

# Toward an Open-Source Foundation Ontology Representing the Longman's Defining Vocabulary: The COSMO Ontology OWL Version

Patrick Cassidy

MICRA, Inc., Plainfield, NJ  
cassidy@micra.com

**Abstract** - The COSMO foundation ontology is being developed to test the hypothesis that there are a relatively small number (under 10,000) of *primitive* ontology elements that are sufficient to serve as the building blocks for any number of more specialized ontology elements representing concepts and terms used in any computer application. Finding evidence for this hypothesis would suggest that a promising tactic to achieve *Semantic Interoperability* among computer applications is to focus effort on the common foundation ontology to that ontology that contains those primitive elements. This will constrain the size of the ontology on which agreement is required, to the minimum that will support accurately relating domain and application ontologies to each other. The rationale, methodology and current status of this project is reported here.

**Index Terms** – Foundation ontology, conceptual primitives, COSMO, semantic interoperability, common ontology, ontology mapping, Longman, defining vocabulary.

## I. INTRODUCTION

Information communicated and analyzed by the intelligence community is highly diverse, including technical, social and psychological concepts. The challenge of using automatic techniques for integrating such information will require adoption of an ontology that is capable of unambiguously representing the full range of knowledge that people communicate. There is as yet no consensus on how to structure that ontology. This paper describes one approach to overcome the lack of agreement caused by multiple fundamentally different approaches to foundation ontology development. The proposed approach depends on three factors: (1) to develop a foundation ontology that is effective as a standard of meaning for communication among many applications, it is not necessary to achieve universal agreement among ontology developers about the structure of the foundation ontology; it is only necessary to build a sufficiently large user group that third-party vendors will have incentive to develop utilities making the ontology easier to use, and applications that demonstrate the usefulness of the ontology for practical purposes. (2) by allowing multiple logically compatible views for representing the same entities, and providing translation utilities between them, many of the differing preferences for representing entities can be accommodated in the same ontology. (3) the number of

different ontology groups that will accept the ontology can be maximized by keeping the foundation ontology as small as possible without compromising its ability to support logical representation of terms and concepts in any application domain. In the COSMO approach, that could be achieved by discovering the smallest inventory of fundamental ontology elements, representing the minimal essential primitive concepts that are needed to build representations of any more complex concept.

## II. BACKGROUND TO THE COSMO APPROACH

### A. The Notion of Conceptual Primitives

The approach proposed here relies on the observation that communication among agents (human or automated) depends on the agents sharing some common set of internally understood concepts, labeled by an agreed set of symbols such as words in human languages, or element names in databases. Wherever a particular community uses concepts not already among the known concepts of other communities, information sharing requires the first community to use a common set of *defining concepts* to construct definitions of the unknown concepts understandable to the other communities. In this manner communicating agents can accurately transfer information on topics familiar or initially unfamiliar to other agents. Information transfer using human languages is facilitated by the existence of a relatively small vocabulary of basic words, representing those commonly understood concepts, that can be used to create linguistic definitions of any specialized concept. Research in Linguistics has explored by experimental techniques the number and identity of the common primitive concepts that are used in linguistic communication among people speaking different languages. Some of that work, summarized by Goddard[1], has suggested that as few as 60 semantic primitives are adequate to construct definitions of a very large number of concepts. A less systematic but more comprehensive demonstration of the power of primitive concepts to suffice for construction of definitions of many words is found in some English-language dictionaries such as

the Longman [2] that use a Defining Vocabulary of basic words with which to define all of the entries in the dictionary. The Longman Defining Vocabulary (hereafter LDV) contains 2148 words, but an investigation [3], [4], [5] has shown that even fewer words are needed to define (recursively) all of the Longman entries. For cases where a proposed definition of a new word uses words not already in the defining vocabulary, the Defining Vocabulary tactic requires that the unrecognized word itself be defined by use of the basic Defining Vocabulary. The answer appears to be that, for the Longman, words recursively defined in such a manner “ground out” using a basic vocabulary of 1433 words representing 3200 word senses.

The success of the linguistic defining vocabulary for dictionaries suggests that a similar tactic could be effective for automated information transfer among computer systems. For automated systems, the “Defining Vocabulary” would take the form of a *foundation ontology* having an inventory of basic concept representations that is sufficient to create representations of any new concept, by combinations of the basic elements. Communities using such a “*Conceptual Defining Vocabulary (CDV)*” (i.e. a common foundation ontology) would be able to pursue their own interests using any local terminology or ontology that suits their purposes, and still communicate their information accurately in a form suitable for automated inferencing, by translating the local information into the terminology of the *common foundation ontology*. Limiting the core foundation ontology to the elements needed for a CDV will minimize the effort required to perform the translations, while ensuring that accurate translations are possible. The question remains whether the linguistic Defining Vocabulary examples can be adapted to the more precise requirements of representing terms and concepts in a logical format, suitable for automated reasoning.

The essential principle of such a tactic for Semantic Interoperability is that, when the separately developed ontologies of two different systems both use the same CDV to specify the structures of their ontology elements, then accurate information sharing can be achieved, even if the two systems each have some separately-defined ontology elements not in the other, by *sharing* the specifications of the ontology elements of each that are not in the other. Since the ontology elements of each system are built from the same primitive elements of the CDV, they will be properly

and accurately interpretable in both systems. The combination of the ontologies of the two systems in effect creates a single merged ontology common to both systems. In that situation, the same input data in both systems will produce the same inferences. Different data in the two systems will create some different inferences, but those will not be logically inconsistent if the data is not inconsistent. For a proper automated merger of the two ontologies, it will be necessary to have utilities that can automatically recognize identical elements created in the two separate local ontologies, and to detect inconsistencies if they exist. But this tactic for interoperability avoids the impossible task of automatically interpreting information in an external ontology that is based on fundamentally different (usually undocumented) assumptions about how to represent the same intended meanings of terms and concepts.

### *B. The Current Absence of a Conceptual Standard*

To function as a conceptual standard that will enable **semantic interoperability**, i.e. permit computers to reason accurately and automatically with transferred information, the syntactic format for a common standard must have at least the expressivity of First-Order Logic (FOL), so as to permit logical inference using rules expressing domain knowledge. Several foundation ontologies, such as OpenCyc[6], SUMO[7], DOLCE[8], and BFO[9], have been developed that have this technical capability. Other knowledge classifications such as NIEM[10] and the DoD Core Taxonomy[11] have less expressiveness. None of these projects has adopted the tactic of creating a CDV, and none has been recognized as a default standard for application builders concerned with specific topics and indifferent to the nuances of representation at the abstract levels. The reasons for lack of wide adoption vary. The complexity of each of the existing foundation ontologies presents a steep learning curve which requires a strong motivation to impel potential users to spend the required time. In the case of Cyc, much of the content (such as the over 1000 specialized reasoning modules) is still proprietary and cannot be part of an open-source project that could include desired components from many non-Cycorp sources. Development of an effective open-source natural-language interface to the ontology is also desirable, to make learning and use convenient. None of the existing foundation ontologies has such an interface. Without publicly available examples showing the benefits of using a complex ontology, a specialized application developer without a need to interoperate

outside the local community is strongly tempted to develop a specialized ontology that is not linked to a foundation ontology. As a result, specialized ontologies with no linkages to any of the major foundation ontologies have proliferated.

The above considerations suggest the following desiderata for a foundation ontology that can be adopted and used by a large enough community to serve as a *de facto* standard of meaning:

- the core set of concept representations required to use the ontology effectively should be as small as possible, but sufficient to support specification of any specialized concept meaning
- the ontology should be fully public and developed by an open procedure, so as to permit alternative logically compatible views of entities; it should be maintained by an open process and allow additions as needed to represent new topics;
- there should be a powerful intuitive natural language interface, capable of determining whether (1) representations of specific concepts are already present in the core foundation ontology or in some public extension, or (2) if not, to list the elements in the ontology closest in meaning
- the ontology format should have the expressiveness of at least FOL
- there should be several open-source substantive applications demonstrating the usefulness of the ontology
- extensions to the core, with logical specifications of concepts based on combinations of the core concept representations, should be maintained and freely available, in the manner of Java library packages, to minimize the need for creating new definitions.

In order to have a *de facto* standard of meaning, it is not necessary to have universal agreement to use only *one* foundation ontology; it is only necessary that *some* foundation ontology have a user community large enough for third-party vendors to have incentive to develop utilities that make the standard easier to use, and to develop applications that demonstrate its utility. It should also have a sufficiently wide community of users that research groups will have an incentive to use it as the standard of meaning through which they can transfer information from diverse separate applications, each using different forms of intelligent information processing.

### III. THE COSMO PROJECT

#### A. Origin

The COSMO ontology [12] is currently being developed to serve as a fully public foundation ontology that contains representations of all of the 2100 words in the LDV, with the intention of serving as a broadly acceptable CDV. COSMO (Common Semantic MOdel) was initiated in 2005 [13] as a project of the Ontology and Taxonomy Coordinating Working Group [14], a working group of the Federal Semantic Interoperability Community of Practice. The origin of COSMO is discussed in more detail in [15]. In early 2008 the project adopted the current goal of representing the LDV. Developing the ontology as a CDV promises to furnish a foundation ontology that has all of the elements (types, relations) needed to build representations of any concept of interest in any application, yet be small enough to be usable without an extended learning period. The goal in effect is to identify the *smallest* foundation ontology that is sufficient to serve as the basis for broad semantic interoperability. Such a foundation ontology will contain representations of the essential units of meaning that can be combined to represent any specialized term or concept of interest in applications.

#### B. Project phasing

COSMO is proceeding in several phases. The first phase, expected to be complete within 3 months, is to create a representation of all of the words in the LDV, in an OWL format [16]. The expressiveness of at least pseudo-second-order logic (a FOL in which variables can represent relations or assertions) is required for some applications such as Natural Language understanding. The plan is therefore to maintain an OWL version, but convert it automatically to a Common-Logic (CL) compliant language such as KIF or IKL. This will require representing rules, functions, and higher-arity relations in the OWL format.

When the COSMO ontology has the full set of LDV words represented, it will be tested for its ability to serve as a CDV, by creating representations of several sets of specialized concepts and discovering how many new fundamental concept representations need to be added to the foundation ontology. It is estimated that this first version will contain over 7500 types (OWL classes), over 700 relations, and over 1000 restrictions that constrain the meanings of the elements.

The COSMO itself is not expected to be adopted without change as a common foundation ontology. The main purpose of this project is to demonstrate the feasibility a Conceptual Defining Vocabulary as an effective basis for semantic interoperability. A CDV that is widely accepted is likely to arise only from a collaborative effort by a broad consortium of ontology builders and users, as well as developers of other knowledge representation constructs such as the NIEM. More than one CDV may eventually find wide use, but the number of such ontologies is likely to be smaller than the number of operating systems, because the greater number and complexity of primitive data structures required for a CDV is larger than those manipulated by operating systems.

### C. Criterion for Success

The criterion for determining whether the COSMO can serve as a starting CDV will be based on the number of new primitive ontology elements that must be added to the COSMO in order to represent groups of new terms or concepts from additional specialized topics. It is expected that *some* additional primitive elements (types, relations) will be need to be added to the COSMO as knowledge in diverse fields is represented. To function as an effective CDV, what is required is that the number of such new primitives added to the ontology will decrease asymptotically as each successive block (e.g. of 500) of new terms is represented using the foundation ontology. Such statistical evidence that there is *some* limit to the number of new terms that must be added will help answer the two questions, of whether there is *any* limit to the number of basic elements required for the CDV, and if so, approximately what is that number.

### D. Allowance for Multiple Viewpoints

Essential to its role in enabling semantic interoperability is that COSMO must be inclusive of all logically compatible views, so as to permit translations among all of the representations used in applications. This means that wherever different ontologists prefer different means of representing a concept, both alternatives are included, with a translation rule (e.g. “bridging axioms”) that automatically converts from one view to the other. An example would be the

concept of “mother” which is represented in some ontologies only as a relation (“isTheMotherOf”), and in others as the type (class) ‘Mother’. The COSMO OWL version can include both representations, but the automatic conversion of such alternative views will often require that rules be used, and will be possible only in the more expressive common-logic format.

Using an ontology representing multiple views could lead to inference that is less efficient than with a more restrictive representation. However, it is expected that multiple alternative representations will be needed only for interoperability among applications, and individual local applications will not use the full ontology, but will select out only those elements required for the local application. In this way, full semantic interoperability can be achieved among applications, without sacrifice of efficiency.

## REFERENCES

- [1] Cliff Goddard, Bad Arguments Against Semantic Primitives, *Theoretical Linguistics*, Vol. 24 (1998), No. 2-3: 129-156. (Available online at: <http://www.une.edu.au/bcss/linguistics/nsm/pdfs/bad-arguments5.pdf>)
- [2] Longman Dictionary of Contemporary English, Longman Group, Essex, England (New Edition, 1987)
- [3] Guo, Cheng-ming (1989) *Constructing a machine-tractable dictionary from "Longman Dictionary of Contemporary English"* (Ph. D. Thesis), New Mexico State University.
- [4] Guo, Cheng-ming (editor) *Machine Tractable Dictionaries: Design and Construction*, Ablex Publishing Co., Norwood NJ (1995).
- [5] Yorick Wilks, Brian Slator, and Louise Guthrie, *Electric Words: Dictionaries, Computers, and Meanings*, MIT Press, Cambridge Mass (1996).
- [6] OpenCyc: <http://opencyc.org/>
- [7] <http://www.ontologyportal.org/>
- [8] See: <http://www.loa-cnr.it/DOLCE.html>
- [9] Pierre Grenon, *BFO in a Nutshell: A Bi-categorical Axiomatization of BFO and Comparison with DOLCE*, IFOMIS report 06/2003 (2003). Available at: [http://www.ifomis.uni-saarland.de/Research/IFOMISReports/IFOMIS%20Report%2006\\_2003.pdf](http://www.ifomis.uni-saarland.de/Research/IFOMISReports/IFOMIS%20Report%2006_2003.pdf). See also : <http://www.ifomis.uni-saarland.de/bfo/>
- [10] See: <http://www.niem.gov/>
- [11] DoD Core Taxonomy: <http://www.dtic.mil/dtic/annualconf/conf05-Dickert.ppt>
- [12] <http://micra.com/COSMO/COSMO.owl>
- [13] [http://semanticcommunity.wik.is/Federal\\_Semantic\\_Interoperability\\_Community\\_of\\_Practice/Work\\_Group\\_Status/Ontology\\_and\\_Taxonomy\\_Coordination/COSMO\\_Common\\_Semantic\\_Model](http://semanticcommunity.wik.is/Federal_Semantic_Interoperability_Community_of_Practice/Work_Group_Status/Ontology_and_Taxonomy_Coordination/COSMO_Common_Semantic_Model)
- [14] [http://semanticcommunity.wik.is/Federal\\_Semantic\\_Interoperability\\_Community\\_of\\_Practice/Work\\_Group\\_Status/Ontology\\_and\\_Taxonomy\\_Coordination](http://semanticcommunity.wik.is/Federal_Semantic_Interoperability_Community_of_Practice/Work_Group_Status/Ontology_and_Taxonomy_Coordination)
- [15] <http://micra.com/COSMO/COSMOoverview.doc>
- [16] The OWL Web Ontology Language Reference: <http://www.w3.org/TR/owl-ref/>