Ontology Construction - An Iterative and Dynamic Task

Holger Wache, Ubbo Visser & Thorsten Scholz
Center for Computing Technologies
University of Bremen, Germany
E-mail: {wache|visser|scholz}@informatik.uni-bremen.de

Abstract

To date the World Wide Web consists mainly of loosely structured data. In order to change this situation and make the Semantic Web successful this vast amount of data needs to be given structure and more importantly, meaning. Ontologies are considered a support in this process as they help us explicitly define concepts. A drawback when using ontologies is that they require a significant modelling effort. On the other hand, semantic inter-correspondences as representation of the semantic relationships between two concepts of different information sources can be additional used as a description. However, they lack formal semantics. In this paper, we discuss a new approach in which we combine these two methods. We will discuss our approach with a look at a minimal ontology, generate hypotheses, and verify these with the semantic inter-correspondences. As a side effect we will use the latter to generate an integration formalism, e.g. in form of rules, which we can use for data integration.

Motivation

The huge amount of electronic data available on the World Wide Web is the main driving factor behind the success of internet applications. However, the majority of the data available is loosely structured and often proprietary. Furthermore, web sites currently contain information which is left mainly to the user for interpretation. The vision of the Semantic Web as described in an article of Burners-Lee et al. (2001) concludes that machines will take over the interpretation of content and meaning on a web site. The key for success for further internet applications is the meaningfulness of the data. Thus, we need semantic knowledge in order to achieve this goal. There is a clear need for methods and technologies that describe semantic knowledge (see also (Fensel *et al.* 2001))

Over the past few years ontologies have be widely discussed (Ciocoiu & Nau 2000; Wache *et al.* 2001; Guarino 1998). Initially, ontologies are introduced as an "explicit specification of a conceptualization" (Gruber 1993). As it turns out, ontologies are useful to support the processes leading towards the Semantic Web because they provide formal semantics. Burners-Lee et al. (2001) emphasizes that the future web will be based on technologies such as ontologies for the description of information sources and for the explicit specification of domain knowledge itself. The future web will also include logical inference mechanisms and others technologies. Unfortunately, discussions have demonstrated that the modelling processes of creating ontologies is just as time consuming and tedious as those knowledge bases created in the late 80's and early 90's. In addition, a domain expert is most unlikely to be able to construct an ontology without the help of a knowledge engineer. Therefore, a lack of user-acceptance is highly probable. At this time a need for methods which simplify the task of constructing ontologies is required.

Requirements & General Idea

In this section we describe the background of ontologies and how they can be used for the semantic description of concepts. Included in this discussion are the requirements needed with regards to the use of ontologies for the Semantic Web. Secondly, we will give an overview about the general idea of our approach.

Requirements: Ontologies can be used to describe the semantics of information sources and to make the content explicit. In regards to the integration of data sources, they can be used for the identification and association of semantically corresponding information concepts. Wache et al. (2001) discusses various roles that ontologies can play within a data integration task. Ontologies are mainly used to describe the content of information sources explicitly. Other roles are based on query mechanisms or verifying models. Here, we will focus on the former.

In nearly all ontology–based integration approaches ontologies are used for the explicit description of the information source semantics. The way in which ontologies are employed, can be diverse. In general, three different approaces can be identified: *single ontology approaches*, *multiple ontologies approaches* and *hybrid approaches*. Figure 1 provides an overview of the three main approaches.

Regardless of which approach is used, the requirements of such ontologies are twofold: first, an ontology should be *minimal*. The demand for this requirement also supports modelling efforts, which should be also minimal. Second, an ontology should be *complete*. For use on the Semantic Web we prefer a practical view on the completeness. We define completeness in the sense that the semantics of the terms within the ontology are precise and sufficient enough

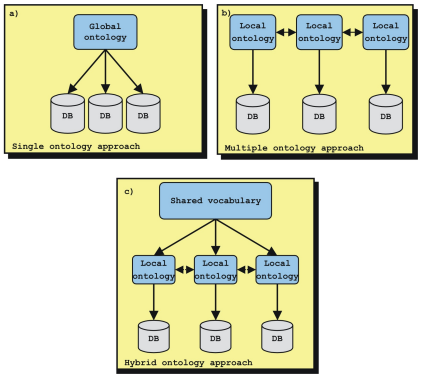

Figure 1: The three possible ways for using ontologies for content explication

for the interpretation requirements of the machines running on the Semantic Web. To have a complete ontology is an extremely difficult task. It can only be verified by domain experts and use over a certain amount of time. Therefore, these two requirements are orthogonal and there is a clear trade-off between minimalism on one hand and completeness on the other.

General idea: The obvious and practical process in the development of a complete and minimal ontology is, to start with a small ontology and extend it to completion. However, how can we determined the completeness of an ontology?

Our basic idea of determining the completeness of an ontology is to apply an additional and different approach for specification of the meaning of concepts. Such a semantic description specifies directly the semantic inter-correspondences of terms from different sources rather than relating each of them to an ontology and then comparing the relationships. Semantic inter-correspondences are the representation of the semantic relationship between two concepts from different information sources. They are means to describe the semantic differences between the concepts. This different semantic description can be used to verify the completeness (and correctness) of the ontology. Examples of how semantic inter-correspondences can be achieved can be found in (Naiman & Ouksel 1995; Parent & Spaccapietra 1998; Kashyap & Sheth 1996).

With respect to the requirements mentioned above we argue that an ontology can be verified if we know the semantic inter-correspondences between concepts described in the ontology. The drawback of only having an ontology is that we have problems with the completeness and the drawback of having only semantic inter-correspondences is that they don't have explicit semantics. Therefore, we argue that a combination of both approaches would relax the problems mentioned.

An Ontology Construction Method

In this section, we describe the method of how we can construct an ontology that holds the requirements we mentioned in the previous section. First we will describe the algorithm and then discuss the statements in more detail.

Construct initial ontology
repeat
  Extend ontology
  Use this ontology to generate hypothesis for semantic inter-correspondences
  Validate and develop semantic inter-correspondences to verify ontology
until ontology is complete

We start this method with a very simple domain terminology which gradually becomes refined during the verification process. This method not only reduces start-up costs, but captures the very subtle differences in the underlying information sources. Application ontologies do not require global agreements from all sources, which further simplifies their development. In order to make the application ontologies comparable, they use a common global vocabulary organized in a terminology (see section 1). This method is reiterated several times until the author of the ontology becomes satisfied with the ontologies degree of completeness.

1. The first step of the process is the acquisition of the semantics of each information source. These sources are described separately without the consideration of other sources. Once this step is completed we can use inference mechanisms of underlying machines to generate a hypothesis which are hidden in the ontology.
2. In the next step we considered the relationship to other sources. For semantic description, a concept from different information sources is labelled with terms from a common domain terminology. This implies the acquisition of its application ontology. In the second step, concepts from different sources are semantically compared. These two concepts are related with so-called semantic inter-correspondences that estimate their semantic difference.

Please note that these two kinds of semantics, the domain semantics and the semantic differences, correlate because they are acquired separately in two steps and can be verified against each other. Figure 2 demonstrates the overall process. Task A represents the development of a minimal ontology whereas task B represents the correspondence in regards to semantic inter-correspondences. The next question is: How do we construct a minimal ontology, which

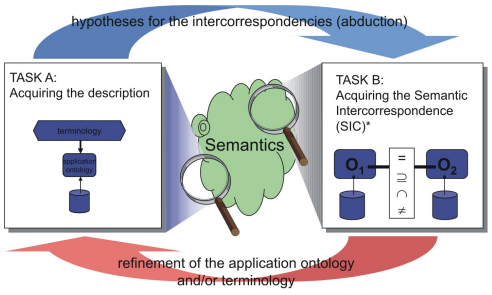

Figure 2: Ontology construction method

Complex terms (l)	ASCII	$\Pi(l) \mapsto$ Concept Term	Interpretation $(\Pi(l))^I$
$l_1 \sqcap l_2$	(AND l_1 l_2)	$l_1 \sqcap l_2$	$l_1^I \cap l_2^I$
$l_1 \sqcup l_2$	(OR l_1 l_2)	$l_1 \sqcup l_2$	$l_1^I \cup l_2^I$
$\neg l_1$	(NOT l_1)	$\neg l_1$	$U \setminus l_1^I$
$l_1 \rhd l_2$	(l_1 OF l_2)	$l_1 \sqcap \exists to.l_2$	$l_1^I \cap \{k^I \mid \exists k^I : (k^I, l_2^I) \in to^I\}$
$l_1 \oplus l_2$	(COMP l_1 l_2)	$\exists$*has-part*.$l_1 \sqcap \exists$*has-part*.l_2	$\{k_1^I \mid \exists k_1^I : (k_1^I, l_1^I) \in$ *has-part*$^I\} \cap$ $\{k_2^I \mid \exists k_2^I : (k_2^I, l_2^I) \in$ *has-part*$^I\}$

Table 1: Complex terms and their corresponding terms in description logics

tools do we use in the process and how do we find semantic inter-correspondences?

MESA

The proposed ontology construction method is implemented in a tool called MEdiator Specification Assistant (MESA). In the following subsections we describe the hybrid ontology approach in MESA and how ontologies can be constructed for information sources. We also discuss which kind of semantic inter-correspondences are supported and how these different kinds of semantic description interact.

A concept defined in an information source represents a set of all possible instances in the database. Each instance of this concept represents a real-world object. For the semantic description, we focus on real-world objects which provide the semantic basis for the concepts. In the follwing the objects will be referred as *extensions* of the concept.

The hybrid ontology approach in MESA

MESA follows the hybrid ontology approach (see figure 1c), which means that an application ontology is constructed for each information source. The terms for the ontologies consists of terms from a global terminology. The global terminology can be seen as a set of primitive terms of a domain. Primitive terms in a financial domain for example can be "stock" or "price". These primitive terms can be combined to complex terms with the help of combination operators (Wache *et al.* 1999). The combination operators (e.g. `AND`, `OR`, and `NOT`) are well-known from description logics but are extended for other purposes, e.g. in the area of information integration. With the help of the `OF`-operator, for example, primitive terms "price" and "stock" can be combined to the complex term "price-OF-stock" indicating that the primitive term "price" is refined to a price of stocks.

All operators can be modelled in description logics (see table 1). Therefore, it is possible to compare complex terms, i.e. it can be inferred if one complex term is a specialization of another or two complex terms are disjoint or overlap. A detailed definition of the combination operators and their semantics can be found in (Wache *et al.* 1999).

In order to explicitly specify the semantics of a concept in an information source, a function L can be established which maps a concept C of a source to its complex term $L(C) = l$. The function can also be viewed as a annotation process of the concept with a semantic description. We call such an annotation $L(C)$ *semantic label* (or label, shortly) in MESA.

Defining labels according to the structure of a concept leads to a combination of complex terms. I.e. the relations of a concept C to other concepts C_i – the structure of the concept – can be used to arrange the labels of C and C_i in the same way. These arranged labels build the application ontology. Please note that each information source can arrange the same complex terms in different ways reflecting the fact that each information source arrange their information in different ways.

In the first step of the ontology construction method the user starts with a small set of the most important primitive terms and annotates the real-world objects with labels, i.e. the combined primitive terms. If more primitive terms are needed the user can extend the terminology. Furthermore, the user only has to annotate one information source without any respect to other information sources. The relationship between concepts of different information sources is the task of the semantic inter-correspondences.

Semantic inter-correspondences in MESA

We define semantic inter-correspondences between concepts C_1 and C_2 based on their extensions as mentioned above. These extensions $E(C)$ of a concept C are defined as the set of real world objects, represented by C. As these sets cannot be captured automatically, the semantic inter-correspondences rely strongly on the specification of the user who should analyze the extensions of the concepts in the corresponding information sources. The reflection of whether two extensions of concepts of different information sources represent the same real-world objects can only be done by a domain expert. Based on these extensions, we classify four semantic inter-correspondences between two concepts C_1 and C_2 (Spaccapietra, Parent, & Dupont 1992):

- *Semantic equivalence* $C_1 \equiv C_2$: The concepts C_1 and C_2 represent the same set of real-world objects, $E(C_1) = E(C_2)$.
- *Semantic subsumption* $C_1 \subset C_2$: The set of real-world objects represented by the concept C_1 is a subset of the set represented by C_2, $E(C_1) \subset E(C_2)$.
- *Semantic intersection* $C_1 \cap C_2$: The sets of real-world objects represented by the concepts C_1 and C_2 overlap partially, while $E(C_1) \not\subset E(C_2)$ and $E(C_2) \not\subset E(C_1)$, so that there is a real intersection between the extensions of the concepts, $E(C_1) \cap E(C_2)$.
- *Semantic incompatibility* $C_1 \neq C_2$: The sets of real-world objects $E(C_1)$ and $E(C_2)$ are completely disjunc-

Labels	Semantic Inter-Correspondences: $C_1 \equiv C_2$	$C_1 \subset C_2$	$C_1 \cap C_2$	$C_1 \neq C_2$
$L(C_1) \equiv L(C_2)$	OK	specialize $L(C_1)$ or generalize $L(C_2)$	specialize $L(C_1)$ and $L(C_2)$	error
$L(C_1) \subset L(C_2)$	specialize $L(C_2)$ or generalize $L(C_1)$	OK	specialize $L(C_2)$ or generalize $L(C_1)$	error
$L(C_1) \cap L(C_2)$	specialize $L(C_1)$ and $L(C_2)$	specialize $L(C_1)$ or generalize $L(C_2)$	OK	disjoint $L(C_1)$ and $L_(C_2)$
$L(C_1) \neg L(C_2)$	error	error	disjoint $L(C_1)$ and $L(C_2)$	OK

Table 2: Verifying labels with semantic inter-correspondences

tive, therefore C_1 and C_2 represent different real-world objects.

These four classes of semantic inter-correspondences are sufficient to describe the relationship between two concepts of different information sources. These classes are defined similar to the common set operators for relations between sets. However, there are also other approaches for modelling relationships between concepts, such as *semantic proximity* by Kashyap & Sheth (1996). The semantic proximity defines the semantic relationship between two concepts based on a context and the level of abstraction between compared concepts. The inter-schema correspondence assertions by Naiman & Ouksel (1995) is another approach, which uses levels of abstraction as well as levels of heterogeneity between compared concepts.

One reason for using the semantic inter-correspondences approach in MESA is its simplicity. The approach also has the lowest modelling effort if compared to the other approaches.

Verify ontology with semantic inter-correspondences

The labels and the semantic inter-correspondences are in close relationship to each other since labels can be verified by the semantic inter-correspondences. The user establishes semantic inter-correspondences considering the extensions of the concepts instead of considering the labels. If the semantic inter-correspondences are established between two concepts C_1 and C_2, there might be differences between the semantic inter-correspondences and the labels of the concepts. Assuming that a previously defined semantic inter-correspondence is correct, table 2 shows which label has to be modified and how. Disjoint labels or semantic incompatibility indicates a modelling error. Therefore, the labels and the semantic inter-correspondences must be checked. In all other cases it is sufficient to refine one or both labels. A refinement is done by specializing or generalizing. A specializing process is done by choosing another primitive term from the terminology or by creating a new primitive term, which is then added to the terminology. The primitive term could either replace the old label of the concept or could be used to extend the label with one of the given combination operators. In case of generalization a label must be simplified by deleting some primitive terms. After a refinement the semantic inter-correspondences has to be validated with respect to the refined label(s). This refinement and validation process re-iterates until the semantic inter-correspondences and the label relationships harmonize (diagonal in table 2).

From a logical point of view, the labels are interpreted as a set in a universe. Comparing the labels of two concepts means comparing their two sets in the universe. The semantic inter-correspondences also compares two sets of extensions. Extensions are the real world objects and therefore can be seen as a "given" universe. If we compare the relationship between two label sets and two extension sets (as in table 2), we verify whether the relationship between the label sets includes the relationship of the extensions. In other words, table 2 compares two relationships of model sets.

Supporting the Knowledge Engineer

One can argue that the additional modelling effort is too tedious within the acquisition phase of an ontology and that the practical relevance is low. However, we can use logical reasoning methods to support the process. The specification of the semantic inter-correspondences can be simplified by using the principle of abduction for finding them. The implication that a semantic inter-correspondence between two concepts requires certain labels for the concepts can be used in the inverse direction by abducting the semantic inter-correspondences from the labels. If the concepts in information sources have been labelled previously, these labels could be used to create a hypothesis on how the relationship between concepts of different sources might be and thus define the semantic inter-correspondences. This process could be supported by a software assistant, which automatically creates the hypothesis for the semantic inter-correspondences. Such an assistant has been developed and implemented in the MESA system, offering the possibility to find semantic inter-correspondences in pre-labelled information sources.

In a financial domain for example two concepts C_1 and C_2 of different sources represent information about securities. Both concepts can be annotated with the label "security". In the real world though C_1 only represents stocks. The knowledge engineer usually focusses on one informa-

tion source while constructing an ontology and would miss the difference. Furthermore, it is practically impossible for the knowledge engineer to have all concepts of all sources in mind. Therefore it is reasonable to annotate C_1 only with "security" omitting further details. MESA, however, can suggest the hypothesis that C_1 and C_2 are semantically equivalent $C_1 \equiv C_2$ because they have the same labels ($L(C_1) \equiv L(C_2)$). In the verification process the domain expert needs to verify the generated semantic inter-correspondence. If he decides that C_1 is a subset of C_2 ($C_1 \subset C_2$) the mismatch becomes obvious and the engineer has to specialize the label of C_1 ($L(C_1)$) or generalize the label of C_2 ($L(C_2)$).

Configuring an Integration system

Additionally we can use the semantic inter-correspondences to achieve interoperability between systems using the developed ontology. Once we have developed semantic inter-correspondences we are able to generate an integration formalism in form of propositional rules. A number of systems in the area of intelligent information integration, are using this kind of rules (Chawathe *et al.* 1994; Wache & Stuckenschmidt 2001).

Conclusion

Both the ontology-driven approach and the semantic inter-correspondences approach can be used to define semantical knowledge needed for the Semantic Web. However, these approaches are labor-intensive and are therefore serious obstacles to their widespread use. One can use both approaches on their own to fulfil the needs of the Semantic Web and both approaches are useful means to support the development of methods towards the activity with respect to the World Wide Web.

In this paper, we have proposed a new approach to gradually develop a domain ontology and to acquire semantic inter-correspondences at the same time. The combination of these two approaches provides us with an ontology, which is complete, an important if not necessary property. We argue that the integration of these two approaches is even better than the single approaches due to the completeness of the ontology. In contrary, one can argue that the demand of minimalization is not fulfilled. The opposite is true: we start with a minimal ontology and use inference mechanisms to derive hidden knowledge. This knowledge can be verified by a domain expert who at the same time is defining semantic inter-correspondences between terms. This will lead to new knowledge of the domain and the process starts again. We think that the demand of minimalization is fulfilled. Given the fact that we would like to develop a complete and minimal ontology this would be one way to achieve the goal.

The approach also allows the support of software assistants, which can help to acquire and verify the semantics of information sources. MESA is such an assistant for the knowledge acquisition process. It supports both the domain expert and the knowledge engineer. This tool is not only a comfortable graphical interface, it also suggests semantic inter-correspondences by an inference mechanism. Both the method and the tool are helpful means in order to make the vision of the Semantic Web a reality.

References

Bernes-Lee, T.; Hendler, J.; and Lassila, O. 2001. The semantic web. *Scientific American* (5).

Chawathe, S.; Garcia-Molina, H.; Hammer, J.; Ireland, K.; Papakonstantinou, Y.; Ullman, J.; and Widom, J. 1994. The tsimmis project: Integration of heterogeneous information sources. In *Conference of the Information Processing Society Japan*, 7–18.

Ciocoiu, M., and Nau, D. S. 2000. Ontology-based semantics. In *Knowledge Representation (KR)*.

Fensel, D.; Harmelen, F. v.; Horrocks, I.; McGuinness, D. L.; and Patel-Schneider, P. F. 2001. Oil: An ontology infrstructure for the semantic web. *IEEE Intelligent Systems* 16(2):38–44.

Gruber, T. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5(2):199–220.

Guarino, N. 1998. Formal ontology and information systems. In Guarino, N., ed., *FOIS 98*. Trento, Italy: IOS Press.

Kashyap, V., and Sheth, A. 1996. Schematic and semantic semilarities between database objects: A context-based approach. *The International Journal on Very Large Data Bases* 5(4):276–304.

Naiman, C. F., and Ouksel, A. M. 1995. A classification of semantic conflicts in heterogeneous database systems. *Journal of Organizational Computing* 167–193.

Parent, C., and Spaccapietra, S. 1998. Issues and approaches of database integration. *Communications of the ACM* 41(5):166–178.

Spaccapietra, S.; Parent, C.; and Dupont, Y. 1992. Model independent assertions for integration of heterogeneous schemas. *VLDB Journal: Very Large Data Bases* 1(1):81–126.

Wache, H.; Scholz, T.; Stieghahn, H.; and König-Ries, B. 1999. An integration method for the specification of rule–oriented mediators. In Kambayashi, Y., and Takakura, H., eds., *Proceedings of the International Symposium on Database Applications in Non-Traditional Environments (DANTE'99)*, 109–112.

Wache, H.; Vögele, T.; Visser, U.; Stuckenschmidt, H.; Schuster, G.; Neumann, H.; and Hübner, S. 2001. Ontology-based integration of information - a survey of existing approaches. In Stuckenschmidt, H., ed., *IJCAI-01 Workshop: Ontologies and Information Sharing*, 108–117.

Wache, H., and Stuckenschmidt, H. 2001. Practical context transformation for information system interoperability. In Akman, V.; Bouquet, P.; Thomason, R.; and Young, R., eds., *Modeling and Using Context*, volume 2116 of *Lecture notes in AI*. Proceedings of the Third International and Interdisciplinary Conference, CONTEXT, Dundee, UK: Springer Verlag. 367–380.

Wiederhold, G. 1992. Mediators in the architecture of future information systems. *IEEE Computer* 25(3):38–49.