Proposal of a Method to Assist The Ontological Study of Domains

Rafael Odon de Alencar, Luis Enrique Zárate

Applied Computational Intelligence Laboratory
Pontifical Catholic University of Minas Gerais
Av. Dom José Gaspar 500, Prédio 34, Sala 111 - Coração Eucarístico
CEP 30535-610 Belo Horizonte - MG - Brasil

odon.rafael@gmail.com, zarate@pucminas.br

Abstract. *Ontologies, since its formal definition in the 90's up to its effective usage with the advent of methods and processes from the Ontology Engineering, have received important attention in projects that demand the formalization of shared knowledge among applications and users. In this context, it is important to popularize and to make people able to create and use them. This article presents a method to assist the ontology capture process, providing apparatus to conceptualize and identify the treated domain trough an ontological study. The proposed method provides metrics and guidelines so that an ontology engineer can identify and organize the elements of a domain, finding fundamental ontological relations among them. A descriptive algorithm is shown to formalize the wished process and some examples are given to better exemplify the utilization of the proposed method.*

1. Introduction

An ontology can be understood as a computational artefact which represents the knowledge of a specific domain in a formal way, expliciting its elements and their attributes, relations and restrictions, thus defining the vocabulary used to describe such domain in a given language. [Gruber 1993] defines ontology as an explicit specification of a conceptualization. Many other authors have commented the terminology around this subject [Guarino and Giaretta 1995, Studer et al. 1998]. During the 90's, the usage of ontology was being consolidated by the proposal of different tools and methods for its construction, evaluation, reusing, sharing, what resulted in the sprouting of the Ontology Engineering [ALMEIDA and BAX 2003, Corcho et al. 2003, Gomez-Perez et al. 2004, Guarino 1998, Jones et al. 1998, Sure et al. 2006].

The Guber definition of ontologies uses the conceptualization idea presented in [Genesereth and Nislsson 1987]. These authors affirm that it is possible to exist more than one conceptualization about the same domain, evidencing the subjective character of this task. It is also said that the conceptualization is as an "invention", existing an ontological commitmentness according to the authors, in such a way that concepts are always accepted if they are useful to their system. The way these concepts will be founded and related can refers to the ontology capture process. [Uschold and King 1995] defined in the first phase of their ontology building method the task of ontology capture, also presented in other works [Uschold and Gruninger 1996]. According to these authors, this task consists in find the key concepts and the existing relations in the domain. [Falbo et al. 1998] say that this task is one of the most important ones and its goal is to capture the domain conceptualization based on the ontology competence. The result of this activity must contain relevant concepts and relations identified and organized.

The proposal of this work is to supply a method to the ontological study of domains called *Sphere-M*, applicable in the ontology capture process, carried through for a human specialist, since this task still does not possess procedures or methods that guide its subjectivity through clear rules and tools. The method's result is a model obtained through the ontological study of a domain. [Welty and Guarino 2001] presented a methodology for ontological analysis of taxonomic relationships that also provide ontological care during the conceptualization process through very solid concepts extracted from philosophy. The *Sphere-M* method intend to work one step before, providing a way to create a more simple model, encouraging people to start the ontology building process by carefully investigating the domain. The proposed method intends to be the basis of a tool to be applied in many problems where ontologies appear.

This article is organized in three sections. In the second one, it is presented the *Sphere-M* method to capture ontology, first showing its fundaments, then a high level approach of the process and an algorithmic approach. After then, metrics to evaluate applications of the method are presented, and then examples of these applications are given. Finally, we discuss the contributions and the conclusions of this work.

2. *Sphere-M* Contributions and Proposals

2.1. Philosophical contributions used in the *Sphere-M*

To elaborate the *Sphere-M*, it was extracted some fundaments from philosophical ideas that were considered important to the process of reality investigation. They were synthesized from readings in Parmenides, Socrates, Descartes and Locke ideas about the knowing, knowledge theory and philosophical ontology. The fundaments are applied in the *Sphere-M* usage, constituting what we call here *Fundamental Ontological Analysis* (FOA). It consist in:

1. *If an element can vary some characteristics, each variation of it corresponds to another more specific element.*
2. *If an element has many characteristics, the abstraction of some of them can give a new more generic element.*
3. *If it's possible to separate an element in its components or characteristics, each single part will represent a new element.*
4. *If it's possible do take a group of elements as components or characteristics of something, defining a consistent set, it will represent another element.*

- Guide-rule: *Check if a generated element is practicable.*

The first fundament will produce relations similar to *is-a*, *is-a-variation-of*, *is-a-type-of*, *is-a-specification-of*. The second one will produce *is-a-generalization-of* relation kind. The third fundament produces relations similar to *is-an-attribute-of*, *is-a-component-of*, *is-an-element-of*, *is-part-of*. The fourth fundament is responsible by relations such as *is-composed-by* or *is-characterized-by*. It is interesting to note that the first fundament opposites to the second, as well the third opposites the fourth. This shows that the investigative process can occur in both directions: specification/generalization, whole/part.

The fundaments are very basic issues that sound obvious. However, it is important to understand them clearly, taking the FOA as basic rules in the reality investigation process. Considering the high subjectivity of the conceptualization task during the ontology capture, this process must be laid over some formal definitions such as FOA does. Later, in the method steps, FOA will be responsible by justify the relations among all domain elements.

2.2. High Level Approach Of The *Sphere-M* Method

The method is based on an middle-out technique, cited as an balanced choice in the [Uschold and King 1995] ontology building method. To keep the Middle-Out characteristics, the *Sphere-M* method does not define an upper or lower limit to the specification and generalization respectively. All the initial elements of the domains can suffer generalizations and specifications.

In order to define a stop condition to the domain analysis process, the *Sphere-M* is based on the definition of key sets that are: *Ceiling*, *Floor* and *Relevant Elements*. The *Ceiling* set is the one that contain the most general elements who define the subject approached in the domain. The *Floor* set is the one that has all the entities and elements that is wished to belong to the domain. The *Relevant Elements* is composed by elements that define the focus of the wished relations among the *Ceiling* and *Floor* sets. These three sets joined, initially, represent the *Domain Elements*, which is the working set from where the process starts. The stop condition proposed by the sets definitions will be explained later in the text.

The way that each set is obtained must be defined according to the characteristic of each applicable situation. Thus, the method has the role of a framework that must be adapted to each source of knowledge or capture model chosen by its user. After establish such method, one follows the process below:

1. *Define the key sets.*
2. *Take the key sets join as the Domain Elements set.*
3. *Submit the Domain Elements to the FOA relating them. Generate new elements if it is necessary.*
4. *Check the link of all Domain Elements with the Ceiling and Floor sets.*
5. *Back to the third step if there is any element not linked to Ceiling and Floor sets.*

To this steps work, it may be understood that the element link to *Ceiling* and *Floor* only exists if there is a path that can connect the element to a *Ceiling* element, and a path that can connect it with a *Floor* element. (See Figure 1)

The FOA intends to supply a way to obtain fundamental ontological relations. In this case, to make the user apply properly the FOA during all the investigation, it is necessary to consider the affirmative: every connection between elements must be associated to a relation justified by a FOA fundament. This affirmative brings an important restriction to the investigative process because it'll guarantee that the user understand the path followed during this ontological investigation. With this rule, it's intended to have a more coherent result, avoiding domain elements of being included in the model without a motivation.

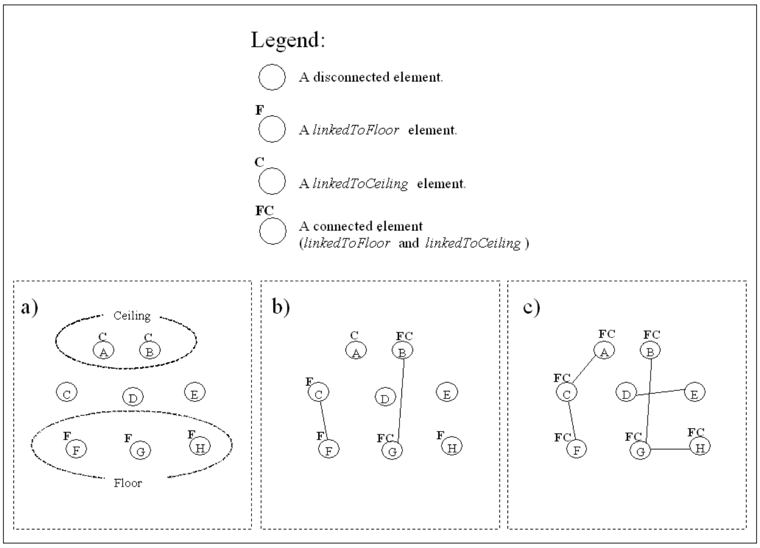

Figure 1. a) Initial state from the elements, showing the *Ceiling* and *Floor* sets. b) The (C,F) relation makes C be linked to the *Floor*. The (B,G) relation makes both B and G to be linked to *Floor* and *Ceiling*. c) After some relations, D and E still don't have a link to *Ceiling* or *Floor*, representing sphere-disconnected elements).

The sphere was chosen to name the method because it is the wished analogy to represent the knowledge relative to a domain according to the methods proposal. It is known that to consider a solid as a sphere, all its surface points must have the same distance from the center. If a single point doesn't follow the rule, there is no sphere, because it will represent an irregular body. To the *Sphere-M*, knowledge will only be obtained when all elements respect the asked relation with the *Ceiling* and *Floor* sets. After all elements reach this related state, is doesn't mean that this sphere cannot change its properties, increasing or decreasing the number of relations and elements. It is still possible to continue the process in order to wide your domain or increase the detail level.

2.3. Algorithmic Approach Of The *Sphere-M* Method

It was suggested the usage of metrics about the described process. To better formalize such metrics, an algorithmic view of the method will be presented, working with the sets and their cardinalities. Consider the set definitions:

- **Ceiling set** (*C*): initial general elements.
- **Floor set** (*F*): initial elements intended to belong the domain
- **Relevant Elements set** (*R*): relevant elements that must guide the *C* and *F* connection.
- **Generated Elements set** (*G*): generated element in a given iteration.
- **Domain Elements set** (*D*): total elements. It starts as ($C \cup F \cup R$)
- **Elements Relations set** (*L*): relations between elements.

A descriptive algorithm to the *Sphere-M* method is showed in Algorithm 1.

Algorithm 1 Sphere-M

$C \leftarrow$ input from user
$F \leftarrow$ input from user
$R \leftarrow$ input from user
$D \leftarrow C \cup F \cup R$
$G \leftarrow$ empty
$L \leftarrow$ empty
for all $c \mid c \in C$ **do**
    Mark c as *linkedToCeiling*
end for
for all $f \mid f \in F$ **do**
    Mark f as *linkedToFloor*
end for
repeat
    $D \leftarrow D \cup G$
    $G \leftarrow$ empty
    repeat
        Apply FOA over all D elements
        Add to L the new relations obtained
        If an element relate to another marked as *linkedToCeiling*, mark it and propagate the mark to its adjacents.
        If an element relate to another marked as *linkedToFloor*, mark it and propagate the mark to its adjacents.
        Add to G the new elements generated
    until (there is nothing to relate or generate with D elements)
until (e is *linkedToCeiling* and *linkedToFloor*, $\forall\ e \in D$)

First, the user must supply the initial sets (*C*, *F* and *R*). The *D* set is initially equivalent to the union of *C*, *F* and *R*, being the used set during the first iteration. As the method goal is to obtain a connected set of elements, where every one has as path to reach the *C* and the *F* sets, a technique is used to check this connection: initially, all the elements from *C* are marked as *linkedToCeiling*, and all the elements from *F* are marked as *linkedToFloor*. When an element is related to other, if one of them has the *linkedToCeiling* or *linkedToFloor* marks, the other must receive the mark and then propagate it to its adjacent using a graphs breadth search. The methods application consists in a repetition loop responsible to promote the *n* iterations needed until reach the connected-state of the sphere, that is when all domain elements are *linkedToFloor* and *linkedToCeiling*. Inside this loop, there are another one responsible by promote the FOA application in all the elements and subsets from *D*. The obtained relations from this analysis are stored in the *L* set, and the new elements generated from the existing ones are stored in the *G* set. On the next iteration the elements of *G* are added to the *D* set, and then *G* becomes an empty set. This guarantees that the generated elements are only used in the next iteration.

2.4. Proposed Metrics to Evaluate the Process

From the algorithm representation of the *Sphere-M*, it is possible to suggest some wished metrics to evaluate the conceptualization process promoted by the method application over a domain. They can be used to quantify and qualify the path used to obtain a final model when the method is applied.

- **Number of iterations**: represents how many steps were walked trying to relate the domain. The number of main loops gives it.
- **Sphere radius:** when the connected-state of the sphere is reached, the number of iterations will represent the sphere radio.
- **Number of elements in the sphere:** in any iteration, the number of elements in the sphere can be obtained by calculating the cardinality of the *D* set (DOMAIN ELEMENTS).

- **Density:** during a given iteration, it is possible to calculate the density of the sphere by checking the reason between the number of connected-elements (both marked as *linkedToCeiling* and *linkedToFloor*) and the total number of elements in the sphere. If the density is 1.0 the connected-state was reached. After then, new iterations will not affect this value.

- **Iteration Efficiency**: the maximum relations that can be obtained in a D set is given by:

$$MAX = \frac{|D| * (|D| - 1)}{2}$$

  The efficiency of a given iteration is the reason of the number of relations obtained until it (cardinality of the L set), and the maximum number of relations (*MAX*).

- **Sphere Productivity**: if the connected-state of the sphere is reached, its productivity can be calculated by the reason of the number of elements on the sphere and the sphere radius.

$$\textit{Sphere Productivity} = \frac{\text{Number of elements in the sphere}}{\text{Sphere radius}}$$

  The productivity is very useful to evaluate a current process, or to compare processes around the same initial sets. However, this metric is not indicated to compare processes from different domains and purposes because different sets of elements can generate different difficulties and conceptualization options.

2.5. Applied Examples of The *Sphere-M* Usage

Some examples of the Sphere-M application are presented next. The examples were developed using the method manually. After some first basic tests, a prototype was implemented to help the execution of the graphs breadth search suggested by the method during the propagation of the *linkedToCeiling* and *linkedToFloor* marks. Details about future implementations are discussed in the Conclusion section.

2.5.1. Example 1: Capturing the ontological model through textual specialist

To present an example of the methods application, it is proposed a class of problem where a human user will act like the domain specialist by extracting the knowledge from a text. Consider then the text extracted from *Wikipedia*, the problem definition and the segments extracted from the text below:

- **Text**: "*The telephone is a telecommunications device which is used to transmit and receive sound (most commonly speech). Most telephones operate through transmission of electric signals over a complex telephone network which allows almost any phone user to communicate with almost any other.*"

- **Problem definition**: it is desired to get a model from the subject Telephone according to the text. It is important to specify the issues: Telephone Working, Telephone Utilization.

- **Text segmentation**: Telephone, Telecommunications Device, Transmit Sound, Receive Sound, Transmission Of Electric Signals, Telephone Network, Phone User, Communicate

The Ceiling set (C) will be the subject of the text. The Floor set (*F*) will be composed by all the elements extracted from the text. The Relevant Elements (*R*) set will be composed by the terms that represent the proposed issues on the problem definition. The Domain Elements set (*D*), initially is $C \cup F \cup R$:

- ***C*** = {*Telephone*}
- ***E*** = {*Telecommunications Device, Transmit Sound, Receive Sound, Transmission Of Electric Signals, Telephone Network, Phone User, Communicate*}
- ***F*** = {*Telephone Working, Telephone Utilization*}
- ***D*** = $C \cup F \cup R$:

Now it is necessary to analyze each element from D, trying to apply the four fundamentals from FOA. If some elements are generated, they can't be analyzed during this iteration. They will be merged to the D set on the next analysis. The resultant relations of the analysis are shown below with their respective FOA fundament number justifying it. It is suggested to execute the algorithm here, helping to control the marks *linkedToCeiling* and *linkedToFloor*.

Telephone (is a) Telecommunications Device. (II) *Transmit Sound (is part of) Telephone Utilization.* (III) *Receive Sound (is part of) Telephone Utilization.*(III) *Transmission Of Electric Signals (is part of) Telephone Working.*(III) *Phone User (is part of) Telephone Utilization.* (III)	*Communicate (is a) Telephone Utilization.* (II) *Telephone Working (is part of) Telephone.* (III) *Telephone Utilization (is part of) Telephone.*(III) *Telephone (is part of) Telephone Network.* (III)

After create this relations, all elements from D are marked as *linkedToCeiling* and *linkedToFloor*, which means that it was obtained a solid sphere. Now it is possible to stop the analysis. However, it's interesting to consider the metrics before stopping.

Number of iterations: 1
Sphere Radius: 1
Number of elements: 10
Density: 10 ÷ 10 = 1
Iteration Efficiency: 9 ÷ 45 = 0.2
Productivity: 10 ÷ 1 = 10

As the efficiency now is 0.2, one concludes that it is possible to continue the analysis, relating the existing elements or creating new elements trying to reach a higher efficiency. Suppose the new relation obtained then:

- *Transmission Of Electric Signals (is a) Transmission*. (II)

This relation generates a new element *Transmission* that can only be related to the others elements on the next iteration, forcing the user to try to full relate the actual *D* set. As the sphere is already connected, new iterations will not affect the density. But the productivity can be increased or decreased depending on the importance of he generated elements to the model. In the second iteration, the initial metrics are:

Number of iterations: 2
Sphere Radius: 1
Number of elements: 11
Density: 11 ÷ 11 = 1
Iteration Efficiency: 10 ÷ 55 = 0.181
Productivity: 11 ÷ 2 = 5, 5

Considering that no more relations could be found among *Transmission* and the existing elements, and there is nothing useful to create or relate, suppose that the process stops here. With this second iteration, the efficiency and the productivity decreased. As the full connectivity was obtained in the first iteration, the generated element *Transmission* could be discarded for example, because it wasn't helpful to improve the metrics and didn't produced a better model.

2.5.2. Example 2: Capturing the ontological model through a website structure

Is it's proposed to apply the *Sphere-M* using as input elements the name of sections, subsections and links from an website. The problem to be solved by this technique is: to present an ontological model able to represent the content of a website, verifying by this model the relation of each collected element with the treated domain. The *C* set is compounded by the title of the document and by other elements which describe the subject of the presented website. The *R* set is compounded by the titles of the main sections of the website. The *F* set is compounded by the titles of the sub-sections and by the hyper-links words founded in the content of these sections. One suggests that the user must filter the collected elements, removing the purely functional elements that only exists to navigate and operate the website. It is also suggested to check the presence of synonymous and abbreviations, or even the terms denoted by alternative forms, generating a list of elements easy to be used.

In this example the site from LICAP (http://www.inf.pucminas.br/projetos/licap, accessed in 01/08/2007) was used to experiment. The website is about the laboratory where this project is being developed, and consists in some main sections presenting the members, projects, publications and other details about the group. The obtained sets are:

- ***C*** = {*Licap*}
- ***E*** = {*Integrantes, Projetos, Publicações, Softwares Desenvolvidos, Ambiente Computacional*}
- ***F*** = {*Historia (do LICAP), Objetivos (do LICAP), Professores, Alunos de Pós Graduação, Alunos de Graduação, Ex-Alunos, Projetos Coordenados Pelo LICAP, Projetos que possuem a colaboração do LICAP, Máquinas, Impressoras, Softwares, Email, Endereço, Instituto de Informática, Puc Minas, CNPQ, FAPEMIG, Currículo Lattes*}
- ***D*** = $C \cup F \cup R$:

In the first interaction, the obtained relations are:

Integrantes (is part of) Licap. (III)	*Impressoras (is part of) Ambiente Computacional.* (III)
Projetos (is part of) Licap. (III)	*Softwares (is part of) Ambiente Computacional.* (III)
Publicações (is part of) Licap. (III)	*Softwares Desenvolvidos (is a variation of) Softwares.* (II)
Softwares Desenvolvidos (is part of) Licap. (III)	*Email (is an attribute of) Licap.* (III)
Ambiente Computacional (is part of) Licap. (III)	*Endereço (is an attribute of) Licap.* (III)
Professores (is a variation of) Integrantes. (II)	*Licap (is part of) Instititto de Informática.* (III)
Alunos de Pós Graduação (is a variation of) Integrantes. (II)	*Instituto de Informática (is part of) Puc Minas.* (III)
Alunos de Graduação (is a variation of) Integrantes. (II)	*Currículo Lattes (is an attribute of) Integrantes.* (III)
Ex-Alunos (is a variation of) Integrantes. (II)	*CNPQ (is a) Órgão de fomento à pesquisa**. (new)*** (II)
Projetos Coordenados (is a variation of) Projetos. (II)	*FAPEMIG (is a) Órgão de fomento à pesquisa. **(new)*** (II)
Projetos Colaborados (is a variation of) Projetos. (II)	*Currículo Lattes (is part of) Plataforma Lattes**. (new)*** (III)
Máquinas (is part of) Ambiente Computacional. (III)	

Some elements were generated:

- ***G*** ={*Órgão de fomento à pesquisa, Plataforma Lattes*}

The metrics at this point of the first iteration are:

Number of iterations: 1
Sphere Radius: (no connected state)
Number of elements: 24
Density: $22 \div 24 = 0.91$

Iteration Efficiency: 20 ÷ 276 = 0.0724
Productivity: (no connected state)

Although the content was easily related, the elements ”CNPQ” and ”FAPEMIG” were not directly related do the domain. To right relate them, it was generated new elements. The connected state of the sphere was not reached yet, so, a new iteration must follow, trying to relate the elements considering now the generated ones. In this new iteration, the new relations founded are:

- *Órgão de fomento à pesquisa (is an attribute of) Projetos.* (III)
- *Plataforma Lattes (is part of) CNPQ.* (III)

The metrics at this point of the second iteration are:

Number of iterations: 2
Sphere Radius: 2
Number of elements: 26
Density: 26 ÷ 26 = 1.0
Iteration Efficiency: 26 ÷ 325 = 0.08
Productivity: 26 ÷ 2 = 13

As it can be seen, the efficiency metric value of the last iteration was better than the first one, because the generated elements could be related to other ones and not only to is generator element. These metrics values could continue to change if the user keep relating and creating elements. It is interesting to note here that the method doesn’t judge the path followed by the domain analysis. The method only helps to organize the knowledge about the domain elements and its relations. The user is responsible by the quality of the knowledge that he is using. The role of the Sphere-M method is to guarantee a minimum coupling level among the elements of the domain, thus justifying the presence of a given element in this domain according to the specialist choices

3. Conclusion

A method to help the ontology capture process, called Sphere-M, was presented in this work. This method defines rules and fundaments to guide the domain conceptualization, helping its user to find basic ontological relations among the elements of the chosen domain. There are many paths to reach a wished coupling level among elements trough the Sphere-M. Once one reaches this level, the method offers apparatus to refine the ontological study and watch the model evolution until the end of the method usage. The method seems to be simple in order to be used in the beginning of the ontology building process, encouraging the ontology engineer and the domain specialist to create a first model to be used later in this building process. One of the most important aspects proposed are the Sphere-M metrics, used to quantify and qualify the path used to obtain a final model. The metrics can be used as a referential to evaluate the capture process when it involves human interaction, allowing comparisons and analysis. The method also presents a framework aspect related to the way that the initial elements are supplied to it.

The presented proposal is the basis for the implementation of a tool that is already being developed at our laboratory. This tool intends to promote the adaptation of *Sphere-M* over existing methods and patterns of the Ontology Engineering. It is wished to implement the tool as a plugin or an external module to export its results to some existing ontology editor.

It is intended also to propose in the future many adaptations of the presented method, applying it in different problems, disseminating the ontology study in domain analysis by providing the software tool based on the present theoric work.

References

Almeida, M. B. and Bax, M. P. (2003). Uma visão geral sobre ontologias: pesquisa sobre definições, aplicações, métodos de avaliação e de construção. *Ciência da Informação*, 32(3):7–20.

Corcho, O., Fernández-López, M., and Gómez-Pérez, A. (2003). Methodologies, tools and languages for building ontologies: where is their meeting point? *Data Knowledge Engineering*, 46(1):41–64.

Falbo, R. D., Menezes, C. S., and Rocha, A. R. (1998). A systematic approach for building ontologies. *In Proceedings of the 6th Ibero-American Conference on Ai: Progress in Artificial intelligence*.

Genesereth, M. R. and Nislsson, N. J. (1987). Logical Foundations of Artifcial Intelligence. Morgan Kaufmann Publishers.

Gómez-Pérez, A., Corcho, O., and Fernandez-Lopez, M. (2004). Ontological Engineering : with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. Springer.

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

Guarino, N. (1998). Formal ontology and information systems. *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98*, pages 3–15.

Guarino, N. and Giaretta, P. (1995). Ontologies and knowledge bases: Towards a terminological clarification. *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, pages 25–32.

Jones, D., Bench-Capon, T., and Visser, P. (1998). Methodologies for ontology development. *In Proceedings of IT&KNOWS Conference, XV IFIP World Computer Congress.*

Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data Knowledge Engineering*, 25(1–2):161–197.

Sure, Y., Tempich, C., and Vrandecic, D. (2006). Ontology engineering methodologies. *In Semantic Web Technologies: Trends and Research in Ontology-based Systems.*

Uschold, M. and Gruninger, M. (1996). Ontologies: principles, methods, and applications. *Knowledge Engineering Review*, 11(2):93–155.

Uschold, M. and King, M. (1995). Towards a methodology for building ontologies. *Workshop on Basic Ontological Issues in Knowledge Sharing, in IJCAI-95.*

Welty, C. A. and Guarino, N. (2001). Supporting ontological analysis of taxonomic relationships. *Data Knowledge Engineering*, 39(1):51–74.