Supporting the decision process for the choice of a domain modeling scheme

Katharina Schwarz
Institute for Information and Computing Sciences,
Utrecht University, The Netherlands
kschwarz@cs.uu.nl

Timo Kouwenhoven
CIBIT, The Netherlands
tkouwenhoven@cibit.nl

Virginia Dignum
Institute for Information and Computing Sciences,
Utrecht University, The Netherlands
virginia@cs.uu.nl

Jacco van Ossenbruggen
CWI, The Netherlands
jacco.van.ossenbruggen@cwi.nl

Abstract Enterprises are experiencing the need to model their domain in order to improve the communication and control over their intellectual assets. There are several different approaches to modeling a domain, some old-fashioned, some trendy; practitioners are at a loss when having to decide which approach best suits their situation. In this research we compare different domain modeling schemes to find out which problem each of them is best suited to solve. The modeling schemes we consider are taxonomy, thesaurus and ontology. We draw relevant information from evaluating case studies and interviewing practitioners in the field of domain modeling. We restrict the scope to those projects that aim at improving the retrieval of information. The aim of this research is to support practitioners when they want to model their domain to choose the right scheme, considering the trade-offs being made between requirements and effort.

Key Words: taxonomy, thesaurus, ontology, information retrieval

1 Introduction

In order to better manage their unstructured and semi-structured data, enterprises develop domain models by which they organize their information resources. Domain models are used for inserting a layer of semantics between the resources and the users so that the chances at finding the required information are increased. This can be in the form of meta data added to resources, or of a navigation structure that reflects the domain model in the graphical user interface. A domain model leads to disambiguation and facilitates communication about the domain [15]. Further, a domain model captures the knowledge within an enterprise and serves as an overview of the domain. We compare different domain modeling schemes to discover their advantages and disadvantages, and to find out which scheme solves which problems best.

Based on the case studies we have evaluated, we consider the schemes *taxonomy*, *thesaurus* and *ontology*. It is difficult to draw a clear line between these schemes. Basically they are all forms of classifications that model relationships between concepts, but they come from different backgrounds and were developed for different purposes.

Taxonomies are related to natural sciences, a well-known taxonomy being Linnaeus' taxonomy of flora and fauna, dating from the 18th century. The traditional purpose of a taxonomy is to establish a hierarchical structure of the domain concepts. A taxonomy inherently does not provide associational or equivalence relationships. In a business context taxonomies can be used to support the quick, consistent and accurate retrieval of resources, and to provide a means for user adaptation [8].

Thesauri are traditionally linked to the world of library sciences. Their purpose is to give each resource a place in the collection so that it can be easily retrieved. A thesaurus consists of terms and their relationships. There are three formally defined types of relationships within a thesaurus; hierarchical, associational and equivalence [1].

We understand the concept of ontology as defined by Thomas R. Gruber [7]: "An ontology is a formal explicit specification of a shared conceptualization for a domain of interest". Ontologies have proved useful for knowledge engineers and the Semantic Web community [2] for describing the knowledge of a domain. In this research we consider ontologies as they are used in the context of the Semantic Web and focus on the simplest W3C recommendation *RDF Schema* for the notation of an ontology. This provides 2 explicitly typed hierarchical relationships; (1) an "is_a" relationship (rdfs:subClassOf) and an "instance_of" relationship (rdf:type). Associational relationships are modeled as properties of concepts.

Much information is available on applying these schemes for developing a domain model, e.g. [1], [5], [8], [12], [13], but no practical overview exists that compares them to each other and indicates the usage, advantages and disadvantages of each. When an enterprise decides to embark on a project that requires a domain model, there are several dependencies that need to be considered for choosing the right scheme. An overview is useful to be able to strike the balance between the invested effort and the gained features. The main factor that influences both the choice of a scheme for organizing information and the construction of the domain model is the problem that needs to be solved.

In this paper we present the preliminary results of our research for supporting the decision making process when choosing a domain modeling scheme. In section 2 we describe our analysis approach and a selection of 3 case studies, one for each of the schemes dealt with. In section 3 we present considerations for making a decision between the possible schemes. We provide conclusions in section 4.

2 Case studies

We have analyzed 12 case studies to learn which types of search-related problems are solved with a domain model and which factors influence the choice of a modeling scheme. We divided each case study in the 3 phases *overview*, *choice* and *development*. In the first phase, we got an overview of the problem, the domain, the resources, the users and

the requirements. In the next phase we established the considerations that had influenced the choice of a domain modeling scheme. In the final phase we gathered information about the development of the project, such as the size of and roles within the development team, the duration and the encountered problems and solutions.

Two typical search problems which we encountered that are solved with the help of a domain model are (1) adaptation to various target groups and (2) search across distributed repositories.

All 3 schemes can solve these problems to some degree. The difference lies in the effort it takes to make the domain structure, and the search features that are gained. The effort is determined by the level of detail in which to model the domain, and the restrictions imposed by the notation of the domain model. The requirements for the search features are determined by the search strategies of the target users. Typical search strategies that we consider here are free text search in a text field, navigation through a tree structure and browsing through a network or different views.

The interplay between these different factors is illustrated with 3 case studies, each one dealing with one of the 3 domain modeling schemes.

2.1 Taxonomy for the Ministry of Transport, Public Works and Water Management [11]

At the ministry, a host of information specialists was responsible for archiving and retrieving the resources produced by policy makers and domain experts. A thesaurus was used for classification of the resources. The policy makers and domain experts were dependent on information specialists to be able to retrieve information they needed, on the one hand because most resources were not available in digital form, on the other hand because the search interface was difficult to use and only understood by the information specialists.

One of the main goals was to enable all target groups to search for information online via a single interface. Three different target groups were identified, the domain experts, the policy makers and the information specialists. These groups have different vocabularies and search strategies. The search strategy depends on the user's prior knowledge, the task, the background and the preferences [4]. An information specialist knows the collection well and is expert at formulating a search query. Therefore an information specialist prefers to search with a free text field. Domain experts and policy makers are less experienced searchers, and they lack the required terminology. This could lead to empty results in a free text search. For example, "bridges" are classified as "artworks" in the terminology of engineers.

For these target groups a tree structure is useful because it provides the context of the concepts that users seek. The solution in this case was to develop a taxonomy that is used twofold. At the backend resources are classified with it. In the interface it is used as a navigation structure, beside a free text search field. This type of interface was characterized as "Google + Yahoo".

A disadvantage of the tree structure is that it is a rigid structure that reflects the conceptual domain model of those who made it, and users with a different conceptual model will still have trouble finding what they search for.

2.2 Thesaurus for the Netherlands Institute for Sound and Vision [10]

The Netherlands Institute for Sound and Vision is situated in the heart of the broadcast mecca of the Netherlands, the Media Park in Hilversum. It records every Dutch public broadcast on a daily basis for preservation of the cultural heritage.

At the beginning of the 90's a project was initiated to unite all governmental audiovisual institutes of the Netherlands in the Media Park. Each institute had a proprietary meta data set and used a controlled vocabulary or thesaurus to keep the descriptions of resources in their individual repositories consistent. It was decided to build a collaborative thesaurus of terms for indexing audiovisual resources, to simplify exchange of material and search across repositories, the GTAA (Gemeenschappelijke Thesaurus

Audiovisuele Archieven).

The domain of the audiovisual archives is boundless. It deals basically with the whole world. For this reason it was decided to model the domain in a thesaurus. A thesaurus can cope with such a diverse and large domain better than a taxonomy, because it provides for more relationships than a taxonomy, and it is not required to press all concepts into a single structure. The domain was divided into 7 main categories ("Subject terms", "Genre", "Corporation names", "People names", "Geographical locations", "Time period" and "Remaining proprietary names"). Of these, only the vocabularies of the first two are structured as real thesauri, the remaining 5 categories contain flat vocabularies. To give an impression of the dimensions in this project, there are 4.500 subject terms and a list of 90.000 people names.

2.3 Ontology for Museum Finland [9]

Museum Finland is a research project carried out at the University of Helsinki. One of the main goals of the project was to make cultural collections available and semantically interoperable on the World Wide Web. This resulted in a portal that gives access to the collections of 3 different museums, situated in 3 different cities. The portal provides a text search field for those who know exactly what they want. Visitors who want to browse the collection are provided with 9 views (called facets) that describe aspects of the collection items (Artifact, Material, Creator, Place of creation, Time of creation, User, Place of usage, Situation, Collection). The visitor can combine any of these views in searching the collections. When a visitor views an individual collection item, besides the regular meta data about the item there are also links to other items from all 3 collections that are related to this one. These relationships are inferred automatically from the underlying ontologies and meta data.

Seven ontologies were developed for this project (Artifacts, Materials, Actors, Locations, Times, Events, Collections), based partly on a Finnish cultural thesaurus that is widely used in Finnish museums. They are described in RDF Schema. All 3 collections are originally based on different organization schemas. They have been transformed to RDF.

Both this project and the project at the Netherlands Institute for Sound and Vision combined distributed repositories for searching across them. At the institute a thesaurus was made, for the musea they chose to use Semantic Web technology and create ontologies. Although semantic search is enabled with both approaches, the added value of the Finnish approach is the inferencing capacity that leads to automatically generated semantic recommendations of related items.

3 Considerations for the choice of a scheme

A comparison of the 3 schemes leads to the following observations.

The focus of ontologies is on modeling the concepts of a domain, whereas the focus of a thesaurus is on structuring the vocabulary of a domain. An ontology is well suited to define and model meta data fields of a domain. A thesaurus is especially suited to structure the allowed values of a specific meta data field, such as "subject" or "genre" in the case study of the Netherlands Institute for Sound and Vision. It is the reason that a thesaurus is the only scheme that provides an equivalence relationship, as vocabulary terms can be synonymous, but concepts are always distinct from each other.

A taxonomy is a generic hierarchical structure, which is a core building block of the other two schemes. Both thesauri and ontologies of domains are pure taxonomies at varying stages of their development. In the case study of the Museum Finland, the "Artifacts", "Materials" and "Times" ontologies are taxonomies. Taxonomies are also being used on their own for structuring domains. There are no standards that indicate how a taxonomy should be built or represented, or what the types of the relationships are. It is therefore simpler to build a taxonomy than to make a thesaurus or ontology, because there are no restrictions or rules. On the other hand,

a taxonomy is more ambiguous and prone to misinterpretation. In practice, thesaurus-like functionality is often added to taxonomies, such as definitions of concepts and equivalent terms in scope notes, and associational relationships.

An ontology defined in RDF Schema has well defined hierarchical relationships that adhere to strict rules of inheritance. A domain model expressed in an ontology is the most precise of the three, because it requires the explicit definition of attributes and properties of concepts. The benefit gained from this effort is the machine readability of meta data described in RDF Schema. A semantic search engine can infer relationships from the annotations, as in the Museum Finland case study.

The size of a domain apparently has little influence on which scheme to choose. Thesauri are well suited for large diverse domains, e.g. the thesaurus of the Netherlands Institute of Sound and Vision or the Getty thesaurus for Arts and Architecture [16] (128.000 terms). There are enormous taxonomies like the Open Directory Project [6] (+ 590.000 categories) and medium taxonomies, such as the Ministry taxonomy (2865 concepts + 1074 synonyms). Ontologies are used for very small domains, like the FOAF ontology [3] (12 classes) and for large domains like the Museum Finland project (4721 classes).

3.1 Rules of thumb

The desired search functionality does influence the choice of a modeling scheme. A taxonomy is best suited as a graphical structure for navigation. This poses some restrictions on the shape of the structure, as it should adhere to basic usability guidelines, such as a limited number of levels and nodes and a largely symmetrical shape.

A thesaurus is best suited for a text search field. The thesaurus can be used for disambiguation of the search query and query expansion. This means that the result set is increased or decreased, depending on the number of hits. If there are too few results, the search engine could also return those resources that are related to a parent node. If there are too many hits, the search engine could refine the search with children nodes.

Property	Description	 Taxonomy	 Thesaurus	 Ontology
Standard	Standard for naming of concepts or terms, establishing relationships, presenting the structure	None	ISO 2788, ANSI/NISO Z39.19-2003	None
Notation	Common representation of structure	Graphical hierarchy	BT, NT, RT, USE, UF	Several W3C recommendations, e.g. RDF Schema, OWL
Modeling constructs	Modeling constructs provided by the notation - this influences how accurately a domain can be modeled	Concepts and untyped relationships	Terms, untyped hierarchical, associational and equivalence relationships, scope notes	Classes (concepts), properties, values, typed relationships, instances
Complexity	Effort required to model a domain	Low	Middle	High

Table 1: Overview of a few comparable properties of the 3 schemes

An ontology is best suited for modeling a domain in great detail. Relationships are explicitly typed, and the properties and value types of concepts are defined. A benefit of this effort in search is the automatic detection of semantically related resources.

Some basic differences between the 3 schemes are summarized in table 1.

4 Conclusions

In this paper we have described the status and aims of our research on supporting the decision process for the choice of a domain modeling scheme. We restrict the scope to those projects that aim at improving the retrieval of information. The result of this research will provide an overview over the possible schemes, support the decision making process of a scheme and give practical guidance in starting the project. So far we have held several interviews with practitioners and evaluated projects for organizing information. We have started to develop a model by which to compare the different schemes.

References

1. Jean Aitchison, Alan Gilchrist, and David Bawden. *Thesaurus construction and use: a practical manual*. Aslib, London, 1997.
2. G. Antoniou and F. Van Harmelen. *A Semantic Web Primer*. The MIT Press, 2004.
3. Dan Brickley, Libby Miller, and active participants of the FOAF mailing list. Friend of a friend. *http://www.foaf-project.org/*.
4. P. Brusilovsky. Methods and techniques of adaptive hypermedia. *Journal on User Modeling and User-Adapted Interaction*, 6:87–129, 1996.
5. B. Chandrasekaran, John R. Josephson, and V. Richard Benjamins. What are ontologies, and why do we need them? *IEEE Intelligent Systems*, pages 20–26, 1999.
6. Web Community. Open directory project. *http://dmoz.org/*.
7. T. R. Gruber. Towards principles for the design of ontologies used for knowledge sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands, 1993. Kluwer Academic Publishers.
8. Rachel Hammond. Taxonomy - the science of classification - 1. *Professional Webmaster*, pages 16–22, 2000.
9. Eero Hyvonen, Miikka Junnila, Suvi Kettula, Eetu Mkel, Samppa Saarela, Mirva Salminen, Ahti Syreeni, Arttu Valo, , and Kim Viljanen. Finnish museums on the semantic web: The users perspective on museumfinland. 2004.
10. Timo Kouwenhoven. Searching + navigating = finding! In *Proceedings of the media management seminar: changing sceneries, changing roles*. Mieke Lauwers FIAT/IFTA, Amsterdam, 2004.
11. Timo Kouwenhoven and Peter Nieuwenhuizen. Presentation: Navigating taxonomies. *ARK-Group Conference on Taxonomies, Amsterdam*, 2005.
12. Natalya Fridman Noy and Deborah L. McGuinness. Ontology development 101: A guide to creating your first ontology. *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05*, 2001.
13. T. Saracevic and Paul B. Kantor. Studying the value of libary and information services. part ii. methodology and taxonomy. *Journal of the American Society for Information Science*, 48(6):543–563, 1997.
14. A.Th. Schreiber, J.M. Akkermans, A.A. Anjewierden, R. de Hoog, N.R. Shadbolt, W. Van de Velde, and B.J. Wielinga. *Knowledge Engineering and Management: The CommonKADS Methodology*. MIT Press, 2000.
15. HA Simon. The architecture of complexity. In *The Sciences of the Artificial*, pages 192–229. MIT Press, Cambridge, Massachusetts, 1981.
16. The J. Paul Getty Trust. Art and architecture thesaurus online. *http://www.getty.edu/research/conducting_research/vocabularies/aat/*.