Grounded ontology – a proposed methodology for emergent ontology engineering

Syed Irfan Nabi  
*Institute of Business Administration, Karachi, Pakistan*

Zaheeruddin Asif  
*Institute of Business Administration, Karachi, Pakistan*

Follow this and additional works at: [https://ir.iba.edu.pk/businessreview](https://ir.iba.edu.pk/businessreview)

Part of the Systems and Communications Commons, and the Systems Architecture Commons

This work is licensed under a Creative Commons Attribution 4.0 International License.

**iRepository Citation**


This article is brought to you by *iRepository* for open access under the Creative Commons Attribution 4.0 License. For more information, please contact irepository@iba.edu.pk.
GROUNDONTOLOGY – A PROPOSEDMETHODOLOGY FOR
EMERGENT ONTOLOGY ENGINEERING

Syed Irfan Nabi
Institute of Business Administration, Karachi, Pakistan

Zaheeruddin Asif
Institute of Business Administration, Karachi, Pakistan

Abstract

This research posits that a domain ontology developed using text-coding technique contributes in conceptualizing and representing state-of-the-art as given by published research in a particular domain. The motivation behind this research is to provide means for creating a better understanding among the researchers through ontology that would present a clearer picture of any domain of interest. However, a general observation on ontology engineering methods is the domination of personal perspective of ontology developer and/or expert in the resultant ontology. Current ontology engineering methods bestow a primary role to ontology developer. Ontology thus developed is heavily biased towards the domain expert’s personal understanding of the domain. However, ontology stands a better chance of being unbiased if it is derived from established research such that it is closely linked to the text of the published research, i.e. entities and their relationships are obtained directly from data through coding. Therefore, this new methodology has been proposed (Grounded Ontology - GO) for deriving an ontology directly from published research texts. An ontology developed using this method can enhance visibility of what others have already done and ensure that research efforts in a domain are directed to new vistas instead of being wasted in duplicating the efforts.

Introduction

An Ontology is a conceptual representation of a domain of interest showing entities and their relationships in the universe of discourse according to Hepp (2007). This research posits that a domain ontology developed using text-coding technique contributes in conceptualizing and representing published research in a particular domain. The term Conceptualize is used as a “simplified view of the world that we wish to represent for some purpose” (Thomas R. Gruber, 1995), provided by “an abstraction over domain of interest in terms of its conceptual entities and their relationships” (Hepp, 2007). To build such ontology a modified ontology engineering approach has been Proposed. In this approach the ontology is derived from the text such that all the entities and relationships can be traced back to the original text. It is based on text coding techniques taken from Grounded Theory Method (GTM) of qualitative research and has been named Grounded Ontology (GO).

It is maintained in literature that one of the possible ways of combining and consolidating domain knowledge is through domain ontology (Chandrasekaran, Josephson, & Benjamins, 1999; A. Gómez-Pérez & Benjamins, 1999; T. R. Gruber, 1991, 1993; N.
Guarino, 1995; Noy & McGuinness, 2001). An agreed-upon ontology may lead to a better understanding by providing a common lexicon (Basile, 2011; Chandrasekaran et al., 1999; Ćosić, Ćosić, & Baća, 2011; Harter & Moon, 2011). Thus, ontology can provide a basis for consolidation of knowledge and shared understanding. However, current ontology engineering methods bestow a primary role to ontology developer. A general observation on ontology engineering methods is the domination of personal perspective of ontology developer and/or expert in the resultant ontology. The resultant ontology is heavily biased towards the domain expert’s personal understanding of the domain.

However, an ontology stands a better chance of being unbiased if it is derived from established research such that it is closely linked to the text of the published research, i.e. entities and their relationships are obtained directly from data through coding (Charmaz, 2006; Strauss, 1987). In-vivo coding is a type of text coding method where exact terms from the text are taken as codes to be used subsequently as entities. Through this coding process the coder’s perspective is reduced (Saldana, 2009, p. 76). In other words, with the use of in-vivo coding, the resultant categorization of entities more closely represents the researchers’ (i.e. authors of the research papers used as corpus) findings. It has been demonstrated in literature that coding data to find entities and their relationships is similar to ontology engineering (Kuziemsky, Downing, Black, & Lau, 2007; Urban, 2009).

The objective of this research is to propose a solution to the criticism of current ontology engineering methodologies. In particular we seek to reduce the personal perspective of the expert getting introduced in the resultant ontology. Simultaneously, it would help in enhancing researcher’s point of view through text coding.

The rest of the paper is organized as follows. The next section discusses lexicon and the notion of mutual understanding among people. In Section 3, domain ontology is discussed, starting with fundamental concept of ontology and concluding at the domain ontology as means of consolidating domain knowledge. The subsequent section is about ontology engineering. It discusses existing methodologies and their limitations. The next section discusses possible solution to overcome these limitations by proposing Grounded Ontology (GO) methodology and describing its main features. The paper concludes with limitations of the proposed solution and future research directions.

1. Lexicon, Conceptualization and Mutual Understanding

While trying to create a better understanding through common lexicon it is noteworthy that conceptualization is very important. “A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualization, explicitly or implicitly” (Thomas R. Gruber, 1995). However, it is imperative to remember that “even if two systems [including ontologies and frameworks] adopt the same vocabulary, there is no guarantee that they can agree on a certain information unless they commit to the same conceptualization” (N. Guarino, 1998).

It needs to be noted that mutual understanding, sometimes referred to as common understanding, does not equate to ‘same’ way of thinking, or agreeing to the other’s viewpoint (Nicola Guarino, 2012). It relates to ‘knowing’ the others’ points of views and their understanding of the domain. Once the interacting entities know others’ understanding, it is easy to find areas of agreement as well as disagreement as shown in

120
Figure 1, where overlapping circles represent the known information about the domain of interest. The overlapped area represent the area of agreement of understanding while the non-overlapping areas represent the areas of disagreement. Thus, it would lead to a better understanding among the interacting people.

The above analysis leads to the conclusion that domain ontology may provide common understanding and language along with consolidation of the domain knowledge as well.

![Diagram showing areas of agreement, disagreement, and common understanding.](image)

**Figure 1: Areas of agreement, disagreement and common understanding**

2. **Domain Ontology**

**Ontology**

Ontology is a “specific artifact expressing the intended meaning of a vocabulary in terms of primitive categories and relations describing the nature and structure of a domain of discourse” Guarino(2012). Infomation scientists use “ontology” to express a shared taxonomy of entities that has been reduced to its simplest and most significant form possible without the loss of generality (Smith, 2003). “An ontology is in this context a dictionary of terms formulated in a canonical syntax and with commonly accepted definitions designed to yield a lexical or taxonomical framework for knowledge-representation which can be shared by different information systems communities” (Smith, 2003).

From the above statements it can be concluded that ontology is a conceptual system of the domain of interest representing entities and their relationships in the universe of discourse.

Researchers from different domains have their own peculiar concepts and terms they use for information representation (Smith, 2003). This leads to exclusiveness and inconsistency when they try combining their efforts. Ontology was introduced as means of resolving such terminological and conceptual incompatibilities (Smith, 2003). Excluding philosophical aspects, ontologies were initially developed to assist knowledge sharing and reuse by Artificial Intelligence community (Fensel, 2001). One of the major challenges that ontology addresses is achieving interoperability between multiple representations of reality (Hepp, 2007). Despite the fact that there is a difference on what exactly ontology is, especially at the intersector of computer science and information systems research (Hepp,
2007), the basic use of ontologies is to enhance “communications between either humans or computers” (Jasper & Uschold, 1999).

**Purpose of Ontology**

The primary purpose of an ontology is a mutual understanding of each other and improved communication among people (Nicola Guarino, 2002; Jasper & Uschold, 1999; Sowa, 2013). The focus of ontologies is on the content, i.e. on the meaning being conveyed by the entities as well as on the structure of the domain they represent (Fensel, 2001; Nicola Guarino, 2002). “The content [that ontologies represent] must be studied, understood, [and] analyzed”, however, it must be remembered that understanding of content is not contingent upon its representation (Nicola Guarino, 2012).

On the other hand for the purpose of human communications an unambiguous but informal specification of ontology would suffice, rather preferred (Jasper & Uschold, 1999; Uschold, 1998). Domain elements (entities and their relationships) specified by well thought out vocabulary with carefully chosen terminology and human readable documentation (or synonym set), can perform better by increasing the user involvement, as this participation does not require knowledge of formal logic (Hepp, 2007).

Information systems perspective of ontologies is focused on meaning and understanding conceptual elements and their relationships. In this context “a collection of named conceptual entities with a natural language definition would count as an ontology” (Hepp, 2007).

Based on the above discussion it can be argued that ontologies are fundamentally for sharing understanding among humans. Formal logic may not be the best way of representation if the purpose is only human-human interaction. Use of informal but unambiguous specification through various other means can achieve better results.

**Use of Ontology**

Ontology has been used for many purposes. Researchers in computer science and information systems have found that ontology is very useful in capturing commonly agreed (Chandrasekaran et al., 1999) relevant information (N. Guarino, 1995). Segregated into domain knowledge and separate from operational knowledge (Noy & McGuinness, 2001) it is available for sharing and reuse (T. R. Gruber, 1993).

Having discussed ontology, let us look at domain ontology.

**What is a Domain Ontology?**

Domain ontology is a type of ontology that has been identified as one of the solutions for an effective and efficient consolidation of domain knowledge and for creating better understanding about it (Čosić et al., 2011; Harter & Moon, 2011, pp. 132–133).

**Domain Ontology – An Efficient Means to Consolidate Domain Knowledge**

Domain ontologies define particular concepts and relationships that form the essential structure of a domain for a specific universe of discourse (Roussey, 2005). This basic structure “describe[s] the concepts in their domain, the relationships between those concepts, and the instances or individuals that are the actual things that populate that structure” (“Lightweight, Domain Ontologies Development Methodology,” 2010). Based on Corcho, Fernández-López, and Gómez-Pérez (2003) it has been stated that domain
ontology provides an accurate picture of the language as well as the entities and their relationships in a particular domain, for the users that work in that domain (“Lightweight, Domain Ontologies Development Methodology” 2010).

From the literature presented it can be ascertained that domain ontologies can provide means for effectively consolidating knowledge. Ontology for a dynamic domain, like information security, needs the ability to remain current for it to be practically usable over an extended period of time. This has not been addressed in those ontologies. Thus, a framework is required to keep current an ontology for such a domain. Further, as stated by Smith and Ceusters (2010), “the most effective way to ensure mutual consistency of ontologies over time and to ensure that ontologies are maintained in such a way as to keep pace with advances in empirical research is to view ontologies as representations of the reality that is described by science”. Assuming that published research represents “reality described by science”, we can base our ontology on concepts and relations extracted directly from published research papers. This can lend an inherent capability of perpetual evolution of such an ontology.

Moreover, these ontologies are highly dependent on experts who select entities required to describe a domain and establish relationships between them. Hence, it is highly desirable to modify current methodologies to make the resultant ontology more closely linked to and firmly grounded in published literature. It will also help in continued evolution of ontology as discussed in the previous paragraph.

3. Ontology Engineering – Existing methodologies and their limitations

At present there are no commonly agreed methods and guidelines for ontology development, which is a problem (A. Gómez-Pérez & Benjamins, 1999). Furthermore, subject to the size of resultant ontology, the development process can be very permissive in the actual implementation of methods and guidelines (A. Gómez-Pérez & Benjamins, 1999).

3.1. Ontology Engineering

Consolidating ontology engineering methods, Casellas (2011) has stated that ontology development could be classified as top-down, bottom-up, and middle-out approach based on where the process begins. It could also be organized on the level of automation: manual, semi-automatic, and fully-automatic. There could be other ways of classification as well. Casellas(2011) further states that generally top-down approach is done manually and bottom-up is automatic, at least initially. He goes on to mention that middle-out approach is typically semi-automatic and is concerned with finding the most important concept, and then completing the hierarchy by specialization and generalization. Choosing a particular methodology is an important decision since among others, one of the ways to characterize an ontology is the methodology used to develop it (Casellas, 2011).

3.2. Selected Current Ontology Engineering Methodologies

Some of the current ontology engineering methods are discussed below. According to Noy & McGuinness (2001), ontology in technological (non-philosophical) sense is derived primarily (or initially) from structured and unstructured text sources. They go on to state that this is predominantly done by employing text mining techniques. Expert opinions are used to define classes and sub-classes, and their properties along with restrictions for a particular domain (Noy & McGuinness, 2001).
Lenat & Guha (1989) gave a multistep process for developing its Cyc ontology and knowledgebase comprising of: (1) Manual extraction of knowledge, (2) Computer aided extraction of knowledge, and (3) Computer managed extraction of knowledge, based on knowledge already extracted in previous steps. Thus, initiating the ontology manually, then augmenting and evolving it automatically. Visser provided a four step methodology, CommonKADS for Legal Knowledge-base Systems (LKBS), that may be used for ontology development, as well (Pepijn R. S. Visser, Kralingen, & Bench-Capon, 1997; P R S Visser, 1998). It includes analysis, conceptual modeling, formal modeling and implementation. CommonKADS has since then become a complete Knowledge Engineering (KE) methodology (Casellas, 2011).

Corcho, Fernández-López, Gómez-Pérez, & López (2005) has described a semi-automatic methodology consisting of specification, conceptualization, formalization, implementation, and maintenance, for the development of legal ontology. It is based on Methontology that was proposed by Gómez-Pérez and Fernández-López, in their various works (Fernández-López & Gómez-Pérez, 2002; Fernández-López, Gómez-Pérez, & Juristo, 1997; Asunción Gómez-Pérez, Fernández-López, & Corcho, 2007) as cited by Casellas(2011). Jarrar and Meersman(2009) proposed DOGMA approach that has three stages; preparatory, domain conceptualization, and application specification.

Milton (2007) came up with ‘47-step guide to knowledge acquisition’ that has many similarities with Common KADS but its ‘generality’ puts it apart from others according to Milton. It is primarily based on manual effort requiring expert input in the form of interviews from initial modeling all the way to final validation. Suárez-Figueroa et al. (2007) created network of ontologies (NeOn); a methodology for developing ontology networks that is a ‘collection of ontologies related together via a variety of different relationships’ (Haase et al., 2006).

3.3. Limitations of Current Ontology Engineering Methodologies

Current ontology engineering methodologies have certain limitations. From common characteristics of the methods described in previous section it may be concluded that almost all the existing ontology engineering efforts are geared towards semantic interoperability of systems. Moreover, meaningful category/concepts are generated by developer/expert based on their personal understanding of the domain. This introduces bias in the ontology.

Another way of deriving entities is using statistical and syntactical techniques coupled with Artificial Intelligence. This requires human expert to filter out the meaningful and relevant entities. As there is no fully-automatic methodology for ontology development that can yield a valid ontology, manual processes have to be used. It increases the development duration. Mostly a semi-automatic/manual methodology is used to incorporate expert opinion, at least to validate the concepts and their relationships, for example SIMOnt by Abulaish et al.(2011). Moreover, the evolution of ontology has not been a major focus.

It may be concluded that the ontologies developed by these methods pertain to a particular point in time, apply in a certain context and are limited to a specific group of people. Sooner or later they are either outdated or require considerable effort to keep them current.

Following characteristics of existing ontology engineering methodologies can be considered as their limitations:
1. Focused primarily on systems interoperability and computer-computer interaction. Essentially required for semantic interoperability of systems, but not aimed at either human-human interaction or common understanding.

2. Reflect ontology engineers’/experts’ personal understanding of the domain.

3. Require human interventions to make the resultant ontology meaningful and useful.


4. Possible Choice of Overcoming These Limitations

Limitations in the existing ontology engineering methodologies have been described above. A possible approach that can help overcome these limitations lies in text coding which is discussed in this section.

4.1. Text Coding

According to Strauss and Corbin (1998), textual data can be coded and analyzed to find concrete description of abstract categories. Among other sources, historical data is used to establish relationships between categories and their descriptions. This technique is based on 1967 work of Glaser and Strauss (1967). It is a “discovery methodology that allows the researcher to develop a theoretical account of the general features of a topic while simultaneously grounding the account in empirical observations or data” (Martin & Turner, 1986). Constant comparison is an important rigorous “tool” for scrutiny of the codes and gathering of analytical insights (Urquhart, Lehmann, & Myers, 2010). It is about discovering concepts, categories and relationships among them (Bryant & Charmaz, 2007). This methodology has clearly defined data analysis procedure, which results in elaborate and novel findings that are substantiated by data (Orlikowski, 1993). Thus, one of the outputs is a list of emergent concepts, categories and sub-categories, and their properties derived directly from the text.

Two important characteristics of this coding methodology as given by Urquhart et al. (2010) are:

1. Joint data collection and constant comparison for analysis and conceptualization.
2. Theoretical sampling to collect all kinds of “slices of data” based on already established categories, concepts and constructs.

The possibility of blending text coding and ontology engineering was initially suggested by Star (1998). It was used by Kuziemsky et al. (2007) to provide richness to “domain relevant model”.

Therefore, based on the above discussion and analysis it seems that ontology development can follow text coding approach. Using research papers from top peer-reviewed journals as corpus for text coding can help in making it more acceptable as well as taking care of ontology evolution. Not only can this approach help in reducing ontology engineer’s bias but it can also help in consolidating domain knowledge. Based on this we have designed an ontology development methodology which is called Grounded Ontology (GO).
5. Proposed GO Methodology

In essence it is proposed that GO be a multi-stage multi-step knowledge summarization and representation process to organize and exhibit knowledge in a simple and concise manner through discovery involving codifying existing knowledge thereby cleanly conceptualizing the emergent core concepts and relationships among them, and building an ontology such that it becomes easy to review the existing knowledge and come to a common understanding.

![Figure 2: Use of Grounded Theory Method and ontology engineering for creating an emergent ontology.](image)

**Figure 3: Stages of ontology development and enhancement**

It is proposed to have four stages shown in Figure 3. Stage 1 is coding of the text in the corpus. Stage 2 is giving a structure to the categories and relationships emergent from the codes and creating seed ontology. Stage 3 is finding other categories and relationship and incorporating them in the seed ontology to form a saturated ontology. Stage 4 is the ongoing enhancement to the saturated ontology. It is done by adding more data (research papers in this case) to the corpus and processing the additional data through Stage 1 coding and merging the additional categories and their relationships to form an enhanced version of the ontology. This stage 4 can be run as and when more data becomes available.

Text coding is a time intensive work that puts high demand on ontologist. Therefore, to make the ontology development manageable and the resultant ontology useful, it is proposed that GO methodology rely on coding most significant portions of the text. This is done to generate the seed ontology through **in-vivo** coding technique. Subsequently, this seed ontology is enhanced to make core ontology through **selective** coding of the relatively less significant sections of the text. Identification of segments in the text with most significant original contributions is not an easy task in unstructured data. Therefore, to use relatively structured text it is suggested that the corpus be composed of published research papers of reputable journals. Research papers have very well defined standard structure i.e. sections...
containing particular type of specific information. The general structure of a research paper has following sections: abstract, introduction, literature review, methodology, results, discussions, limitations and conclusions. Here, the original contribution of the paper is primarily mentioned in the abstract in a concise manner. Other significant sections include conclusions, discussions, and results. Therefore, for seed ontology the abstracts are coded using in-vivo technique. Conclusions are coded using selective coding technique. Discussions and results may also be coded subsequently through selective coding technique if deemed necessary.

5.1. Comparison of GO with other methodologies using GTM

This approach is different from both Kuziemsky et al.’s (2007) and Urban’s (2009) that have used GTM, as it is a multi-step multi-stage methodology. There is a difference in application of GT method as well, as given in Table 1.

Table 1. Comparison of GT methodology applied by Kuziemsky et al. [15] and GO

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Coding Technique</td>
<td>Open, Axial and Selective</td>
<td>In-vivo and Selective</td>
</tr>
<tr>
<td>2</td>
<td>Purpose</td>
<td>Better understanding of domain</td>
<td>Presenting state-of-the-art in domain.</td>
</tr>
<tr>
<td>3</td>
<td>Information Sources</td>
<td>Practice experience of health care professional, patients’ health management charts, and research literature</td>
<td>Research papers from journals</td>
</tr>
</tbody>
</table>

Kuziemsky et al. (2007) have used open, axial and selective coding techniques. They used the grounded theory methodology as a means of analyzing relevant information sources for better understanding of the domain of interest. Similarly Urban (2009) has suggested the use of blended approach for understanding the unstructured information. GO is different. It uses in-vivo and selective coding techniques aiming to present stare-of-the-art in domain of interest. Selective coding is a second cycle text coding technique to ‘compare, reorganize and “focus” codes into categories, prioritize them ... and synthesize them to formulate a central or core category that becomes the foundation’ (Saldana, 2009, pp. 51–51).

5.2. GO Approach to Overcome Limitations of Existing Ontology Engineering Methodologies

The proposed GO approach is designed to address four limitations of current approaches mentioned in Section 4.3.

Overcoming the Limitation of Computer-Computer Interaction

As opposed to computer-computer interaction, GO approach is aimed at developing and representing ontology not only for better understanding but also for communicating that understanding among humans. Considering that domain experts are not necessarily also experts in philosophical and mathematical logic, simple notation with natural language
expressions is used in this ontology. The purpose is to convey the intended meaning while balancing precision with ease of understanding.

**Overcoming the Limitation of Personal Understanding**

The use of *in-vivo* text-coding technique would ensure that entities are taken from the text and capable of being traced back to the original text. This would help to convey the intended meaning of the author of the research paper while reducing personal opinion of the ontologist.

**Overcoming the Limitation of Human Intervention**

Human intervention is required for any ontology to be meaningful and this requirement cannot be eliminated. However to reduce the human effort in development of ontology using GO method, *in-vivo* coding is restricted to most significant section of a research paper that specifically describes the contribution of that particular research. For the other sections selective coding with constant comparison is employed. The background to research, presented as literature review, is not to be coded.

**Overcoming the Limitation of Evolution of Ontology**

As the ontology developed using GO method is based on published research therefore, it can be taken as reality presented by science. This in itself would effectively ensure continued evolution of the ontology. Further, for a constant evolution and maintenance of ontology in dynamic domains, the FocalPoint framework has been proposed by Nabi et al (Nabi, Asif, Iradat, Arain, & Ghani, 2013) can be implemented in future.

The GO methodology has following advantages:

1. State-of-the-art of the domain will be readily known as it has the advantage of using published research as the basis of ontology development. Also a mechanism of continual evolution (FocalPoint) shall account for the dynamism of the domain(Nabi et al., 2013).
2. Non-replication of research would help reduce the chances of re-inventing of wheel. The efforts thus saved can be directed to extending the frontiers of research.
3. It would ensure resolution of any confusion that might exist within research community as it would provide not only common understanding but also common lexicon for better understanding.

**6. Limitations and Future Research**

One of the limitations of this methodology is the possibility of development of various codes leading to different ontologies. To overcome this it is recommended that the principle of mutual understanding be enforced i.e. different understanding of the same text can exist and the author of the research paper may be consulted to find the intended meaning. Also, a group of leading professionals of the domain can debate and decide upon any category or a relationship in the ontology. This would also cater for the legitimacy and evolution of the ontology as well.

Another limitation of this methodology is the use of structured text in corpus. This is an inherent limitation of GO methodology. Perhaps in future this limitation can be relaxed by applying Artificial Intelligence text classification algorithms that use naïve Bayes classifier.
As is the case with any new methodology, the impact of GO can only be ascertained if the methodology is made widely available to researchers and practitioners for use. The use and acceptability of the resultant ontologies can then form the basis to assess the efficacy of this proposed methodology. As of now it presents a potentially valuable addition to the many other available ontology engineering methodologies.

For future research, this methodology may be applied to generate an ontology for a specific domain.

References:


“...The question now is how truly individual – as in bold, original, unique - you can be if you never step back from the crowd. When we think and write from within our busyness, surrounded by countless other voices, too often the result is reactive, derivative, short-shelf-life stuff. The greatest gifts one can give to the outward world lie within. To reach them, you have to go there.”