Semantic Interoperability on the Web: Case Finnish Museums Online

Eero Hyvönen[1,2], Suvi Kettula[2,3], Vilho Raatikka[2,1], Samppa Saarela[2,1], and Kim Viljanen[2,1]

[1] University of Helsinki, Department of Computer Science
firstname.lastname@cs.helsinki.fi,
http://www.cs.helsinki.fi/group/seco/
[2] Helsinki Institute for Information Technology (HIIT)
[3] Espoo City Museum

Abstract. This paper discusses the problem of making the contents of heterogeneous databases semantically interoperable on the web. The problem is addressed through a real-world case study by considering the new possibilities and challenges that the WWW gives to museums when publishing their collections online. We argue that interoperability at syntactic, semantic, and pragmatic levels is needed in order to combine the collections logically and to provide the museum visitor with useful search and navigation services into the semantically rich cultural contents. A technical solution to the problem is proposed based on semantic web technologies and a demonstrational implementation that combines two Finnish museum databases is discussed.

1 Museums on the web: a new challenge

The computer screen cannot replace the experience of visiting a museum itself, but can offer several benefits when studying the collections:

Access to larger collections Large collections that cannot be put on display physically can be shown to the public using the computer.

Usability of information Information technology enables the retrieving, combination, and display of information in ways that are technically or economically impossible in traditional exhibition spaces.

Interactivity and audiovisuals Information technology provides interactivity and can be used to extend and enrich the collection material with images, sound, and video material.

The World Wide Web (WWW) [2] can offer still another complement to the traditional museum visit. In order to study the collections, the visitor does not have to visit the museum during certain opening hours at a specific location.

There are several technical challenges in developing the WWW channel for transmitting collection contents to the public. Cataloging, annotating, and maintaining the data contents for WWW publication already poses a challenge to any single museum. This problem is enhanced further when using the WWW for combining collection data from the databases of several museums. To the user such a global system would, however, be very helpful for many reasons:

- The distribution of collection objects in museums at different locations creates an obstacle to information retrieval, for both the public and for researchers.
- One does not have to bother how the different museums store their collection data and what particular vocabularies (e.g., MASA[4] [13] or museum specific word lists) and classification systems (e.g., Outline of Cultural Materials[5] [18]) are used.
- The collections of different museums can be studied through a single interface and access point, which simplifies usage.

From the information retrieval viewpoint, it should be possible to use the collections as if they were in a single database.

The problem of making heterogenous databases homogeneous is an old one. A difficulty there is how to make the different database schemas mutually interoperable. In schema integration [1, 17], the idea is to construct a global homogeneous view of independently developed database schemas. The global repository is created before querying it. Another possibility is to keep the schemas as they are and integrate the database systems by a central application that uses a general information retrieval protocol. The data integration is done in a lazy fashion during the query evaluation time. For example, the Z39.50[6] protocol is used by museums and libraries for searching and retrieving information from remote databases.

Using these approaches in providing the user with a homogeneous view to collection data means that the databases are integrated either in a tight fashion physically or through the application using the network. In this paper, a more loosely coupled approach will be proposed instead. The collection contents are published in the local public directories of the museums in the same way as ordinary web pages. A web crawler is used to fetch and index the contents and a search facility is provided to the users as a server-side application. For example, search engines such as Google[7] and AltaVista[8] are based on this idea. A major benefit of such a loosely coupled system is that that the content providers can publish and maintain their contents independently from the service provider. The idea of publishing one's data on the web independently in this fashion is one of the key points underlying the whole WWW.[9]

[4] http://www.nba.fi/DEVELOP/asiasana.htm
[5] http://www.yale.edu/hraf/ocm_list.html
[6] http://lcweb.loc.gov/agency/
[7] http://www.google.com
[8] http://www.altavista.com
[9] A recent development in this direction on the web is Web Services [5], the idea of publishing functional services on the web for the machines to use.

In addition to making the database schemas interoperable, semantic interoperability with respect to content is needed. For example, the different vocabularies used in the collection data must be made mutually coherent. This issue is not addressed by the traditional schema-based approaches above in which explicit semantic descriptions of the content do not exist.

In this paper we address the problem of making existing distributed collection databases mutually interoperable on the semantic level. We argue that semantic web technologies offer a promising approach to facilitating homogeneous, semantic information retrieval based on heterogenous databases on the web.

In what follows, the problem of logical combination of heterogeneous collections at syntactic, semantic, and pragmatic levels is first discussed. Application of the web languages XML [3], RDF [12], and RDF Schema [4] in data and knowledge representation is discussed. After this, a system architecture for addressing the semantic interoperability problem is proposed, and the case implementation of our prototype, "Finnish Museums on the Semantic Web" (FMS), is discussed. To begin with, this system will integrate collection data from the heterogeneous collection databases of the Espoo City Museum[10] and the National Museum of Finland[11].

2 Logical combination of collections

In order to combine the collection data of different museums into one logical virtual entity, the data must be combined according to both syntax and semantics.

2.1 Syntactic interoperability

The museum databases are distributed (exist in physically different places) and heterogeneous, i.e., they 1) use different database systems by different manufacturers, and 2) their logical structure (schema, tables, fields, names, etc.) may vary. 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 presenting collection data is agreed upon. What is needed is a shared, mutually agreed presentation language for collection data.

In web applications, XML (eXtensible Markup Language) [3] has quickly been adopted as a formal way to agree on data representation languages. XML is a meta-language, a language to define languages. With the help of XML, different communities can easily and accurately define their own domain specific presentation languages. XML is penetrating the museum information systems,

[10] http://www.espoo.fi/museo
[11] http://www.nba.fi

as well. For example, the SPECTRUM cataloging system of the mda[12] has an XML specification that is being tested during 2002 by the Consortium for the Interchange of Museum Information CIMI[13] (CIMI).

Using XML gives us the following advantages:

Open standards XML data is simple text. The format is not restricted to the products of certain manufacturers, although major manufactures and communities are committed to using and supporting it. There are and will be several tools available for XML data including free public domain software. The use of such open standards makes it easier to change from using one commercial museum system to another and to combine different systems with each other.

Transporting collection data Collection data is to be preserved and collected for a long time. This poses a conflict to the hectic world of information technology, where enterprises and products come and go at a fast rate. Thus, museums have to change and transfer their databases to new systems time and time again. With open standards, this can usually be accomplished both flexibly and economically.

Combining the collections Open standards enable the combination of data from different collections. Depending on the situation, several collections can be physically joined into one database, or the use of several collections can be integrated over the web with the help of common data representation languages and data transmission practices.

Multi-channel publishing With XML it is easy to produce different manifestations of the collection data, e.g. both for the WWW and in print (XSLT transformations).

With the help of XML, museums can agree on a mutual standard for recording collection data, regardless of the collection system in use. When the museums have agreed on the common language, the transmission, combination, and WWW publishing of the collections becomes significantly easier.

In the FMS system, the combination of museum data is based on a common XML language at the structural level. This language is used to express the collection data to be published on the WWW. As a first step towards such an XML language, an FMS XML Schema specification has been designed [16].

2.2 Semantic interoperability

XML is a modern solution for combining data on the level of syntax. For a deeper combination and integration of data, the terminology used in the collection data has to be united in meaning, i.e., with respect to semantics. For example, the same kind of piece of furniture might be classified as a "footstool" in one museum, while it is called a "bench" in another. Thus, a search for "bench" would leave out the items classified as "footstools".

[12] http://www.mda.org.uk, previously known as the Museum Documentation Association.

[13] http://www.cimi.org/wg/xml_spectrum

Mapping terminologies with ontological concepts One of the main semantic relations that is used in controlled vocabularies and thesauri [9], such as the WordNet[14] [6] is the relation between super- and subordinate meanings (hyponymy), e.g. the fact that "footstools" are a kind of "furniture". If this semantic relationship is known, a search for "furniture" should find "tables" as well as "chairs," "cabinets" and other types of furniture. Meronymy, for its part, is the semantic relationship between parts and a whole, e.g., in the way that a "table top" is part of a "table," or "Helsinki" is part of "Finland." In addition, the collection data contains many straightforward semantic relationships. For example, the "creator" relationship may connect a collection object to a certain "manufacturer" or "artisan", and the relationship for the "place of production" to a certain "location". Such relationships can be defined formally by ontology languages that are being developed by the World Wide Web Consortium W3C[15] and others.

In order to combine collection data semantically, an exact definition, the ontology, of physical object classes and other concepts, as well as the relationships between them, is needed. For example, the International Council of Museums (ICOM)[16] develops a general ontology called CIDOC object-oriented Conceptual Reference Model (CRM)[17] for the documentation of cultural heritage. In addition to such an ontology, one needs general thematic ontologies of place, time, style etc. and application domain specific ontologies focusing on particular areas of collections, such as textiles or furniture.

When defining ontologies in WWW applications, a natural "standard" choice is the RDF recommendation [12] and RDF Schema specification [4] of the W3C. These languages are syntactically based on XML[18] but their semantic data model is a labeled directed graph, a semantic network of relationships between web resources and literal data items. RDF is a way to present metadata as named properties of the data. With RDF Schema we can agree on the terminology and constraints to be used in the RDF descriptions. The first version of the FMS system is based on RDF(S) semantics when presenting ontologies and metadata.

In practice, the semantic combination of data in the FMS means that the linguistic expressions that stand as data field values on the structural level, i.e., as the values of the XML elements, are mapped onto the classes of the general ontology. There are two basic problems involved: 1) Mapping synonymous terms on the same concepts. 2) Disambiguating polysemous terms. Both tasks are handled in the FMS by a metadata editor tool called Meedio [16] before the data is published. Meedio gets as input collection records conforming to the FMS XML Schema, and transforms them into RDF format (to be read in by the web crawler of the FMS system). During the interactive Meedio transformation, the values in the XML data elements are mapped on the corresponding semantic concepts

[14] http://www.cogsci.princeton.edu/~wn/
[15] http://www.w3.org
[16] http://www.icom.org/
[17] http://cidoc.ics.forth.gr/
[18] Other syntaxes could be specified as well.

defined by an RDFS ontology. The transformation is based on associating ontological classes with linguistic synonymous words and expressions.

In synonym mapping, terms are simply mapped on concepts. For example, the class `Benches`[19] may have the property `synonyms` and the set {`footstool`, `bench`} as its value. Assume then that the FMS-system finds a collection data record of a "bench" of the form

```
<collectionObject>
   <objectName>bench</objectName>
   ...
</collectionObject>
```

from the web site of a museum and a "footstool" record

```
<collectionObject>
   <objectName>footstool</objectName>
   ...
</collectionObject>
```

from the web site of another museum. With the help of the ontology, these two objects can be recognized as instances of the same ontology class `Benches` resulting into two RDF instances of the same `type`:

```
<rdf:Description rdf:about='URI of the collection object'>
   <rdf:type rdf:resource='http://www.fms.fi/furniture#Benches'/>
   ...
</rdf:Description >
```

In this way the different terminologies used in different museums by different catalogers can be integrated (assuming that the terms have the same meaning in different organizations).

In a polysemous situation the same term refers to several different concepts. For example, the word "chip" refers to a squirrel species, a semiconductor component, a kind of coin, and a piece of wood. Meedio handles semantic disambiguation by asking the user to make the choice whenever it finds a disambiguous term-class mapping according to the ontology.

Enriching the semantic content When a data value in an XML element, e.g. term "footstool" above, is attached with a class in the ontology, the collection record is merged into the semantic RDF graph defined by the ontology and other collection instance data. An important side effect of this process is semantic enrichment where new meaning is automatically added to collection items in two ways. Firstly, the generic ontological relations defined once by the ontologist are automatically inherited from the ontological definitions to instance

[19] We use English names and symbols in our examples, but the actual names in the FMS system are in Finnish.

data. For example, if the class "Benches" is given a category according to the OCM classification [18] in the ontology, then all footstools and benches will have this classification as well. As a result, the museum database or the museum cataloger does not have to provide this piece of information for each item in the class. Secondly, semantic associations emerge automatically with other related collection item instances. For example, a particular bench from a museum may have the same manufacturer as a footstool in another museum, which may be an important piece of information to the user.

The enriched semantic RDF graph forms the underlying database on which services for the end-user can be created, an approach similar to some earlier semantic web systems, such as [7, 8]. In the WWW there are two basic ways of retrieving information: by browsing pages through links and by generating hit lists by search engines. In the FMS system, a semantic WWW browser for such data is being developed as well as an ontological search engine for the Finnish language.

2.3 Pragmatics of collection objects

During cataloging, information on the collection objects such as object type and description, history of usage, provenience etc. is added into the database. With the help of controlled vocabulary keywords and classifications, contextual metadata is added to the item in order to enhance information retrieval facilities later on. On the structural level, data fields like object type, donator etc. can be identified and connected to the item. On the semantic level, names that stand as the values of the data fields can be connected with the semantic concepts defined by the ontologies. In linguistics, the next level after semantics is pragmatics describing the usage and purposes of language meanings. Museum collection data have their pragmatic dimension, too.

In our view, pragmatics of collection objects tell how the exhibits are used and in what roles they participate in different processes, events, and skills of life. For example, let us consider the process of "fishing". This process is associated with different instruments, agents, places, and time. A "fish net" and a "boat", for example, are connected with "fishing" as instruments. The "fishing" event or skill thus offers the possibility to relate different objects, agents, locations, and time with each other. From the point of view of the museum visitor, this kind of pragmatic level is especially interesting, because such context data "brings the objects to life".

Pragmatic processes are intangible abstract events. Until now they have not been considered as collection objects of their own to be stored in databases. A fishing event can be recorded on video, or a documentary can be written on the production of linen but they are stored as pieces of documentation, not as abstract processes. With semantic web technologies it is possible to design ontologies that describe the processes, and data can be represented and stored according to those ontologies. By storing processes in addition to objects in museum databases, it would become possible to create a new approach to retrieving and perceiving exhibit data. The processes where a physical collection object is used in different roles could be found automatically through the process ontology. This gives the user the possibility to find semantic links to other concepts and items related by the processes. For example, from "fishing net" we can move forward to "fishing processes" and from there to "fishing boats", "fishermen", and other fishing topics. In the user interface, these associations can be visualized, for example, as normal hypertext links when browsing the collection on the WWW in the same way as is Topic Maps [14].

3 Finnish museums on the semantic web

In order to realize the goal presented in the previous section, a research and development project[20] is being carried out at the Helsinki University Computer Science Department and the Helsinki Institute for Information Technology (HIIT). As a case study, the collection databases of the Espoo City Museum and the National Museum of Finland are used. (Other museums may join the system later.) These databases use different DBMSs (SQL Server and Ingres), have different relational schemas, and are located physically in different cities. The systems are representatives of two major camps of Finnish museum information systems: the Antikvaria group (the Espoo City Museum) and Musketti (the National Museum). For the sake of simplicity, the implementation of the system is first restricted to only one part of exhibit collections, the textiles. The technical solution is, however, applicable to any object domain and the exhibition can be complemented later with other classes of museum objects.

3.1 Viewing collections as a virtual WWW space

The traditional approach in publishing collections on the web is to use keyword-based search for the collection database. The user types in keywords and the system generates hit lists matching with the given words. This approach is used, e.g., in the Australian Museums Online system[21]. In our view, such functionality is useful especially to an expert user looking for some specific objects. The goal the ordinary museum visitor is, however, usually something quite different from trying to find certain objects. One would rather want to learn about the past and experience it with the help of the collections. In physical exhibitions the cognitive museum experience is often based on the interesting semantic and thematic combination of exhibits and their contextual information. The same principle applies to WWW exhibitions, as well.

To take a step towards this ambitious goal, the FMS system transforms collection databases into a virtual semantic web space (see figure 1). Its pages are linked with each other with semantic links that are useful for finding information based on its content. The idea is to offer to the user a semantic browsing and searching facility in the combined collection knowledge base [11]. This facility

[20] http://www.cs.helsinki.fi/group/seco/
[21] http://www.amonline.net.au/

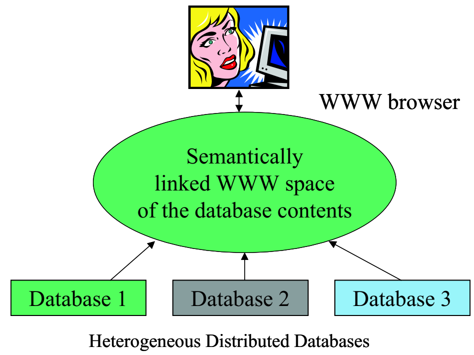

Fig. 1. The idea of the Finnish Museums on the Semantic Web system from a user's point of view.

is implemented by a piece of server-side software, called Ontogator. When the user views the exhibition entry page (URL) with a web browser, Ontogator dynamically generates WWW pages with links to other pages of interest. The FMS home page is the single entry point through which the user enters the virtual museum collections' WWW space.

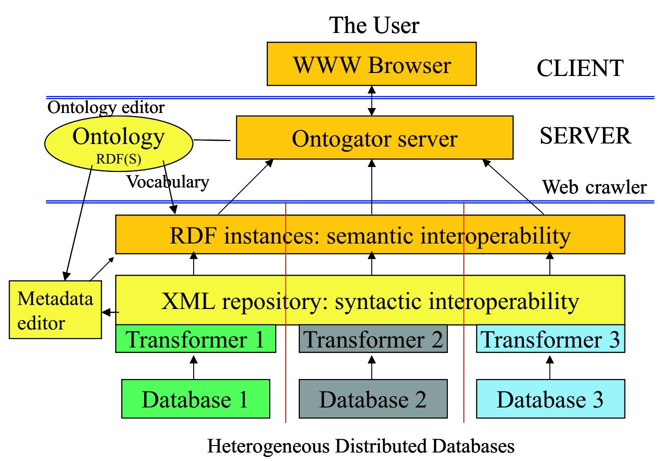

Fig. 2. The architecture of the Finnish Museums on the Semantic Web system.

The general architecture of the FMS system is presented in figure 2. Museums join the system by producing exhibit cards according to the FMS XML Schema, and by checking the semantic validity of the results with the metadata editor

Meedio [16]. The metadata in RDF form produced by Meedio is placed in a public directory on the museum's WWW server. In this way, the museum can easily and completely control of the information it wants to and can publish. The museum does not need to allow the FMS system access their internal database system through the web. Data security, firewall, and related problems do not arise.

The collection data of different museums is read in by the web crawler of the FMS system and combined into a global RDF database, a semantic graph that consists of the ontology and the collection data. Based on this graph, Ontogator dynamically generates the collection WWW space (figure 1) for the user's web browser.

Ontogator will provide the user with the following semantics-based facilities.

View-based filtering Ontogator shows the multiple ontologies used in annotating collection data, such as ObjectType, Material, etc. By selecting ontological classes from these hierarchies, the user can express the search profile easily in the right terminology. For example, by selecting `ObjectType=carpet` and `Material=silk`, silk carpets are found. This view-based idea to information filtering is adapted from the HiBrowse system developed for a bibliographical information retrieval system [15]. Using the system is based on the metaphor of opening directory folders – the idea used in Windows Explorer and in many other systems.

Topic-based navigation Ontogator supports topic-based navigation according to the underlying idea of Topic Maps [14]. The creation of semantic links between topics of interest is based on 1) the collection domain ontologies (classes and their relations) and 2) on actual collection data (instance data). The links give the user contextual and pragmatic information about the objects in the collection.

Ontological search engine for Finnish A search engine is being developed for generating hit lists in the same fashion as search engines on the WWW. However, our engine will understand and make use of the semantic relationships between keywords. The conjugation of Finnish words is also taken into account of.

The actual implementation of the system is underway. To illustrate the use of the FMS system, figure 3 shows the user interface of our first experimental implementation of Ontogator. This system [11] was originally created for an image database of the Helsinki University Museum, but illustrates our current idea to be used in FMS as well. The system implements view-based filtering and facilitates topic-based navigation in a restricted sense by recommending semantically related images. On the left in the window, the user may choose different ontological views of the images related to the promotional ceremonies of the University of Helsinki. Here the ontologies of "People", "Places", and "Events" are open. The ontologies show the user the concepts that will be needed when searching for images. This is necessary, because the end-user might not be familiar with the promotion-related concepts. By opening the ontology hierarchies, the user

Fig. 3. A semantic browser for the image database of the Helsinki University Museum.

chooses the ones that are of interest, and the browser shows all the images that fit all the chosen concepts. On the right, an image found by this method is shown. Underneath it, the system displays a number of related recommended images that are connected semantically with the main image in some way. The system finds such images automatically based on the underlying RDF database. For example, in the recommended images, the same persons may appear as in the main image, but in another context. By browsing the collections according to the associations, the user may make a journey into the world of promotional events and images.

4 Conclusions

This paper discussed the problem of making heterogeneous databases semantically interoperable on the web. It was argued that in addition to syntactic interoperability semantic interoperability is needed for combining the data logically. Ontology techniques of semantic web research were proposed to solving the interoperability problem on the semantic level. The Finnish Museums on the Semantic Web system was used as a case application.

Using ontologies has several benefits.

Terminological coherence The different terminological conventions of different museums and individual catalogers can be made mutually interoperable.

Enriching collection data Ontological class definitions enrich collection data semantically by, for example, property inheritance.

Pragmatic contexts Object ontologies can be complemented with pragmatic level immaterial contextual patterns, such as descriptions of processes and skills. They can provide the user with further insight on how the objects were used, manufactured etc.

The resulting enriched database can be used as a basis for implementing WWW exhibitions richer in content. Two ways of doing this were proposed. Firstly, ontological classes can be exposed to the user in order to facilitate view-based information filtering. Secondly, the ontological relations between collection data elements can be used as an associative structure that facilitates semantic browsing between related concepts.

Our practical goal is to test and demonstrate feasibility of these ideas in practice by implementing the FMS systems as a prototype. The architecture of the system was presented and first implementational experiments of some parts of the whole system were discussed.

Acknowledgements

Our work is funded mainly by the National Technology Agency Tekes, Nokia, TietoEnator, the Espoo City Museum, and the Foundation of the Helsinki University Museum, and was supported by the National Board of Antiquities. Fruitful co-operation and discussions with Sten Björkman, Kati Heinämies, Anneli Rantanen, Petteri Saarinen, and Sirkka Valanto is acknowledged.

References

1. C. Batini, M. Lenzerini, and S. B. Navathe. A comparative analysis of methodologies for database schema integration. *Computing Surveys*, 18(4):323–364, 1986.
2. T. Berners-Lee, M. Fischetti, and M. Dertouzos. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*. Harper Business, 2000.
3. N. Bradley. *The XML Companion*. Addison-Wesley, 2002.
4. 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/.
5. P. Cauldwell, R. Chawla, V. Chopra, G. Damschen, C. Dix, T. Hong, F. Norton, U.Ogbuji, G. Olander, A. Richman, K. Saunders, and Z. Zaev. *Professional XML web services*. WROX Press, Birminghan, UK, 2001.
6. C. Fellbaum, editor. *WordNet. An electronic lexical database*. The MIT Press, Cambridge, Massachusetts, 2001.
7. 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.
8. 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 tools for knowledge management. In *eBusiness and eWork, Madrid, Spain*, 2000.
9. D. J. Foskett. Thesaurus. In *Encyclopaedia of Library and Information Science, Volume 30*, pages 416–462. Marcel Dekker, New York, 1980.
10. 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.
11. E. Hyvönen, A. Styrman, and S. Saarela. Ontology-based image retrieval. In Hyvönen and Klemettinen [10], pages 43–45. http://www.hiit.fi.
12. 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/.
13. R. L. Leskinen, editor. *Museoalan asiasanasto*. Museovirasto, Helsinki, Finland, 1997.
14. Steve Pepper. The TAO of Topic Maps. In *Proceedings of XML Europe 2000, Paris, France*, 2000. http://www.ontopia.net/topicmaps/materials/rdf.html.
15. 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.
16. V. Raatikka and E. Hyvönen. Ontology-based semantic metadata validation. In Hyvönen and Klemettinen [10], pages 30–42. http://www.hiit.fi.
17. E. Rahm and P. A. Bernstein. On matching schemas automatically. Technical Report MSR-TR-2001-17, Microsoft Research, Microsoft Corporation, Redmond, WA, USA, 2001.
18. P. Sihvo, editor. *Kulttuuriaineiston luokitus. Outline of cultural materials.* Museovirasto, Helsinki, Finland, 1996.