From Thesaurus to Ontology

B. J. Wielinga **A. Th. Schreiber** **J. Wielemaker** **J. A. C. Sandberg**
University of Amsterdam, Social Science Informatics
Roetersstraat 15, NL 1018 WB Amsterdam, The Netherlands
{wielinga,schreiber,jan,sandberg}@swi.psy.uva.nl

Abstract

Thesauri such as the Art and Architecture Thesaurus (AAT) provide structured vocabularies for describing art objects. However, if we want to create a knowledge-rich description of an (image of an) art object, such as required by the "semantic web", thesauri turn out to provide only part of the knowledge needed. In this paper we look at problems related to capturing background knowledge for art resources. We describe a case study in which we attempt to construct an ontology for a subset of art-object descriptions, namely antique furniture, using AAT as well as metadata standards as input. We discuss the representation requirements for such an ontology as well as representational problems for our sample ontology with respect to the emerging web standards for knowledge representation (RDF, RDFS, OIL).

Keywords

Ontology construction, thesaurus, web standards, image indexing

INTRODUCTION

In this paper we address the problem of capturing knowledge needed for indexing and retrieving image information using highly structured semantic descriptions. Such structured descriptions can be much richer than the traditional "set of terms approach". In fact they come nearer to a description in natural language, often considered to be the ideal way of describing and indexing pictorial material. In order to circumvent the problems of ambiguity in natural language descriptions and queries, structured descriptions should be limited to a fixed set of predefined structures and a closed vocabulary. In this paper we assume that the structured descriptions are created by a human annotator using specialized tools. Two related problems arise in this approach: (1) how can a human be supported during the annotation process, and (2) where does the vocabulary or ontology for filling in the structured descriptions come from? The solution to these problems that we will pursue in this paper is to extend an existing thesaurus with additional knowledge such that is becomes an ontology suitable to support rich structured descriptions. The paper is structured as follows. First we will discuss various alternative approaches to image indexing and retrieval and the requirements that they pose on the vocabulary. Then we will discuss the properties of a particular thesaurus, the Art and Architecture Thesaurus (AAT) in the light of these requirements. We then discuss the construction of an ontology for antique furniture using AAT and existing metadata standards. With respect to knowledge representation we have tried to adhere to the new web standards as much as possible and we discuss problems arising in the pursuing this objective.

K-CAP'01, October 22-23, 2001, Victoria, British Columbia, Canada.

IMAGE RETRIEVAL

There are several paradigms for image retrieval currently in use:

- Content-based image retrieval (CBIR)
- Text-based image retrieval
- Field-based image retrieval
- Structure-based image retrieval

We will discuss each of these approaches in turn.

The content-based image retrieval paradigm indexes images on their intrinsic and primary features, which are computed by various image analysis algorithms. These features include color structure, shape properties, textures etc. This paradigm will not be discussed here since the link with the more semantically oriented other methods is very difficult to make given the current state of the art in mage processing.

There are a number of different forms that text-based image retrieval can take:

- Keyword search with free vocabulary
- Keyword search with a closed vocabulary
- Thesaurus-based search, where not only the vocabulary is closed but also hierarchical (broader and narrower terms) and other relations can be taken into account in the search process.

The general characteristic of this method is that the query is composed of a (possibly Boolean structured) set of terms. The index usually consists of an unordered set of terms. The indexing and retrieval process can both be supported by tools to browse and select terms from the vocabulary. Such browsers are available for large thesauri such as AAT, LCSH and ICONCLASS.

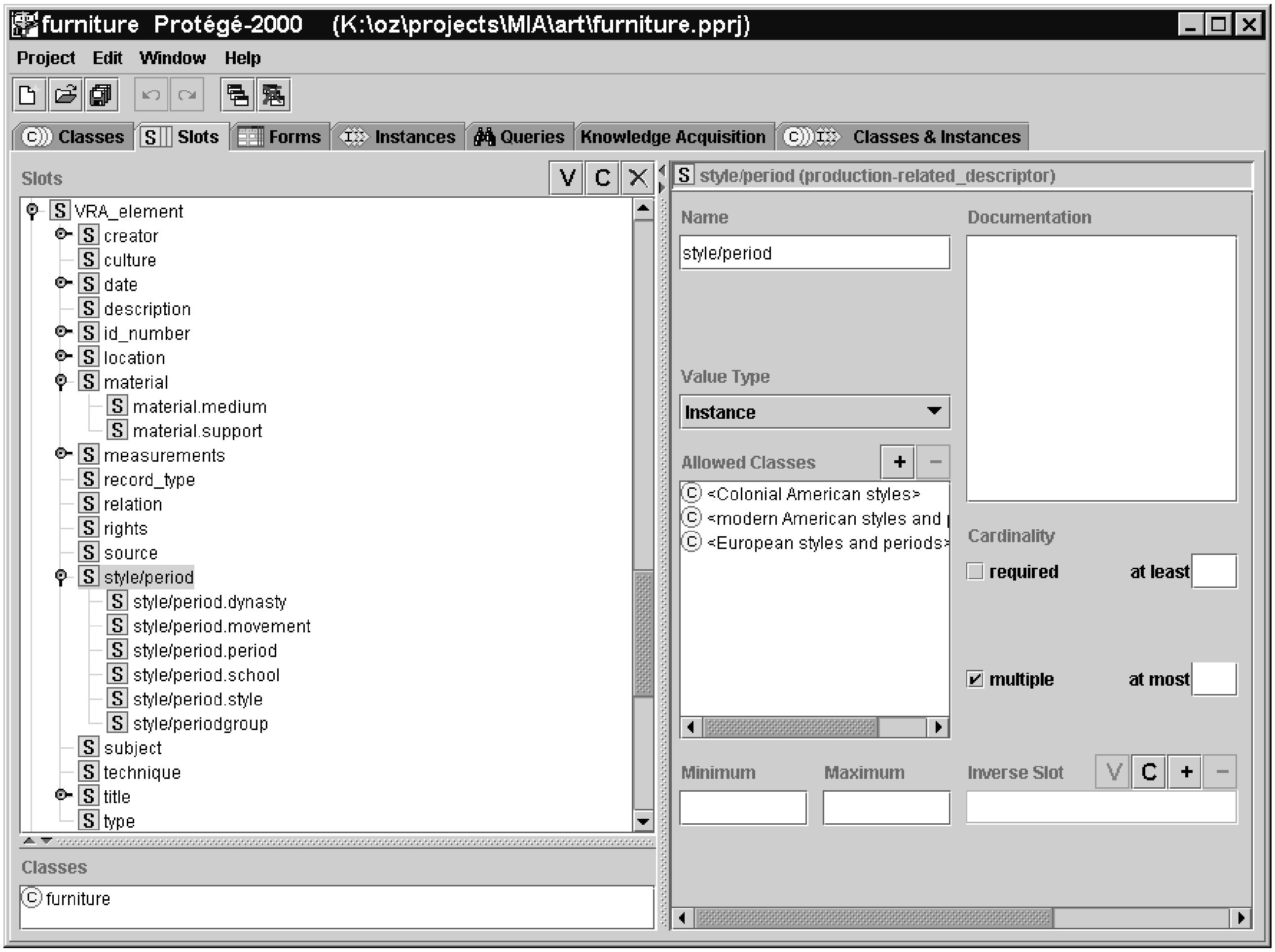

Figure 1: VRA element set defined as Protégé slots. The element qualifiers are defined as subslots, which translate to RDFS subproperties. For a particular visual object multiple instances of a data element can be defined. For example, one can define multiple styles for a piece of furniture.

The field-based approach describes or retrieves an item not by a set of keywords, but by a set of attribute-value pairs. Typically, a metadata schema is defined that describes the elements (fields) and some indication is given what values can be assigned to a particular field. The most widely used schema is the Dublin Core metadata template (DC) [7] for describing documents in general. For specialized domains such as the description of art objects in museums, qualified versions of DC have been created, such as the Visual Resource Association (VRA) Core Categories. VRA version 3.0 [3] defines 17 data elements for describing visual resources. Some data elements have qualifiers which can (optionally) be used to specify more detailed semantics of the data value. For example, VRA defines a data element style/period with qualifiers such as style, period and school. The data elements are linked to one or more corresponding DC elements. For example, style/period is linked to the DC elements coverage and subject. For a particular visual object multiple instances of a data element can be defined. For example, one can define multiple styles for a piece of furniture.

Fig. 1 shows a representation of the set of VRA elements. This representation was developed with the help of the Protégé-2000 tool [6]. The data elements are represented as Protégé slots; the qualifiers as subslots, allowing one to specialize the value set of the element for the qualifier. We come back to this in more detail in the discussion about the ontology.

Many of the field-based initiatives recommend the use of closed vocabularies such as AAT [10], but do not associate particular parts of a thesaurus with a field. As a consequence the only support that a human indexer has is the thesaurus browser. To improve the support for indexing a mapping is required from the fields to particular parts of the thesaurus, such that the indexer is only presented with terms that are relevant for a particular field. As we will argue in subsequent sections of the paper, this is not always easy to do.

Where the field-based approach essentially uses a flat structure of attribute-value pairs, the structure-based approach allows more complex descriptions involving relations. For example, a description of a piece of furniture can include a description of its components, e.g. a drawer of a chest. The components are again objects that can be described using a number of attributes such as material, size, shape. Components can even have components themselves, e.g. drawers can have handles. The structure-based approach introduces a large degree of complexity in the indexing process. Relational descriptions can vary widely between different categories of objects. Furniture can have components, but paintings in general do not have components, they can be described by a complex subject matter structure. A solution to the problem of complexity of the indexing process is to use contextual information to constrain the relations and terms presented to the indexer. We first discuss what the knowledge requirements are with respect to existing thesauri in order to create knowledge-rich art-object descriptions.

ANALYSIS OF EXISTING THESAURI

A first requirement for a thesaurus to be useful in the field- and structure-based approach is that it provides a hierarchical structure that has an unambiguous interpretation. Some hierarchically organized thesauri, such as ICONCLASS [12], mix the sub/super class relation with a part-of relation [1]. AAT uses a strict sub/super class relation in a single inheritance hierarchy. The single inheritance limits the amount of information about a term that can be derived from its position in the hierarchy, as terms can be classified in multiple ways, e.g., material by form or by origin. AAT attacks this problem through qualification of certain terms. For example, the concept **landscape** is represented by two terms: **landscape (representation)** and **landscape (environment)**. This solution has some drawbacks: the distinction between the two qualified terms may not be clear to a user and it is difficult to decide where subclasses of the concept should be placed.

A second requirement is that fields in a description can be linked to particular parts of the thesaurus. For example, the field material should be linked to the part of a thesaurus that contains a hierarchy of material types. In some cases this is straightforward. AAT for example has a hierarchy **Materials**, which clearly defines the terms that can be used as value for the material field. However, there are many cases where values to be assigned to a field are scattered over several parts of the thesaurus. In AAT certain types of porcelain (e.g. five-colored porcelain or Wucai) are situated under <**Chinese ceramics styles**>, while **ironstone** (a semi-porcelain) is located in the **Object Genres** hierarchy. This is not only a problem when the user is presented with a hierarchy or list of values from which a selection has to be made, but also a problem for search processes that use inheritance. Searching for a **ceramics** object requires knowledge about the various parts of the thesaurus hierarchy.

A third requirement follows from the complexity of the indexing space. A human indexer who uses the structure-based approach, will be confronted with large sets of possible values to choose from. For example, the **Materials** hierarchy in AAT contains several hundreds of terms. A solution to this problem is to constrain the value-sets for a particular field, based on a partial description of the image or object. For example, when it is known that an object is a piece of furniture, the possible materials, styles and periods of that object are highly constrained. In some cases various fields can be inferred from information available in other fields. If an object is described as a **Ming vase**, the material is **porcelain**, the region of origin is **China** and the period is between 1368 and 1644.

EXTENDING THE AAT

As the basis for building an ontology for indexing images, we have used the Art and Architecture thesaurus. The AAT is the most elaborate and most standardized body of knowledge concerning the classification of art objects. It contains about 28.000 main terms and 120.000 terms in all, including synonyms and related terms. Besides it offers scope notes: textual definitions of AAT concepts for a major part of main terms. The AAT concepts are represented in 33 hierarchies. A particular concept occurs only once in the full AAT hierarchy, following the ISO 5964 standard (Guidelines for the Establishment and Development of Multilingual Thesauri). AAT uses intermediate concepts ("guide terms") to group concepts lower in the hierarchy. For example, <**ceramic and ceramic products**> is such an intermediate concept.

In an early attempt to use the AAT thesaurus as an ontology [13] we treated the main terms as concept names in the knowledge base. Although this is possible since each main term in AAT is unique, it causes problems when a concept can also be identified by its synonyms, as is the case in WordNet synsets [9]. Searching AAT for the term wood returns as first concept **woods** (area with trees) rather than **wood** (material). It was decided to represent each concept in the knowledge base by a unique identifier, derived from the AAT record number.

The full AAT hierarchy was converted into a hierarchy of concepts, where each concept has a label slot corresponding with the main term in AAT and a synonyms slot where alternate terms are represented. The knowledge base is represented in RDFS [2]. We constructed an RDFS browser to inspect and browse the hierarchy. A snapshot of the browser is shown in Fig. 2.

A second step was to augment a number of concepts with additional slots and fillers. For example, concepts representing a style or period were augmented with slots time period from, time period to, general style and region. The values for these slots were partly derived using explicit tables of periods, and partly by using the intermediate concepts in AAT. For example, the British furniture style **George IV** (1820-1830) is augmented with **Regency** as a more general style indication. A third step was to add knowledge about the relation between possible values of fields and nodes in the knowledge base.

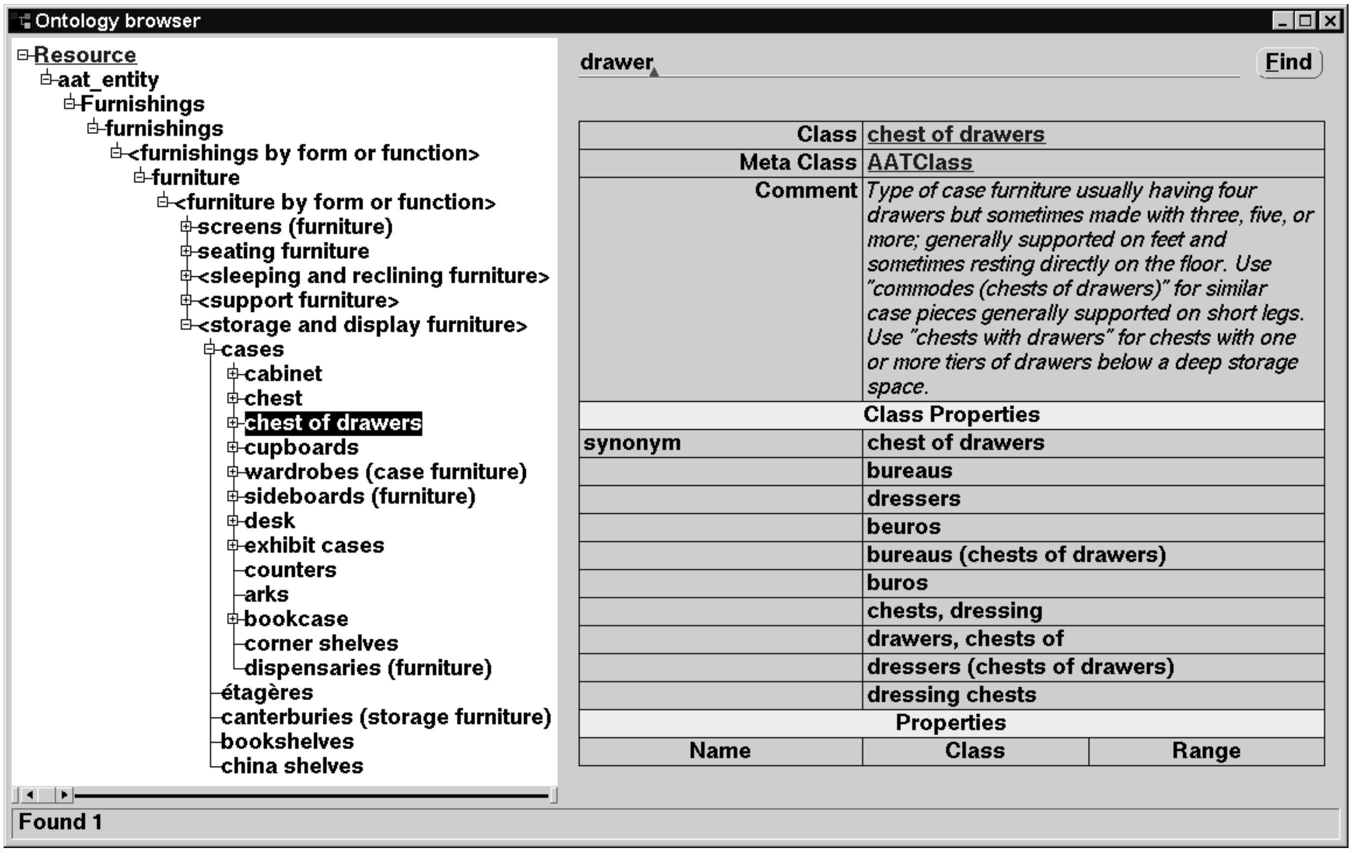

Figure 2: Part of the AAT hierarchy. The snap shot is of our RDFS browser in which an RDFS version of AAT has been loaded

The nature of this knowledge is discussed below.

AN ONTOLOGY FOR FURNITURE

We developed an ontology for a subset of art objects, namely antique Western furniture. This ontology was developed in three steps:

1. Construction of a description template for antique furniture: what kind of information does one want to record for a particular furniture item?
2. Linking the furniture properties to specific subsets of AAT that can be used as values for furniture properties.
3. Describing additional domain knowledge, in particular about constraints between furniture-property values.

Furniture description template

Fig. 3 shows the template we developed for describing a piece of antique furniture. A piece of furniture can be described through 25 "descriptors".[1] Of these 25 descriptors, 17 are derived from the VRA Core Categories [3] (see Fig. 1). The other descriptors are based on the results of the European GRASP project [13]. This project developed an ontology for describing and retrieving stolen art objects. The following "GRASP" slots were added to the VRA elements: **functional context** (e.g., `religious`), **intended location**, **form**, **color**, **color cardinality** (e.g., `monochrome`), **color type** (e.g., `primary colors`), **marking**, and **component**. This last descriptor allows for describing subparts of a piece of furniture (e.g., the feet or drawers of a chest). AAT provides a special hierarchy of terms for this, namely <**furniture components**>. Qualifiers of the data elements were defined as subslots.

We used Protégé-2000 [6] as ontology editor with RDFS as the underlying representation language. The furniture concept is represented as a Protégé class and the descriptors as template slots of this class. Protégé slots are translated into RDFS properties; the qualifiers are translated into subproperties.

This simple representation leads to a long unstructured list of furniture descriptors. In addition, we also wanted to represent natural groups of descriptors. We distinguished four descriptor groups:

1. *Production-related descriptors*: e.g., creator ("maker"), style/period, technique.
2. *Physical descriptors*: e.g., measurements, color, material, etc.

[1] The term descriptor is sometimes used to indicate an attribute value, but we use it here in the "attribute" sense.

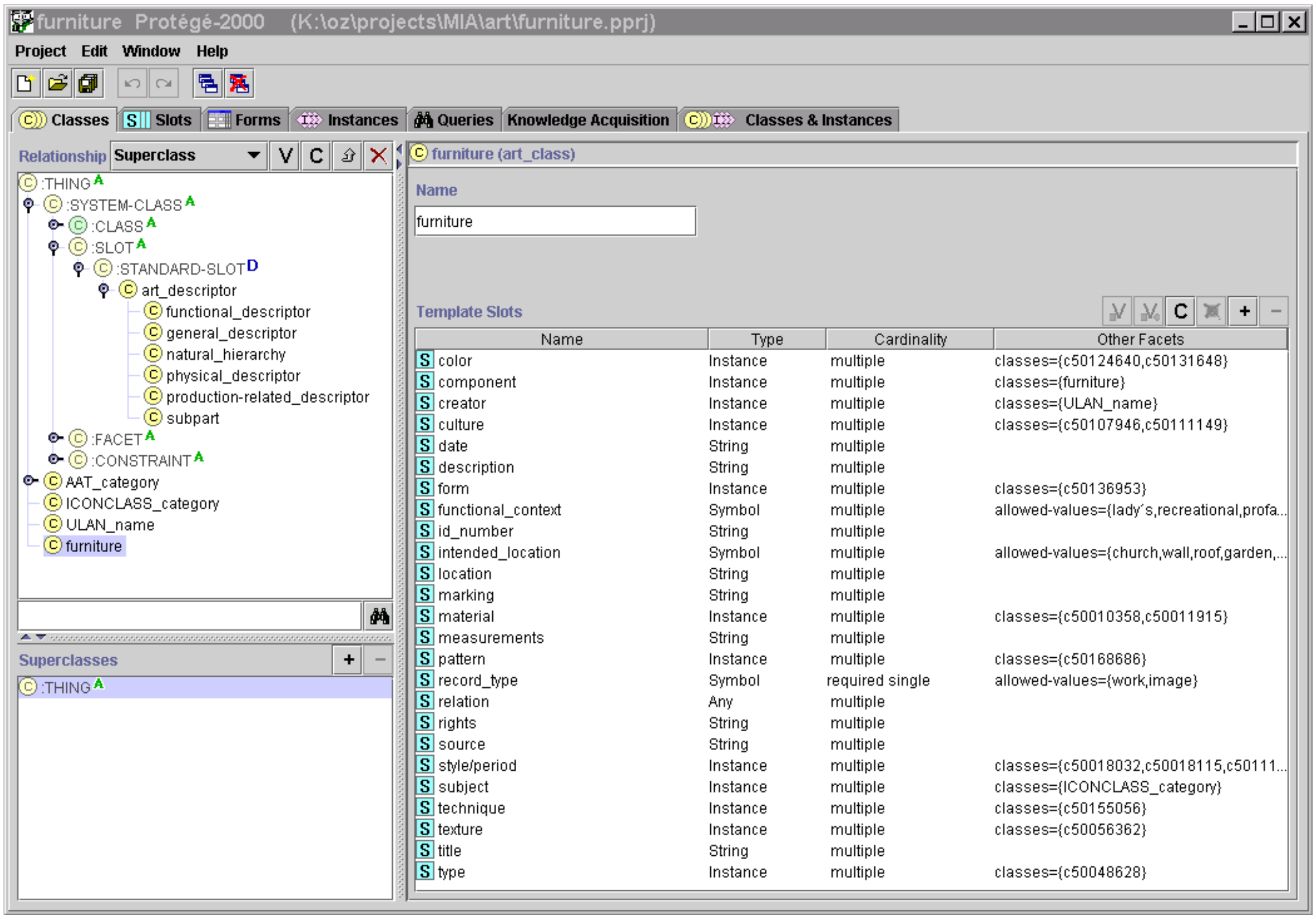

Figure 3: Furniture description template. This template containis the 17 VRA data elements plus 8 additional elements

3. *Functional descriptors*: related to the intended usage of the furniture item, e.g., intended location functional context.

4. *Administrative descriptors*: e.g., collection ID, rights, current location.

It is tempting to represent these descriptor groups as an aggregation: a furniture description has four subparts, one for each descriptor group. However, one requirement we had with respect to the use of RDFS/RDF was that a general RDF-aware browser should be able to interpret as much as possible the resulting furniture-item description. From this point of view the representation of a furniture template as consisting of subparts with their own set of descriptors is cumbersome. It would mean that in the RDF representation there is only an indirect link from the furniture instance to the descriptor triple:[2]

```
<rdf:Description about="furniture36">
  <physical_description rdf:resource="phdesc53"/>
</rdf:Description>
```

```
<physicalDescription rdf:about="phdesc53">
  <material rdf:resource="&aat;mahogany"/>
</physicalDescription>
```

We therefore refrained from using a part-of organization of descriptors. Instead, we defined a metaclass **art descriptor** with the descriptor groups as subclasses. Subsequently, the furniture slots were defined as instances of the appropriate art-descriptor subclass. For example, the property technique is an instance of a **production-related descriptor**. The descriptor metaclasses are listed in Fig. 3 (see the "class" tab at the left).

One of the reasons we prefer Protégé as RDFS editor is that it supports, as RDF/RDFS does, treating instances as classes and vice versa. Not allowing this is in fact a weakness of many description-logic languages, which adhere to a strict separation. Martin [8] considers class/instance flexibility as a central requirement for adequate conceptual modelling.

The VRA element type plays a special role. This descriptor is used to represent the natural category to which the furniture item belongs, e.g. a **case**. For furniture we used the

[2] As we will see further on, the specification of material in the example using a class is problematic.

AAT hierarchy under the "guide term" <**furniture by form or function**> as the value set for the type element. Part of the furniture hierarchy can be found in Fig. 2. Additional domain knowledge is typically centered around these categories. For indexing purposes the furniture category is crucial because the categorization can be used during retrieval for query generalization (e.g., **case** → <**storage and display furniture**>) or query specialization (e.g., **case** → **chest-of-drawers**).

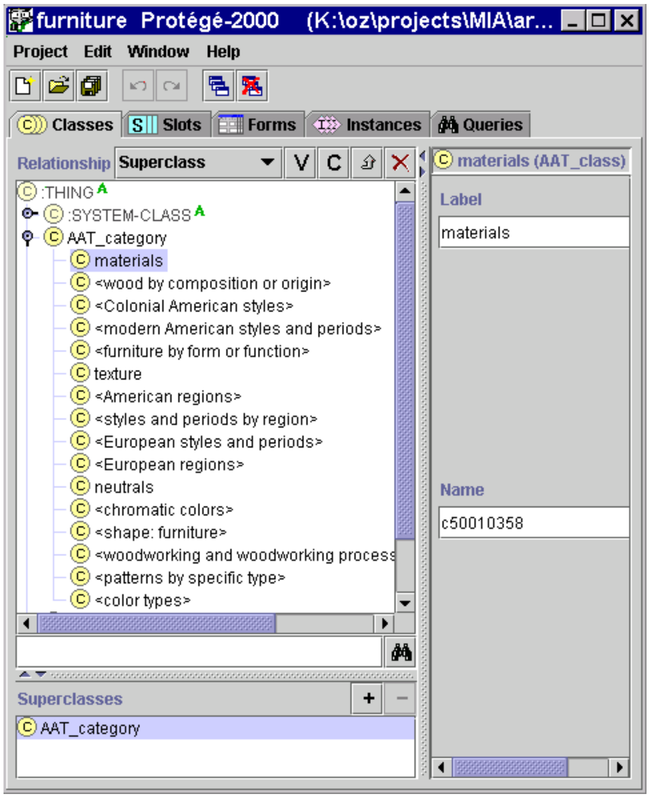

Figure 4: AAT categories used for furniture descriptions. All AAT terms in the hierarchy below the category can act as a value for a particular furniture descriptor

Linking to AAT

For nine slots in the furniture template, parts of the AAT hierarchy could be identified as slot value sets. Fig. 4 shows the AAT categories we used. In some cases, multiple parts of AAT act as alternative value sets for a single slot. For example, the AAT categories **neutrals** and <**chromatic colors**> provide the controlled vocabulary for the color slot. Both are subclasses of the AAT category **colors**, to which also a hierarchy of **color types** belongs which do not represent legal values color slot. Fig. 1 shows another example: the slot style/period can be filled with a term from three alternative AAT hierarchies (e.g., <**European styles and periods**>).

What we frequently wanted to do is to specify a class in the AAT hierarchy where all subclasses in this subpart of the hierarchy are possible slot values for a descriptor. Representing this kind of value types is not straightforward with the current (web) representation methods. Protégé allows the specification of a "class" as the value type for a slot and asks for one or more superclasses for the allowed class values. However, this information is lost in the translation to RDFS. Property ranges are defined in RDFS through the class of the RDF instance, which in this case is just class **class**. Using the OIL language [5] instead of RDFS would not have helped us here. OIL allows classes as slot-value types, but only when explicitly enumerated in a disjunction:

```
class-def furniture
  slot-constraint color
    has-value (black OR grey OR white OR ...)
```

We could have solved this problem by a different mapping from AAT to RDFS/RDF. Currently, we map all terms in the AAT hierarchy to RDFS classes. One could take the view that leaf terms in AAT should be considered as instances. However, this is not a realistic solution. Often, there are many subtle term specializations in the AAT hierarchy. For example, in the color hierarchy there is a term "pink", which also acts as a superclass for a whole range of pink colors (e.g.., variants of "purplish pink"). Both `pink` and and its specializations should be available as a value for the color descriptor. Even the AAT "guide terms" can be useful as a descriptor value, in situations where an indexer does not know to which subcategory an item belongs.

We finally decided to represent descriptor values as instances of RDFS classes representing AAT concepts. For example, we defined the value of the descriptor color as an *instance of* the AAT categories **neutrals** or **chromatic colors**. This means that an RDF annotation of a piece of furniture cannot have a property "color" with value "pink". Instead, the property value should be some instance of "pink". With "pink" represented by the AAT record **aat:c50124707**, the RDF for the annotation becomes:

```
<rdf:Description about="furniture34">
  <color>
    <aat:c50124707/>
  </color>
</rdf:Description>
```

This expression is the RDF serialization of two relations. The first defines that the property "color" of `furniture34` has the value `pink22` and the second defines that `pink22` is an instance of **aat:c50124707**, a class labeled "pink". In these relations `pink22` is an *anonymous* resource generated by the RDF parser.

From a philosophical point of view something can be said in favor of this representation: "pink" can be considered to be an idealization (in the Platonic sense) of a color, of which the particular color of a piece of furniture is only an approximation. Still, the representation feels somewhat awkward.

Figure 5: Sample furniture piece: an 18th century chest-of-drawers, Late Georgian style, made of mahogany. "Chest of drawers" is a main term in AAT; the AAT description can be found in the right-hand part of Fig. 2

Adding domain knowledge

In addition to the furniture descriptors and their value sets, there is also a considerable amount of domain knowledge about relationships between descriptor values. To illustrate this we look at an example piece of antique furniture (Fig. 5, taken from [4, p. 28]). The figure shows an 18th-century chest-of-drawers in Late Georgian style (1760–1811), made primarily of mahogany.

Several types of art-historic background knowledge can be distinguished here:

- Knowledge about the relationship between a style period ("Late Georgian") and a time period. Sometimes, the period of a style is dependent on the "culture", e.g. the British Queen Anne style is shorter (1702–1714) than its American pendant (1702–1727).
- Knowledge about the relationship between style periods and furniture characteristics. For example, Late Georgian chests-of-drawers were typically made of mahogany.

This kind of domain knowledge can be extremely useful for supporting both the image indexing and retrieval. During indexing domain knowledge can be used to suggest descriptor values, which puts less burden on the task of the annotator. During retrieval, domain knowledge can be used to make semantic matches, e.g. to retrieve images of Late Georgian chests when a person is looking for "chest mahogany".

However, there are a number of problems in representing this domain knowledge. Firstly, there is no way in RDFS to extend a set of class/property definitions with this kind of inter-property constraints. The same holds for the OIL language. The OIL slot constraints only apply to a single slot and cannot be used to specify constraints between slots.

Protégé has a constraint language based on KIF and therefore expressing these constraints in Protégé is possible. However, we are then confronted with a second problem. The domain knowledge does not consist of absolute statements about the state of affairs in antique furniture, but provides us mainly with elaborate default knowledge. For example, a Late Georgian chest-of-drawers can be made from oak, but if we have no knowledge to the contrary we can assume it is made from mahogany. This default nature of domain knowledge is also true for time periods of furniture styles, although the period specification (Queen Anne: 1702–1714) may suggest otherwise. The period borders are treated by art historians as indicative only. The semantics of a first-order language are therefore hardly appropriate for expressing the art-historic domain knowledge in this domain. In an earlier case study concerned with indexing photographs of apes [11] we were confronted with similar problems (e.g., urang-utans typically live in Indonesia and have an orange color).

DISCUSSION

One can view this paper as a case study in "real-life" knowledge representation. Many of the issues raised have been discussed and solved in knowledge representation theory. However, in the context of web standards and existing knowledge corpora severe constraints are placed on the representational vehicles. One cannot just redfine the representation of a thesaurus or define a new knowledge-representation standard for the web.

The goal of the Semantic Web initiative is to annotate large amounts of information resources with knowledge-rich metadata. In this paper we have argued that such annotations, in particular of non-textual material such as images, should be based on a rich metadata structure in connection with an ontology. Building ontologies for large domains, such as medicine or arts, is a costly affair. However, in many domains thesauri have been built that can be a basis for the construction of an ontology. A thesaurus should satisfy a number of criteria: it should have a strict sub/superclass hierarchical structure, it should be based on unique concepts rather than on natural-language terms and it should be representable in a format that is compliant with emerging web standards. In the ontology construction process additional knowledge should be added to the basic hierarchical structure of concepts derived from the thesaurus. This knowledge can come from different sources: the location of a concept in the hierarchy, additional sources such as Wordnet, or special purpose documents. Through this process we have created a knowledge base derived from the Art and Architecture Thesaurus (AAT) represented in RDFS. In a case study we have used this ontology as basis for an annotation tool for describing (images of) art objects, in particular antique furniture. The basis of the tool is a metadata structure which is a highly qualified and extended Dublin Core structure. Each of the descriptor elements could be linked to one or more parts of the AAT, thus providing constraints on the values that can be assigned to the elements. An ever better support for annotation and retrieval can be given when additional constraints are added to the ontology, which essentially consist of complex relations between partial descriptions of objects or images. While the basic metadata knowledge can be represented within the semantic framework of RDFS, the constraint relations require additional representational contructs not available in RDFS and other semantic Web oriented languages, such as OIL. For the time being we have designed a format for representing these constraints that can be used in our own tools, but which is meaningless to the average RDFS application.

Acknowledgments

This work was supported by the ICES-KIS project "Multimedia Information Analysis" (MIA) funded by the Dutch government.

REFERENCES

1. S. Bechofer and C. Goble. Thesaurus construction through knowledge representation. *Data & Knowledge Engineering*, 37:25–45, 2001.

2. D. Brickley and R. V. Guha. Resource description framework (RDF) schema specification 1.0. Candidate recommendation, W3C Consortium, 27 March 2000. See: http://www.w3.org.

3. Visual Resources Association Standards Committee. VRA core categories, version 3.0. Technical report, Visual Resources Association, July 2000. URL: www.gsd.harvard.edu/ staffaw3/vra/vracore3.htm.

4. R. Davidson. *Miller's Antique Checklist: Furniture*. Reed, London, 1991.

5. D. Fensel, I. Horrocks, F. van Harmelen, S. Decker, M. Erdmann, and M. Klein. OIL in a nutshell. In *Knowledge Engineering and Knowledge Management: 12th International Conference EKAW2000, Juan-les-Pins*, volume 1937 of *Lecture Notes in Artificial Intelligence*, pages 1–16, Berlin/Heidelberg, 2000. Springer-Verlag.

6. N. Fridman Noy, R. W. Fergerson, and M. A. Musen. The knowledge model of Protégé-2000: combining interoperability and flexibility. In *Knowledge Engineering and Knowledge Management: 12th International Conference EKAW2000, Juan-les-Pins*, volume 1937 of *Lecture Notes in Artificial Intelligence*, pages 17–32, Berlin/Heidelberg, 2000. Springer-Verlag. Also as: Technical Report Stanford University, School of Medicine, SMI-2000-0830.

7. Dublin Core Metadata Initiative. *Dublin Core Metadata Element Set Version 1.1: Reference Description*, July 1999. Url: http://dublincore.org/documents/1999/07/02/dces/.

8. J. Martin. *Object-Oriented Methods – A Foundation. UML edition*. Prentice Hall, Upper Saddle River, NJ, 1997.

9. G. Miller. WordNet: A lexical database for english. *Comm. ACM*, 38(11), November 1995.

10. T. Peterson. *Introduction to the Art and Architecture Thesaurus*. Oxford University Press, 1994. See also: http://shiva.pub.getty.edu.

11. A. Th. Schreiber, B. Dubbeldam, J. Wielemaker, and B. J. Wielinga. Ontology-based photo annotation. *IEEE Intelligent Systems*, May/June, 2001.

12. H. van der Waal. ICONCLASS: An inconographic classification system. Technical report, Royal Dutch Academy of Sciences (KNAW), 1985.

13. B. J. Wielinga, J. A. C. Sandberg, and A. Th. Schreiber. Methods and techniques for knowledge management: What has knowledge engineering to offer? *Expert Systems With Applications*, 13(1):73–84, 1997.