C. Baker and K.-H. Cheung, ed., *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*, New York: Springer Verlag, 2006, 139-158.

Chapter 6

THE EVALUATION OF ONTOLOGIES

Toward Improved Semantic Interoperability

Leo Obrst[1], Werner Ceusters[3,] Inderjeet Mani[1], Steve Ray[4], Barry Smith[3]

[1]MITRE, [2]Bashpole,Inc., [3]State University of New York at Buffalo, [4]US National Institute of Standards and Technology

Abstract: Recent years have seen rapid progress in the development of ontologies as semantic models intended to capture and represent aspects of the real world. There is, however, great variation in the quality of ontologies. If ontologies are to become progressively better in the future, more rigorously developed, and more appropriately compared, then a systematic discipline of ontology evaluation must be created to ensure quality of content and methodology. Systematic methods for ontology evaluation will take into account representation of individual ontologies, performance (in terms of accuracy, domain coverage and the efficiency and quality of automated reasoning using the ontologies) on tasks for which the ontology is designed and used, degree of alignment with other ontologies and their compatibility with automated reasoning. A sound and systematic approach to ontology evaluation is required to transform ontology engineering into a true scientific and engineering discipline. This chapter discusses issues and problems in ontology evaluation, describes some current strategies, and suggests some approaches that might be useful in the future.

Key words: ontology; evaluation; alignment; semantic interoperability; semantic similarity; validation; certification.

1. INTRODUCTION

Recent years have seen rapid progress in the development of ontologies intended to capture and represent aspects of the real world. Because ontologies explicitly represent domains – constituted by the entities, properties, and relationships that exist in the real world – they can be used to provide heterogeneously structured databases and multiple systems with comparable semantics. Ontologies thus support semantic interoperability and integration in organizations in many domains, with notable successes thus far in the life sciences.

There is, however, great variation in the quality of ontologies. Prospective users of these ontologies typically have no insight as to their coverage, their intelligibility to human users and curators, their validity and soundness, their consistency, the sort of inferences for which they can be used, or their ability to be adapted and reused for wider purposes.

In addition, there are systems such as controlled vocabularies, thesauri and terminologies that in the best case exhibit some ontological features or that are developed using ontology tools, but that are not ontologies in their own right. The pervasive use of the term 'ontology' for such resources is unfortunate.

Users are unsure whether particular ontologies can help them solve their particular data, application, or service problems. Enterprises and communities are not confident that large ontologies formed from the merging or mapping together of smaller ontologies will enable wider semantic operability for their aggregated data and complex applications, or merely result in greater conceptual confusion.

If ontologies are to become progressively better in the future, more rigorously developed, and more appropriately compared, then a systematic discipline of ontology evaluation must be created to ensure quality of content and methodology. Ideally it will ensure also that an evolutionary path towards improvement in ontologies is created, analogous to the paths to improvement with which we are familiar in the traditional domains of science and engineering.

2. ISSUES IN ONTOLOGY EVALUATION

An ontology can be evaluated against many criteria: its coverage of a particular domain and the richness, complexity and granularity of that coverage; the specific use cases, scenarios, requirements, applications, and data sources it was developed to address; and formal properties such as the consistency and completeness of the ontology and the representation language in which it is modeled. Ontologies can also be evaluated per questions such as the following: Is the ontology mappable to some specific upper ontology, so that its evaluation will be at least partially dependent on the evaluation of the latter also? What is the ontology's underlying philosophical theory about reality? Theory perspectives include *idealist*: reality is dependent on mind or is ultimately mental in nature, *realist*: universals or invariant patterns really exist independently of minds (and thus of observers), *conceptualist*: universals are neither independently existing nor just names but exist only in human and possibly other animal minds as abstractions from particulars, *nominalist*: only particulars exist and universals do not exist in reality or in our minds but only as general terms; *3-dimensionalist*: space and time exist independently and material objects are extended in space and endure through time, *4-dimensionalist*: only a combined spacetime exists; etc. [for realist perspective in life sciences, see 14]? Finally, what kinds of reasoning methods can be invoked on the ontology, i.e., by the inference engine that uses it? The latter question highlights the importance also of the evaluation of ontology tools, though this chapter will not directly address that topic.

Ontology evaluation includes aspects of ontology validation and verification relating to structural, functional, and usability issues. [28, 29] develop a theoretical framework and a formal model for evaluating ontologies, including a meta-ontology of semiotics (the study of signs and signification, i.e., the bearing of meaning by those signs, a generalization of linguistics to other sign systems beyond natural language) called O^2 and an ontology of ontology evaluation and validation called oQual [29, p. 2]. oQual uses the evaluation matrix of [36] to answer general evaluation questions on the goals, functions, use cases, stage of development, methodology employed in the ontology development process, and usability of the ontology.

One issue in evaluating ontologies is whether to perform glass box (component-based) vs. black box (task-based) evaluation, the latter usually applied to ontologies that are tightly integrated with an application performing specific tasks [36]. An example of such an application might be a semantic search engine that uses a domain specific ontology to search over a collection of documents.

2.1 Knowledge Representation

Of importance in evaluating an ontology is the expressivity of the knowledge representation (KR) language the ontology is represented in, in light of the trade-off between the value of *high expressivity* and the cost of *computation*. Emphasis on high expressivity is manifested by First-Order Logic (FOL)-based languages such as Common Logic (CL) [18], the Interoperable Knowledge Representation for Intelligence Support (IKRIS) language [38], and the Web Ontology Language's (OWL) most expressive dialect OWL Full [1, 19]. Emphasis on minimizing the cost of computation is currently manifested by OWL-Lite, OWL-DL (description logic) and other description logics.

Two ontologies, both covering the same domain, one expressed in OWL-Lite and one expressed in CL, necessarily will be evaluated differently, say, for a given domain application that requires fine model precision, e.g., fully automated selling and purchasing as envisioned for a range of semantic Web services. For a less precise task, say, for classifying documents in a loose topic hierarchy, either one may be sufficient.

The KR language defines the syntax and the semantics for the ontology models expressed in that KR language. Figure 1 [54] displays the three levels that are involved: the meta-language, i.e., the KR language, the ontology concept or type level, and the instance level. The lowest level instantiates the generic properties described by the middle level.

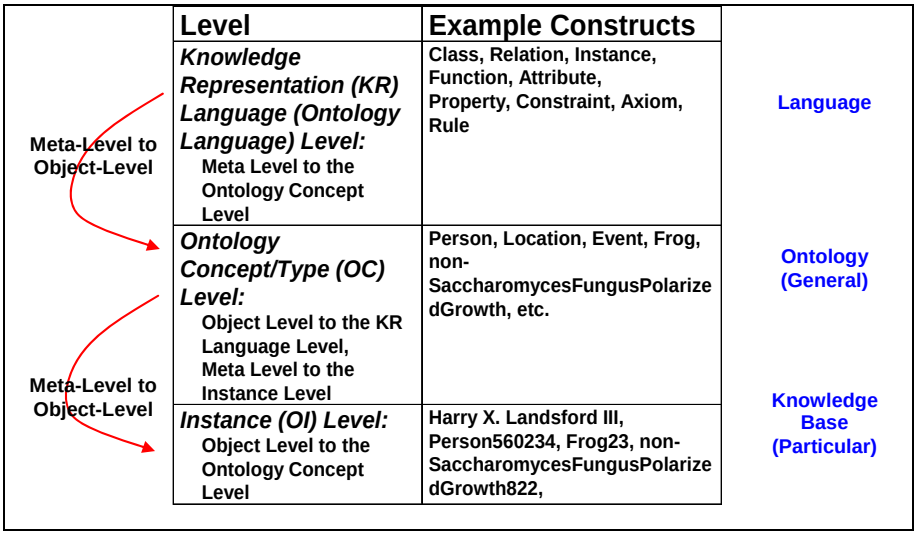

Figure -1. Ontology Representation Levels

Constructs defined in the KR language can be arbitrarily different. For example, description logics such as OWL are quite different from FOL languages such as CL. Some first-order languages such as IKRIS have non-standard extensions, e.g., *quotations and contexts*. OWL-Full allows classes to also be individuals (instances). Finally, OWL also has been extended with the Semantic Web Rule Language SWRL, which combines description logic constructs with a Horn rule-like capability as found in logic programming (a generalized Modus Ponens proof form syntactically restricted to permit efficient automated inference).

Any evaluation of an ontology has to account for the expressivity of the KR language in which it is modeled. One way to level the playing field for evaluation therefore is to translate the ontology to be evaluated to a canonical KR language, typically a very expressive language such as CL, which can be problematic insofar as there will likely not be a fully automated translation from the less expressive to the more expressive language.

The ontology to be evaluated may also be mapped to an upper ontology that defines constructs that are not in the KR language. For example, an upper ontology may define class, relation, property, attribute, facet, quality, or trope. More

commonly, an upper ontology will define notions of space and time (3-D), or spacetime (4-D) [63], and endurants, perdurants, or both [34], and parts, wholes which lower ontologies use [65, 75]. The given ontology thereby can use these object-level assertions. Thus, ontology evaluation must also consider the mapping to an upper ontology.

Finally, the formal properties of the KR language will be significant for evaluating ontologies and reasoning methods on those ontologies. Formal properties include *soundness* (any expression that can be derived from the knowledge base (KB) of the ontology and its instances is logically implied by that KB), *completeness* (any expression that is logically implied by the KB can be derived), and *decidability* (being both sound and complete). All of these will correlate with the formal *complexity* (time of execution, space of memory needed to compute an answer). One can consider *undecidability* as meaning that a query may never terminate, since an inference engine will be searching an infinite space. A very expressive language such as FOL is *semi-decidable*: it is decidable in that if a theorem is logically entailed by a FOL theory, a proof will eventually be found, but undecidable in that if a theorem is not logically entailed, a proof of that may never be found. Decidability of a language or logic does not mean tractability of the automated reasoning on that language, but there is a relationship. Expressivity and complexity are typically inversely proportional to the tractability of reasoning.

A related property having to do with the ontology represented in the KR language is *consistency* (if contradictions can be proven from a given proposition, then the theory is inconsistent). Inconsistent theories have no formal models (interpretations of those theories, the semantics). Inconsistency may manifest itself by circularity, disjoint partition errors, and other semantic inconsistencies, e.g., incorrect classifications. Similarly, there are other ontology-level correlates of the formal properties. Ontology incompleteness is indicated by imprecisely defined or missing concepts, partially defined disjointness properties, redundancy of class, instance, or relation [61].

2.2 Use Cases and Domain Requirements of Ontologies

In early ontology engineering, methodological considerations were introduced that remain significant today. One is the use of competency questions to drive out requirements [33]. Competency questions are those an ontologist frames prior to the development of the ontology. These consist of bottom-up questions one would like answered concerning the data sources the ontology would encompass and also top-down questions one would like answered considering the nature of the domain. Such questions tend to push the ontologist to construct specific use cases and modeling requirements – sound software engineering practices – to drive and constrain the ontology development. Once an inference engine can give reasonably complete and coherent answers (consider them queries or theorems) to the competency questions, as gauged by a domain expert, the development effort is completed. These competency questions thus also act as a test suite, providing value during both analysis and validation.

The domain requirements driven out by competency questions and use cases are ontology evaluation criteria. The requirements can focus on aspects such as physical vs. functional properties (the latter is more important for human artifacts), which will vary for the same entities depending on the intent of the model. Consider, for example, a supply chain ontology of chemicals. Raw manufacturers may focus on physical chemical properties such as valency, Ph factor, volatility, human toxicity, purity level, etc., while down-stream supply chain vendors such as paint manufacturers may focus on properties such as drying time, light reflectability, heat resistance , etc.

2.3 Semantic Agreement and Consensus Building

Measurement of human agreement on classification tasks has been well-studied. Similar measurement can be applied to the problem of classifying instances in terms of an ontology or mapping a concept to candidate classes in one or more ontologies. Researchers developing linguistic classification schemes for annotating corpora have measured inter-

annotator agreement using the Kappa statistic [64, 9], defined as

$$K = (P(A) - P(E)) / 1 - P(E)$$

where "P(A) is the proportion of times that the coders agree and P(E) is the proportion of times that we would expect them to agree by chance." [9]

Such measurements have played a crucial role in the evolution of such annotation schemes, some of which have resulted in successful solutions to problems in computational linguistics. Such metrics are appropriate when the categories involved are already defined and where annotators are required to choose between possible categories.

Inter-annotator agreement studies have been carried out in the course of Gene Ontology annotation of terms in documents [7], in the context of the BioCreative information extraction task. It was found here that expert annotators (EBI GOA project curators) [23] were generally correct in their annotations, but missed a few, and that the specificity of the annotation varied depending on their biological knowledge.

Semantic agreement is highly influenced by the degree to which humans are trained in a set of guidelines for how to label examples in terms of categories, and the richness of these guidelines. For certain problems, guidelines may have to be refined to arrive at more agreement; where there is eventual disagreement, adjudication may have to be used. The process of arriving at the right categories involves a variety of factors that include aspects of group collaboration. Delphi methods [50] have a role here, but have been relatively underexplored for use in ontology evaluation.

2.4 Semantic Similarity and Semantic Distance

The majority of ontologies exist, or can be represented in, a graph-based form. *Semantic distance* and *semantic similarity* are two measures used in graph representations to capture to what extent two nodes in a graph are related. Whereas semantic distance measures how closely two nodes are topologically related in a graph, semantic similarity captures to what extent two nodes might represent the same entity in

reality. Obviously, the two notions are closely related, but there are some important differences. In a fracture ontology, for example, a node representing a "fractured arm" should have a very short semantic distance from one referring to an "arm fracture"; yet the semantic similarity would still be low: a fracture cannot *be* an arm. It is now a measure of a high-quality ontology that it should be possible to compute the semantic distance of post-coordinated terms such as "patient-WITH-arm fracture" and "patient-WITH-fractured arm" as being minimal, and the semantic similarity as being maximal.

Various approaches to the calculation of these values have been proposed. They tend to fall into two categories. Edge-based methods exploit mainly the idea of path-length in a network (with or without additional weights according to the type of link traversed). Node-based methods also take into account contextual factors, such as the degree to which cognate terms are to be found in a large corpus [58], the idea being that the information content associated with nodes related to terms that occur often in a corpus is lower than of nodes that occur rarely, and that information-low nodes tend to appear higher in an ontology hierarchy. Still more sophisticated edge-based methods are described in [80] which is based entirely on the hierarchical *is_a*-relationship, and in [74], where this idea is expanded to take account also of other sorts of relationships between nodes.

2.5 Alignment with Other Ontologies

Ontology alignment (mapping, articulation) attempts to compare two ontologies, where one ontology is the 'reference' ontology against which a candidate ontology should be compared. Arriving at a suitable reference ontology can be challenging; preferably, it should be one that was created under similar conditions, with similar goals, to the candidate ontology. This issue is less a problem when, say, comparing different versions of the same ontology.

Ontology alignment can provide some information about the relative quality of the ontologies aligned. It falls short of providing full evaluation metrics, however, since we do not as yet have gold standard reference ontologies. In [15] an attempt is made to base such a metric on using reality as the gold standard.

Alignment is usually described as an activity that, given two arbitrary ontologies O1 and O2, aims to find for each 'concept' in ontology O1 a corresponding 'concept' in ontology O2 that has the same intended meaning [43, 22, 40]. To say that two concepts have similar semantics, on this account, means roughly that they occupy similar places in their lattices. A problem with the above is, however, clear: ontology alignment is defined in terms of the correspondence (equivalence, sameness, similarity) of concepts. But how, precisely, do we gain access to concepts in order to determine whether they stand in a relation of correspondence?

One option is via definitions, but then these definitions themselves, supplied by the ontologies to be matched, will likely employ different terms (or 'concepts'), so that the problem of matching has merely been shifted to another place. Another option, as suggested in [22], is to establish correspondence by looking at the positions of given concepts in their surrounding concept lattices. But how, unless we have already matched some single concepts, can we compare 'places' in distinct lattices (these 'lattices' may have very different mathematical forms)? This leaves only some statistically-based algorithms involving lexical term-matching, the results of whose application have thus far proved uneven, to say the least.

When [24] surveyed ontology alignment methods, they found that the majority are based on analyzing either the vocabulary used to label concepts or the structure in terms of which the latter are organized. Term-based comparison is, as mentioned above, problematic because of term synonymy (multiple terms may have very similar meanings) and term ambiguity, i.e., polysemy (a given term may refer to multiple distinct referents). In addition, term comparisons require a degree of morphological normalization, and complex multi-word terms need to be handled. The use of structure-based comparison is, however, applicable in the restricted case where the ontologies being aligned are very similar, as in version comparison [20].

One can use coarse-grained methods for comparing ontologies in terms of distance, while paying lip-service to the term-matching problem. Research on ontology induction for biology has followed such an approach in comparing system-generated ontologies with human ones. For example, [52]

limited the terms to those in the reference ontology, comparing relations closed among those terms in each of the ontologies. Their relation precision measures the proportion of relations a distance D1 apart in one ontology that are at most a distance D2 apart in the other, subject to a variety of constraints (e.g., the direction and type of the links being the same, similar, different, etc.) The disadvantage of such distance-based measures is their over-sensitivity to small changes in node ordering; also, the 'conceptual' salience of particular nodes is not taken into account. In related work, [41] measures the percentage of times terms in a parent-child relationship appear in an immediate or transitive parent-child relationship in the other.

3. ONTOLOGIES FOR THE LIFE SCIENCES: EVALUATION TECHNIQUES

In the life sciences, widely-used ontologies such as the Gene Ontology, UMLS, BioPAX, etc. are being used primarily to perform 'associative' query expansion during search or to reconcile annotations, rather than for deep reasoning. A number of ontologies used in biology have been developed or enhanced with description logic representations to permit richer inferential use, including the Gene Ontology Next Generation Project (GONG) [77], SNOMED-Clinical Terms [73], the Unified Medical Language System (UMLS) [57, 42, 17], the Generalised Architecture for Languages, Encyclopaedias and Nomenclatures in Medicine (GALEN) [59], the Foundational Model of Anatomy (FMA) [79], and the National Cancer Institute (NCI) Thesaurus [30]. The use of description logics here provides a degree of evaluation in terms of error-checking of the terminological structure.

The deeper reasoning tasks that ontologies have been used for include: classification, e.g., finding the most specific protein family for an entity in a protein database [76], answering queries related to process models of a vaccinia virus life cycle [37], and reasoning about part-whole models of anatomy [35]. However, there are a number of problems with such ontologies, of the sorts described in [10, 11] which demonstrate that the error-checking mechanisms provided by description logic tools do not suffice to find all errors.

Many techniques are being used for ontology evaluation in the life sciences and more generally. For fairly exhaustive summaries of current practice, see [4, 5]. In this section, we look at a number of the techniques: evaluation with respect to the use of an ontology in an application, with respect to domain data sources, assessment by humans against a set of criteria, natural language evaluation techniques, and the use of reality itself as a benchmark. The section concludes with a discussion of prospects for the future: accrediting and certifying ontologies that have passed some evaluation criteria, and the notion of an ontology maturity model.

3.1 Evaluate Use Of Ontology In An Application

Task-based evaluations offer a useful framework for measuring practical aspects of ontology deployment, such as the human ability to formulate queries using the query language provided by the ontology, the accuracy of responses provided by the system's inferential component, the degree of explanation capability offered by the system, the coverage of the ontology in terms of the degree of reuse across domains, the scalability of the knowledge base, and the ease of use of the query component. Such task-based evaluations can leverage use-cases or scenarios to characterize the target knowledge requirements. In the DARPA High-Performance Knowledge Bases project [16], the evaluation included a crisis management scenario, where evaluators formulated parameterized test questions and answer keys, and subjectively graded question formulation, answers, and system explanations regarding inferential steps. In the case of the qualitative assessment of CYC [48] for use by the Internal Revenue Service [60], the use-cases were drawn from FAQs and topics at the IRS web site. The questions could include statements, and were selected to be complex enough to require ontology-based inference. Another assessment of CYC [51] was focused on its use for word-sense disambiguation and coreference in natural language processing. Here the queries chosen were taxonomic queries as well as queries that examined distances between pairs of concepts.

Another task-based evaluation scheme involves using textbooks and other found material to guide task-specific

knowledge capture requirements. In the Rapid Knowledge Formation (RKF) project [37], subject-matter experts added knowledge about DNA transcription to two ontological systems, Cycorp's CYC and SRI's SHAKEN, based on ten pages from a standard textbook. Independent judges carried out subjective grading of the accuracy of the answers obtained to test questions as well as the degree of reuse (old vs. new axioms used). Further, comparisons of performance of subject-matter experts were carried out against knowledge engineers from the developer institutions. A particularly interesting feature of RKF was the use of challenging 'explanation' questions, e.g., '*Can transcription be performed on either strand of a given DNA gene segment with equivalent effects? Explain.*' A similar approach was taken in the HALO pilot project [27], which used a chemistry domain and involved CYC, SHAKEN, and Ontoprise's Ontonova. In HALO, both the test questions and the assessment were modeled on Advanced Placement chemistry tests.

Task-based evaluations, however, can be expensive to carry out and the results cannot be used to test systems whenever the need arises. Further, measurements of reuse face the problem of counting concepts or axioms, which depends on what sorts of concepts are reified in a particular ontology.

3.2 Comparison of Ontology Against a Source of Domain Data

Coverage of the ontology can be evaluated with respect to other ontologies and databases representing a particular domain. For example, the Gene Ontology has been automatically mapped to a number of other classifications as well as to databases. However, such coverage estimates are subject to noise in the mapping, of the sort discussed earlier for term-matching methods. In addition, entity normalization (mapping from attested names to database ids) is non-trivial in biological domains, as shown in [2], where increased length of the names and ambiguity in the vocabulary was tied to substantially poorer performance for mouse genes compared to yeast genes.

Ontologies can also be mapped automatically to a corpus of documents representative of a particular domain, and this mapping can be used to assess or compare ontologies. The

approach of [6] compares ontologies by examining only the concepts which are common to the ontology and the corpus. Each ontology is represented by a feature vector, and the distance between the ontologies is represented by the distance between the vectors. The approach also provides a method for estimating the probability of the ontology given the corpus. The approach ignores relationships between concepts, and is subject to the standard problems with term-matching.

3.3 Assessment by Humans Against a Set of Criteria

Assessment by humans against a set of criteria had been used extensively by Ceusters and Smith in a series of studies of ontologies and terminologies in biomedicine:

- The Gene Ontology [69, 72, 68]
- Systematized Nomenclature of Medicine (SNOMED) Clinical Terms (CT) [31, 10, 11, 3]
- The National Cancer Institute Thesaurus [13, 46]
- The Unified Medical Language System [71, 45]
- ICF (International Classification of Functioning, Disability and Health) [49]
- HL7-RIM [67] and ISO terminology and data integration standards [66, 70].

The principles in question are derived largely from common sense: provide clear documentation, use terms in a consistent (and consistently non-ambiguous) way, provide updating and versioning procedures, and procedures for users to propose corrections and additions to the ontology. Some are derived from basic (philosophical) logic, including the theory of definitions – for example: avoid circular definitions; do not give a new meaning to a term with an already established use in the domain in which the ontology is intended to be used; define the principal relations in the ontology (for example *is_a* and *part_of*) and used them in consistent ways. Yet others are derived from the tradition of philosophical realism: see section [Using Reality as Benchmark] below. For a general overview see [12, 44], which describe also how the application of some of these principles to the evaluation of ontologies can be implemented in automated reasoning systems.

3.4 Natural Language Evaluation Techniques

Natural language processing tasks such as information extraction, question-answering, and abstracting are knowledge-hungry tasks. It is therefore natural to consider evaluation of ontologies in terms of their impact on these tasks. Information extraction in biomedical text has made heavy use of the Gene Ontology; it is possible to subtract out or substitute other ontologies such as UMLS to see the impact on performance. Further, in the BioCreative evaluation [76], one of the tasks was to find evidence in a paper for the GO code provided for a given protein. The best systems for this task had around 30% accuracy, in part because of the difficulty of the inference involved. For example, the text passage "The p21waf/cip1 protein is a universal inhibitor of cyclin kinases and plays an important role in inhibiting cell proliferation." is evidence for the GO annotation of that protein as having a molecular function of "negative regulation of cell proliferation (GO code: 0008285)", which requires a system to make the difficult inference that inhibition is equivalent to negative regulation. The impact of ontologies on such an 'entailment' task could be measured.

Question-answering is another technology where ontologies can play a useful role in bridging the gap between a natural language question and a candidate passage in a document. Current systems use WordNet along with ad-hoc taxonomies rather than full-fledged ontologies. Accuracy on question-answering tasks can provide a task-based measure of the impact of an ontological resource and its components. Such applications also present challenging requirements in terms of performance efficiency. Question-answering systems for the life sciences are still in their infancy, however.

3.5 Using Reality as Benchmark

The authors of [14] propose a technique for ontology evaluation based on determining the semantic correspondences between nodes in two ontologies identified during ontology matching and subsequent mapping or merging, and in particular by the examination of the changes made in subsequent versions of an ontology by its curators. They build a metric resting on a distinction between three

levels which have a role to play where ontologies are used as artifacts for annotation and automated reasoning as for example in the field of biomedicine: (1) the reality on the side of the patient; (2): the cognitive representations of this reality embodied in observations and interpretations on the part of clinicians and others, and (3) the publicly accessible concretizations in representational artifacts of various sorts, of which ontologies are examples. To establish the metric it is necessary first of all to specify the features by which an eventual gold standard must be marked. Each node in such an ontology would need to designate (1) a single portion of reality (POR) (denoting instances, universals, and the simple and complex combinations these form through interrelationships of various types [14]), which is (2) relevant to the purposes of the ontology, and such that (3) the authors of the ontology intended to use this node to designate this POR. Moreover, (4) no PORs objectively relevant to these purposes would be missed by the ontology. We can now obtain a measure of the quality of an ontology (and of the work, and competence, of its developers) by determining the degree to which successive versions of the ontology approximate ever more closely to this ideal, something which can be quantified by documenting the different kinds of changes in an ontology, reflecting for example (1) changes in the underlying reality (does the appearance or disappearance of an entry in a new version of an ontology relate to the appearance or disappearance of entities or of relationships among entities in reality?); (2) changes in our scientific understanding; (3) reassessments of what is relevant for inclusion in an ontology, or (4) encoding errors introduced during ontology curation (for example through erroneous introduction of duplicate entries reflecting lack of attention to differences in spelling). We can measure the degree of improvement along each of these dimensions in each successive version of the ontology by tracking the history of revisions. The metric can be used also with measures of the performance of an ontology in applications; a divergence between the two is once again a sign that the ontology does not line up with the reality it is supposed to represent.

3.6 Ontology Accreditation, Certification, Maturity Model

Once validation, verification, and evaluation of ontologies become standard practice, a further evolution toward more rigor is to issue accreditation or certification (to a given ontology or to a team of ontology developers or an organization) based on a set of recognized evaluation criteria by an accrediting body (top-down) or an accrediting process (bottom-up) similar to the trustworthiness, reputation, and feedback mechanisms of online services and communities such as E-Bay and Amazon [21]. This kind of "Good Ontology-keeping" seal of approval would compute and assign a quality rating of the ontology [55, 53]. An alternative approach might include ontology repositories that have some entrance requirements, e.g., an open-rating system extended with topic-specific trust [49, 56]. The emerging Extended Metadata Repositories (XMDR) project [78], based on the ISO/IEC 11179 Metadata Registries standard [39], represents another repository paradigm that includes ontology registration, mapping services, and prospectively certification.

As discussed throughout this paper, additional measures associated with an ontology accreditation score could be domain, breadth of application or coverage within that domain, average taxonomic depth and relational density of nodes, completeness of axiomatic specification, adherence to principled methodologies such as Methontology [25, 26] and OntoClean [34].

Creation of an ontology maturity model may also be useful [55], like the Software Engineering Institute's Capability Maturity Model Integration [8]: a process of subprocesses in a full ontology lifecycle model, with gradations and decision procedures for maturity of ontologies by which organizations and ontologies could be gauged in terms of rigor of the ontology engineering process. Levels of maturity in the model could be defined by many of the properties discussed in this chapter, including degree of logical formalization, axiomatizability and satisfiability measures; strictness and properties of the ontology development process; quality of ontology; linkage to reference, utility, middle, and upper ontologies; domain of application usage; and tool support,

including KR language, development, and reasoning assistance.

4. NEXT STEPS AND RECOMMENDATIONS

The ultimate evaluation of an ontology is in terms of its adoption and successful use, rather than its consistency or coverage. The Gene Ontology, while clearly impoverished in many representational aspects, is a fundamental success story.

In the long run, rigorous ontology evaluation must evolve in support of a broader engineering discipline of semantics and ontologies, which itself would be part of an information engineering discipline. A rigorous engineering discipline in semantics and ontologies must therefore include certain attributes in common with other engineering disciplines:

- A formal, verifiable science base
- Tested theories that allow prediction
- Defined units of measure
- Well-defined engineering practices

If as a society we hope to reliably build complex information systems incorporating ontologies, these foundational elements must be available to engineering practitioners. This will not be an easy undertaking. A measurable science of semantics or ontologies requires some fundamental questions to be answered, such as what are meaningful, theoretically grounded units of measure in this new science. Beyond the early work performed by [62] on information entropy as a measure for uncertainty in a message, little progress has been made. And yet, intuitively we deal with notions such as 'semantic proximity' in our daily lives. In other words, we satisfy ourselves, usually through dialogue, that our own conceptualization of some notion is 'close enough' to that of another to allow meaningful discourse. Just how to characterize the dimension in which 'close enough' is evaluated, much less what the unit of measure is, remains an unsolved problem.

Therefore, as a community we need to approach ontology evaluation as part of a larger endeavor to systematize the construction of information systems. In this way, we can

realistically hope to succeed in building ever more complex systems without drowning in complexity.

ACKNOWLEDGEMENTS

This work was supported in part by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant 1 U 54 HG004028.

The views expressed in this paper are those of the authors alone and do not reflect the official policy or position of The MITRE Corporation or any other organization or individual.

REFERENCES

1. Bechhofer, Sean; Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, Lynn Andrea Stein. 2004. OWL Web Ontology Language Reference. W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-ref/.
2. Blaschke, C., L. Hirschman, A. Valencia, and A. Yeh. 2004. A critical assessment of text mining methods in molecular biology. BMC Bioinformatics (22-article special issue), Volume 6, Supplement 1. http://www.biomedcentral.com/1471-2105/6?issue=S1.
3. Bodenreider, Olivier; Barry Smith; Anand Kumar; and Anita Burgun, 2004. Investigating subsumption in DL-based terminologies: a case study in SNOMED-CT, in: U. Hahn, S. Schulz and R. Cornet (eds.), *Proceedings of the First International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004*), 12–20.
4. Brank, Janez; Marko Grobelnik; Dunja Mladenić. 2005a. Ontology evaluation, deliverable D1.6.1, EU-IST Project IST-2003-506826 Semantically Enabled Knowledge Technologies (SEKT), Jožef Stefan Institute, Ljubljana, Slovenia, May 8, 2005.
5. Brank, Janez; Marko Grobelnik; Dunja Mladenić. 2005b. A survey of ontology evaluation techniques. SiKDD05.
6. Brewster, C., Alani, H., Dasmahapatra, S. and Wilks, Y. 2004. Data driven ontology evaluation. In *Proceedings of International Conference on Language Resources and Evaluation,* Lisbon, Portugal.
7. Camon E., Barrell D., Dimmer E., Lee V., Magrane M., Maslen J., Binns D., Apweiler R. 2005. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics* 6(1): S17 (2005).
8. Capability Maturity Model Integration (CMMI). Software Engineering Institute, Carnegie-Mellon University. http://www.sei.cmu.edu/cmmi.

9. Carletta, J.1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.
10. Ceusters, Werner, Barry Smith and Jim Flanagan, 2003, Ontology and medical terminology: why description logics are not enough, in *Proceedings of the Conference: Towards an Electronic Patient Record (TEPR 2003)*, San Antonio 10-14 May 2003, Boston, MA: Medical Records Institute (CD-ROM publication).
11 Ceusters, Werner, Barry Smith, Anand Kumar, Christoffel Dhaen, 2004a. Ontology-based error detection in SNOMED-CT, in M. Fieschi, et al. (eds.), *Medinfo 2004*, Amsterdam: IOS Press, 482–486.
12. Ceusters W, Smith B, JM Fielding, 2004b. LinkSuite: formally robust ontology-based data and information integration. In Rahm E (Ed.): *Data Integration in the Life Sciences: DILS 2004,* (Lecture Notes in Computer Science 2994) Springer 2004, p. 124-139.
13. Ceusters W, Smith B. 2005. A terminological and ontological analysis of the NCI thesaurus. *Methods of Information in Medicine 2005;* 44: 498-507.
14. Ceusters W, Smith B. 2006a. A realism-based approach to the evolution of biomedical ontologies. Forthcoming in *Proceedings of the AMIA 2006 Annual Symposium,* Washington DC, November 11-15, 2006.
15 Ceusters W, Smith B. 2006b.Towards A realism-based metric for quality assurance in ontology matching (forthcoming in *Proceedings of FOIS-2006*).
16. Cohen, P., Schrag, R., Jones, E., Pease, A., Lin, A., Starr, B., Easter, D., Gunning, D., and Burke, M. 1998. The DARPA High Performance Knowledge Bases project. *Artificial Intelligence Magazine,* vol. 19, no. 4, 1998, pp.25-49.
17. Cornet R, Abu-Hanna A. Usability of expressive description logics--a case study in UMLS. *Proceedings of AMIA Symp 2002*:180-4.
18. Common Logic Standard, June 21, 2006 version.. http://cl.tamu.edu/docs/cl/24707-21-June-2006.pdf.
19. Daconta, M., K. Smith, L. Obrst. *The Semantic Web: The Future of XML, Web Services, and Knowledge Management*. John Wiley, Inc., 2003.
20. Daude, J., Padro, L. and Rigau, G. 2001 A Complete WN1.5 to WN1.6 Mapping. NAACL-2001 Workshop on WordNet and Other Lexical Resources: Applications, Extension, and Customization, 83-88.
21. Dellarocas, Chrysanthos. 2006. Reputation mechanisms. Forthcoming in *Handbook on Economics and Information Systems*. (T. Hendershott, ed.), Elsevier Publishing. http://www.rhsmith.umd.edu/faculty/cdell/papers/elsevierchapter.pdf.
22. Ehrig M, Sure Y. 2004. Ontology mapping - an integrated approach. In *Proceedings of the First European Semantic Web Symposium, ESWS 2004,* volume 3053 of Lecture Notes in Computer Science, pages 76–91, Heraklion, Greece, May 2004. Springer Verlag.
23. European Bioinformatics Institute (EBI) Gene Ontology Annotation (GOA) Project. http://www.ebi.ac.uk/GOA/.
24. Euzénat J; Thanh Le Bach; Jesús Barrasa; Paolo Bouquet; Jan De Bo; Rose Dieng; Marc Ehrig; Manfred Hauswirth; Mustafa Jarrar; Ruben Lara; Diana Maynard; Amedeo Napoli; Giorgos Stamou; Heiner Stuckenschmidt; Pavel Shvaiko; Sergio Tessaris; Sven Van Acker; Ilya Zaihrayeu. 2004.

KnowledgeWeb deliverable D2.2.3: State of the art on ontology alignment. V1.2, August 2004. http://www.starlab.vub.ac.be/research/projects/knowledgeweb/kweb-223.pdf.

25. Fernández, M. 1999. Overview of methodologies for building ontologies. *Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends. (IJCAI99).* August. 1999. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-18/4-fernandez.pdf.
26. Fernandéz, Mariano; Gómez-Pérez, Asunción; and Juristo, Natalia. 1997. METHONTOLOGY: from ontological art to ontological engineering. *Workshop on Ontological Engineering. Spring Spring Symposium Series. AAAI97*, Stanford.
27. Friedland, N. S., Allen, P. G., Witbrok, M., Matthews, G., Salay, N., Miraglia, P., Angele, J., Staab, S., Israel, D., Chaudhri, V., Porter, B., Barker, K., and Clark, P. 2004. Towards a quantitative, platform-independent analysis of knowledge systems. *Proceedings of KR'2004.*
28. Gangemi, A.; Catenacci, C.; Ciaramita, M.; Lehmann, J. 2005. A theoretical framework for ontology evaluation and validation. In *Proceedings of SWAP2005.*
29. Gangemi, Aldo; Carola Catenacci; Massimiliano Ciaramita; and Jos Lehmann. 2006. Modelling ontology evaluation and validation. To appear in *Proceedings of ESWC2006,* Springer.
30. Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B. 2003. The National Cancer Institute's Thesaurus and Ontology. *Journal of Web Semantics 2003* 1(1). http://www.websemanticsjournal.org/ps/pub/2004-6.
31. Goldberg LJ, Ceusters W, Eisner J, Smith B 2005. The significance of SNODENT, *Stud Health Technol Inform.* 2005;116:737-742.
32. Grenon, P. 2003. Spatio-temporality in basic formal ontology: SNAP and SPAN, upperlevel ontology, and framework of formalization (part I). *Technical Report Series 05/2003, IFOMIS*, 2003.
33. Gruninger, M. and Fox, M.S. 1995. Methodology for the design and evaluation of ontologies. In: *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95*, Montreal.
34. Guarino, N; C. Welty. 2002. Evaluating ontological decisions with OntoClean. *Communications of the ACM.* 45(2):61-65. New York: ACM Press. http://portal.acm.org/citation.cfm?doid=503124.503150.
35. Hahn, U. and Schulz, S. 2003. Towards a broad-coverage biomedical ontology based on description logics. *Pacific Symposium on Biocomputing 8*, 2003, 577-588.
36. Hartmann, Jens; Peter Spyns; Alain Giboin; Diana Maynard; Roberta Cuel; Mari Carmen Suárez-Figueroa. 2005. D1.2.3 Methods for ontology evaluation. EU-IST Network of Excellence (NoE) IST-2004-507482 KWEB Deliverable D1.2.3 (WP 1.2).
37. IET. RKF Y1 evaluation report, October 2001. http://www.iet.com/Projects/RKF/IET-RKF-Y1-Evaluation.ppt.
38. Interoperable Knowledge Representation for Intelligence Support (IKRIS). http://nrrc.mitre.org/NRRC/ikris.htm.
39. ISO/IEC 11179 specification. http://metadata-standards.org/.

40. Kalfoglou Y, Schorlemmer M. 2003. Ontology mapping: the state of the art. *Knowledge Engineering Review,* 18(1):1--31, 2003.
41. Kashyap, V., Ramakrishnan, C., Thomas, C., and Sheth, A. 2005. TaxaMiner; an experimental framework for automated taxonomy bootstrapping. *International Journal of Web and Grid Services,* Special Issue on Semantic Web and Mining Reasoning, September 2005.
42. Kashyap V, Borgida A. 2003. Representing the UMLS semantic network using OWL: (Or "What's in a Semantic Web link?"). In: Fensel D, Sycara K, Mylopoulos J, editors. *The SemanticWeb – ISWC 2003*. Heidelberg: Springer-Verlag; 2003. p. 1-16
43. Klein, M. 2001. Combining and relating ontologies: an analysis of problems and solutions. In A. Gomez-Perez, M. Gruninger, H. Stuckenschmidt, and M. Uschold, editors, *Workshop on Ontologies and Information Sharing, IJCAI01*, Seattle, USA, 2001.
44. Köhler J, Munn K, Rüegg A, Skusa A, Smith B. 2006. Quality control for terms and definitions in ontologies and taxonomies, *BMC Bioinform*, 2006;7:212-220.
45. Kumar A, Smith B. 2003. The Unified Medical Language System and the Gene Ontology, *KI 2003*:;135-148.
46. Kumar A, Smith B. 2005. Oncology ontology in the NCI Thesaurus, *Artificial Intelligence in Medicine Europe (AIME)*, (Lecture Notes in Computer Science 3581), 2005;:213-220.
47. Kumar A, Smith B. 2006. The Ontology of processes and functions: a study of the international classification of functioning, disability and health. http://ontology.buffalo.edu/medo/ICF.pdf.
48. Lenat, D. B. 1995. Cyc: a large-scale investment in knowledge infrastructure. *Communications of the ACM* 38, no. 11.
49. Lewen, Holger; Kaustubh Supekar; Natalya F. Noy; and Mark A. Musen. 2006. TopicSpecific Trust and Open Rating Systems: An approach for ontology evaluation, *Workshop on Evaluation of Ontologies for the Web EON 2006*, WWW2006, May 22–26, 2006, Edinburgh, UK.
50. Linstone, H. A. and Turoff, M., Editors 2006. The delphi method: techniques and applications. http://www.is.njit.edu/pubs/delphibook/. Electronic reproduction of: Linstone,H.&Turoff,M.*The Delphi Method: Techniques and Applications*. Reading, Ma.: Addison-Wesley, 1975.
51. Mahesh, K., S. Nirenburg and S. Beale. 1996. KR requirements for natural language semantics: a critical evaluation of Cyc. *Proceedings of KR-96*.
52. Mani, I., Samuel, S., Concepcion, K., and Vogel, D. 2004. Automatically inducing ontologies from corpora. Proceedings of *CompuTerm 2004: 3rd International Workshop on Computational Terminology,* COLING'2004, Geneva.
53. Open Biomedical Ontologies (OBO) Foundry. http://obofoundry.org.
54. Obrst, L., H. Liu, R. Wray. 2003. Ontologies for corporate web applications. *Artificial Intelligence Magazine, special issue on Ontologies*, American Association for Artificial Intelligence, Chris Welty, ed., Fall, 2003, pp. 49-62.
55. Obrst, Leo; Todd Hughes; Steve Ray. 2006. Prospects and possibilities for ontology evaluation: the view from NCOR. *Workshop on Evaluation of Ontologies for the Web (EON2006)*, Edinburgh, UK, May 22, 2006.

56. Patel, C.; K. Supekar, Y. Lee, and E. Park. 2003. Ontokhoj: a semantic web portal for ontology searching, ranking and classification. In *Proc. of the Fifth ACM Workshop on Web Information and Data Management,* pages 58–61, New Orleans, Louisiana, USA, 2003. ACM Press.
57. Pisanelli D.M., Gangemi A, Steve G. An ontological analysis of the UMLS Methathesaurus. Proceedings of *AMIA Symp* 1998:810-4.
58. Polcicová, G., and Návrat, P. 2002. Semantic similarity in content-based filtering: In *Proc. of ADBI2002 Advances in Databases and Information Systems,* Manolopoulos, Y. and Návrat, P. (Eds.), Springer LNCS 2435, 2002, 80-85.
59. Rogers, J.E., and Rector, A. L. GALEN's Model of parts and wholes: experience and comparisons *Annual Fall Symposium of American Medical Informatics Association*, Los Angeles CA Hanley & Belfus Inc Philadelphia PA;:714-8
60. Sanguino, R. Evaluation of Cyc. 2001. *LEF Grant Report, CSC*, Miami, FL, March 2001, http://www2.csc.com/lef/programs/grants/finalpapers/sanguino_eval_cyc.pdf.
61. Seipel, Dietmar; Joachim Baumeister. Declarative methods for the evaluation of ontologies. University of Wurzburg. 2004.
62. Shannon, C.E. 1948. A mathematical theory of communication, *Bell System Technical Journal*, vol. 27, pp. 379-423, 623-656, July, October, 1948
63. Sider, T. 2002. *Four-Dimensionalism: An Ontology of Persistence and Time*. Oxford: Oxford University Press.
64 Siegel, S. and N. J. Castellan, N. J. 1988. *Nonparametric Statistics for the Behavioural Sciences*. McGraw-Hill, 2nd edition, 1988.
65. Simons, P. 1987. *Parts: A Study in Ontology*. Clarendon Press, Oxford, 1987.
66. Smith B, Ceusters W, Temmerman R. 2005. Wüsteria, *Stud Health Technol Inform.* 2005;116:647-652.
67. Smith B, Ceusters W. 2006. HL7 RIM: An incoherent standard, *Stud Health Technol Inform,* 2006, in press.
68. Smith B, Kumar A. 2004. On controlled vocabularies in bioinformatics: a case study in the Gene Ontology, *BIOSILICO: Drug Discovery Today*, 2004;2:246-252.
69. Smith B, Williams J, Schulze-Kremer S. 2003. The ontology of the Gene Ontology, *Proc AMIA Symp*. 2003;:609-613.
70. Smith B. 2006. Against idiosyncracy in ontology develoment, under review.
71 Smith B. 2006a. From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies, *J Biomed Inform*, 2006; 39(3): 288-298.
72. Smith, Barry; Jacob Köhler; Anand Kumar. 2004. On the application of formal principles to life science data: a case study in the Gene Ontology. Erhard Rahm (Ed.): *Data Integration in the Life Sciences, First International Workshop, DILS 2004*, Leipzig, Germany, March 25-26, 2004, Proceedings, pp. 79-94. Lecture Notes in Computer Science 2994 Springer 2004.
73. Systematized Nomenclature of Medicine (SNOMED) Clinical Terms (CT). http://www.snomed.org/.

74. Van Buggenhout C, Ceusters W. A novel view on information content of concepts in a large ontology and a view on the structure and the quality of the ontology. *International Journal of Medical Informatics* 2005;74(2-4):125-32.
75. Varzi, A.C. 1998. Basic problems of mereotopology. In: Guarino, N., ed. *Formal Ontology in Information Systems,* Amsterdam: IOS Press, pp. 29-38.
76. Wolstencroft, K., R. McEntire, R., Stevens, R., Tabernero, L. and Brass, A. 2005. Constructing ontology-driven protein family databases. *Bioinformatics* 2005 21(8):1685-1692
77. Wroe CJ, Stevens R, Goble CA, Ashburner M. 2000. A methodology to migrate the Gene Ontology to a description logic environment using DAML+OIL. *Pac Symp Biocomput* 2003:624-35.
78. Extended Metadata Registry (XMDR). http://xmdr.org/.
79. Zhang, S., Bodenreider, O., Golbreich, C. Experience in reasoning with the Foundational Model of Anatomy in OWL DL. Pacific *Symposium on Biocomputing 2006*: World Scientific; 2006. p. 200-211.
80. Zhong J, Zhu H, Li J, Yu Y. 2005. Conceptual graph matching for semantic search. In Priss U, Corbett D, Angelova G (eds.) *Conceptual Structures: Integration and Interfaces (ICCS2002)*, 2002, 92-106.