systems are made by different manufacturers, and their logical structure (schemes, tables, fields, etc.) may vary.

The first step of combining domain data from multiple sources is, thus, gaining syntactic interoperability. This task is highly system dependent. For example on the level of structure, combining collection data means that the collection record data fields meaning the same thing but under different labels in different databases, such as ”name of object” and ”object name”, are identified as the same, common labels are given to the fields, and a common way of representing collection data is agreed upon.

The combining can be done by agreeing on a shared presentation language for collection data. When the museums have agreed on this, the transmission, combination, and WWW publishing of the collections becomes significantly easier. In the MUSEUMFINLAND system, the combination of museum data at the structural level is based on a common XML schema. This schema is used to express the collection data to be published on the WWW. A simplified example of the XML can be found in [7].

The syntactic homogenization into XML makes the other steps of the process system independent, so that these steps don’t have to be changed at all when new museums join MUSEUMFINLAND or old museums change their databases.

The transformation procedure from database to XML depends on the database schema and system at hand, and is described more in detail in [12]. For the portal version currently on the web, we created database to XML transformers for three different database systems used in three different museums.

3 Terminology Creation

A terminology defines a mapping between terms and concepts. This makes automation of the annotation process possible. Figure 2 illustrates the role of terminology as a mediating layer between the conceptual layer and the data layer. On the top is the concept layer that is described by a set of global domain ontologies. Under that is the terminology layer that contains all the terms used for describing different things that relate

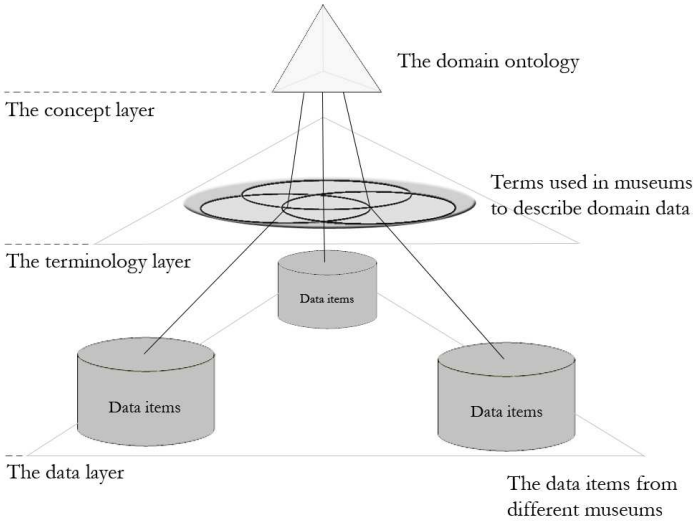

Fig. 2. The mapping of data items to the domain ontology through the terminology layer

to the domain. The terminology layer is broader than the concept layer, since concepts can be expressed in various ways. Under the terminology layer is the largest of the layers, the data to be annotated. Terminologies used in different databases intersect on the terminology layer but may have non-overlapping parts as well.

A term on the terminology layer is usually used as a value in several data items at the data layer. It is therefore easier to map data items to concepts by using the terms than by mapping data items directly to concepts. When terms have been excessively annotated, the data itself can be annotated almost automatically.

In MUSEUMFINLAND a terminology is represented by a term ontology, where the notion of the term is defined by the class Term. The class Term has six properties: `concept`, `singular`, `plural`, `definition`, `usage` and `comment`. They are inherited by the term instances called *term cards*. A term card associates a term as a string with an URI in an ontology represented as the value of the property `concept`. Both `singular` and `plural` forms of the term string are stored explicitly for two reasons. First, this eliminates the need for Finnish morphological analysis that is complex even when making the singular/plural distinction. Second, singular and plural forms are sometimes used with different meaning in Finnish thesauri. For example, the plural term “operas” would typically refer to different compositions and the singular “opera” to the abstract art form. To make the semantic distinction at the term card level, the former term can be represented by a term card with missing singular form and the latter term with missing plural form. Property `definition` is a string representing the definition of the term. Property `usage` is used to indicate obsolete terms in the same way as the USE attribute is used in thesauri. Finally, the `comment` property can be filled to

store any other useful information concerning the term, like context information, or the history of the term card.

Two different methods were used in terminology creation:

1. **Thesaurus to Taxonomy Transformation**
   Some 6000 new term instances were created based on the Finnish cultural thesaurus MASA [9] that was converted into a domain ontology (taxonomy). A term card for each thesaurus entry was created and associated with the ontology class corresponding to the entry. For obsolete terms, the associated ontology resource can be found by the USE attribute value. The morphological tool MachineSyntax[3] was used for creating the missing plural or singular forms for the term cards.
2. **Term Ontology Population from Databases**
   New term cards are created automatically for unknown terms that are found in artifact record data. The created term cards are automatically filled with contextual information concerning the meaning of the term. This information helps the human editor to fill the `concept` property. For example, assume that one has an ontology M of materials and a related terminology T. To enhance the terminology, the material property values of a collection database can be read. If a material term not present in T is encountered, a term card with the new term but without a reference to an ontological concept can be created. A human editor can then define the meaning by making the reference to the ontology and also create new entries for own terms if needed.
   For efficiency reasons, the new terms are ranked by their frequency of use, so that the human editor can annotate the most used terms, and leave the most infrequent terms unannotated. This way the editor's work amount in relation to the coverage of the term ontology is optimized.

Figure 3 depicts the general term extraction process in MUSEUMFINLAND. The process involves a local process at each museum and a global process at MUSEUMFINLAND. The tool Terminator extracts individual term candidates from the museum collection items presented in XML. The entity of one item is called an *XML card*. A human editor annotates ambiguous terms or terms not known by the system. The result is a set of new term cards. This set is included in the museum's local terminology and terms of global interest can be included in the global terminology of the whole system for other museums to use.

The global terminology consists of terms that are used in all the museums. It reduces the workload of individual museums, since these terms do not need to be included in local terminologies. The global terminology can be extended when needed. On the other hand, the local terminology is important because it makes it possible for individual museums to use and maintain their own terminologies.

The problem of the term creation approach described above is how to deal with free text descriptions. It is not very useful to regard field values that consist of long textual descriptions as single terms. For example "art poster" is a good term, but the term "A time-worn middle sized poster of a painting by Van Gogh" is not. This term probably wouldn't have any duplicates in the rest of the data, and annotation of the data item on

[3] http://www.conexor.fi/m_syntax.html

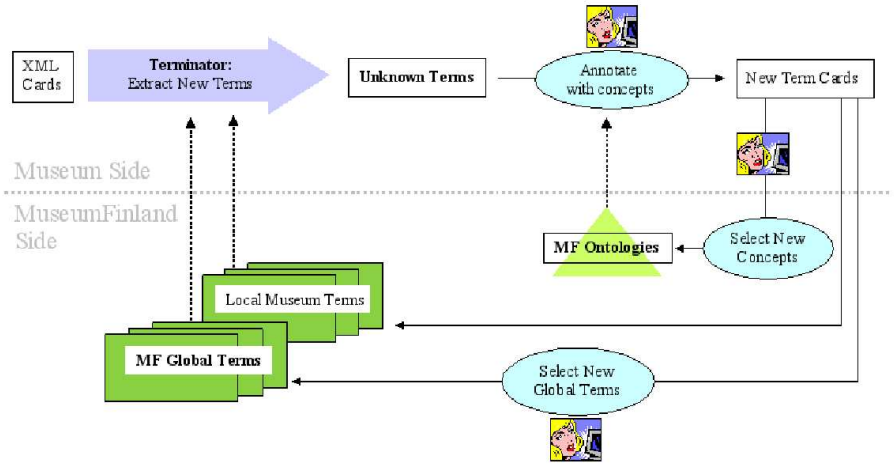

Fig. 3. Creating new term cards in MUSEUMFINLAND.

the data layer (cf. figure 2) instead of annotating it as a term would be as simple and more natural.

Sometimes the data in the databases is erroneous. For example spelling errors were common. In these cases a term card can be created for an erroneous term that has been excessively used, so that the semantic enrichment makes the right ontological links, even though the database data is not corrected.

4 Annotation Creation

The last step in the content creation process is the semi-automatic annotation, which makes the data semantically interoperable. This can be done when the database contents have been transformed into coherent XML form, and the terminology mappings have been created.

In this paper, semantic interoperability means that the terms used in describing the data have to be interpreted semantically in a mutually consistent way. This is done by linking literal data values on the XML level, called *features*, to the ontological concepts on the RDF level. In practice, the string-valued features that are expressed in the shared XML syntax are transformed into the Uniform Resource Identifiers (URI) of the corresponding classes and individuals in the ontologies.

The features of the data items fall in two categories: *literal features* and *ontological features*. Literal features are to be represented only as literal values on the RDF level. They are, for example, used in the user interface. Ontological features are values that need to be linked not only to literal values but also to ontological concepts (URI).

The XML to RDF transformation can be done by algorithm 1. Each ontological feature is associated with a separate domain ontology by the property-domain mapping. For example, the material values of artifacts are found from a domain ontology of

Let X be a set of XML cards with literal features L and ontological features P, having values V (terms) ;
Let O be a set of ontologies ;
Let *Property-domain mapping* $d : P \rightarrow O$ map each ontological property to a domain ontology ;
Let *Terminology mapping* $t : V, O \rightarrow S$ map the XML card feature values V of the ontological property P to the classes and individuals S in O ;
Result: A set R of RDF triples.

$R := \varnothing$;
foreach *XML card* $x \in X$ **do**
  Create an RDF card instance i;
  foreach *feature* $f \in P \cup L$ *having value* v **do**
    $R := \{< i, f\text{-literal}, v >\} \cup R$;
    if $f \in P$ **then**
      $R := \{< i, f, s >\} \cup R$, where $s = t(v, o)$ is a collection of resources in the underlying domain ontology $o = d(f)$ so that s is found through terminology mapping;
    end
  end
end

Algorithm 1: Creating ontological annotations.

materials, place of usage feature values are found from a location ontology, and so on. This mapping can be used for disambiguating homonymous terms referring to resources in different ontologies. The algorithm creates for each XML card feature *f*, represented as an XML element, a corresponding RDF triple with a corresponding predicate name *f-literal* and a literal object value. For ontological features, an additional triple is created whose predicate name is the name of the feature and the object value consists of URIs to the possible resources that the literal feature value may refer to according to the terminology *t*.

Algorithm 1 is the basis of the semi-automatic annotation creation tool Annomobile (cf. figure 1) in MUSEUMFINLAND. Annomobile gets XML cards as input and produces the corresponding annotations in RDF format as output. The annotations follow an annotation schema that is expressed by an RDF Schema.

We have chosen fifteen different fields from the museum collection data records to be shown in the portal to the end-user. Nine of these features are ontological and hence linked to domain ontologies during the annotation process. The nine ontological features and their ranges, i.e. the seven domain ontologies to which the features are linked to, are presented in table 1. The ontologies (ranges) define the domains on which the term disambiguation is based on. The ontological features and domain ontologies are described in some more detail in [8].

When mapping ontological feature values to URIs in domain ontologies, two problem situations may occur:

Ontological feature	Ontology/Range	Ontological feature	Ontology/Range
Object type	Artifacts	**Material**	Materials
Creator	Actors	**Location of creation**	Locations
Time of creation	Times	**User**	Actors
Location of usage	Locations	**Situation of usage**	Situations
Collection	Collections		

Table 1. The nine ontological features of collection items and seven ontologies used in MUSEUMFINLAND.

Unknown values. The feature value may be unknown, i.e. there are no applicable term card candidates in the terminology. The solution to this is to map the feature value either to a more general term, e.g. to the root of the domain, or to an instance that represents all unknown cases. For example, if one knows that an artifact is created in some house in the city of Helsinki, but the address is unknown, one can create an instance called "unknown house" which is part of Helsinki and annotate the item with this instance.

Homonyms. The problem of homonymous terms occurs only when there are homonyms within the content of one domain ontology. The simple solution employed in our work is to fill the RDF card with all potential choices, inform the human editor of the problem, and ask him to remove the false interpretations on the RDF card manually. Our first experiments seem to indicate, that at least in Finnish not much manual work is needed, since homonymy typically occurs between terms referring to different domain ontologies. However, the problem still remains in some cases and is likely to be more severe in languages like English having more homonymy.

Table 2 shows some statistical results were obtained from the annotation process of building MUSEUMFINLAND. The content material came from three heterogeneous collection databases in three different museums. The number of collection items in the material totaled 6046, and every item had nine fields on average that needed to be linked to ontological concepts through the annotation process. All these nine fields could contain multiple literal values, all of which should be linked to different ontological concepts. For example, the place of usage field could contain several location names.

The table indicates that homonyms do not occur too often in the data. It can be seen also that in most cases the homonyms belong to different domains. Hence, the simple disambiguation scheme based on feature value domains worked well in practice and not much human editing was needed after using Annomobile.

	Museum 1	Museum 2	Museum 3
Total of annotated museum items	1354	1682	3010
Items with homonyms (total)	567	388	448
Items with homonyms disambiguated	424	332	334
Items with homonyms not disambiguated	143	56	114

Table 2. Results from annotating data with Annomobile in MUSEUMFINLAND.

5 Discussion

5.1 Lessons Learned

A general problem encountered in the content work was that the original museum collection data in the databases was not systematically annotated. Various conventions are in use in different museum systems and museums. Automatic annotation was relatively easy when descriptions in the database tables are done in a consistent manner using thesauri and without inflecting words. However, the descriptions in many cases were given in more or less free text. For example, use of free text was common in the data fields describing the techniques by which the artifacts were created. Furthermore, individual catalogers have used different terms and notations in cataloging. To handle these cases, the free text was tokenized into words or phrases which were then interpreted as keywords. This approach works, if term cards with ontological links are created from these keywords, and was adopted to both Terminator and Annomobile. The drawback here is, that if the vocabulary used in the free text is large, also the number of new term cards will be high and the manual workload in their annotation will be considerable. The vocabulary used in the MUSEUMFINLAND case, however, mostly conforms to the entries in the Finnish cultural thesaurus MASA, and this approach seems to be feasible. The homonymy problem is most severe in general free text description fields, since they are most prone to consist of conceptually general data where disambiguation cannot be based on the ontology to which the text field is related. Nonetheless, the Terminator and Annomobile tools proved out to be decent programs, annotating the data well enough for the purposes of the project.

5.2 Related Work

Lots of research has been done in annotating web pages or documents using manual or semi-automatic techniques and natural language processing. CREAM and Ont-O-Mat [1] and the SHOE Knowledge Annotator [3] are examples of such work.

Stojanovic et al. [14] present an approach that resembles ours in trying to create a mapping between a database and an ontology, but they haven't tackled the questions of integrating many databases or using global and local terminology to make the mapping inside a domain. Also [2] addresses the problems of mapping databases to ontologies, but their way of doing the mapping is very different from ours; in deep annotation the data is kept in the database, and the data is dynamically fetched from the database. Also, in our process we annotate the data through terminology, while deep annotation uses the database structure.

Also others have used the distinction of different layers of domain data and knowledge (figure 2). In [13] the concepts-terms-data model has been used to define different elements used for creating an ontology out of a thesaurus.

The idea of annotating cultural contents in terms of multiple ontologies has been explored also, e.g. in [4]. Other ontology-related approaches used for indexing cultural content include Iconclass[4][15] and the Art and Architecture Thesaurus[5] [11].

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

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

As far as we know, our annotation process is the first one to provide semantic enrichment through terminological interoperability among several content providers, and to the semantic extent described in this paper. The output of the process, i.e. the annotated museum collection items, have been published in a semantic web portal MUSEUMFINLAND for all Internet users to enjoy.

References

1. S. Handschuh, S. Staab, and F. Ciravegna. S-cream - semi-automatic creation of metadata. In *Proceedings of EKAW 2002, LNCS*, pages 358–372, 2002.
2. S. Handschuh, S. Staab, and R. Volz. On deep annotation. In *Proceedings of International World Wide Web Conference*, pages 431–438, 2003.
3. J. Hefflin, J. Hendler, and S. Luke. Shoe: A knowledge representation language for internet applications. Technical report, Dept. of Computer Science, University of Maryland at College Park, 1999.
4. L. Hollink, A. Th. Schreiber, J. Wielemaker, and B.J. Wielinga. Semantic annotation of image collections. In *Proceedings KCAP'03, Florida*, October, 2003.
5. E. Hyvönen, S. Saarela, and K. Viljanen. Application of ontology-based techniques to view-based semantic search and browsing. In *Proceedings of the First European Semantic Web Symposium, May 10-12, Heraklion, Greece*, pages 92–106. Springer–Verlag, Berlin, 2004.
6. E. Hyvönen, M. Junnila, S. Kettula, E. Mäkelä, S. Saarela, M. Salminen, A. Syreeni, A. Valo, and K. Viljanen. Finnish Museums on the Semantic Web. User's perspective on MuseumFinland. In *Selected Papers from an International Conference Museums and the Web 2004 (MW2004), Arlington, Virginia, USA*.
7. E. Hyvönen, M. Junnila, S. Kettula, S. Saarela, M. Salminen, A. Syreeni, A. Valo, and K. Viljanen. Publishing collections in the Finnish Museums on the Semantic Web portal – first results. In *Proceedings of the XML Finland 2003 conference. Kuopio, Finland*.
8. E. Hyvönen, M. Salminen, S. Kettula, and M. Junnila. A content creation process for the Semantic Web, 2004. Proceeding of OntoLex 2004: Ontologies and Lexical Resources in Distributed Environments, May 29, Lisbon, Portugal.
9. R. L. Leskinen, editor. *Museoalan asiasanasto*. Museovirasto, Helsinki, Finland, 1997.
10. E. Mäkelä, E. Hyvönen, S. Saarela, and K. Viljanen. OntoViews — a tool for creating semantic web portals. In *Proceedings of the Third International Semantic Web Conference, Nov 7-11, Hiroshima, Japan*. Springer–Verlag, Berlin, 2004.
11. T. Peterson. Introduction to the Art and Architechure thesaurus, 1994. http://shiva.pub.getty.edu.
12. V. Raatikka and E. Hyvönen. Ontology-based semantic metadata validation. Number 2002-03 in HIIT Publications, pages 28–40. Helsinki Institute for Information Technology (HIIT), Helsinki, Finland, 2002. http://www.hiit.fi/publications/.
13. D. Soergel, B. Lauser, A. Liang, F. Fisseha, J. Keizer, and S. Katz. Reengineering thesauri for new applications: the agrovoc example. *Journal of Digital Information*, (4), 2004.
14. L. Stojanovic, N. Stojanovic, and R. Volz. Migrating data-intensive web sites into the semantic web. In *Proceedings of the ACM Symposium on Applied Computing SAC-02, Madrid, 2002*, pages 1100–1107, 2002.
15. J. van den Berg. Subject retrieval in pictorial information systems. In *Proceedings of the 18th international congress of historical sciences, Montreal, Canada*, pages 21–29, 1995.