

Terminology Integration for the Management of distributed Information Resources

Ubbo Visser, Heiner Stuckenschmidt, Christoph Schlieder, Holger Wache, Ingo Timm

1. Introduction

Efficient information management and the processes therein become more and more important within enterprises or when enterprises are merging together. Most information systems use specific data models and databases for this purpose. This implies that making new data available to the system requires, that the data be transferred, into the system's specific data format. This is a process, which is very time consuming and tedious. Data acquisition, automatically or semi-automatically, often makes large-scale investment in technical infrastructure and/or manpower inevitable. These obstacles are some of the reasons behind the concept of information integration.

Problems that might arise due to heterogeneity of the data are already well known within the distributed database systems community (e.g. [Kim and Seo, 1991]). In general, heterogeneity problems can be divided into three categories:

- Syntax (e. g. data format heterogeneity)
- Structure (e. g. homonyms, synonyms or different attributes in database tables)
- Semantic (e. g. intended meaning of terms in a special context or application)

For information management problems on the structural and semantic level with regards to terminologies are important. Terminologies are important because they contain the companies' knowledge. The IT manager is confronted with the task of how to map one terminology to another terminology. Lately, approaches based on formal ontologies have shown that they are promising.

We discuss an ontology-based approach for the solution to this problem. This approach has been developed within the BUSTER (Bremen University Semantic Translator for Enhanced Retrieval) project which addresses the above mentioned categories by providing a common interface to heterogeneous information sources in terms of an intelligent information broker (see www.semantic-translation.com).

2. Ontology-Based Information Integration

In order to achieve semantic interoperability across information system using different terminologies, the *meaning* of the information that is interchanged has to be understood across the systems. Semantic conflicts occur whenever two contexts do not use the same interpretation of the information. The use of ontologies for the explication of implicit and hidden knowledge is a possible approach to overcome the problem of semantic heterogeneity.

In nearly all ontology--based integration approaches ontologies are used for the explicit description of the information source semantics. But there are different ways of how to

employ the ontologies. In general, three different directions can be identified: *single ontology approaches*, *multiple ontologies approaches* and *hybrid approaches*. Figure 1 gives an overview of the three main architectures.

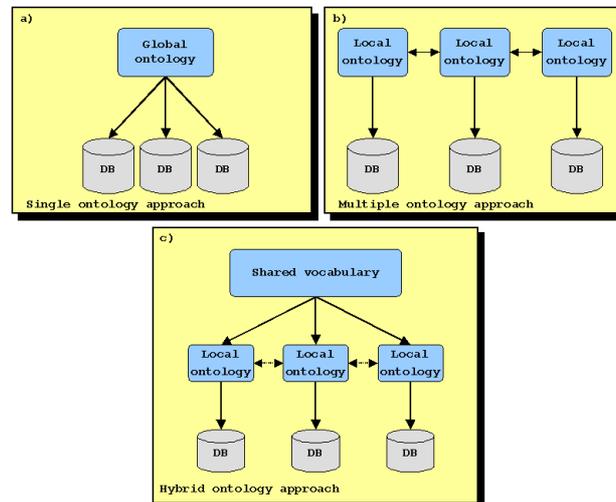


Figure 1: The three possible ways for using ontologies for content explication

- **Single Ontology approaches:** Single ontology approaches use one global ontology providing a shared vocabulary for the specification of the semantics (see fig. 1a). All information sources are related to one global ontology. A prominent approach of this kind of ontology integration is SIMS [Arens et al., 1996].
- **Multiple Ontologies:** In multiple ontology approaches, each information source is described by its own ontology (fig. 1b). For example, in OBSERVER [Mena et al., 1996] the semantics of an information source is described by a separate ontology.
- **Hybrid Approaches:** To overcome the drawbacks of the single or multiple ontology approaches (e.g. finding the minimal ontological commitment), hybrid approaches were developed (fig. 1c). Similar to multiple ontology approaches the semantics of each source is described by its own ontology. But in order to make the source ontologies comparable to each other they are built upon one global shared vocabulary [Goh, 1997]. The shared vocabulary contains basic terms (the primitives) of a domain. In order to build complex terms of a source ontology the primitives are combined by some operators. Sometimes the shared vocabulary is also an ontology [Stuckenschmidt et al., 2000b]. This approach is used within the BUSTER system.

3. Context-based Approach for Terminology Integration

Semantic conflicts occur, whenever two contexts do not use the same interpretation of the information. Goh [1997] identifies three main causes for semantic heterogeneity.

- *Confounding conflicts* occur when information items seem to have the same meaning, but differ in reality, e.g. due to different temporal contexts.
- *Scaling conflicts* occur when different reference systems are used to measure a value. Examples are different currencies or marks.
- *Naming conflicts* occurs when naming schemes of information differ significantly. A frequent phenomenon is the presence of homonyms and synonyms.

It has been argued that semantic heterogeneity can be resolved by transforming information from one context into another (see for example [Sciore et al., 1994]). We investigated two different types of context transformations and their application to the terminology integration problem.

- Rule-based functional transformation [Wache, 1999]
- Classification-based transformation [Stuckenschmidt and Visser, 2000].

We argued that these two kinds of context transformation supplement each other in the sense that functional transformation is well suited to resolve scaling conflicts while classification based transformation can be used to resolve non-trivial naming conflicts [Stuckenschmidt and Wache, 2000].

A conceptual model of the context of each information source builds a basis for integration on the semantic level. We call this process context transformation, because we take the information about the context of the source providing a new context description for that entity within the new information source. Here, we focus on context-transformation by classification. We refer to [Wache, 1999] for further details about the context-transformation with rules.

Context Transformation by Classification

We represent conceptual context with description logic. The idea is to use the inference capabilities of the description logics to derive type transformation rules. The main inference mechanism used in description logics is subsumption checking. A concept is said to subsume another concept, if the membership of the latter implies membership in the former. Following the semantics defined in [Stuckenschmidt and Wache, 2000] the subsumption relation between two concepts is equivalent to a subset relation between the extensions of the concept definition. Subsumption checking can be seen as a special classification method, because it returns a list of classes B_i (concepts) a member of a given concept A belongs to. In terms of subsumption reasoning a context transformation task can be defined as follows:

Let S and T be two terminological contexts represented by sets of concept definitions with subsumption relations $\sqsubseteq_S, \sqsubseteq_T$ and concept membership relations \in_S, \in_T . Let further S be a concept from one terminological context S ($S \in S$). Then the transformation of a data set s from context S into the context T is described by $(s \in_S S \Rightarrow s \in_T T) \Leftrightarrow (S \sqsubseteq_T T)$. In general, it is not decidable, whether the condition $(S \sqsubseteq_T T)$ holds, because the subsumption relation is only defined for the context T while the concept definition S is

taken from context S and is therefore used by a different subsumption relation. At this point, the shared vocabulary plays an important part. Provided, that the concepts from both contexts are defined using the same basic vocabulary, we get a unified subsumption relation defined as $\sqsubseteq = \sqsubseteq_S \cup \sqsubseteq_T$. We can compute \sqsubseteq using available subsumption reasoner that support the language. The result is a set of elements, which belong to the computed class. We can use these to define a set of new context transformation rules. These rules supplement the rule base for context transformation integrate classification-based and rule-based transformation.

4. Application of Terminology Integration

Both approaches, context transformation with rules and context transformation by classification have been developed and implemented within the BUSTER system. One application of the BUSTER system is the integration of catalogue systems within a geographical domain. We describe how the terminology of one catalogue system can be transformed into a different standard classification.

Catalogue Systems

Geographical information systems normally distinguish different types of spatial objects. Different standards exist specifying these object types. These standards are also called catalogues. Since there is more than one standard, these catalogues compete with each other. To date, no satisfactory solution has been found to integrate these catalogues. In our evaluation we concentrate on different types of areas distinguished by the type of use.

We use two catalogue systems, namely the German ATKIS-OK-1000 [AdV, 1998] and the European CORINE (Co-ordination of Information on the Environment) land cover catalogue [EEA, 1999]. The ATKIS catalogue is an official information system in Germany. It offers several types of objects including definitions of different types of areas. Figure 2 (left) shows the different types of areas defined in the catalogue.

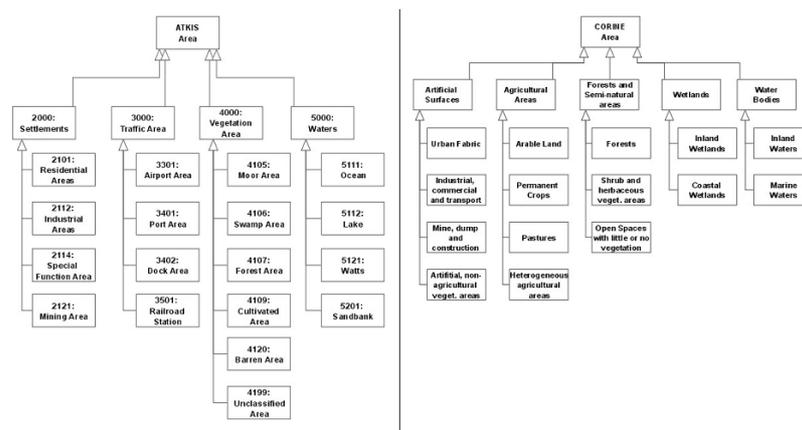


Figure 2: Taxonomy of land-use types in the ATKIS-OK-1000 catalogue (left) and in the CORINE land cover nomenclature

CORINE land cover is a deliverable of the CORINE program the European Commission carried out from 1985 to 1990. The nomenclature as one result developed in the CORINE program can be seen as another catalogue, because it also defines a taxonomy of area types (see figure 2 right) with a description of characteristic properties of the different land types.

The taxonomies of land-use types in figures 2 and 3 illustrate the context problem mentioned in the introduction. The set of land types chosen for these catalogues are biased by their intended use: while the ATKIS catalogue is used to administrate human activities and their impact on land use in terms of buildings and other installations, the focus of the CORINE catalogues is on the state of the environment in terms of vegetation forms. Consequently, the ATKIS catalogue contains fine-grained distinctions between different types of areas used for human activities (i.e. different types of areas used for traffic and transportation) while natural areas are only distinguished very roughly. The CORINE taxonomy on the other hand contains many different kinds of natural areas (i.e. different types of cultivated areas), which are not further distinguished in the ATKIS catalogue. On the other hand, areas used for commerce and traffic is summarized in one type.

Despite these differences in the conception of the catalogues the definition of the land-use types can be reduced to some fundamental properties. We identified six properties used to define the classes in the two catalogues. Beside *size* and *general type of use* (e.g. production, transportation or cultivation) the *kinds of structures* built on top of an area, the *shape of the ground* and *natural vegetation* as well as kinds of *cultivated plants* are discriminating characteristics. Using these properties we performed successful experiments with the identification of forest areas [Visser et al., 2001].

5. Discussion

We have seen that context transformation by re-classification is powerful enough to solve semantic heterogeneity problems within terminological integration tasks. However, we would like to address two issues at this point: (a) modeling effort and (b) domain independency.

One might argue that the amount of time spending on the modeling task in order to get a proper ontology is too much. In section 4 we have seen that our approach can be used to solve the terminological integration task of the ATKIS/CORINE scenario. The amount of time that we invested in the ontology-modeling task was acceptable.

The proposed context transformation approach should be generic enough to solve problems within different domains. We successfully applied our approach to a second application area (supply chain management) and conclude that the approach is domain independent. This domain independency is important to note and one of the key issues of our approach.

The discussed approaches are developed to support both the retrieval and integration of distributed and heterogeneous data based on thematic relevance criteria. However,

thematic relevance is not the only criteria, often data objects refer to some kind of geographical space. Gazetteers can be used for spatial reasoning. In [Schlieder et al., 2001] we have shown that existing gazetteer approaches are not sophisticated enough to serve the users queries and presented a new geographic footprint based on connection graphs. Our newest ideas point to the integration of terminological and spatial reasoning.

References:

- [1] Arens, Y., C.-N. Hsu, and C.A. Knoblock: Query Processing in the SIMS Information Mediator, in *Advanced Planning Technology*. 1996, AAAI Press: California, USA.
- [2] Goh, C.H.: Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Sources, in *Sloan School of Management*. 1997, MIT: Boston. p. 112.
- [3] Kim, W. and J. Seo: Classifying schematic and data heterogeneity in multidatabase systems. *IEEE Computer*, 1991. 24(12): p. 12-18.
- [4] Mena, E., Kashyab, V., Sheth, A., and Illarramendi, A.: Managing Multiple Information Sources through Ontologies: Relationship between Vocabulary Heterogeneity and Loss of Information, in *Proceedings of the 3rd Workshop Knowledge Representation Meets Databases (KRDB '96)*, F. Baader, et al., Editors. 1996.
- [5] Schlieder, C., T. Vögele, and U. Visser. Qualitative Spatial Representation for Information Retrieval by Gazetteers. in *Conference of Spatial Information Theory COSIT*. 2001. Morrow Bay, CA: Springer, to appear.
- [6] Schuster, G. and H. Stuckenschmidt: Building shared terminologies for ontology integration. In: *Künstliche Intelligenz (KI)*. 2001. Wien, to appear
- [7] Sciore, E., M. Siegel, and A. Rosenthal: Context Interchange Using Meta-Attributes. *ACM Transactions on Database Systems*, 1994. 19(2): p. 254-290.
- [8] Stuckenschmidt, H., F.v. Harmelen, D. Fensel, M. Klein, and I. Horrocks, Catalogue Integration: A case study in ontology-based semantic translation. 2000, Computer Science Departement: Amsterdam, IR 474. (b)
- [9] Stuckenschmidt, H. and U. Visser. Semantic Translation Based on Approximate Re-Classification. in *Workshop on Semantic Approximation, Granularity and Vagueness, Workshop of the Seventh International Conference on Principles of Knowledge Representation and Reasoning*. 2000. Breckenridge.

- [10] Stuckenschmidt, H. and H. Wache, Context Modelling and Transformation for Semantic Interoperability, in Knowledge Representation Meets Databases (KRDB 2000). 2000.
- [11] Stuckenschmidt, H., H. Wache, T. Vögele, and U. Visser: Enabling Technologies for Interoperability, in Workshop: Information Sharing: Methods and Applications at the 14th International Symposium of Computer Science for Environmental Protection. 2000. Bonn: TZI.
- [12] Visser, U., H. Stuckenschmidt, G. Schuster, and T. Vögele: Ontologies for Geographic Information Processing. Computers & Geosciences, 2001, to appear