A Content Creation Process for the Semantic Web

Eero Hyvönen, Mirva Salminen, Miikka Junnila, Suvi Kettula

Helsinki Institute for Information Technology (HIIT), University of Helsinki
P.O. Box 26, 00014 UNIV. OF HELSINKI, FINLAND
{firstname.lastname}@cs.helsinki.fi
http://www.cs.helsinki.fi/group/seco/

Abstract

This paper discusses the creation of terminologies, ontologies, and annotations when publishing semantic web content. The problem is approached by presenting the content creation processes of the semantic portal MUSEUMFINLAND that is intended for publishing collections of Finnish museums on the web.

1. Introduction

The key idea of the Semantic Web (Berners-Lee et al., 2001) is to annotate web resources with machine interpretable metadata. Based on the metadata, intelligent applications such as semantic portals (Maedche et al., 2001) can be created. Metadata creation includes two major parts. First, the ontologies (Fensel, 2004) and vocabularies used as the basis in metadata descriptions are defined. Second, the web resources are annotated with metadata conforming to the definitions.

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 and resource-consuming and usually economically infeasible with larger datasets. Automation of the annotation process is therefore needed. This paper addresses the problem of metadata creation for the Semantic Web through a real life case study. We describe the content creation process developed for the MUSEUMFINLAND[1] (Hyvönen et al., 2004a) semantic portal. This application publicizes cultural collection data from several heterogeneous distributed museum databases in Finland. we define what kind of data is needed in bringing the heterogenous cultural collections into one uniform semantically linked space and focus on how this process can be done with minimal human intervention.

2. Specification for Content Need

MUSEUMFINLAND provides the user with two services: 1) a multi-facet (Pollitt, 1998; Hearst et al., 2002) search engine based on ontologies and 2) a recommendation system for semantic browsing[2].

In order to provide the semantically interlinked and machine understandable inter-museum exhibition and the facets underlying the services, four kinds of content creation processes are needed:

1. Ontology Creation. The core of the system is the set of seven domain ontologies listed in table 1.

2. Terminology Creation. The museums have heterogenous contents and use different vocabularies, so a term ontology is needed to define linguistic words and expressions and their relation to ontological concepts. A separate term ontology makes MUSEUMFINLAND flexible with respect to variance in terminologies used at different museums and by different catalogers. The museums can keep their local terminological conventions as long as they tell the meaning of their own terms by a (URI) reference to the ontologies.

3. Annotation Creation. During the annotation creation process the data from the museum databases is annotated semantically. The prosess makes the heterogeneous collection data syntactically and semanticlly interoperable.

4. Recommendation Creation. Rules that define more associative relations between different metadata items need to be created. These rules are based on the domain ontologies, the collection item annotations, and expert knowledge.

Figure 1 depicts the corresponding content creation processes in MUSEUMFINLAND. The final result of the process is the MUSEUMFINLAND RDF(S)[3] Knowledge Base. It consists of the ontologies, the annotated collection data, and an additional Rule Base that is used for enriching the metadata. With the rules new implicit relations are inferred from the explicit metadata.

In the following the sub-processes of figure 1 are explained in more detail.

3. Ontology Creation

In the ontology creation process, three main methods were needed: *manual editing*, *thesaurus transformation*, and *ontology population*. These methods are discussed next.

3.1. Manual Editing

Ontologies are typically created or enhanced by hand using an ontology editor. This is feasible, e.g., with small ontologies, semantically complex ontologies, or if there are no thesauri or other data repositories available for

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

[2] The idea of these services is explained in (Hyvönen et al., 2004b).

[3] http://www.w3.org/RDF/ and http://www.w3.org/2000/01/rdf-schema

Ontology	Content	Classes	Instances
Artifacts	Classes for tangible collection objects	3227	0
Materials	Substances that the artifacts are made of	364	0
Situations	Situations, events, and processes in the society	992	0
Actors	Persons, companies, organization, and other active agents	26	1715
Locations	Continents, countries, cities, villages, farms etc.	33	864
Times	Eras, centuries, etc. as time intervals	57	0
Collections	Museum collections included in the system	22	24

Table 1: Ontologies in the MUSEUMFINLAND portal.

View category	View	Underlying ontology
Object	Object type	Artifacts
	Material	Materials
Creation	Creator	Actors
	Location of creation	Locations
	Time of creation	Times
Usage	User	Actors
	Location of usage	Locations
	Situation of usage	Situations
Museum	Collection	Collections

Table 2: View facets in the MUSEUMFINLAND portal.

computer-based ontology creation. In our case, the Collections ontology classifying the collections in MUSEUMFINLAND and the Times ontology that represents a taxonomy of different time eras and periods by time intervals were created in this way. All ontologies have been enhanced manually to some extent even if much of the creation work could be automated. In this work the Protégé-2000[4] editor with its RDF plug-in was mostly used.

3.2. Thesaurus Transformation

Controlled vocabularies and thesauri are usually used when indexing collection items in a database. A thesaurus employs a small number of relationships to organize the terms, such as those listed in table 3 (Foskett, 1980). Also references to synonyms, antonyms, and homonyms may be explicitly presented.

In Finland, the most notable and widely used thesaurus for cultural content in Finnish is MASA (Leskinen, 1997) maintained by the National Board of Antiquities[5]. MASA consists of over 6000 terms and employs the relational structure of table 3. This repository was available as a database and its terms could be used as a basis for creating ontologies.

When transforming a thesaurus into an ontology, the NT/BT relations can be used as a first approximation for the subsumption taxonomy. However, lots of manual corrections are needed for several reason. First, the semantics of the NT/BT relation typically includes different forms of both hyponymy and meronymy, which may not be desirable. Second, the relations are often defined locally without considering a larger global context. For example, the entry Make-up mirror can be a narrower term (NT) of Mirror and the entry Mirror can be a narrower term of Furniture. However, one should not infer from this transitively that a make-up mirror is a piece of furniture like one could with a proper subsumption (subClassOf) hierarchy. Third, the NT/BT relations are not systematically developed in thesauri. For example, in the case of MASA it turned out that there were about 2600 roots that had no broader term among the 6000 terms. The thesauri may also contain some errors that have not been detected by the term bank system used for editing the thesaurus. In our case, some missing reciprocal links and even circularity in the NT/BT relation was detected.

MASA thesaurus was transformed into a new taxonomic ontology called MAO in three steps:

1. A meta-level for MAO-ontology was created using Protégé-2000. This meta-level consists of meta-classes that describe the properties of the ontological classes to be created as MAO-classes. The meta-properties fall into two categories: 1) Semantic relations of the thesaurus as they are, such as BT, NT, etc. 2) Metadata documenting the meaning and creation history of the classes, such as creator, date-of-creation, etc.

2. An RDF Schema structure conforming to the RDFS representation conventions of Protégé-2000 was created automatically from the database. This structure represented the entries of the thesaurus as classes organized into an initial subClassOf taxonomy corresponding to the NT/BT relation.

3. A human editor, museum curator, edited the hierarchy further with Protégé-2000 into a proper taxonomy by introducing new concepts and by re-organizing the classes. Some 600 new classes were created during this phase.

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

[5] http://www.nba.fi

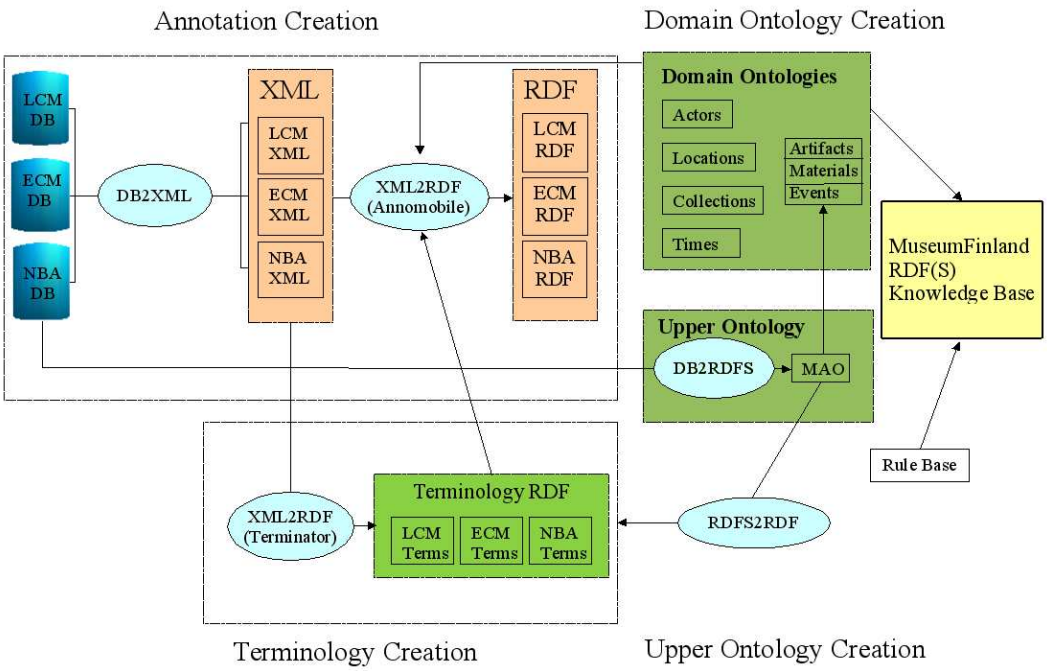

Figure 1: Content creation process in MUSEUMFINLAND.

Symbol	Relationship
USE	Equivalent to "see" reference
UF	Use for, reciprocal of USE
SN	Scope note
BT	Broader term, in a hierarchical array
NT	Narrower term, in a hierarchical array; the reciprocal of BT
RT	Related term, expressing any useful relation other than BT/NT

Table 3: Typical relationships and their symbols used in thesauri (Foskett, 1980).

The transformation in step (2) can be done easily by an algorithm that created RDF(S) classes for thesaurus entries and an initial subsumption hierarchy. For each entry a *term card* mapping the term to a class URI on the ontology was created. Obsolete terms identified by the USE property were omitted from the taxonomy in order to prevent creation of multiple classes for a single concept. However, term cards were created for these entries since obsolete terms are encountered in databases that have evolved during long time periods, and thus need to be mapped to ontology concepts.

In this way, three domain ontologies, Artifacts, Materials, and Events in table 1 emerged as sub-ontologies of MAO. These ontologies were later on extended based on collection item data from the collections of the National Museum[6], Espoo City Museum[7], and Lahti City Museum[8].

3.3. Ontology Population

By ontology population we refer to a process, where the class structure of the ontology already exists and is extended with instance data (individuals). This can be done either by a computer or by a human editor. In our case, the Actors and Locations ontologies in table 1 were created in this way by a semi-automatic process.

The class structure of the Locations ontology is small and could be created by hand. The main content in the ontology is its individual location instances (e.g., Helsinki or Finland) and their mutual meronymy relations (e.g., Helsinki is a part of Finland). An initial set of individual countries and cities (a couple hundred individuals) was generated automatically from official data sources, such as the list of Finnish cities and counties. However, most of the instance data had to be populated from the collection databases, since the museum databases include specific location information — for example specific estates or historic locations —that were not available in the official data sources. For these locations some meronymy relations could identified automatically. This is because many collection data entries contain both a general and a more particular location term (e.g., Paris in Texas or Paris in France), from which the meronymy relation could be deducted. For ambiguous location names, the rdf:type and part-of properties had to be edited by a human editor.

[6]http://www.nba.fi
[7]http://www.espoo.fi/museo/
[8]http://www.lahti.fi/museot/

As in Locations, the class structure of the Actors ontology is small (Person, Company, etc.) and could be created by hand. Most of the resources in the ontology are instances, such as particular persons. The individuals were populated from the databases. In some cases, the class of the instance could be deduced from the original data. If not, the computer made a guess and let the human editor check the result. For example, it may be known that a certain string, say "John Doe", is a person's name but the sex has not been represented explicitly. The computer can then create an instance of class Person and let the editor change the class to either Woman or Man.

4. Terminology Creation

A thesaurus organizes words. This is in contrast with conceptual ontologies that organize concepts underlying the words. For example, a single conceptual ontology can manifest itself as a set of thesauri in different languages. An ontology is — in principle — language independent in nature, but in practice many concepts are language dependent. The distinction between terms and concepts has many practical consequences also within one language. It is possible to define and use different terminologies as long as a mapping from the terms to concepts is provided. In this way, for example, old collection metadata containing obsolete terms can be used and different terminologies of different museums and of different persons can be made interoperable.

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 the properties of table 4. They are inherited by the term instances, 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 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 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.

A terminology ontology is represented by a Protégé-2000 project that consists of the Term class as an RDF Schema, term instances in RDF, and the referenced ontology represented as an included project. Three different methods were used in terminology creation:

1. **Manual development**

   The terminology ontology can be enhanced and new individual terms created by hand with the ontology editor.

2. **Thesaurus to taxonomy transformation**

   New term instances can be created when transforming a thesaurus into an ontology. Here a term card for each thesaurus entry is 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. For entries in singular form (e.g., abstract concepts such as "opera" and materials) the plural form is empty. For those entries in plural form whose singular form represents some other concept, the singular form should be empty. For other entries, both singular and plural forms are created. The morphological tool MachineSyntax[9] was used for creating the missing plural or singular forms the term cards.

3. **New term generation**

   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 help 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 association to an ontological concept can be created. A human editor can then define the meaning by making the association to the ontology.

Figure 2 depicts the general term extraction process in MUSEUMFINLAND. The process involves a local process at the museum and a global process at MUSEUMFINLAND. There are four different term ontologies: one for terms related to MAO concepts, one for Locations, one for Actors, and one for Collections. For the museum side, we created a tool called Terminator. It extracts individual term candidates from the collection data records. 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 and local term bases have a clear distribution of work: The global terminology consists of terms that are usefull for all the museums. It reduces the workload of individual museums, since these terms need not be included in local terminologies. The local term base, on the other hand, is important for it makes possible for individual museums to maintain their own terminologies.

The global term base can be extended when needed: For example when creating new terms, it may occur that there is no appropriate concept in the ontologies that a new term can be associated with. In this case, the term is associated with a more general concept and a suggestion is made to MUSEUMFINLAND for extending the ontology later on with a more accurate concept.

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

Property	Meaning
`singular`	Singular form of the term as a string
`plural`	Plural form of the term
`concept`	URI of the concept in an ontology
`definition`	Definition of the term or info from a data source
`usage`	Value that tells whether the term is obsolete or in use
`comment`	Any additional information concerning the term

Table 4: Term card properties.

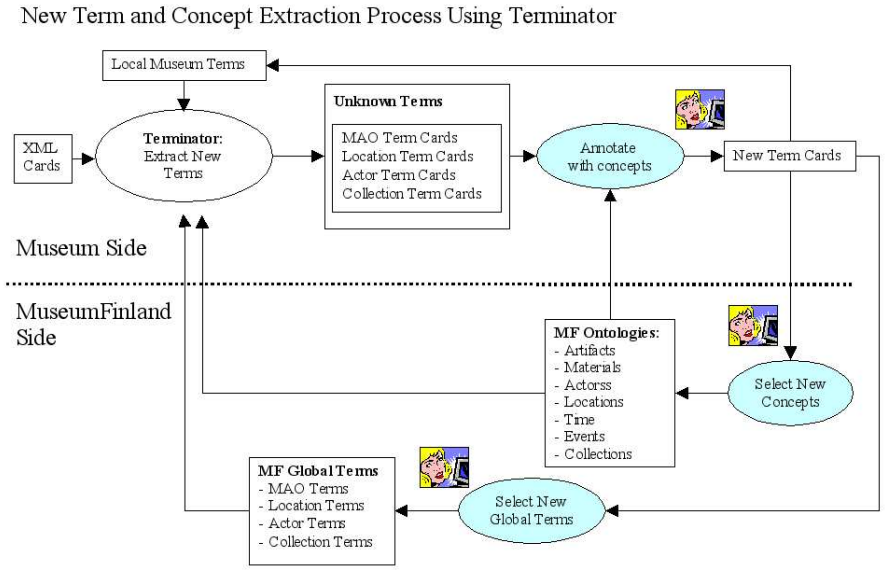

Figure 2: Creating new term cards in MUSEUMFINLAND.

5. Annotation Creation

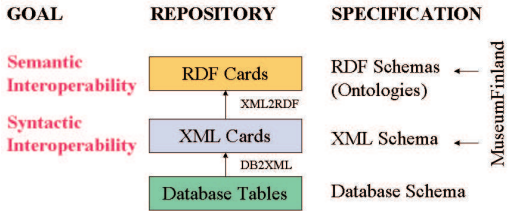

Figure 3: Transforming museum collection data from database into RDF.

Figure 3 depicts the process of transforming collection data records into RDF format in MUSEUMFINLAND. The museum collections locate in heterogenous and distributed databases. The first step towards semantic interlinkage is to attain syntactic interoperability among all the collections. This is done by transforming the collections into XML that is shared by the co-operating museums. As the database schemas of museums are not conforming, the XML card lets every museum to deside which of their database fields to use in filling the XML cards.

Next, the XML is transformed into the final RDF metadata form used by the portal. The RDF conforms to the RDF Schema ontologies of table 1, which guarantees semantic interoperability. The XML to RDF transformation is essentially based on the terms cards by which string values at the XML level, such as "Finland", are transformed into corresponding concept URIs of the ontologies, such as `http://www.fms.fi/locations#Finland`. A semi-automatic tool called Annomobile has been implemented to perform the transformation. The XML to RDF process is discussed and its algorithm is described in more detail in (Hyvönen et al., 2003).

The XML to RDF transformation cannot be done fully automatically due to unknown and homonymous terms. The problem of unknown terms can, in principle, be solved by generating all needed term cards before running the XML2RDF transformation. The problem of homonymous terms occurs when there are homonyms within the context of a data field (e.g., material, location, etc.) each of which refers to one domain ontology (Material, Location, etc.). Homonymous terms that belong to different domains (e.g. term "Malmi" that refers to both a material and a location concept) can be distinguished without human intervention. Our first experiments indicate that, at least in Finnish, homonymy typically occurs between terms referring to different domain ontologies, and the problem of semantic disambiguation is smaller than initially expected. For example there are only 29 homonymic concepts in MAO-ontology which is 0,4% of the total number of classes in MAO.

6. Discussion

6.1. Contributions

This paper presents an overview of content creation process for a Semantic Web application MUSEUMFINLAND. In our work the process is evaluated through a real life case and it has proved out to be usefull in many ways:

Terminological interoperability. The terms used in different institutions can be made mutually interoperable while still maintaining the museum's own terminologies by mapping the terms onto common shared ontologies.

Terminology sharing. Terms that are commonly used in all the museums can be shared by all the museums, which lowers the number of local terms needed.

Ontology sharing. Ontologies provide means to make exact references to the external world. For example, the Locations ontology and actors ontology are shared by the museums in order to make correct and interoperable references.

Automatic content enrichment. Artifact descriptions can be automatically annotated based on term ontologies. In addition, ontological class definitions, rules, and consolidated metadata enrich collection data semantically.

6.2. Related work

The idea of annotating cultural contents in terms of multiple ontologies has already been explored, e.g. in (Hollink et al., 2003). Other ontology-related approaches use for indexing cultural content include Iconclass[10] (van den Berg, 1995) and Art and Architecture Thesaurus[11] (Peterson, 1994). As far as we know, MUSEUMFINLAND is the first one to provide semantical enrichment through terminological interoperability among a number actors and to the extent described int this paper.

Computer based ontology creation and ontology population can be done using domain texts as has been discussed e.g. in (Velardi et al., 2001). Mining of taxonomical relations and instances from text is more error prone but obviously feasible if no other data is available. Our approach of using data-to-be-annotated as source for ontology population ensures that we create only those instances that we need. The transformation process thesauries into presentations with semantic web languages ontology has been discussed also in (Wielinga et al., 2004).

6.3. Further work

Practical problems were encountered when transforming the database contents into RDF. For example, the museum collection data used as the input for Annomobile includes not only terms, but also complex phrases, such as value *case: case for a prize spoon, competition at Salpausselka, 1924, 10 km skiing*, and free text. To handle these cases, the free text and complex phrases were tokenized into words or phrases which were then interpreted as keywords. This approach works, when term cards with ontological links are created from these keywords, and is 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 and thus also the manual workload in their annotation will be high. In MUSEUMFINLAND case it however proved out, that the keyword approach works, since the number of new terms created falls considerably after the intitial term creation.

The annotation cannot be fully automated due to problems of homonymy. The homonymy problem is most severe in free text fields, since they are most prone to consist of conceptually general data where disambiguation cannot be based on the facet/ontology to which the text field is related. To completely solve this problem, museum cataloging systems should be enhanced with ontology support.

The semantic portal which used data produced by the described content creation process was opened on the web in March 2004[12]. In near future we plan to extent the collections of the system with paintings and graphics from the Finnish National Gallery and also with data from the National Museum describing the most valuable cultural sites in Finland. Later on, we may also have the opportunity to incorporate moving images from the Finnish Broadcasting Company. These lay new challenges for content creation process and MUSEUMFINLAND: Our goal is to show that RDF can be used as the basis for making very different kind of contents semantically interoperable.

Acknowledgments

Our work is funded mainly by the National Technology Agency Tekes, Nokia Corp., TietoEnator Corp., 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

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

Fensel, D., 2004. *Ontologies: Silver bullet for knowledge management and electronic commerce (2nd Edition)*. Springer-Verlag.

Foskett, D. J., 1980. Thesaurus. In *Encyclopaedia of Library and Information Science, Volume 30*. Marcel Dekker, New York, pages 416–462.

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

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

Hyvönen, E., M. Junnila, S. Kettula, E. Mäkelä, S. Saarela, M. Salminen, A. Syreeni, A. Valo, and K. Viljanen, 2004a. 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*.

[10] http://www.inconclass.nl

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

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

Http://www.archimuse.com/mw2004/papers/hyvonen/hyvonen.html.

Hyvönen, E., M. Junnila, S. Kettula, S. Saarela, M. Salminen, A. Syreeni, A. Valo, and K. Viljanen, 2003. Publishing collections in the Finnish Museums on the Semantic Web portal – first results. In *Proceedings of the XML Finland 2003 conference. Kuopio, Finland*. Http://www.cs.helsinki.fi/u/eahyvone/publications/xmlfinland2003/FMSOverview.pdf.

Hyvönen, E., S. Saarela, and K. Viljanen, 2004b. 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, (forthcoming)*. Springer–Verlag, Berlin.

Leskinen, R. L. (ed.), 1997. *Museoalan asiasanasto*. Museovirasto, Helsinki, Finland.

Maedche, A., S. Staab, N. Stojanovic, R. Struder, and Y. Sure, 2001. Semantic portal — the SEAL approach. Technical report, Institute AIFB, University of Karlsruhe, Germany.

Peterson, T., 1994. Introduction to the Art and Architechure thesaurus. Http://shiva.pub.getty.edu.

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

van den Berg, J., 1995. Subject retrieval in pictorial information systems. In *Proceedings of the 18th international congress of historical sciences, Montreal, Canada*. Http://www.iconclass.nl/texts/history05.html.

Velardi, P., P. Fabriani, and M. Missikoff, 2001. Using text processing techniques to automatically enrich a domain ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems, 2001, Ogunquit, Maine, USA*.

Wielinga, B., J. Wielemaker, G. Schreiber, and M.van Assem, 2004. Methods for porting resources to the semantic web. In *the 1st European Semantic Web Symposium, 2004, Heraklion, Greece (forthcoming)*.