Principles for Knowledge Engineering on the Web

Guus Schreiber and **Lora Aroyo**
VU University Amsterdam, Computer Science
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
E-mail: `schreiber@cs.vu.nl`, `l.m.aroyo@cs.vu.nl`

Abstract

With the advent of the Web and the efforts towards a Semantic Web the nature of knowledge engineering has changed drastically. In this **position paper** we propose four principles for knowledge engineering on a Web scale. We illustrate these principles with examples from our research in developing a Semantic Web application targeted at cross-collection search in virtual cultural-heritage collections.

Changes in the nature of knowledge engineering

In the eighties the typical knowledge-engineering task was to model and formalize knowledge in one particular application domain. Knowledge engineering was carried out in the context of the construction of knowledge systems that were targeted at problem-solving tasks such as diagnosis, assessment and planning. Knowledge engineering was performed in a relatively small, closed area. In the nineties we saw the advent of ontologies as a vehicle for integration of knowledge bases built for different applications. The term "distributed" is used, but it refers to a set of physically distributed well-understood knowledge bases.

During the last decade the Web has resulted in a dramatic change of the nature of knowledge engineering. One can view Semantic Web applications as the knowledge systems of this new era, but their knowledge bases have very different properties compared to the closed-world knowledge systems. Knowledge engineers are confronted with a multitude of knowledge sources, multi-lingual, often shallow and heterogeneous. In this paper we propose a number of principles that can guide knowledge engineering in a Web context (Section). We illustrate these principles with examples from knowledge-engineering practice in the E-Culture project (briefly introduced in the next section).

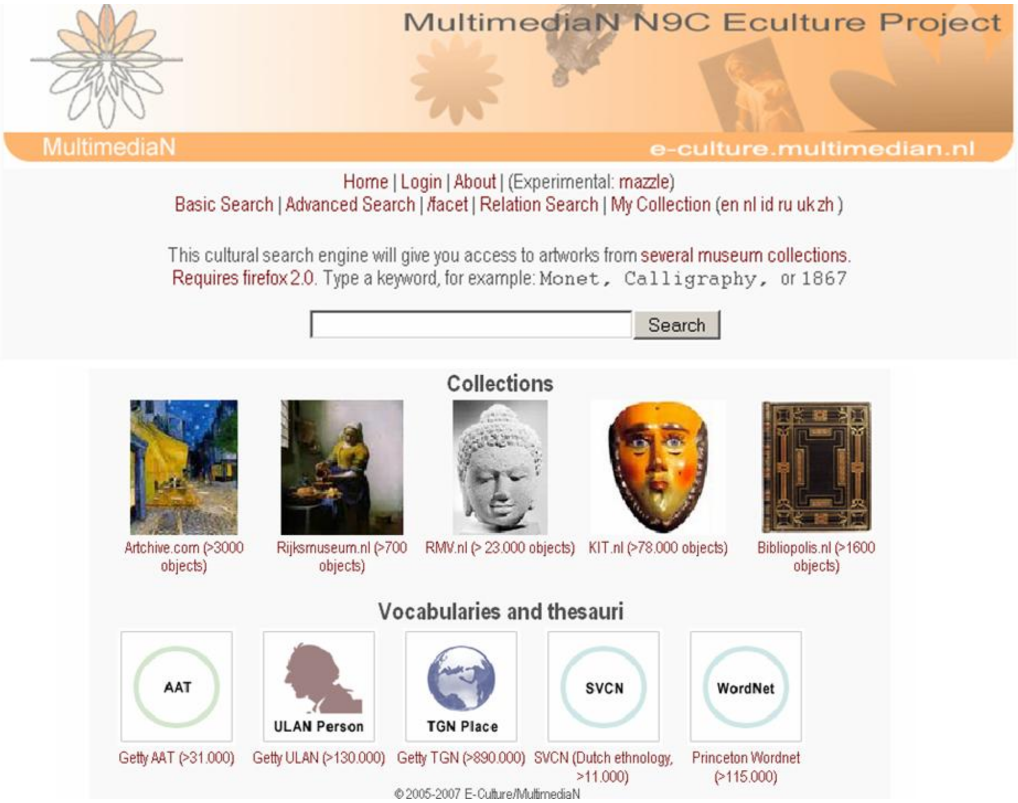
E-Culture: a sample application

The E-Culture demonstrator (Schreiber *et al.* 2006) (winner of the Semantic Web Challenge at ISWC'06) deals with data and metadata of a range of cultural-heritage collections. The objective of the demonstrator is to show how semantic-web technology can be used to provide semantic search in a large virtual collection of cultural-heritage resources. The knowledge base contains a number of vocabularies, which range in size from small (1,000 entries) to large (300,000 entries). Collection data (i.e. images of artworks) should have a URL at the local site of the collection owner. The metadata are harvested in a central server. The server provides semantic search algorithms for answering queries. For example, a query for an artwork that depicts "Paris" returns paintings depicting Paris, but also those depicting Montmartre, despite the fact that "Paris" is not part of the metadata of this work. A screen shot of the basic search interface is shown in Figure 1. Detailed technical information about the demonstrator, such as details about the search algorithms and about scalability issues, can be found on the project website[1]. The demonstrator is continuously updated and currently contains metadata of more than 100,000 web-accessible cultural-heritage objects from a variety (mainly Dutch) collections, including also the vocabularies used to index these collections.

For including a particular collection into the virtual collection of the demonstrator, the E-Culture team carries out four knowledge-engineering tasks (Tordai, Omelayenko, & Schreiber 2007). Firstly, the collection-specific vocabularies are made available in RDF/OWL format. Secondly, the metadata scheme is represented as a specialization of Dublin Core elements, typically using constructs such as `rdfs:subPropertyOf`. Thirdly, the metadata are enriched with additional concepts from other vocabularies in the collection. For example, we try to replace a string 'Amsterdam' with a corresponding **Amsterdam** concept from a geographical vocabulary. Fourthly the collection vocabulary is, where possible, aligned with existing vocabularies in the virtual collection by introducing semantic links such as owl:sameAs. For example, the the art style **Edo** in the vocabulary of Dutch ethnographic musea is aligned with the art style **Edo/Tokugawa** in Getty's Art & Architecture Thesaurus (AAT). The first two steps are relatively easy. The enrichment and alignment steps are more complex and require extensive use of both manual and automatic knowledge-acquisition techniques.

In the next section we use the E-Culture application to illustrate the proposed "knowledge-engineering for the Web"

[1] http:e-culture.multimedian.nl



MultimediaN N9C Eculture Project

MultimediaN

e-culture.multimedian.nl

Home | Login | About | (Experimental: mazzle)

Basic Search | Advanced Search | /facet | Relation Search | My Collection (en nl id ru uk zh)

This cultural search engine will give you access to artworks from several museum collections. Requires firefox 2.0. Type a keyword, for example: `Monet, Calligraphy,` or 1867

Search

Collections

Artchive.com (>3000 objects)

Rijksmuseum.nl (>700 objects)

RMV.nl (> 23.000 objects)

KIT.nl (>78.000 objects)

Bibliopolis.nl (>1600 objects)

Vocabularies and thesauri

Getty AAT (>31.000)

Getty ULAN (>130.000)

Getty TGN (>890.000)

SVCN (Dutch ethnology, >11.000)

Princeton Wordnet (>115.000)

© 2005-2007 E-Culture/MultimediaN

Figure 1: Basic search screen of E-Culture application (`http://e-culture.multimedia.ml/demo/search`)

principles.

Principles for knowledge engineering on the Web

Principle 1: Be modest!

The Web contains a wealth of knowledge sources developed by domain experts over decades or even centuries. In cultural heritage this knowledge is contained in extensive thesauri with knowledge about artists, styles, periods, represented in a semi-formalized way. In other domains such as medicine we find a similar situation. Also, general knowledge sources such as lexical thesauri (WordNets for different languages) and geographical databases (TGN, Geonames) provide key knowledge structures for intelligent Web applications in specific application areas.

As an example, let's take a look at the Union List of Artist Names (ULAN), one of the cultural-heritage thesauri of the Getty Foundation[2]. The term "list" is a somewhat misleading: the thesaurus contains extensive biographical information in the form of a associative semantic network. A fragment of the ULAN entry for the French painter Henri Matisse is shown in Figure 2. The information given is useful in semantic-search applications. For example, when a person is searching for works by Matisse, the E-Culture application uses links to related artists (e.g. student, teacher, worked With) to suggest other possibly interesting art works. ULAN also contains spelling variations for artist names (e.g., 18 name variants of `Rembrandt`), which makes preprocessing of artist names in queries almost superfluous.

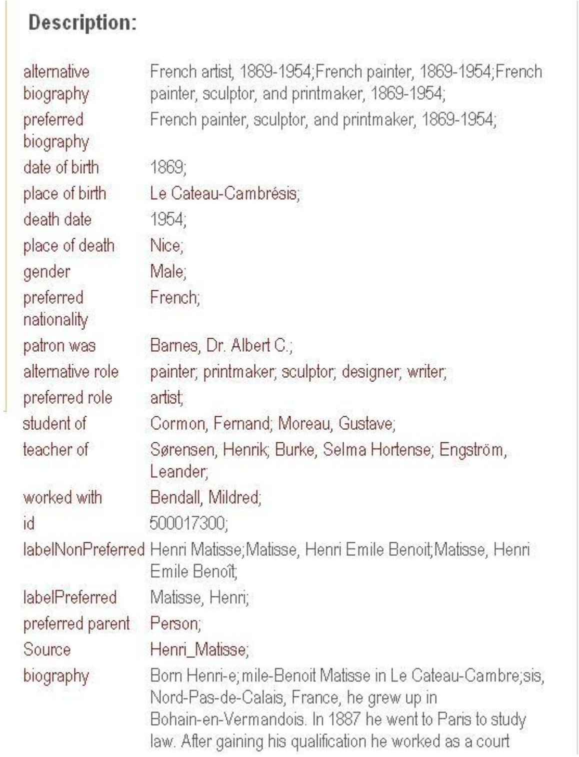
As knowledge engineers we should strive to control our eagerness to discard such a knowledge source just because it contains some errors. Computer scientists have criticized the English-American WordNet for its ambiguous semantics of the hyponym relation, which can have the meaning of subclass, part-of and instance-of[3]. Although this is true, it is no reason to discard the WordNet altogether. Errors in large knowledge sources are a fact of life. The E-Culture demonstrator is an example that shows how WordNet, despite its shortcomings, can be used effectively as a knowledge source.

Principle 2: Think large!

When Lenat gave his famous invited talk "On thresholds of knowledge" (Lenat & Feigenbaum 1991) at IJCAI'87 in Milan about the need to formalize encyclopedic knowledge, he found few supporters in the audience. With the benefit of hindsight we must now admit that he was right in the sense that Semantic Web applications require large amounts relatively simple domain knowledge. Geographical databases,

[2] http://www.getty.edu/research/conducting_research/vocabularies/ulan/

[3] Me recent versions of WordNet have more refined semantics, but that is beside the point.



Description:

alternative biography	French artist, 1869-1954;French painter, 1869-1954;French painter, sculptor, and printmaker, 1869-1954;
preferred biography	French painter, sculptor, and printmaker, 1869-1954;
date of birth	1869;
place of birth	Le Cateau-Cambrésis;
death date	1954;
place of death	Nice;
gender	Male;
preferred nationality	French;
patron was	Barnes, Dr. Albert C.;
alternative role	painter; printmaker; sculptor; designer; writer;
preferred role	artist;
student of	Cormon, Fernand; Moreau, Gustave;
teacher of	Sørensen, Henrik; Burke, Selma Hortense; Engström, Leander;
worked with	Bendall, Mildred;
id	500017300;
labelNonPreferred	Henri Matisse;Matisse, Henri Emile Benoit;Matisse, Henri Emile Benoît;
labelPreferred	Matisse, Henri;
preferred parent	Person;
Source	Henri_Matisse;
biography	Born Henri-e;mile-Benoit Matisse in Le Cateau-Cambre;sis, Nord-Pas-de-Calais, France, he grew up in Bohain-en-Vermandois. In 1887 he went to Paris to study law. After gaining his qualification he worked as a court

Figure 2: ULAN thesaurus entry for Henri Matisse (screen shot of the E-Culture demonstrator)

WordNets and thesauri such as the AAT and ULAN are examples of this. Although the level of formal semantics of these sources might be lower that the ontologies in research papers, their sheer size makes them valuable (if not essential) knowledge repositories. Our knowledge-engineering techniques should scale to the level where the methods can work with such large corpora.

Publishing and accessing large vocabularies on the Web is not always an easy job. Detailed knowledge-engineering issues arise with respect to, for example, URI naming and access policies. The publication of WordNet 2.0 on the Web (van Assem, Gangemi, & Schreiber 2006) provides a case study of how to tackle these problems. The W3C has also published technical recipes for Web publication of vocabularies (Berrueta & Phipps 2008).

Principle 3: Develop and use patterns!

A key outcome of knowledge-engineering research in the eighties and early nineties (typically for stand-alone knowledge systems) was the use of patterns, such as patterns for the structure of knowledge-intensive tasks like diagnosis and assessment (Eriksson *et al.* 1995; Schreiber *et al.* 1994). This insight is still a useful guiding principle.

SKOS[4] is a pattern for representing vocabularies on the Web in an interoperable way. SKOS has already attracted a large user community. Gangemi has described a library of ontology-engineering patterns (Gangemi 2005). Other sample patterns have been described by participants of the W3C Semantic Web Best Practices Group[5], e.g. for n-ary relations (Noy & Rector 2006), value spaces (Rector 2005), and part-whole relations[6].

Pattern development is a way to describe good practices of knowledge engineering, while still preserving the possibility for the pattern user to adapt the pattern to the needs of the domain. Patterns should contain the minimal set of ontological commitments required for modeling a certain type of knowledge. In a sense, patterns are the alternatives for top-level ontologies, which often are ontologically over-committed and therefore hardly used in knowledge-engineering practice (with the possible exception of more or less universally agreed ontologies about notions like time and units of measure). In the cultural-heritage domain CIDOC-CRM (Doerr 2005) is often quoted as a top-level ontology that should be used, but in practice it hardly is. It is our conjecture that top-level ontologies have not reached the level of maturity to be used for modeling adequately the semantics of an application area. Using either patterns or partial alignments (see the next principle) is a more realistic option. In E-Culture we found the patterns for thesauri (SKOS), for n-ary relations and for value spaces to be useful ontology-engineering aids.

Principle 4: Don't recreate but enrich and align!

There is still a tendency in the ontology-engineering community to think that we need to re-engineer existing knowledge sources because these are "wrong", or at least contain mistakes, This approach is unrealistic and will mean the Semantic Web will never come about.

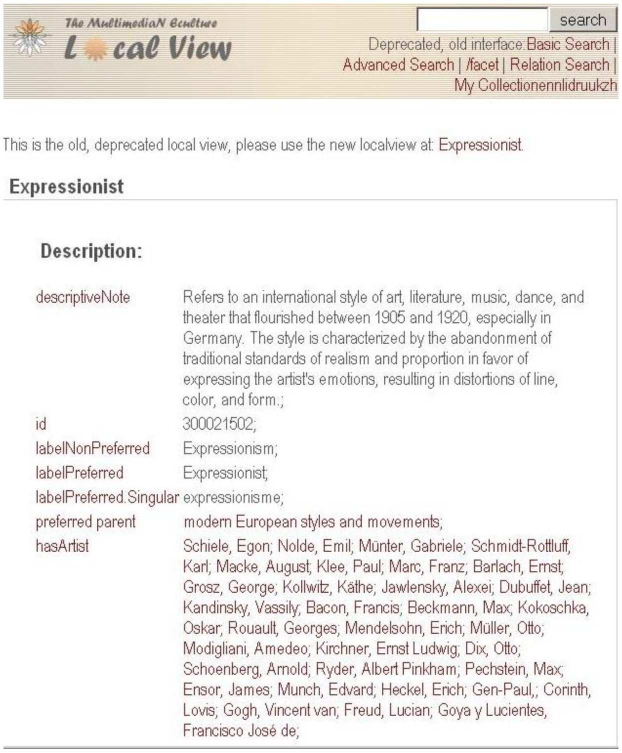
One nice feature of the Web is that it allows us to add knowledge to existing knowledge sources. By simply using the URL mechanism we can easily create a knowledge base that contains additional information about a vocabulary published elsewhere on the Web. Instead of recreating ontologies, we should be content with *enriching* and *aligning* existing knowledge sources. For example, concepts defined in thesauri often have scope notes which contains lots of implicit semantics about the concept. Knowledge-engineering techniques should be employed to make these implicit meanings explicit.

For example, take the description of the concept "Expressionist" in the Getty Art & Architecture Thesaurus, of which a part is shown in Figure 3. The scope note contains information about start and end time of this art style. We can also make explicit that it is mainly a German art style. Knowledge engineers should use their information-extraction techniques to make such semantic information explicit and in this way enrich the thesaurus. We don't need to change the thesaurus; we don't even require permission of the AAT owners: as long as the AAT is published on the Web and the concept has a URI, we can define our enrichments and publish those also on the Web for all to use.

[4] http://www.w3.org/2004/02/skos/

[5] http://www.w3.org/2001/sw/BestPractices/

[6] http://www.w3.org/2001/sw/BestPractices/OEP/SimplePartWhole/index.html


The MultimediaN Eculture
Local View
search
Deprecated, old interface: Basic Search | Advanced Search | /facet | Relation Search | My Collectionennlidruukzh

This is the old, deprecated local view, please use the new localview at: Expressionist.

Expressionist

Description:

descriptiveNote	Refers to an international style of art, literature, music, dance, and theater that flourished between 1905 and 1920, especially in Germany. The style is characterized by the abandonment of traditional standards of realism and proportion in favor of expressing the artist's emotions, resulting in distortions of line, color, and form.;
id	300021502;
labelNonPreferred	Expressionism;
labelPreferred	Expressionist;
labelPreferred.Singular	expressionisme;
preferred parent	modern European styles and movements;
hasArtist	Schiele, Egon; Nolde, Emil; Münter, Gabriele; Schmidt-Rottluff, Karl; Macke, August; Klee, Paul; Marc, Franz; Barlach, Ernst; Grosz, George; Kollwitz, Käthe; Jawlensky, Alexei; Dubuffet, Jean; Kandinsky, Vassily; Bacon, Francis; Beckmann, Max; Kokoschka, Oskar; Rouault, Georges; Mendelsohn, Erich; Müller, Otto; Modigliani, Amedeo; Kirchner, Ernst Ludwig; Dix, Otto; Schoenberg, Arnold; Ryder, Albert Pinkham; Pechstein, Max; Ensor, James; Munch, Edvard; Heckel, Erich; Gen-Paul,; Corinth, Lovis; Gogh, Vincent van; Freud, Lucian; Goya y Lucientes, Francisco José de;

Figure 3: AAT concept **Expressionist** with a scope note. The hasArtist links are an example of enrichment. Snapshot of the E-Culture demonstrator

Another important form of enrichment is ontology alignment. Over the past few years the Semantic Web community has made a considerable research effort to develop new alignment techniques. In our view alignment techniques are essential. Complete unification of ontologies is infeasible, at least within the foreseeable future. But we do not want to end up with a collection of unconnected ontologies. Therefore, we need to find alignments that give us a partial unification. The Ontology Evaluation Alignment Initiative[7] (OAEI) is an important activity in this area in which ontology-alignment methods are compared based on their performance against real-life data sets. The E-Culture demonstrator makes heavy use of alignments between the different vocabularies used in annotating the collections. The hasArtist values in Figure 3 are an example alignments: they link AAT styles to artists in ULAN (these alignment were established semi-automatically, see for details (de Boer, van Someren, & Wielinga 2006)).

Principle 5: Beware of ontological over-commitment!

Each statement in an ontology commits the user of this ontology to a particular view of the domain. If a definition in an ontology is stronger than needed, than the ontology is over-committed. For example, if we state that the name of a person must have a first name and a last name we are introducing a western bias into the ontology and may not be able to use the ontology in all intended cases (think of cultures with different naming conventions). Ontology engineers should aim to define an ontology with a minimal set of ontological commitments. You can translate this into an (oversimplified) slogan: "smaller ontologies are better!". The article of Gruber (Gruber 1994) gives some principles for minimal commitments.

Principle 6: Specifying a data model in OWL does not make it an ontology!

A question often asked by people with a database background is: "what is the difference between data models and ontologies?". One first thing to note is that the difference between ontologies and data models does *not* lie in the language being used. One can define an ontology in a basic ER language (although you will be hampered in what you can say); similarly, one can write a data model with OWL. Writing something in OWL does not make it an ontology! The key difference is not the language, but the intended use. A data model is a model of the information structure in some restricted well-delimited application domain, whereas an ontology is intended to provide a set of shared concepts for multiple users and applications. To put it simply: data models live in a relatively small closed world; ontologies are meant for an open, distributed world (hence their importance for the Web). So, defining a name as consisting of a first name and a last name might be perfectly OK in a data model, but may be viewed as incorrect in an ontology.

As an aside, it must be added that there is a tendency to extend the scope of data models, e.g. in large companies, and thus there is an increasing tendency to "ontologize" data models.

Principle 7: The required level of formal semantics depends on the domain!

The quality of an ontology is by some measured through the number of OWL constructs used. This is a serious misconception and does not take the differences between domains into account. As an example, let's consider two knowledge-rich areas with large amounts of open knowledge sources (and therefore ideal for development of Semantic Web application): biomedicine and cultural heritage. The need for a highly formal OWL language mainly comes from the biomedical area. There is a good reason for this: biomedical people are typically working with large models (e.g., chemical, anatomical, genetic) which have precise formal knowledge models. Describing the physical and functional characteristics of such models requires a high degree of formal expressivity. The cultural-heritage domain, on the other hand, is by its nature less formal. Many of the links in the semantic network do not lend themselves well for a detailed semantic description. For example, the relation between a gene and a DNA strand can be defined much more precisely than the relation between an artists and his/her teacher. The use of OWL in the E-Culture project was extremely useful, but it was limited to precisely four features of OWL, namely the individual-equality relation `owl:sameAs` (for align-

[7] http://oaei.ontologymatching.org/

ment and disambiguation) and the logical property characteristics `owl:inverseOf`, `owl:symmetricProperty` and `owl:transitiveProperty` (for traversing the semantic-search graph). For this reason debates about the required expressivity of OWL (e.g. OWL DL versus OWL Full) should not take the form of theological debates, but rather be approached as a dialectic discussion in which differences in application perspectives have to be acknowledge and taken into account.

Post mortem

This position paper was written from the perspective of application developers in one particular field, and should be understood in this context. However, we strongly belief that only through creating "data points" by building and using realistic Semantic Web applications, we, as a research community, can move the Semantic Web field above the level of an academic exercise. When we (the "we" here refers to the team of people listed in the acknowledgments) constructed the demonstrator and submitted it to the Semantic Web Challenge, we got comments like "oh, but this is the traditional Semantic Web system". Yes, that's true. But the problem is that hardly anybody has, as yet, actually built this "traditional Semantic Web system", at least not with a stable and enduring user community.

Acknowledgments

The E-Culture work is supported by the MultimediaN project funded by the BSIK program of the Dutch government. The project is a joint effort of VU University Amsterdam, Centre of Mathematics and Computer Science CWI, University of Amsterdam, Digital Heritage Netherlands DEN and the Netherlands Institute of Cultural Heritage ICN. Other people involved in the MulimediaN subproject E-Culture are Alia Amin, Mark van Assem, Victor de Boer, Lynda Hardman, Michiel Hildebrand, Laura Hollink, Marco de Niet, Borys Omelayenko, Marie-France van Orsouw, Jacco van Ossenbruggen, Jos Taekema, Annemiek Teesing, Anna Tordai, Jan Wielemaker, and Bob Wielinga.

References

Berrueta, D., and Phipps, J. 2008. Best practice recipes for publishing RDF vocabularies. W3c working draft, World-Wide Web Consortium. http://www.w3.org/TR/swbp-vocab-pub/.

de Boer, V.; van Someren, M.; and Wielinga, B. 2006. Extracting instances of relations from web documents using redundancy. In *Proc. Third European Semantic Web Conference (ESWC'06), Budvar, Montenegro*. Accepted for publication. http://staff.science.uva.nl/ vdeboer/publications/eswc06paper.pdf.

Doerr, M. 2005. The cidoc crm, an ontological approach to schema heterogeneity. In Kalfoglou, Y.; Schorlemmer, M.; Sheth, A.; Staab, S.; and Uschold, M., eds., *Semantic Interoperability and Integration*, number 04391 in Dagstuhl Seminar Proceedings. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany.

Eriksson, H.; Shahar, Y.; Tu, S. W.; Puerta, A. R.; and Musen, M. A. 1995. Task modeling with reusable problem-solving methods. *Artificial Intelligence* 79(2):293–326.

Gangemi, A. 2005. Ontology design patterns for semantic web content. In *The Semantic Web – ISWC 2005: 4th International Semantic Web Conference, Galway, Ireland, November 6-10, 2005. Proceedings*, volume 3729 of *Lecture Notes in Computer Science*, 262–276.

Gruber, T. R. 1994. Towards principles for the design of ontologies used for knowledge sharing. In Guarino, N., and Poli, R., eds., *Formal Ontology in Conceptual Analysis and Knowledge Representation*. Boston: Kluwer.

Lenat, D. B., and Feigenbaum, E. A. 1991. On the thresholds of knowledge. *Artificial Intelligence* 47(1-3):185–251.

Noy, N., and Rector, A. 2006. Defining n-ary relations on the Semantic Web. W3c working group note, World-Wide Web Consortium. http://www.w3.org/TR/swbp-n-aryRelations/.

Rector, A. 2005. Representing specified values in OWL: "value partitions" and "value sets". W3c working group note, World-Wide Web Consortium. http://www.w3.org/TR/swbp-specified-values/.

Schreiber, A. T.; Wielinga, B. J.; de Hoog, R.; Akkermans, J. M.; and Van de Velde, W. 1994. CommonKADS: A comprehensive methodology for KBS development. *IEEE Expert* 9(6):28–37.

Schreiber, G.; Amin, A.; van Assem, M.; de Boer, V.; Hardman, L.; Hildebrand, M.; Hollink, L.; Huang, Z.; van Kersen, J.; de Niet, M.; Omelayenko, B.; van Ossenbruggen, J.; Siebes, R.; Taekema, J.; Wielemaker, J.; and Wielinga, B. 2006. Multimedian e-culture demonstrator. In *The Semnantic Web – ISWC 2006, Athens, Georgia*, volume 4273 of *LNCS*, 951–958. Springer Verlag. Winner Semantic Web Challenge 2006.

Tordai, A.; Omelayenko, B.; and Schreiber, G. 2007. Thesaurus and metadata alignment for a semantic e-culture application. In *Proc. 4th Int. conf. on Knowledge Capture, Whistler, Canada*, 199–200. ACM.

van Assem, M.; Gangemi, A.; and Schreiber, G. 2006. RDF/OWL representation of WordNet. Technical report, W3C Working Draft. http://www.w3.org/TR/wordnet-rdf/.