Publishing Collections in the "Finnish Museums on the Semantic Web" Portal – First Results

Eero Hyvönen*
University of Helsinki and HIIT

Miikka Junnila
HIIT and University of Helsinki

Suvi Kettula
HIIT and Espoo City Museum

Samppa Saarela
HIIT and University of Helsinki

Mirva Salminen
HIIT and University of Helsinki

Ahti Syreeni
HIIT and University of Helsinki

Arttu Valo
HIIT and University of Helsinki

Kim Viljanen
HIIT and University of Helsinki

ABSTRACT

This paper presents a scheme for publishing museum collections on the Semantic Web. It is shown how museums with their semantically rich and interrelated collection content could start creating large, consolidated semantic collection portals together on the web. By semantic web techniques, it is possible to make collections semantically interoperable and provide the museum visitors with intelligent content-based search and browsing services to the global collection base. The idea and its challenges are addressed through a real-world application, the first prototype of MUSEUMFINLAND, a semantic portal for Finnish museums to publish their collections on the Semantic Web.

1. THE VISION OF MUSEUMS ON THE SEMANTIC WEB

This paper argues that publication of museum collections on the web [8] is a very promising application on the Semantic Web[1] [1, 4]. Museums possess large amounts of digitized data and metadata in their collection databases. A special characteristic of cultural collection contents is semantical richness with an interconnected nature. Collection items have a history and are related in many ways to our environment, to the society, and to other collection items. For example, a chair may be made of oak and leather, may be of jugend style, was designed by a famous designer, was manufactured by a certain company during a time period, was used in a certain castle together with other pieces of furniture, and so on. Other collection items, locations, time periods, designers, companies etc. can be related to the chair through their properties and implicitly constitute a complicated semantic network and a knowledge base [20] of associations. Cultural, art, and other kinds of museums constitute an interlinked memory of our society covering all aspects of life. Semantic web metadata and ontology standards and tools, such as RDF [13] and RDF Schema (RDFS) [2] can be used for managing this versatility of concepts and associations in a novel way and for making it explicit for the museum visitors and researchers. Furthermore, the web makes it possible to consolidate data from distributed heterogeneous collections and publish it for anybody who has an Internet access.

To realize these ideas, we have developed a demonstrational prototype of a semantic web portal called "MUSEUMFINLAND — Finnish Museums on the Semantic Web". This system contains collections of cultural artifacts, such as textiles, pieces of furniture, tools etc. The contents come from the collections of the National Museum[2], Espoo City Museum[3], and Lahti City Museum[4]. These museums use three different relational database schemas, data base systems, and collection management systems (called Musketti, Escoll, and Antikvaria, respectively).

The main goals of developing the system are the following:

Global view to distributed collections It is possible to use the heterogeneous distributed collections of the museums participating in the system as if the collections were in a single uniform repository.

Content-based information retrieval The system supports intelligent information retrieval based on ontological concepts, not on simple keyword string matching as is customary with current search engines. For example, since Helsinki is a part of Finland, the concept "Finland" in the locational sense matches not only "Finland" but also "Helsinki".

Semantically linked contents A most interesting aspect

*Contact information for all authors:
Email: firstname.lastname@cs.helsinki.fi
Mail: P.O. Box 26, 00014 UNIV. OF HELSINKI, FINLAND
http://cs.helsinki.fi/group/seco/

[1] http://www.w3.org/2001/sw/,
http://www.semanticweb.org

Paper presented at the Salzburg Research Symposium, Arts and Humanities in the Digital Domain, Oct 6-7, 2003, Salzburg Research, Salzburg, Austria. Published in the Proceedings of XML Finland 2003, Kuopio, Finland.

[2] http://www.nba.fi
[3] http://www.espoo.fi/museo
[4] http://www.lahti.fi/Kulttuuri/museot

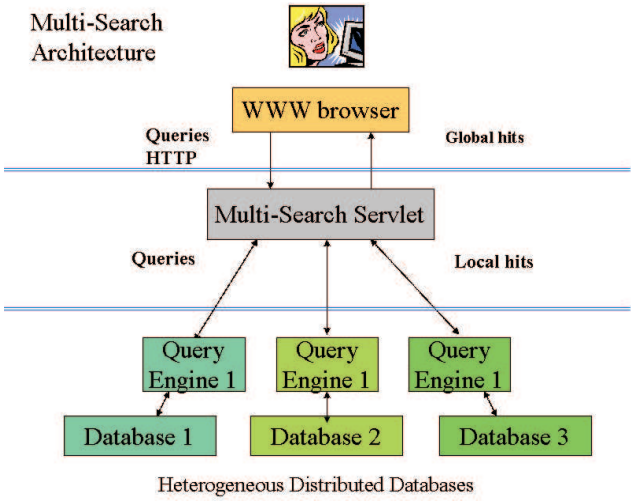

Figure 1: Multi-search architecture. The global query is answered independly at each local database.

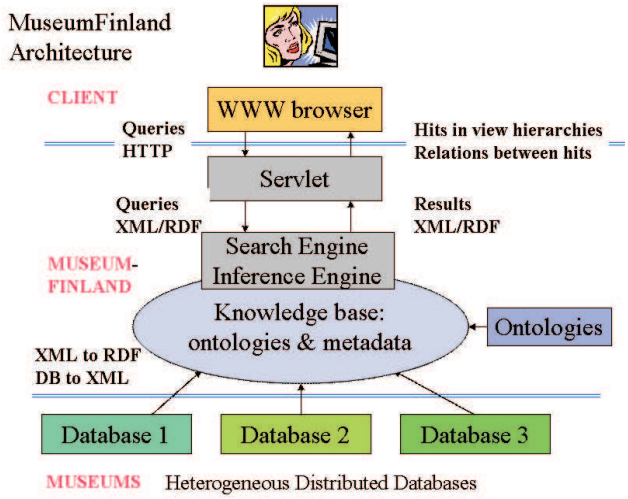

Figure 2: Information retrieval in MuseumFinland. Local database contents are first merged and the query is evaluated with respect to the global interrelated data.

of the collection items to the end-user are the implicit semantic relations that relate collection data with their context and to each other. For example, assume that a painting in an art collection depicts a castle and the collection of another museum includes artifacts from there. If the user views the painting, then (s)he is likely to be interested in having a look at the actual artifacts stored in the other museum's database. In MUSEUMFINLAND, such associations can be exposed to the end-user by defining them in terms of logical predicate rules that make use of the underlying ontologies and collection metadata.

Easy local content publication One of the key drivers behind the success of the WWW is the ease of content publication. Everybody can publish content easily and independently by just maintaining (HTML) files in a local public directory. This simple publishing practice through public directories is used in MUSEUMFINLAND, too.

In the following, these goals and solutions developed in our work are described by discussing them one after another. After this, main results of the work are summarized, lessons learned discussed, and directions for further research outlined.

2. GLOBAL VIEW TO DISTRIBUTED COLLECTIONS

Museums and their collection databases are usually situated at different locations. This creates an obstacle to information retrieval for both the public and for researchers. To address the problem, the web can be used for creating a single interface and access point through which a search query can be sent to distributed local databases and the results combined into a global hit list. This "multi-search" approach, as depicted in figure 1, is widely applied and there are many cultural collection systems on the web based on it, such as the portals Australian Museums Online[5] and Artefacts Canada[6].

A problem of multi-search is that by processing the query independently at each *local database*, the *global* dependencies, associations between objects in different collections are difficult to found. Since exposing semantic associations between collections items is one of our main goals, MUSEUMFINLAND cannot be based on the multi-search paradigm. Instead, the local collections are first consolidated into a global repository, and the queries are answered based on it (cf. figure 2). Mutually shared conceptual models, ontologies, are used for enriching the content and for making the collections interoperable. To show and make use of the associations, the collection items are represented as web pages interlinked with each other through the semantic associations. The MUSEUMFINLAND home page is the single entry point through which the end-user enters the virtual museum collections' WWW space. The system provides the user with a view-based search engine and a semantic searching and browsing facility in the combined collection knowledge base, as presented in [11, 10].

The architecture of MUSEUMFINLAND is depicted in figure 3. Museums join the system by producing collection metadata in RDF format. The metadata is placed in a public directory on the museum's WWW server, or is sent to the service provider. In order to combine the collection data of different museums into one logical virtual entity, the data must be made interoperable in both syntax and semantics.

2.1 Syntactic Interoperability

The database contents of the museums are first transformed into XML. Its syntax is defined by an XML schema that is shared by the co-operating museums. This guarantees mutual syntactic interoperability of museum collection data.

[5] http://www.amonline.net.au/
[6] http://www.chin.gc.ca

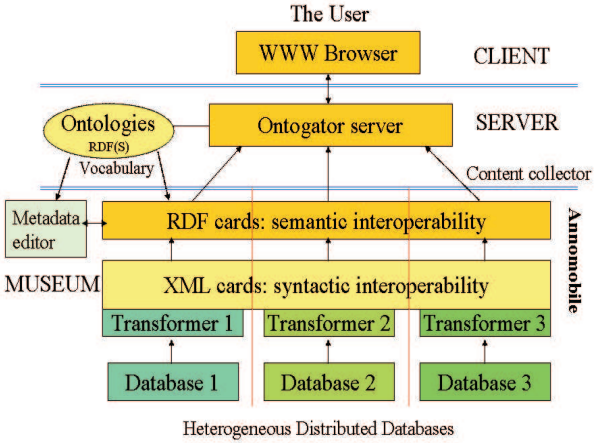

Figure 3: The architecture of the Finnish Museums on the Semantic Web system.

The XML schema tells what (meta)data must be provided for describing collection items. Based on the schema, each collection item has an XML description of its own called the *XML card*. For example, the XML card representing a calendar is presented below[7].

```
<artifactCard created="2003-7-29 10:43:16">
  <artifactId> ECM:22461:1 </artifactId>
  <artifactType> Christmas calendar,
        Finland's Scouters Assoc. </artifactType>
  <museum> Espoo City Museum </museum>
  <material> cardboard </material>
  <keywords>
    <keyword> Christmas </keyword>
    <keyword> calendar </keyword>
    <keyword> scouts </keyword>
  </keywords>
  <placeOfUsage> Tapiola, Espoo <placeOfUsage>
  <creator> Ulla Vaajakoski </creator>
  ...
  <photo> photos/image3451.jpg </photo>
</artifactCard>
```

The idea of the XML card is to present the main *features* of a collection item in a simple XML form. The features are represented in an XML card by subelements. For example, elements `artifactId`, `artifactType` etc. indicate features in the above XML card. The values of the features, such as the string "Espoo City Museum" in <museum> Espoo City Museum </museum>, are strings read form the underlying database tables.

From the syntactic viewpoint, there are several difficulties in creating the feature values. Below some problems are listed and the solution approaches taken in MUSEUMFINLAND outlined.

- Imprecise data. The information available is often imprecise in different ways. One has be able to make the distinction between the following cases: 1) The value is missing but existing, i.e., *unknown*. For example, the creator of a painting may be unknown. 2) The *value does not exist*. For example, a telephone machine may not have the artistic style feature at all. 3) The value is *uncertain*. For example, the manufacturing time of a chair may be during 1850-1870.

  In MUSEUMFINLAND, unknown values are represented by a special symbol, missing features are identified by empty (string) values, and uncertainty is represented by time intervals and by using more general classes as values. For example, class "metal" can be used for uncertain metallic material. A problem is that the distinctions between different forms of imprecise data have not been made systematically in collection data-databases.

- Complex values. The value of a property is often a combination of facts that may be stored in different database tables. For example, an artifact may have a genus name (e.g., "toy") with a species name ("Donald Duck toy"), additional colloquial names, and names in different languages. Such detailed information should not be lost. Complex values can be represented as collections of values but the problem is that the meaning of complex values is not always easy to interpret algorithmically.

- Dealing with typing errors. The information available is in many cases syntactically erroneous due to typing errors. This is a problem that should of course be solved already when cataloging the items, but errors occur and have to be dealt with. In MUSEUMFINLAND erroneus and unknown names can be identified later when transforming the XML cards into RDF. A log file is created indicating the problems encountered and a human editor makes the needed modifications.

2.2 Semantic Interoperability

In this paper, semantic interoperability means that the terms used in describing the collection data in different collections, i.e., the feature values of the XML cards such as "cardboard" as the material above, has to be interpreted semantically in a mutually consistent way. An idea widely employed in semantic web research is to use an ontology for defining the underlying concepts and then to define a semantic interpretation mapping from the terms and descriptions to concepts.

Semantic interpretation means in practice, that the XML card with its string-valued feature values is transformed into an *RDF card* with similar RDF properties, but where the string values are transformed into the Uniform Resource Identifiers (URI) of the corresponding classes and individuals in the ontologies. For example, the XML card above in RDF form is:

```
<rdf:RDF
  xmlns:rdf=
   "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:card=
   "http://www.fms.fi/RDFCard#">
 <card:RDFCard
   rdf:about="http://www.fms.fi/rdfCard#card11023">
   card:artifactId="16851"
   card:artifactType-www="calendar"
```

[7]The example is translated and slightly simplified from the original version in Finnish.

```
   card:artifactType=
    "http://www.fms.fi/artifacts#calendar"
   card:museum-www="Espoo City Museum"
   card:museum=
    "http://www.fms.fi/agents#EspooCityMuseum"
   card:material-www="cardboard"
   card:material=
    "http://www.fms.fi/materials#cardboard"
   ...
 </card:RDFCard>
...
</rdf:RDF>
```

The features of collection items fall in two categories: *literal features* and *ontological features*. The value x of each feature p in the XML card (e.g., that material is "cardboard") is represented by the corresponding *literal property* p-www=x in the RDF card (e.g., material-www="cardboard"). Literal property values will be shown in the user interface. In addition, each ontological feature in the XML card will be represented by an additional *ontological property* with same name in the RDF card. Its value is a URI that relates the card to the ontological RDF resources in the underlying knowledge base. For example, the feature `artifactId` is literal and is not connected with the ontology resources in the above RDF card. In contrast, the ontological feature "material" is represented with a literal property `www-material` and the ontological property `material` that has an RDF resource (URI) as its value. This URI connects the card resource with the material ontology and through it with other card resources.

The classes and individuals referred to in the RDF card are defined by a set of RDFS ontologies. Each ontological feature, such as "material", is associated with a *domain ontology* of its own. For example, artifact, material, and technique ontologies have been defined based on the Finnish MASA thesaurus [14] of keywords used in several museums for indexing data. This thesaurus was first transformed into RDF(S) from a database and has then been extended and edited by hand. The ontology now contains some 6600 classes organized in a taxonomy. There is also a location ontology that defines concepts, such as "country", "town" etc. Their instances are individual areas and locations (some 840 mostly Finnish places) and are related with each other by a part-of meronymy. Each place has a unique URI that can be used to disambiguate places with the same name. In the same way, an agent ontology defines concepts such as "person", "company", "museum" etc. whose instances are active individuals (some 1300 instances at the moment). Each agent has a unique URI that can be used to disambiguate agents with the same name. This ontology is used for the properties "creator" and "user" in the RDF card. There is also an ontology for time periods and an ontology of collections in different museums. Still another ontology of "activities and processes" contains a taxonomy of concepts such as "wedding", "fishing", and "sports". It is used to provide the end-user with an event-based view to cultural artifacts by associating them with corresponding events (e.g., "ring" with "wedding" and "net" with "fishing") through annotations and rules.

More formally, the XML to RDF transformation can be done by the following procedure:

Procedure XML2RDF

1. Input:
   (a) A set C of XML cards with literal features L and ontological features P.
   (b) A set O of ontologies.
   (c) *Property-domain mapping* $d : P \rightarrow O$ that maps each ontological properties to a domain ontology.
   (d) *Terminology mapping* $t : V, O \rightarrow S$ that maps the XML card feature values (terms) V of the ontological properties P to the classes and individuals S in O.
2. Output: A set R of RDF triples.
3. $R := \varnothing$
4. For each XML card $c \in C$ do
   (a) Create an RDF card instance i.
   (b) For each $p \in P \cup L$ having value v do
      i. $R := \{< i,$www-$a,v >\} \cup R$
      ii. If $p \in P$, then $R := \{< i,p,r >\} \cup R$, where $r = t(v, o)$ is a collection of resources in the underlying domain ontology $o = d(p)$.

The procedure is based on the terminology mapping t that defines how strings on the syntactic XML level are mapped onto semantic classes and individuals on the RDF level. In MuseumFinland, the terminology mapping is defined using a separate term ontology whose instances are term definitions, *term cards*. The properties of the term card include "concept", whose value is a resource of an ontology, and word forms, whose values tell the string forms of the term, such as the singular, plural, and appreviated forms. The term cards contain also other information related to term definition, maintenance, usage, etc. From the XML to RDF transformation view point, the term card is a tuple $< c, w_1, w_2, ...w_n >$ where an ontology resource c is associated with a set of word forms w_i.

The terminology mapping is applied in the XML2RDF procedure to determine the resource $r = t(v, o)$, where $o = d(p)$. This can be done by using the term cards and the following procedure:

1. Let $T = \{< c, w_1, w_2, \ldots w_n > |\ v \in \{w_1, w_2, \ldots w_n\} \wedge c \in o \wedge o = d(p)\}$ be the set of term cards containing the word form v.
2. Let $C = \{c\ |< c, w_1, w_2, \ldots w_n >\in T\}$ be the set of possible ontology resources corresponding to v.
3. Let $n = |C|$ be the number of members in $C = \{c_1, c_2, \ldots, c_n\}$.
4. If $n = 0$, then inform the user in a log file that an unknown term was found and that a new corresponding term card must created.
5. Else if $n = 1$, then $r = c_1$.
6. Else if $n > 1$, then $r = c$, where c is the URI of a collection of the RDF resources in $C = \{c_1, c_2, \ldots, c_n\}$. Inform the user in a log file file that the property p is homonymous with repect to the domain ontology and must be disambiguated by human intervention by editing the RDF card.

The procedure shows that the XML2RDF transformation cannot be done fully automatically due to unknown and homonymous terms. The problem of unknown terms can be solved by generating all needed term cards before running the XML2RDF transformation. This can be done as follows. First, generate an initial set of term cards. In our work, for example, we first generated some 6000 term cards from the MASA thesaurus by a special script implemented for the purpose. Second, given the XML cards of a museum database, sort out the unknown terms along the different ontological features p. Third, for each p generate partly filled term cards (with empty ontological URI values c). Fourth, a human editor maps the unknown terms with ontology concepts using the term cards in an ontology editor. In our work, we have used the Protégé-2000[8] ontology editor for editing the ontologies, term cards, and RDF cards. Its user interface is simple enough to be used by museum personnel that usually do not have programming skills.

The problem of homonymous terms occurs when there are homonyms within the contect of one domain ontology. The simple solution emloyed 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.

During the XML2RDF transformation, the XML cards are merged into the semantic RDF graph defined by the ontologies and the collection metadata. An important side effect of this process is *semantic enrichment* where new meaning is automatically added to the collection data in two ways. Firstly, the generic ontological relations defined once by the ontologist are automatically inherited from the ontological definitions to instance data. For example, if the class "benches", with related terms "bench" and "footstool", is given a category resource according to the Outline of Cultural Materials (OCM) classification [19] in the ontology, then all items cataloged as benches or footstools will have this classification as well. As a result, the museum cataloger does not have to provide the OCM classification or change the term convention she is using. Additional implicit semantic associations can be added into the knowledge base by logical rules. Secondly, semantic associations emerge automatically with other related collection item instances. For example, a particular bench from a museum A may have the same manufacturer as a footstool in another museum B, which may be an important piece of information to the user.

3. CONTENT-BASED INFORMATION RETRIEVAL

The enriched semantic RDF graph forms the underlying knowledge base on which intelligent services for the end-user can be created [3, 5]. MuseumFinland provides the user with two major services.

1. A *view-based search engine* that is based on the underlying concepts and ontologies instead of simple keywords.
2. A *semantic recommendation system* by which the user can find out explicit and implicit semantic associations within the global collection data, and use the associations for browsing the collections.

In this section the search engine is shortly discussed. Semantic recommendations are considered after this.

The view-based search engine used in MuseumFinland is a new server-based version of Ontogator [10]. It is based on the *view-based* search paradigm[9] [17, 6]. In view-based search, the idea is to organize the terminological keywords of the underlying database into orthogonal category taxonomies called *views*. Views are used extensively in the user interface in helping the user to formulate the queries. The user can express the query easily in the right terminology by selecting (sub)categories from the views. For example, by selecting "carpet" from an object type taxonomy and "silk" from a material taxonomy, silk carpets are found. Views are also useful in presenting the results in terms of categories and in navigating the underlying database. Ontogator is written in Java and Jena 2.0[10] and it is used by XML/RDF messages over HTTP protocol.

At the moment, there are nine view hierarchies in use. They are grouped under the four headings of "Artifact Characteristics" (object type and object material views), "Artifact Creation" (manufacturer, manufacturing time, and manufacturing location views), "Usage" (user, place of usage, and usage event views), and "Maintenance" (museum collection view). The view hierarchies are projected from the underlying ontologies used in annotating the collection data. The projection is based on logical predicates that define how the hierarchy is formed. This view projection method is described in [12]. For example, in the object type view the relations rdfs:subClassOf and rdf:type are used, while in the manufacturing location and place of usage views the part-of relation is used. SWI-Prolog[11] with its RDF parser [23] is used for creating the projections.

To illustrate the use of the MuseumFinland search system, figure 4 shows the user interface of our first experimental implementation. On the left in the window, the user makes selections from the nine view hierarchies. In the figure, two selections related to the usage ("Käyttö") has already been made: the place "Finland" was selected from the place of usage view ("Käyttöpaikka") and event fishing ("kalastus") was selected from the usage event view. The paths selected in the corresponding hierarchies are indicated by the symbol >. The items found, i.e., artifacts used for fishing in Finland, are seen on the right. Further refinement of the selections can be made by the dropdown lists that show the next level in each view hierarchy. The user curretly views the choices using the dropdown list in the object type view ("Esinetyyppi"). The possible choices here are: handicrafts ("käsityöt"), 8 hits; tools and hunting equipment, 27 hits; containers and their parts, 1 hit.

The views show to the user the concepts that can be used for searching. This is useful, because the end-user might not be familiar with the choices available. By showing proactively the number of hits for each possible next choice the user never ends up in dead-ends, where no hits are found—a

[8] http://protege.stanford.edu

[9] See http://www.view-based-search.com for a historical review of the idea.

[10] http://www.hpl.hp.com/semweb/jena.htm

[11] http://www.swi-prolog.org

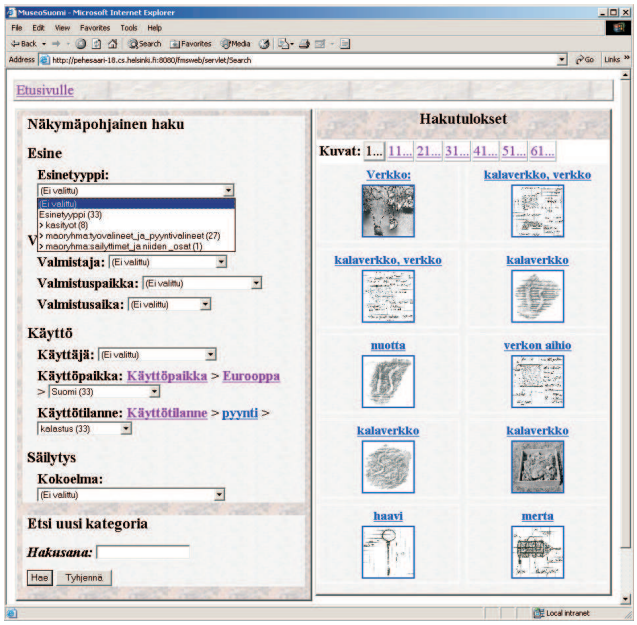

Figure 4: View-based search in MuseumFinland.

typical situation in conventional search engines. By opening the view hierarchies, the user chooses categories that are of interest, and the browser shows immediately all the collection items that fit the chosen constraints, and updates all category hit counts.

Finding relevant categories becomes a search problem of its own when dealing with thousands of categories. To help the user, the user interface also contains a search engine for finding view categories (header "Etsi uusi categoria" in figure 4. This engine is part of Ontogator as well but is based on matching keyword strings with view category labels. The labels may be given in alteranative forms (synonyms) and in alternative languages. The categories found are represented as a hit list of paths from view roots. By selecting a category from the list the corresponding selection in the view is made for view-based search.

4. SEMANTICALLY LINKED CONTENTS

One of the main goals of the MuseumFinland portal is to reveal the rich semantic linkage connecting the collection objects with each other. The links can be *explicit* or *implicit*. Explicit links correspond to the RDF statements (triples) in the underlying knowledge base and are based on the collection domain ontologies (classes and their properties) and the actual collection data (instance data). For example, an instance of a painting may have the RDF property dc:creator linking the art work to an individual artist. Implicit links can be defined in terms of explicit ones but are not present in the RDF graph. For example, if there are explicit links linking children with their mothers and fathers, then implicit links such as "grandfather" or "cousin" can be defined.

In MuseumFinland, implicit links are defined declaratively in terms of logic by using Prolog predicates. Each predicate defines a semantic association and gives it an explanatory label, such as "cousin of". By applying such a predicate to a collection item resource, implicitly related

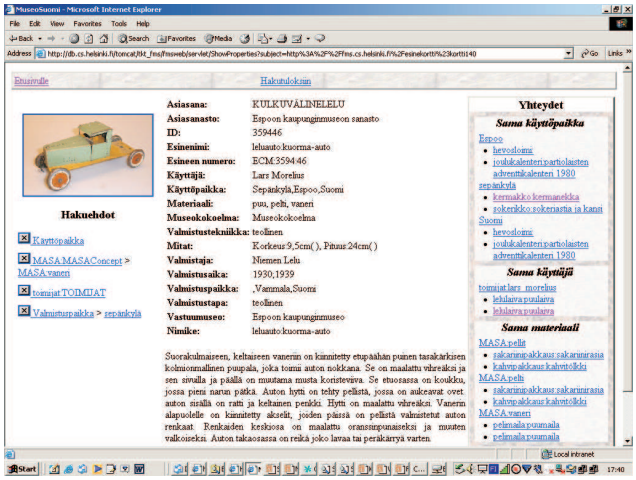

Figure 5: Collection item metadata with semantic recommendations.

other resources with respect to the semantic association can be found. On the HTML level in the user interface, the label of the association is used as the name for the link and the found resource as the target. For example, if the family relations of artists are known in the ontology, then such a predicate could infer links to other pages depicting paintings whose creator is of the same family.

An illustration of the idea is shown in the screen shot of figure 5. On the left, a collection item found with its metadata is shown. On the right, the system displays links to other recommended collection items. For example, collection items made out of the same material and used by the same person are recommended. Such links can be found by specifying the corresponding logical predicates based on the underlying RDF model. By clicking the association links, the user can browse seamlessly between the collection items from the different museums that provide the content.

The semantic recommendation system of MuseumFinland is implemented as a logic server called "Ontodella". This system is based on the SWI-Prolog HTTP server version. The MuseumFinland system itself is a Java servlet that queries with the Ontogator and Ontodella servers with XML/RDF messages over HTTP. The user interface is constructed using XSL transformations from the query results to HTML.

5. EASY LOCAL CONTENT PUBLICATION

A practical goal of our work is to design a *process* for Finnish museums to publish their collections on the Semantic Web. Museums are usually not very competent with information technology, do not necessarily have their own servers and software on the WWW, and are not willing to invest their money in expensive server side applications. The publication scheme should therefore be simple.

The responsibilities between the museums and the MuseumFinland portal maintainer are divide as follows:

Museum side Each museum is responsible for creating the data and metadata descriptions of the published col-

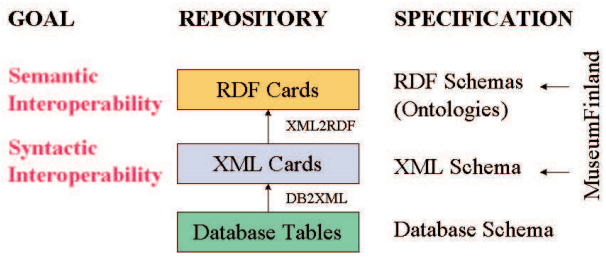

Figure 6: Data transformations from a museum database into RDF.

lection items. The descriptions are put in a public directory in the same way as conventional web pages. Of course, the contents can also be sent to MuseumFinland by other means, e.g., on a CD. In this way the internal museum database systems can be made independent from the MUSEUMFINLAND system. No special servers (other than a WWW server) or holes in firewalls are needed. The museum can see and control what data is being published in a very transparent way.

Portal side MUSEUMFINLAND system collects the data and metadata from the local museums, creates the global RDF knowledge base, and provides the end-users with the searching and browsing facilities on the web.

Figure 6 depicts the transformation process from a museum collection database into RDF from the museum perspective. The museum first transforms its collection data into XML cards. This transformation is dependent on the museum database in use and needs custom programming. The transformation can be made relatively easily in a few weeks. For our first prototype, we implemented the transformations for three different database systems. Once implemented, the transformation can be repeated later with new data. This transformation is discussed in more detail in [18].

After this, XML cards are transformed into RDF as discussed in the previous sections. For the XML to RDF transformation, we have implemented a special tool called "Annomobile". Before publication, the RDF cards output by Annomobile are checked by using Protégé-2000 and by checking the warnings in the log files produced during the transformation.

6. DISCUSSION

This paper presented an overview of MUSEUMFINLAND, a system and publication channel for making heterogeneous Museum collection databases semantically interoperable on the web. Techniques and procedures for ensuring syntactic and semantic interoperability by data transformations were outlined.

6.1 Benefits of the ontology approach

The power of the system lies in the use of ontologies:

Exact definitions By using ontologies, museums can define the concepts used in cataloging in a precise, machine understandable way.

Terminological interoperability The terms used in different institutions can be made mutually interoperable by mapping them onto common shared ontologies. The ontologies are not used as a norm for telling the museums what terms to use, but rather to make it possible to tolerate terminological variance as far as the terminology mapping from the local term conventions to the global ontology is provided.

Ontology sharing Ontologies provide means for making exact references to the external world. For example, in MUSEUMFINLAND, the location ontology (villages, cities, countries, etc.) and the actor ontology (persons, companies, etc.) is shared by the museums in order to make the right and interoperable references. For example, two persons who happen to have the same name should be disambiguated by different URIs, and a person whose name can be written in many ways, should be identified by a single URI to which the alternative terms refer.

Automatic content enrichment Ontological class definitions, rules, and consolidated metadata enrich collection data semantically.

Intelligent services Ontologies can be used as a basis for intelligent services to the end-user. In MUSEUMFINLAND, the view-based search engine is based on the underlying ontological structures and the semantic link recommendation systems reveals to the end-user the underlying semantical context of the collection items and their mutual relations.

6.2 Related Work

MUSEUMFINLAND is a novel adaptation of the idea of semantic portals to solving interoperbility problems of museum collection databases when publishing their content on the Semantic Web.

The novelty of the view-based search engine of MUSEUMFINLAND with respect to other view-based systems [17, 6] lies in its capability of using RDF(S) ontologies as the basis of search. The main benefits obtained are: 1) Ontogical logical inference can be employed in projecting the views from the ontology (e.g., the location meronymies and hyponymies). 2) The implicit complicated relations between view categories and the underlying data resources to be searched for can be specified flexibly in terms of logical predicates. The idea of Ontogator is to combine virtues of the view- and ontology-based search paradigms [10].

This idea of linking collection items with semantic associations is related to Topic Maps [15]. However, in our case the links are not given by the topic map but are determined by logical inference using the underlying RDFS ontology and RDF metadata. Another application of this idea to generating semantically linked static HTML sites from RDF(S) repositories is presented in [12]. In the HyperMuseum [21], collection items are also semantically linked with each other. Here linking is based on shared words in the metadata and their linguistic relations, such as synonymy and antonymy. In contrast, our system is not based on words but on ontological references in the underlying RDF(S) knowledge base and the links can be defined freely in terms of logical rules. The idea of annotating cultural artifacts in terms of multiple ontologies has been explored e.g. in [7].

6.3 Lessons learned

Several practical problems were encountered in transforming the database contents into RDF. Even if the XML card is syntactically well-formed, several semantic interpretation problems have to be addressed during the XML to RDF transformation.

The values of the features in XML cards may be complicated expressions. For example, value "Christmas calendar, Finland's Scouters assoc." is not a term but a complex phrase. The same concept may be referred to with different syntactic expressions (e.g., "Scouters' Christmas calendar") depending on the cataloger and notational conventions used. Using standard terminology in cataloging would help in solving this problem but in practice this is impossible, and there will be variation in desciptions. In MUSEUMFINLAND, each terminological variant expression has to be defined as a separate term definition, term card, that essentially maps a string with a concept. No deeper syntactic analysis is performed.

The XML feature value may also have several components. For example, the type of collection item may have many descriptive components. When determining the right URI reference, all data components must be consired together. Human intervention may by needed to check the right or best interpretation depending on the situation.

Difficult semantic problems arise when two terms are only partly synonymous or homonymous or map only partly on the ontology concepts. An example of this is the English term "river", meanings a large stream, and the French term "riviere", meaning a large or small river running to another river. In MUSEUMFINLAND, the meaning of the terms is not analyzed beyond determining references to ontological concepts. Each semantically different concept should be defined in an ontology and have a unique URI. This approach means in practice that semantic accurary must sometimes be compromized.

In many classification systems, such as ICONCLASS[12] [22] and Art and Architecture Thesaurus (AAT) [16], the terms are given as property values of the ontology classes. We believe that by using separate term cards the separation between linguistic word forms and concepts is clearer, and it is easier to extend an ontology with new terms, be they local terms used in different museums or complete terminologies in other languages. Modifying the terminology can be done at the museums locally by modifying term cards alone and without changing the shared global ontology, which would affect the other museums work as a side effect.

The RDF(S) knowledge base of our first prototype of MUSEUMFINLAND includes 4100 RDF cards from the collections of the National Museum, Espoo City Museum, and Lahti City Museum. This prototype shows that MUSEUMFINLAND concept in general is feasible and that the technology scales up at least to the order 10.000 of cards and view categories. The response times for search queries have typically been under 2 seconds on an ordinary PC server.

6.4 Further work

However, further research is still needed. More content analysis work is needed in developing a set of recommendation predicates that would be of most interest to the users. It is possible that their implementation would require changes in the ontologies and better annotated content. The XML2RDF transformation cannot be fully automated due to problems of homonymy and emergence of new terms and concepts with new collection items. To solve the problem, the cataloging systems should be enhanced with ontology support. Ways of collaboration between museum content providers and portal maintenance people need to be developed in order to develop MuseumFinland from an application into a continuos publication process for the participating museums. For example, protocols for adding, modifying, and retracting RDF cards and ontology resources according to the wishes of the museums need to be developed. The current user interface is intended for demonstrating the ideas behind MUSEUMFINLAND and is being developed further for a public service.

In the near future we plan to extend the collections of the system with paintings and graphics from the MUUSA database of the Finnish National Gallery[13]. MUUSA is capable of exporting the contents in an RDF format conforming to the CIDOC CRM[14] ontology. We also plan to incorporate in the system a database from the National Museum describing the most valuable cultural sites in Finland. Our goal is to show how RDF can be used as the basis for making very different kind of contents semantically interoperable. This requires development of art ontologies and aligning them with the other ontologies.

[12] http://ww.iconclass.nl

Acknowledgements

Thanks to Vilho Raatikka for help in the database to XML transformations and discussions. The Espoo City Museum, National Museum, Lahti City Museum, and Finnish National Callery made their collections available for our research. Our work is funded mainly by the National Technology Agency Tekes, Nokia, TietoEnator, the Espoo City Museum, the Foundation of the Helsinki University Museum, the National Board of Antiquities, and the Antikvaria Group consisting of some 20 Finnish museums.

7. REFERENCES

[1] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.

[2] D. Brickley and R. V. Guha. *Resource Description Framework (RDF) Schema Specification 1.0, W3C Candidate Recommendation 2000-03-27*, February 2000. http://www.w3.org/TR/2000/CR-rdf-schema-20000327/.

[3] D. Fensel, J. Angele, S. Decker, M. Erdman, H. Schnurr, S. Staab, R. Studer, and A. Witt. On2broker: semantic-based access to information sources at the WWW. In *World bonference on the WWW and Internet (WebNet 99)*, 1999.

[4] D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, editors. *Weaving the Semantic Web*. The MIT Press, 2002.

[5] D. Fensel, F. van Harmelen, M. Klein, H. Akkermans, J. Broekstra, C. Fluit, J. van der Meer, H. Schnurr, R. Studer, J. Hughes, U. Krohn, J. Davies, R. Engels, B. Bremdal, F. Ygge, T. Lau, B. Novotny, U. Reimer, and I. Horrocks. On-to-knowledge: ontology-based

[13] http://www.fng.fi

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

tools for knowledge management. In *eBusiness and eWork, Madrid, Spain*, 2000.

[6] M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, and K.-P. Lee. Finding the flow in web site search. *CACM*, 45(9):42–49, 2002.

[7] L. Hollink, A. Th. Schereiber, J. Wielemaker, and B.J. Wielinga. Semantic annotations of image collections. In *Proceedings KCAP'03, Florida*, October, 2003.

[8] E. Hyvönen, S. Kettula, V. Raatikka, S. Saarela, and Kim Viljanen. Semantic interoperability on the web. Case Finnish Museums Online. In Hyvönen and Klemettinen [9], pages 41–53. http://www.hiit.fi.

[9] E. Hyvönen and M. Klemettinen, editors. *Towards the semantic web and web services. Proceedings of the XML Finland 2002 conference. Helsinki, Finland*, number 2002-03 in HIIT Publications. Helsinki Institute for Information Technology (HIIT), Helsinki, Finland, 2002. http://www.hiit.fi.

[10] E. Hyvönen, S. Saarela, and K. Viljanen. Ontogator: combining view- and ontology-based search with semantic browsing. In *Proceedings of the XML Finland 2003 conference. Kuopio, Finland*, 2003. http://www.cs.helsinki.fi/u/eahyvone/publications/xmlfinland2003/yomXMLFinland2003.pdf.

[11] E. Hyvönen, A. Styrman, and S. Saarela. Ontology-based image retrieval. In Hyvönen and Klemettinen [9], pages 15–27. http://www.hiit.fi.

[12] E. Hyvönen, A. Valo, K. Viljanen, and M. Holi. Publishing semantic web content as semantically linked html pages. In *Proceedings of XML Finland 2003, Kuopio, Finland*, 2003. http://www.cs.helsinki.fi/u/eahyvone/publications/xmlfinland2003/swehg_article_xmlfi2003.pdf.

[13] O. Lassila and R. R. Swick (editors). Resource description framework (RDF): Model and syntax specification. Technical report, W3C, February 1999. W3C Recommendation 1999-02-22, http://www.w3.org/TR/REC-rdf-syntax/.

[14] R. L. Leskinen, editor. *Museoalan asiasanasto*. Museovirasto, Helsinki, Finland, 1997.

[15] Steve Pepper. The TAO of Topic Maps. In *Proceedings of XML Europe 2000, Paris, France*, 2000. http://www.ontopia.net/topicmaps/materials/rdf.html.

[16] T. Peterson. Introduction to the Art and Architechure Thesaurus, 1994. http://shiva.pub.getty.edu.

[17] A. S. Pollitt. The key role of classification and indexing in view-based searching. Technical report, University of Huddersfield, UK, 1998. http://www.ifla.org/IV/ifla63/63polst.pdf.

[18] V. Raatikka and E. Hyvönen. Ontology-based semantic metadata validation. In Hyvönen and Klemettinen [9], pages 28–40. http://www.hiit.fi.

[19] P. Sihvo, editor. *Kulttuuriaineiston luokitus. Outline of cultural materials*. Museovirasto, Helsinki, Finland, 1996.

[20] J. Sowa. *Knowledge Representation. Logical, Philosophical, and Computational Foundations*. Brooks/Cole, 2000.

[21] Peter Stuer, Robert Meersman, and Steven De Bruyne. The HyperMuseum theme generator system: Ontology-based internet support for active use of digital museum data for teaching and presentations. In D. Bearman and J. Trant, editors, *Museums and the Web 2001: Selected Papers*. Archieves & Museum Informatics, 2001. http://www.archimuse.com/mw2001/papers/stuer/stuer.html.

[22] 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. http://www.iconclass.nl/texts/history05.html.

[23] J. Wielemaker, A. Th. Schereiber, and B. J. Wielinga. Prolog-based infrastructure for RDF: performance and scalability. In *Proceedings ISWC'03, Florida*. Springer–Verlag, Berlin, October, 2003.