Applied Ontology 3 (2008) 191–200
DOI 10.3233/AO-2008-0052
IOS Press

Communiqué

Ontology Summit 2007 – Ontology, taxonomy, folksonomy: Understanding the distinctions

Michael Gruninger, Olivier Bodenreider, Frank Olken, Leo Obrst and Peter Yim

1. Introduction

Under the appellation of "ontology" are found many different types of artifacts created and used in different communities to represent entities and their relationships for purposes including annotating datasets, supporting natural language understanding, integrating information sources, semantic interoperability and to serve as a background knowledge in various applications.

The Ontology Summit 2007 "Ontology, taxonomy, folksonomy: Understanding the distinctions",[1] was an attempt to bring together various communities (computer scientists, information scientists, philosophers, domain experts) having a different understanding of what is an ontology, and to foster dialog and cooperation among these communities.

In practice, ontologies cover a spectrum of useful artifacts, from formal upper-level ontologies expressed in first order logic, such as Basic Formal Ontology (BFO), Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE), Suggested Upper Merged Ontology (SUMO), and Process Specification Language (PSL), to folksonomies (the simple lists of user-defined keywords to annotate resources on the Web). In between these two extremities of the ontology spectrum are taxonomies, conceptual models and controlled vocabularies such as Medical Subject Headings (MeSH), often used for information indexing and retrieval, and whose organization is mostly hierarchical. Finally, there are ontologies which represent not only subsumption, but also other kinds of relationships among entities (e.g., functional, physical), often based on formalisms such as frames or description logics. Examples of such ontologies in the biomedical domain include the Foundational Model of Anatomy, SNOMED CT and the NCI Thesaurus.

The goal of the Ontology Summit was not to establish a definitive definition of the word "ontology", which has proven to be extremely challenging due to the diversity of artifacts it can refer to. Rather, the results of the Summit identified a limited number of key dimensions along which ontologies can

[1]The Ontology Summit took place April 22–23, 2007 in Gaithersburg, Maryland, at the National Institute of Standards and Technology (NIST). It was co-organized by NIST and the Ontolog Forum and co-sponsored by some 50 institutions.

be characterized and to provide operational definitions for these dimensions. The relative position of ontologies in the space defined by these dimensions, the "Framework", is indicative of the similarities and differences between these ontologies. The Framework has been applied to the characterization of a dozen ontologies, whose descriptions were collected through a survey.

2. Ontology framework

Ontologies in computer science were originally proposed to enable sharable and reusable representations of knowledge. Nevertheless, the sheer range of current work in ontologies (including taxonomies, thesauri, topic maps, conceptual models, and formal ontologies specified in various logical languages) raises the possibility of ontologies being developed without a common understanding of their definition, implementation and applications. Our objective is to provide a framework that supports diversity without divergence, ensuring that we can maintain sharability and reusability among the different approaches to ontologies.

One major goal of the Ontology Summit 2007 was to bring together the various communities working on ontology-related activities to encourage cooperative efforts. Toward this end the summit attempted to provide a characterization by constructing a typology of ontologies. The framework of dimensions is comprised of two groups: semantic dimensions and pragmatic dimensions (see Fig. 1). Semantic dimensions include expressiveness, structure and representational granularity. Pragmatic dimensions include intended use, use of automated reasoning, identifying whether the ontology is prescriptive or descriptive, the design methodology, and governance.

An important application of the ontology framework will be to serve as the basis for specifying metadata for different ontologies, which will include the properties and characteristics we use to describe an ontology. This will allow different ontologies to be compared, particularly when they are developed using different approaches. Ultimately, the ontology metadata can be used to characterize the conditions under which ontologies can be shared and reused.

Dimensions typically carry some metric that allows comparisons. Within the Framework, these ideas are generalized. Some dimensions consist of an explicit set of possibilities (such as the *identification* of ontology representation languages), while other dimensions allow a partial ordering (such as the *comparison* of ontology representation languages). It is envisioned that metrics for comparing ontologies along the various dimensions will arise from the population of the Framework with the entire range of ontologies being developed within the community.

2.1. Semantic dimensions

One commonality shared by all approaches is that an ontology includes a vocabulary together with a specification of the intended interpretations (meanings) of the terms in the vocabulary. This specification includes:

- Identification of the fundamental categories in the domain;
- Identification of the ways in which members of the categories are related to each other;
- Constraints on the ways in which the relationships can be used.

Semantic dimensions characterize how a given approach specifies the meanings of terms, which includes the expressiveness of the ontology representation language, structural properties, and the representational granularity of the ontology's specification.

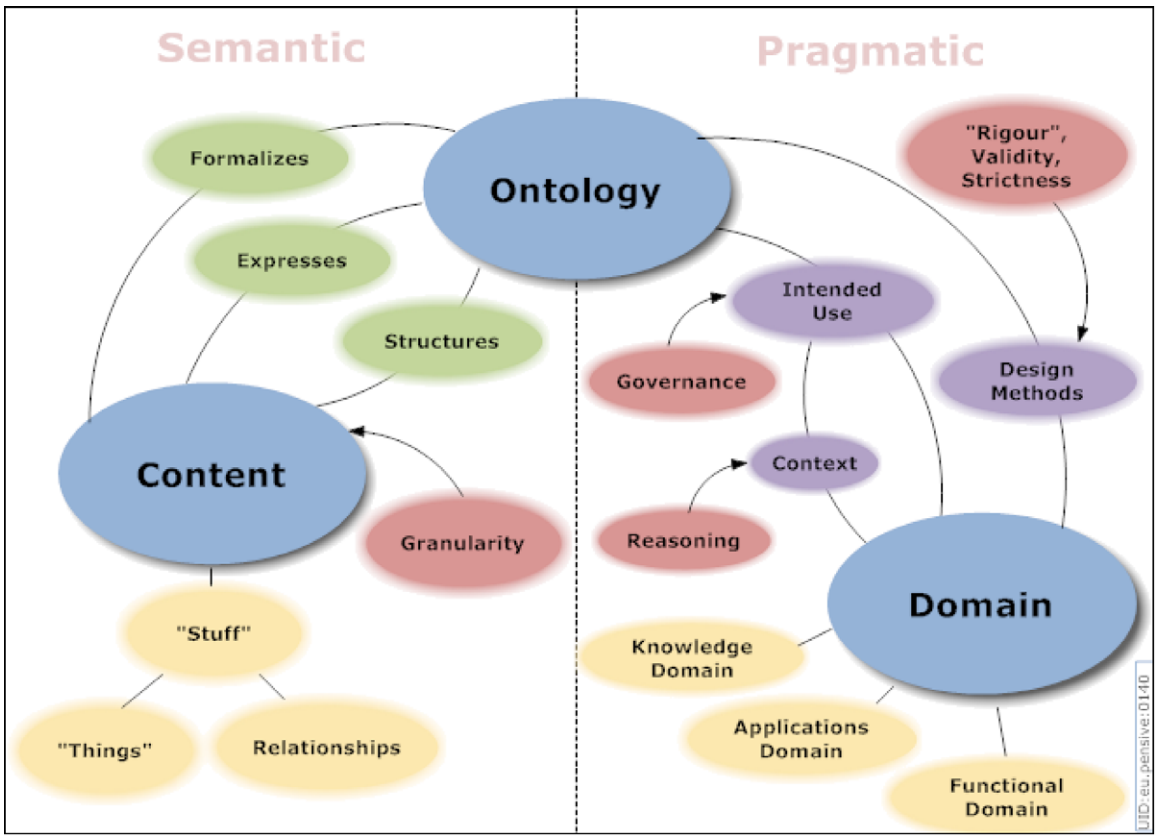

Fig. 1. Ontology framework.

3. Expressiveness of the ontology representation language

Although an ontology is not a definition of form (e.g., the syntax of a language), an ontology defines its vocabulary *in* some representational form. Consequently, ontologies differ in the expressive power of the language used in the specification. Some conceptualizations require a highly expressive language to define the concepts, whereas others can be specified with a less expressive language. An informal ontology may be expressed only as a list of terms and definitions in a natural language such as English. This often leads to some confusion between ontologies and the languages used to represent the ontologies. The oft-cited "semantic spectrum", for example, is a comparison of languages rather than ontologies themselves.

Comparison of ontologies along the language expressiveness leads to two distinctions. The first distinction rests on the specification of the language itself. Logical languages have both a formal syntax and a model-theoretic semantics; examples include RDF, OWL and Common Logic. Semiformal languages, such as XML and EXPRESS, have a formal syntax but lack a model-theoretic semantics. Finally, there are also numerous ontologies whose terms and definitions are specified only in natural language.

The second distinction is the notion of expressiveness, which can be rigorously defined for logical languages. One representation is as *expressive* as another if it can encode all the meanings of the other language. For example, higher order logics are more expressive than first order logic, which is more

expressive than description logic. Expressiveness gives a partial order on languages, because some languages might encode some of the statements of one language but not others.

Ontologies themselves can be compared with respect to the languages with the minimal expressiveness required to define their vocabularies. For example, although a taxonomy can be specified in a highly expressive language such as Common Logic, it only requires a much more restricted language to specify the subclass relationship between classes.

It is often the case that a semiformal language can be mapped into a logical language in such a way that the implicit semantics is captured by the semantics of the logical language. For example, a taxonomy written in XML can be represented in a logical language that includes classes and a subclass of relation.

4. Level of structure

This semantic dimension is akin to the notion of structured and unstructured data in computer science. In ontologies that are specified in a logical language, the level of structure often corresponds to formality of the definitions for the terms in the vocabulary. An ontology that specifies formally defined concepts such as mathematical abstractions have many structural properties, while an ontology that specifies very loosely defined concepts such as document and hyperlink have few structural properties. Many ontologies are semistructured, containing a mix of formal and informal definitions of concepts and relationships. For example, a bibliographic ontology for data about books may contain the concept of date, with formal constraints on the notion of time envisioned (high structure), and the concept of book title, which is only known as a string of text (low structure).

To make these ideas more precise, we can characterize an ontology with respect to the extent to which the intended interpretations of the vocabulary are defined in a logical language. In a *structured* ontology, the intended interpretations for all terms in the vocabulary are defined by sentences in a logical language. In a *semi-structured* ontology, the intended interpretations of some terms in the vocabulary are captured in a semiformal language. Semistructured ontologies require extralogical conditions or special implementations to specify the intended interpretations of some of the terms in their vocabularies. Finally, the intended interpretations of all terms in the vocabulary of an *unstructured* ontology are represented in a semiformal or informal language.

5. Representational granularity

While expressiveness is a characteristic of the language in which an ontology is given, granularity is a property of the content of ontology itself. An ontology with coarse granularity is specified using only very general representational primitives, such as concepts and subsumption in a taxonomy, whereas an ontology with fine granularity specifies much more detail about the properties of concepts and how they can relate to each other. This characterization is independent of the ontology representation language, since even an ontology expressed in a formal language may contains few classes and properties in its vocabulary. On the other hand, an informal ontology may be specified by a very detailed English description of biological classes and discriminating properties that includes many restrictions concerning how terms can relate to each other. Likewise, a very formal ontology expressed in CL, Cyc-L or OWL + SWRL may contain thousands of classes, thousands of properties, thousands of rules, and billions of instances and individuals.

Some quantifiable metrics that may give indications of the representational granularity include the average subclass/subproperty depth, average density of terms within a hierarchy and average number of axioms per term.

5.1. Pragmatic dimensions

Many of the differences among the variety of approaches to ontologies arise from conditions that go beyond the logical properties of the ontologies themselves. The Framework addresses these differences by the identification of pragmatic dimensions that cover the context in which an ontology is designed and used. In many cases, these pragmatic factors influence the semantic properties of the ontology.

6. Intended use

Ontologies are typically designed with respect to some intended application, which include:

- sharing knowledge bases;
- enabling communication among software agents;
- integration of disparate data sets;
- decision support;
- semantic frameworks for enterprise architectures;
- representation of a natural language vocabulary;
- representation of semantics for services and complex software applications;
- helping provide knowledge-enhanced search;
- providing a conceptual framework for indexing content.

The intended use often means that there is some application that is envisioned for which the ontology is being developed. For example, one might want to situate documents within a framing topic taxonomy that roughly characterizes the primary content of the document: this is categorization and helps one semantically loosely organize document collections. Or one might want to use a thesaurus and especially its synonyms and narrower-than terms to enhance a search engine that can employ query term expansion, expanding the user's text search terms to include synonyms or more specific terms and thus increasing the recall (i.e., total set of relevant items) of retrieved documents.

7. Role of automated reasoning

Many applications of ontologies focus on automated reasoning, which imposes strong constraints on both the language and the content in the specification of the ontology. In fact, an alternative way to think of level of structure is the degree to which the ontology can support computation: structured data such as relational database of numbers has strong constraints and supports powerful computation; raw text documents with hyperlinks have few constraints on what can be said or inferred from the data. Also, logical ontology representation languages provide characterizations of notions such as consistency and entailment that provide a means of evaluating reasoning systems. We can therefore define a dimension that compares applications of ontologies according to properties of the kinds of reasoning that are used in conjunction with the ontologies. Simple automated reasoning requires machine semantic interpretability of the content; this can be a theorem prover that implements the deductive rules of a logic or possibly

a special inference engine has been constructed that knows how to interpret the content of the ontology. The former approach has the advantage that the implementation of the reasoner is independent of the content of the ontology, although the disadvantage is that the algorithms implemented by the reasoner may not be efficient. On the other hand, the construction of a special interpreter is often ad hoc and is based on specific aspects of the content of the ontology, which limits the reusability of the reasoner to other ontologies.

Following these observations, we can begin to characterize ontologies by the kinds of reasoning supported in software applications that use the ontology:

- Simple automated reasoning, for example, making inferences using the subclass relation by which properties defined at the parent class are inherited down to the children classes.
- Special automated reasoning, such as classificational reasoning in description logics, in which the reasoner may be able to take an arbitrary assertion in the KR language and classify it with respect to the taxonomic backbone of the ontology.
- General automated reasoning, which involves the use of *deductive rules* that combine information from across the ontology. Such inference rules characterize dependencies much like if-then-else statements in programming languages or business rules that try to characterize things that have to hold in an enterprise but which cannot typically be expressed in relational databases or object models.

8. Descriptive vs. prescriptive

The third pragmatic dimension concerns the source of the intended interpretations of the ontology's vocabulary. In descriptive approaches, the content of the ontology describes the intended interpretations by characterizing the entities and the relations among entities as a user or an expert might characterize those objects. In prescriptive approaches, the content prescribes the intended interpretations by explicitly mandating the way that those entities and their relationships are characterized.

Descriptive approaches often take a looser notion of characterization, perhaps allowing arbitrary objects into the model, which might not exist in the real world but which are significant conceptual items for the given user community. Prescriptive approaches often take a stricter notion of characterization, stating that only objects which actually exist or that represent natural kinds or types of things in the real world should be represented in the content of the engineering model.

9. Design methodology

The methodology employed in the construction of the ontology constitutes another pragmatic dimension within the Ontology Framework. A bottom-up (sometimes called "empirical") methodology places strong emphasis on either solely analyzing the data sources so that the resulting ontology covers their semantics, or on enabling arbitrary persons to characterize their content as they personally see fit. There is often an auxiliary notion or assumption that by doing so, patterns of characterizations may emerge or be preferred by a large group or community of persons.

A top-down (sometimes called "rationalist") methodology places strong emphasis on developing the ontology using known notions about the world or domain. The resulting ontology is independent of existing data sources whose semantics will be covered by the resulting ontology. The scope of the ontology

is often determined by considering a range of competency questions that a domain expert might want to ask about the domain.

There are of course many other ontology design methodologies that incorporate aspects of both bottom-up and top-down approaches. After considering the large number of intended uses for ontologies, it is easy to see that there is also a spectrum of design methodologies. They vary from a strong software engineering design lifecycle with requirements, evaluation and verification, all the way to a "no-design" methodology in which folksonomies emerge from the local behavior of thousands of individual users. Design methodologies are strongly related to the intended use. For instance, methodologies for managing controlled vocabularies and taxonomies are very social in nature and are intended to capture generalities about the meanings of words in a culture or domain. Also, the "verification" of ontologies is related to the role of reasoning or types of computational services to be enabled by the ontologies. For example, if an ontology is used for data integration, the verification of the consistency and completeness of data metamodels are important, whereas in a domain where the meanings of terms has legal consequences, verification is more about capturing the provenance of the design choices, rooted in authority or appropriate process. Consequently, any Ontology Framework will not only need to be populated with the ontologies themselves, but also with some indication of the various design methodologies as well.

Some members of the community have argued that ontologies should be considered a type of designed artifact, and that ontological engineering should be thought of as a discipline complementary to software engineering and to virtually any discipline dealing with data and information exchange. It is expected that this discipline will increasingly become a standard part of the relevant curricula.

9.1. Governance

The governance dimension addresses how decisions concerning the structure and (especially) content of an ontology are made, including the specification of quality criteria and certification. There was agreement at the summit that ontology with legal or regulatory implications will need to defer to existing legal, regulatory, and professional organizations concerning the natural language definitions of entities and semantic relationships. Ontology development should be viewed as an effort to organize and formalize concept definitions and relationships which are conventionally defined by existing institutions, not as an attempt to replace existing definitions with de novo definitions generated by autonomous computer scientists. This makes it necessary to record the provenance of every definition that is incorporated into an ontology (e.g., the controlling legislation, regulation, or standard from which a definition is taken).

10. Additional findings

10.1. Folksonomies and formal ontologies

One of the issues discussed during the Summit was the relationship between social tagging and folksonomies on the one hand, and more traditional structured formal ontologies such as taxonomies and axiomatized ontologies on the other. Until recently these efforts have been viewed as competitive approaches. The consensus of the Ontology Summit was that social tagging efforts should be viewed as large scale corpora to be used for inferring and validating more formal ontologies, akin to the use of large text corpora in computational linguistics studies. In addition, more formal ontologies can be used to inform social tagging by providing improved tag sets and faceted tagging.

10.2. Survey

In order to elicit the distinctions between various kinds of ontologies, an interactive survey was designed and posted on the Web in order to engage various communities. The respondents were invited to identify the community of which they are representatives and to describe the value of ontologies, as well as issues with ontologies in this community. The last section of the survey invited the respondents to describe and characterize the ontologies or related artifacts in use in this community.

Over fifty respondents from forty-two communities submitted entries to the survey. The best represented communities were Formal ontology, Applications development, Standards development, Web 2.0 and Biomedicine. Forty-one terms were identified as closely related to ontology, including formal ontology, upper ontology, concept system and controlled vocabulary. Some 70 ontologies from a variety of domains were characterized in the survey, including formal ontologies (e.g., BFO, DOLCE, SUMO, PSL), biomedical ontologies (e.g., Gene Ontology, SNOMED CT, UMLS), thesauri (e.g., MeSH, National Agricultural Library Thesaurus), folksonomies (e.g., Social bookmarking tags), general ontologies (WordNet, OpenCyc). The list also includes markup languages (e.g., NeuroML), representation formalisms (e.g., Entity-Relation model, OWL, WSDL-S) and various ISO standards (e.g., ISO 11179). Overall, this sample clearly illustrates the diversity of artifacts collected under "ontology".

11. Summary

The Ontology Summit 2007 "Ontology, taxonomy, folksonomy: Understanding the distinctions" was an attempt to collaboratively identify important dimensions whereby ontologies could be characterized. It resulted in a Framework including six major dimensions, as illustrated in the diagram. An interactive survey also was realized to contribute to the description of existing ontologies. In the face-to-face meeting, the Framework was put to a stress test: the participants used it to position a dozen ontologies along its dimensions.

We recognize that the current Framework is still preliminary. In particular, work needs to be pursued on refining the dimensions and establishing operational definitions for populating the Framework. We encourage professionals from various disciplines to contribute to this work by joining the Ontolog Forum Community of Practice.

12. Ontology summit convenors and co-sponsors

Organizing committee:

- Olivier Bodenreider (NLM, NIH)
- Tom Gruber (TagCommons)
- Nicola Guarino (ISTC–CNR)
- Ivan Herman (W3C)
- Deborah McGuinness (Stanford KSL)
- Mark Musen (NCOR, NCBO, Stanford SMI)
- Leo Obrst (Ontolog, MITRE)
- Frank Olken (NSF)
- Steve Ray (NIST)

- Barry Smith (NCOR, SUNY-Buffalo)
- Chris Welty (IBM Research)
- Peter Yim (Ontolog, CIM3)

Co-sponsors (organizations who are providing technical or funding support, and/or endorsing the objective of the Ontology Summit 2007):

- Accuracy & Aesthetics (Deborah MacPherson)
- The *Applied Ontology* journal (Nicola Guarino & Mark Musen)
- The Boeing Company (Michael Uschold)
- Bremen Ontology Research Group at Bremen University, supported by the faculties of Informatics and Linguistics and the Collaborative Research Center for Spatial Cognition (John Bateman/Till Mossakowski)
- Department of Philosophy, University at Buffalo (Barry Smith)
- CNR Institute for Cognitive Sciences and Technologies, Laboratory of Applied Ontology, Trento (Nicola Guarino)
- EPISTLE – European Process Industries STEP Technical Liaison Executive, the consortium responsible for the development of ISO 15926 (Matthew West)
- Conmergence (Ed Dodds)
- CIM3 (Peter Yim)
- Cycorp (Doug Lenat)
- Essential Strategies (David Hay)
- Humanmarkup.org (Rex Brooks)
- IBM Research (Chris Welty)
- Institute for Human and Machine Cognition (IHMC) (Pat Hayes)
- Indian Statistical Institute (Nabonita Guha)
- InfoCloud Solutions (Thomas Van der Wal)
- ISO TC 184 SC 4 JWG8 – the ISO working group that developed the ISO 18629 Process Specification Language (PSL) standard (Michael Gruninger)
- ISO/IEC JTC 1/SC 32 N 1498 – ISO 24707 Common Logic (CL) Working Group (Harry Delugach (editor), Pat Hayes, Chris Menzel, John Sowa, Murray Altheim, Elisa Kendall et al.)
- Java Professionals (Doug Holmes)
- Mathet Consulting (Matthew Hettinger)
- The Department of Accounting and Information Systems, Michigan State University (Bill McCarthy)
- MITRE (Leo Obrst, Pat Cassidy)
- NCBO – The National Center for Biomedical Ontology (Mark Musen)
- NCOR – (US) National Center for Ontological Research (Barry Smith & Mark Musen)
- The NeOn project (Aldo Gangemi & Enrico Motta)
- NIST – (US) National Institute of Standards and Technology (Steve Ray, Alan Bond et al.)
- NLM, NIH – (US) National Library of Medicine of the National Institute of Health (Olivier Bodenreider)
- The College of Computer and Information Science, Northeastern University (Larry Finkelstein, Ken Baclawski)
- OASIS Business-Centric Methodology (BCM) TC (Carl Mattocks)
- OASIS Universal Business Language (UBL) TC (Jon Bosak & Tim McGrath)

- Ontolog (Peter Yim/Leo Obrst/Kurt Conrad)
- Ontology Works (Bill Andersen)
- Pensive.eu (Peter Brown)
- Project10X (Mills Davis)
- Semantic Arts (Dave McComb)
- Shell International Petroleum (Matthew West)
- Chair of System Integration and Management, Moscow Institute of Physics and Technology (SIM, MIPT) (Leonid Ototsky)
- Stanford Medical Informatics (SMI) (Mark Musen)
- Stanford Knowledge Systems, AI Laboratory (KSL) (Deborah McGuinness)
- TagCommons (Tom Gruber)
- Department of Philosophy, Texas A&M University (Chris Menzel)
- Tall Tree Labs (Bob Smith)
- UCLA (Alan Bond)
- Universidad Distrital (Jorge Enrique Saby Beltran)
- University of Toronto (Michael Gruninger)
- Virginia Modeling Analysis and Simulation Center (Charles Turnitsa)
- VivoMind Intelligence (John Sowa)
- W3C (Ivan Herman & Ian Jacobs)