NISTIR 7774

NIST Workshop on Ontology Evaluation

Ram D. Sriram
Conrad Bock
Fabian Neuhaus
Evan Wallace
Mary Brady
Software and Systems Division
Information Technology Laboratory

Mark A. Musen
Stanford University

Joanne S. Luciano
The MITRE Corporation

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 7774

NIST Workshop on Ontology Evaluation

Ram D. Sriram
Conrad Bock
Fabian Neuhaus
Evan Wallace
Mary Brady
Software and Systems Division
Information Technology Laboratory

Mark A. Musen
Stanford University

Joanne S. Luciano
The MITRE Corporation

October 2011

U.S. Department of Commerce
John E. Bryson, Secretary
National Institute of Standards and Technology
Patrick D. Gallagher, Under Secretary of Commerce for Standards and Technology and Director

Meeting Report: "NIST Workshop on Ontology Evaluation"

Ram D. Sriram[1], Mark A. Musen[2], and Joanne S. Luciano[3], Conrad Bock[1], Fabian Neuhaus[1], Evan Wallace[1], and Mary Brady[1] (editors)

[1] National Institute of Standards and Technology
[2] Stanford University
[3] The MITRE Corporation (currently with Rensselaer Polytechnic Institute)

Speakers: Vinay Chaudhri, Mark A. Musen, Natasha Noy, Barry Smith, Robert Stevens, Michael Uschold, Chris Welty

Abstract

The National Institute for Standards and Technology sponsored a workshop in October, 2007, on the subject of ontology evaluation. An international group of invited experts met for two days to discuss problems in measuring ontology quality. The workshop highlighted several divisions among ontology developers regarding approaches to ontology evaluation. These divisions were generally reflective of the opinions of the participants. However, the workshop documented a paucity of empirical evidence in support of any particular position. Given the importance of ontologies to every knowledge-intensive human activity, there is an urgent need for research to develop an empirically derived knowledge base of best practices in ontology engineering and methods for assuring ontology quality over time. This is a report of the workshop discussion and brainstorming by the participants about what such a research program might look like.

Introduction

Over the last decade, the World Wide Web has become an essential tool in research and development in all areas of science. However, as the Web has grown, finding, integrating, and using the right information has become a critical problem hindering research. The most often stated solution is to provide "meaning" to the information, creating what is termed the Semantic Web [1][2]. Dr. Elias Zerhouni, former Director of the United States National Institutes of Health, said in an October 2003 *Science* article that, "the research community needs wide access to … scientific resources that are more sensitive, more robust, and more easily adaptable to researchers' individual needs." The Semantic Web will provide the technological tools that add semantics to information and that allow information to be shared, reused, and integrated across application and community boundaries. We believe that ontologies will play a crucial role in the realization of the Semantic Web.

Ontologies are becoming increasingly relied upon as the preferred approach for data aggregation and integration, semantic search, and knowledge representation in a broad range of disciplines. However, while a main motivation for the use of ontologies has been to enable knowledge reuse—in large part stemming from a community derived common set of vocabulary terms and their inter-relationships—not much attention has been given to providing users with methods or metrics that enable them to assess the utility of extant ontologies or even ontologies they develop themselves for particular purposes. There are several barriers to the effective utilization of

ontologies: lack of a systematic method for evaluating ontologies, inadequate techniques for verification and validation, lack of standard methods for comparing ontologies, and paucity of real-world applications demonstrating effectiveness of ontologies.

To address the issues above, a workshop was held at the National Institute of Standards and Technology on October 26th and 27th, 2007 to generate a research plan for the development of systematic methods for evaluating ontologies. The co-chairs of the workshop were Ram D. Sriram (National Institute of Standards and Technology), Mark A. Musen (Stanford University), and Carol A. Bean (National Institutes of Health). The topics for the workshop included the following:

- *Representation.* The language in which an ontology is expressed (its meta-language) should be used according to its intended syntax and semantics, to ensure that the ontology is properly understood by the user community and by computer-based tools. This topic addresses how to check that an ontology is using its meta-language properly.
- *Accuracy.* A well-constructed ontology is not very useful if its content is not accurate. This topic concerns methods to ensure that an ontology reflects the latest domain knowledge.
- *Reasoners.* An ontology can support automatic computation of the knowledge that is otherwise not obvious in the ontology. This topic addresses how to determine that automatically deduced information is consistent and valid.
- *Performance metrics*. Reasoners and other computational services are not very useful if they consume too many resources, including compute time. This topic concerns the bounds that users should expect from various kinds of computational services.
- *Tools and Testbeds.* Ontology evaluation is a complex task that can be facilitated by testing environments, graphical tools, and automation of some aspects of evaluation. This topic addresses computer-aided ontology evaluation.
- *Certification.* Ontologies that pass rigorous evaluation should be recognized by the community, to encourage the development and adoption of those of higher quality. This topic concerns the methods for official recognition of ontologies meeting high standards. Of particular concern is the role of social engineering to develop practices and tools that support the routine assessment and review of ontologies by the people who use them.

The workshop had several presentations and breakout sessions. This report summarizes these presentations and breakout sessions. In our report of the discussions following each presentation, we use the abbreviation *AM* to connote an audience member, unless otherwise specified. Additional resources related to the workshop, including slides from each of the presentations are available at http://sites.google.com/a/cme.nist.gov/workshop-on-ontology-evaluation/Home/.

Presentation Summaries

Summaries of presentations, except Michael Uschold's talk entitled "Evaluating Ontologies based on Requirements," are provided below.

1. Use Cases for Ontologies - Mark Musen, Stanford University

Musen discussed why the very basic question "What is an ontology?" is difficult to answer. He argued that the answer depends on the perspective of the ontology developer. He reviewed four common views of ontologies:

- Ontologies are models of reality
  - The Foundational Model of Anatomy is one example
- Ontologies are models of information
  - This notion, often called Information Model, is not clearly defined
  - HL-7 Reference Information Model (RIM) purports to describe all of the information necessary to model the clinical enterprise [3]. However, the HL-7 RIM is a good example of an information model that isn't a model of any reality.
- Ontologies are compendia of controlled terms
  - The emphasis is on the terms and relationships that these models may include, where relationships are designed more as a convenient way of making sure that terms become accessible and useful when they are retrieved.
  - The Gene Ontology (GO) was developed originally as a set of related terms to bridge databases of model-organism genomics [4].
- Ontologies are software
  - The issue that is probably most problematic for purposes of evaluation is the notion that ontologies are software.
  - This view becomes an important notion as the Object Management Group pushes forth its Model Driven Architecture (MDA) platforms [5]. The MDA approach incorporates models that are reusable across software products, providing standard descriptions of domain entities that are going to be used ultimately to drive software engineering. Musen argued that the notion of ontology as applied within MDA is going to be very confusing to traditional ontology engineers because, in this setting, the ontology is meant to be software.

An ontology developer often had one of these uses in mind. One of the problems of adapting ontologies for new uses is that we don't necessarily know which of these views drove the development of an ontology and thus what assumptions the developers made.

Musen then discussed differences in scope and purpose that make certain reuses problematic, citing examples of the GALEN ontology [6] versus the Foundational Model of Anatomy [7], and CYC for general Web search. Finally, the following were listed as implications for ontology evaluation:

- Ontologies are built with particular purposes in mind
- Attempts to co-opt ontologies for new purposes often are not successful
- Ontologies built for one purpose may be inadequate in other contexts
- Ontologies need to be evaluated within their context of use, but that context is often at best implicit

2. Dimensions of Ontology Evaluation - Robert Stevens, University of Manchester

Stevens started with the following opening remarks: "When asked at a meeting last year in Manchester about major issues in building ontologies, two things that no one mentioned at all were encoding the ontology and evaluation. Perhaps one of the reasons for this lack of interest in evaluating the ontology is that producing the ontology is so difficult that the last thing they want

to find out is that it wasn't any good. Or perhaps that just by using the ontology, they were evaluating it."

Stevens then presented a number of issues that make ontology evaluation difficult.

Evaluating What? Evaluation must ask a question. When evaluating software, we need to pin down a question on what we are going to evaluate. What are the criteria for evaluation?

Dimensions of Evaluation. Example dimensions might include:

- Does the ontology conform to some notion of reality?
- Does it conform to ontological principles? Does it conform to philosophical principles?
- Does it have community buy-in? Does it agree with the consensus about the understanding of the domain that my community has? If not, is the community wrong or is my ontology wrong?
- Is it logically consistent?
- Can people understand the ontology (is it comprehensible)?

In the Small and in the Large. Stevens went over questions of scope to be asked when evaluating an ontology:

- Completeness of coverage (large)
- Can I ask the questions I need to ask (large)
- Can I not ask the questions that I should be able to ask (large)
- Is each statement in the ontology correct (small)? If so, is the whole ontology correct (large)
- Does the logical definition match the textual definition (small)

Methods

- Do domain and/or ontological experts agree with the ontology?
- What does the community think of the ontology?
- Can one use tools, such as Web Ontology Language Description Logic (OWL DL) classifiers, for testing consistency [9]? Can one use text resources to support looking at coverage, comparing the ontology with text corpora? Can one use query tools to help ask questions?

Evidence (and assorted ideas)

- Metrics - difficult, many are qualitative, counting as "hard" evidence
- Expert opinion to provide peer review
- Results of application of tools
- Does the ontology allow for interoperation?
- Does it drive my application?
- Is it re-usable?

Discussion

Stevens wanted to know if there were any glaring omissions in the dimensions that he presented or whether he missed something.

AM: There are a couple of kinds of processes here. One is a kind of quality-control process like OntoClean [10] and another is a construction process. There you might distinguish between individually built ontologies versus community curated ones (ones that had an open commentary process versus those that had a closed body of experts).
Chris Welty: I enumerate some dimensions in my talk. I think you covered most of them, but let me say what they are: coverage, richness, correctness, commitment, and organization. Coverage is how much of the domain your ontology captures and richness is how dense the detail is modeled (simple taxonomy *versus* relationships among things, from Alan Rector). Richness is not necessarily a good thing. Commitment is what things in reality does your ontology commit to the existence of. Additionally, organization and modularity should be considered. Organization is how the ontology organizes elements in the domain, and modularity (from Nicola Guarino) involves making meaning clear (making crisp distinctions).
A discussion about comparability of consistent meaning followed. For example, discussion about metrics for comparing two ontologies came up. It was also pointed out that the evaluation should be done within a context. There were discussions on the boundaries of ontologies, as they should be crisp and unambiguous, allowing people to share information. Ontology designers should be able to answer the questions: "who cares?" and "how does this matter?"

3. Ontology Engineering (It's About the Instances) - Chris Welty, IBM Research

Welty began by stating: "Ontologies are the primary component of knowledge-based systems. You cannot build good knowledge-based systems without good knowledge. Software must be anchored in basic concepts of the domain, for example, customers, products, etc."

Ontologies are not about the classes[1], they are about things that are classified (instances). A superclass does not describe its subclasses, it describes common aspects of all the instances of its subclasses. Identity criteria are for telling instances apart (boundary conditions), telling when two things are actually one thing, or telling when one thing cannot be two. For example, a class of diseases is related to classes of people who have those diseases, but each instance of a disease occurs in a single individual person. Leaf nodes of a class hierarchy are not instances.

Common pitfalls in ontology building include errors in the following areas:

- Composition (e.g., Arm is written as a *subclassof* Body, instead of *partOf*)
- Constitution (e.g., Statue *subclass* Marble, which should be really Statue is *madeOf* Marble)
- Disjunction (e.g., V8-engine is written as a *subclassof* CarPart, but some engines are used in boats)
- Polysemy (e.g., Book *subclass* PhysicalObject, when the meaning refers to the abstract narrative)
- Purely organizational nodes (e.g., FictionalBookbyLatinAmericanAuthor *subClass* FictionalBook, when reasoners should deduce these kinds of relationships)
- Instantiation (e.g., PinotNoir *instanceOf* Grape, when it should be a subclass),
- Temporality (e.g., YoungElvis *instanceOf* Elvis).

[1] Please refer to the work on OntoClean [10].

Linguistic tests enable saying things about instances in ways that make sense, for example:

- If P *subclass* Q, you should be able to say "P is a kind of Q" (CB: or "P's are Q's")
- If a *instanceof* P, you should be able to say "a is a P"
- If a *instanceof* P *subClassof* Q, you should be able to say "a is a Q"
- For every instance, there should be a class of which it is always (rigidly) an instance that is its natural label
- You should *not* find it natural to say, if P *subclassof* Q, "P has Q", "P might be Q", "P was Q"

There needs to be a distinction made between ontology and classification. Ontology is supposed to be about objective reality, whereas classifications are entirely subjective. For example, the Dewey decimal system – a classification system - is designed just to find a place for the book. It should not be confused with an ontology. (AM: Even biological classifications are subjective.) Other comments made by Welty:

- It is important to distinguish backbone types that are rigid and unchangeable from changeable types. For example, Book is rigid, whereas whether it's an action novel or a romance novel is subjective.

- Ontologies by nature are not suited to automatic evaluation.

- Other criteria include having example instances for every class. Developers should check that each instance is also an instance of all the superclasses.

- Automated metrics are useful, but people tend to abuse them. There are many automated metrics that are based on linguistics. Evaluators should avoid purely name-based techniques, such as evaluating the names in the ontology against other names.

4. Dynamic Evaluation of Ontologies - Vinay Chaudhri, SRI

This talk focused on the lessons learned during the development of an Intelligent Personal Assistant – CALO (Cognitive Assistant that Learns and Organizes) [11]. CALO helps a user to manage and organize information, in addition to helping the user perform various tasks. CALO has a component that learns the user's ontology (i.e., user's model of his or her work) for executing tasks. CALO is a major software effort involving more than 20 groups.

Building the knowledge-base: The knowledge-base development followed a software engineering process, which involved the following stages: 1) requirements gathering, 2) knowledge reuse, 3) knowledge extension, 4) implementation, 5) evaluation, and 6) refinement. For requirements gathering, the various information objects, tasks, properties, and other information needed to build the system were obtained by the developers and users. An extensive code analysis was performed to find out the ontologies that were needed. In the knowledge reuse stage, two issues had to be addressed: 1) picking a knowledge-representation language, and 2) selecting various ontologies. The developers of CALO used Knowledge Machine (for object-oriented representation of tasks, meetings, and entities) [12], SPARK Process Modeling Language (for detailed representation of executable procedures) [13], Weighted Max-Sat Rules (for learned knowledge) [14], and

OWL/Resource Description Framework (RDF) (for user's world) [15]. The knowledge-base had about 1000 classes and 1000 relations.

In the CALO project, system builders:

- engineered a large knowledge-based for a personal office assistant
- used existing ontologies, and extended them to model meetings
- leveraged state-of-the-art Semantic Web standards and tools
- deployed in a system with over 100 modules written in seven different languages

5. Evaluation through an Editorial Process: Using Editors who have Interiorized Ontoclean - Barry Smith, SUNY Buffalo

Peer reviews are widely used by editors of scientific journals and funding agencies to ensure scientific quality. This method of evaluation is applicable in any given scientific domain provided that: (1) the scientific community agrees about the criteria for good science in their domain; (2) the peer reviewers adhere to these criteria; and (3) the peer reviewers are trusted by the community.

Science-based ontologies are semantically structured controlled vocabularies that are developed by open scientific communities with the goal of improving the interoperability and accessibility of scientific data. Although scientific ontologies represent their content in a very different way from scientific journals, ontologies represent scientific knowledge and thus can be subjected to the same editorial processes as journal articles or textbooks. For this reason the Open Biomedical Ontologies (OBO) Foundry decided to evaluate ontologies by peer reviews [16][17].

The OBO Foundry is a collaborative effort with the goal of developing a set of interoperable, humanly-validated reference ontologies for all major domains of biomedical research. It involves a group of ontology developers who have agreed in advance to the adoption of a growing set of principles specifying best practices in ontology development. These principles are designed to ensure three key features: interoperability, openness, and orthogonality.

Authors who want their ontology to be part of the OBO Foundry submit their ontologies to the OBO Foundry board. The ontologies are evaluated by the community; this process usually involves discussions in publicly accessible email forums. The members of the OBO Foundry Board decide whether a given ontology adheres to the Foundry principles, and adjudicate in areas of overlap between ontologies.

The primary advantages of this evaluation process are:

- This evaluation method has been proven successfully. Currently, there are 17 ontologies in the OBO Foundry that cover a vast range of biomedical topics.
- The evaluation process involves a large scientific community, which contributes to the widespread use of the ontology.
- The community-driven evaluation process encourages cooperation between members of a community with similar interests.
- The orthogonality principle allows a division of labor. As a result, the author of a single ontology can decide almost all design decisions about his or her ontology.

The key concerns raised by attendees at the workshop included:

- The review process relies to a large degree on human experts and, thus, is expensive.
- The evaluation process does not result in a quantifiable result.
- By making the evaluation process open, a key benefit of double-blind reviews is lost. The behavior of people in publicly accessible forums is influenced by their social status, personal relations, and political considerations; this might have an impact on the evaluation.
- The evaluation is a one-time process that determines whether the ontology is added to the list of OBO Foundry ontologies. However, ontologies are updated constantly and there seems to be no process in place to ensure the quality of the ontologies over time.
- Since the evaluation process does not happen in regular intervals and does not lead to quantifiable results, this evaluation method does not provide answers to questions such as "How much did my ontology improve in the last three months?" Thus it is not useful for the ongoing management of ontologies.

Discussion

AM: What are your plans for validation?
BS: This will evolve over time, and may become more algorithmic. Publications in journals and other people using it may provide empirical validation.
AM: The validation process should itself be formally validated. Saying that we are better than other groups is not good enough.
AM: Is there some kind of metric that we can use to find out if my ontology is better today than yesterday, assuming I made some changes to the ontologies based on your principles?
Several people chimed in here and suggested that it looks as if validation is subject to human interpretation.
AM: Do journals accept any of the current ontologies for validation?
BS: Not right now, but I believe that in the future OBO will play this role.

6. Ontology Evaluation through Community Contributions (Ontologies and Web 2.0) - Natasha Noy, Stanford University

Natasha Noy started her talk with the following statement: "This work looks at ontology evaluation from the perspective of those who use them, compared to those who develop them. You need people to make these evaluations, rather than machines. In particular, you need a community of users."

The need for a community-based evaluation is based on several premises:

1. *Not all criteria are measurable*. How do you answer the question, "Is this ontology good for my task?" We will need input from other users doing similar tasks, as well as from the authors of the ontology.
2. *There is no single perfect ontology in a domain*. A "good" feature in one setting can be a "bad" feature in another setting. For example, units are different in the United States and Burma than in other countries. Abbreviations are not the same in all languages.

Axiomatization might be good or bad depending on the level of precision that is needed. Even if the best ontology existed, we would not agree on what it is.

3. *Some part of ontology design is closer to art than science*, even for scientific ontologies. Different points of view must be accounted for. Different contexts require ontologies to place different emphasis on different aspects of the same reality.

We can have closed or open rating systems on the Web. In a closed system, only a group of "editors" can provide ratings (e.g., the Open Directory Project [18]), whereas in an open system, anyone can publish reviews and ratings (e.g., Amazon and Epinions). We can have a similar strategy for ontologies. We can have usage information, such as applications that have successfully used the ontology, and a list of problems encountered. We can answer coverage questions such as: Does the ontology cover the domain properly? Are there major gaps? Are some parts developed better than others?

Discussion

Community-Based Popularity metrics (e.g., Amazon rating type evaluations). *Motivation* for participation by the community was thought to be: egotistical (i.e., recognition by the community) and altruistic (i.e., helping others within the community make better decisions). An ontology in a specific application area should motivate users to comment on their experience.
Concerns:
Applicability: The approach is not always applicable because ontologies change; books do not. Thus, reviews would become outdated when problems are repaired. Furthermore, aggregation of negative reviews would not add much value. The technique also depends on a large audience of potential reviewers, which may not be the case for many ontologies.
Quality: Review quality at Amazon is mixed.
Cost: Good reviews are very expensive to produce.
Participation: The approach will be successful only initially, as people will lose motivation if they did not get the personal return on their investment.
Empirical evidence that community-based evaluation works has not been established. There only is evidence that Web-based reviews are currently in vogue.
Requirements: A critical mass (sufficient number of users) for open review of ontologies to work.

7. Moving Toward the Next Generation of Information Standards – Steve Ray, NIST

Day 1 ended with Steve Ray's talk. After giving an overview of NIST's mission and the role of measurements in industrial practice, Steve Ray indicated his expectations of the workshop's outcome.

During the second day of the workshop, the participants were divided up into several break-out groups with the goal of proposing specific strategies for ontology evaluation. We present the summaries of these discussion as provided by the different groups.

Breakout Session Summaries

Breakout 1: Problems with Identifying Ontology Defects

Group Members: Mark Musen, Ram Sriram, Robert Stevens, and Joanne Luciano

The group started with a hypothesis that "people are not effective and are not efficient at identifying ontology defects by inspection." The following taxonomy of errors was proposed:

- Ontological/Modeling errors identified by an ontological engineer
- Domain content errors (e.g., every cell has a nucleus)
- Usability errors: The users of ontology make errors in applying the ontology
- Missing information
    a. Inappropriate subsumption
    b. Missing axioms
    c. Disjointness information
    d. Straightforward errors: Things classified in the wrong sub-branch horizontally or vertically
- Scope errors: The ontology should not contain information on topic X
- Style errors
    a. Inconsistent naming conventions
    b. Same name for a class and for a property, distinguished only by case
- Circular references in the ontology

Techniques for measuring the effectiveness of peer review of ontologies were outlined, as follows:

- Take a well known ontology in a domain. Insert errors in it, and measure how accurately and quickly the users are able to identify the errors.
    a. Different classes of users
    b. Different classes of errors
- Are there candidate ontologies on which this way of reviewing can be tried out?
- Motivation
    a. Why would users do it?
    b. How can they be enticed?
- When we ask for a review, what exactly are we asking users to do?
    a. Are they reviewing term definitions?
    b. Class–subclass relationships?
    c. Granularity of modeling?
    d. Whether the ontological distinctions are made correctly?
- Can the users suggest changes to the ontology?
- What are the false positive and false negative rates for users' suggested changes?

One suggestion was to develop an ontology usage experiment, where a set of users apply the same ontology to a specific problem (e.g., annotating a research article). The quality of an ontology can be determined by measuring the number of steps one has to go through in order to get to the term one is looking for (location metric), and whether the same term gets used for more than one task (multiplicity metric).

Breakout 2: Requirements-based Ontology Evaluation

Group members: Jim Brinkley, Leo Obrst, Barry Smith, Evan Wallace

Normally, the developers of an ontology have a specific use for the ontology in mind. This use generates questions such as: What are the types of objects which must be included in the coverage of the ontology? What kinds of reasoning should the ontology support?

Ontologies ought to be evaluated on the basis of the degree to which they answer questions such as these (which reflect dimensions of evaluation):

- *Bottom up*: Does the ontology extend to cover all the data that we have?
- *Top down*: Does the ontology draw on standard resources, for example to support temporal reasoning? Does it respect best practices (as determined by static evaluation)?

Evaluation of ontologies versus evaluation of systems using ontologies

Typically, an ontology is part of a larger software system built to achieve a specific purpose. This situation creates the quandary in that it can be difficult to know if the source of a problem is the ontology or some other part of the system.

Black box evaluation: How well does the system work as a whole?

Study: Using TREC[2] Natural Language Processing (NLP) data, evaluate ontology-based technology against other technologies (e.g., statistical pattern recognition). What kinds of benefits are seen when ontologies are used?

Glass box evaluation: How does the ontology help a given system to achieve its purpose? How well does it work within the system?

Experiment: Build a system (e.g., for NLP purposes), with a plug-and-play facility for hot-swapping of ontologies, and compare the results. For instance, one might use the GALEN anatomy ontology and test the system [6], and hot swap in the FMA anatomy ontology and test the system [7], and then hot swap the NCI [19] or SNOMED [20] anatomy ontologies, and compare the results. Assessment is non-trivial.

Evaluation in terms of costs and benefits

How do we evaluate the cost of ontology development, and how do we estimate the economic benefit? How can we track and measure such benefits?

Experiment: How do we classify and measure the types of economic benefits ontology technology can bring?

Describe existing success stories of ontologies and ontology-like things—that is, widely used schemes for referring to things or concepts (e.g., DUNS Numbers[3], UNSPSC[4], MeSH [21], GO [4]).

---

[2] Text REtrievel Conference, see http://trec.nist.gov.

[3] A Data Universal Numbering System (DUNS) number is a unique identification number provided by Dun and Bradstreet, see http://smallbusiness.dnb.com.

[4] United Nations Standard Products and Services Code (UNSPC) is a global classification hierarchy owned by the United Nations Development Programme, see http://www.unspsc.org.

Classification of costs and benefits

Costs: ontology development, developer training, user lobbying, user training, help desk, user support, feedback, dissemination, maintenance, etc.

Benefits: effectiveness + user satisfaction (e.g. in search and retrieval), decreased development time, decreased level of software maintenance, increased interoperability, increased re-usability, and increased leverage for single user.

Experiment to test re-usability

Develop an ontology for one purpose (e.g. search), and attempt to apply it to a range of different purposes (e.g. data integration) employing common information systems, keeping track of:

1. How much needs to be changed and how much remains constant to serve the new purpose?
2. Effectiveness for the new purpose.
3. What added value is brought by use of common upper level/reference ontologies?

Life-cycle-based evaluation

Can existing methods for evaluating software artifacts be applied to ontologies?

Experiment:

- Survey standards for best practices in development, deployment and maintenance which already exist in other domains (e.g. object oriented software) and which might be applied to ontology.
- Measure the degree to which given ontology efforts are meeting these standards, and infer recommendations as to how they can be improved.
- Against this background, summarize the state of the art of ontology evaluation. How well do we understand the relevant domain of evaluation itself? Study the papers submitted to/at this meeting.

Meta-study

As an initial step towards mapping the space: create a three or four column table covering the following categories: types of ontology, use cases, evaluation approaches and possibly experiments. Some fillers for this would be:

- *Types of ontology*: representations of reality, models of information, controlled vocabularies, software
- *Use cases*: search and retrieval, query-answering, question answering, annotation, data integration, system interoperability, semantic web services, natural language processing, decision support, training, enhanced software development
- *Evaluation*: requirements-based, static, dynamic, process-based, editorial-based, community-based, cost-benefit analysis

Breakout Summary 3: $10 Million Dollar Question (Framework for Ontology Measurement)

Group members: Conrad Bock, Xenia Fiorentini, Michael Gruninger, Natasha Noy, Mike Uschold, Chris Welty

Goal: Define a program to spend 10 million dollars on ontology evaluation.
The result would be:

1) A framework for performing controlled experiments regarding ontology-evaluation metrics and methods.
2) A set of benchmark ontologies evaluated from various domains, evaluated within this framework.

The framework would involve at least three dimensions:

1) Alternative application tasks (for example, search, data/system, integration, decision support, data modeling, in silico, scientific).
2) Alternative methods of ontology evaluation.
3) Alternative domains, such as biomedicine or engineering.

Experiments might determine where tasks are independent of the other dimensions (or dependent on them)—for example, tasks that are independent of domain, metrics that are task specific, biases of methods for particular ontology languages.

Experiments might be conducted from at least two viewpoints. In both cases, evaluation criteria might be combined, with tradeoffs, as follows:

1) For ontology developers, providing critiques or suggestions to improve a particular ontology.
2) For ontology users, providing relative ranking of candidate ontologies for particular tasks and domains.

Breakout Summary 4: Ontology and Ontology Process Measurement

Group members: Larry Hunter, Onard Mejino, Fabian Neuhaus, Peter Good, Cliff Joslyn

The breakout group focused on the evaluation of science-based ontologies. It became quickly obvious that the evaluation of an ontology as a knowledge artifact needs to be part of a bigger evaluation that covers the whole management and community process in which it is embedded. (For example, for a potential user, it is important to know whether an ontology is only available in a proprietary language, whether it is widely adopted, or whether the authors no longer maintain the ontology.)

We distinguish among four groups of evaluation dimensions. Each of these dimensions is formulated below as a question. This step would be to develop measurements that enable quantification in answering these questions.

1. Content-related dimensions of evaluation

An ontology is a representation of some knowledge in a particular domain. The following evaluation dimensions are concerned with the content of an ontology:

- Is the content of the ontology true or does it contain errors?
    - What is the distribution of the errors?
    - What kind of errors does it contain?
- Does the ontology contain inconsistently defined terms?
- Is the ontology well-structured?
- How complete does the ontology cover the domain it is supposed to cover?
- Do users of the ontology independently choose the same terms for the same purpose?

2. Platform-related dimensions of evaluation

While discussing "content-related dimensions of evaluation," we considered an ontology as an abstract representation of knowledge. When discussing the platform-related dimensions, we consider an ontology to be a piece of software. Ontologies are always expressed in a given language and embedded in an environment of editing and visualization tools. We considered questions such as:

- How expressive is the language that is used?
- Is the content represented in a standard format?
- Does the platform support inferences?
- Does the platform support ontology maintenance (e.g., version control)?
- Does the platform support visualization?
- Does the platform support reporting (e.g., some statistics and quality metrics)?

3. Applicability and Utility dimensions of evaluation

Another group of evaluation dimension is concerned with the use of ontologies. For any potential use case of an ontology one can ask:

- Is the ontology widely in use?
- Is the ontology used for different types of applications?
- Has the ontology been used for tasks such as:
    - Annotating data
    - Integrating data
    - Reasoning (abductive, deductive, inductive)
    - Natural-language processing
    - Describing Web services (ontological signatures)
    - Describing scientific experiments

4. Management and editorial process dimensions of evaluation

The value of an ontology for a given project can be strongly influenced by the way the ontology is managed.

- Is the ontology well-documented?
- Is the ontology publicly available or is the content proprietary?
- Are the authors of the ontology responsive to comments and suggestions?
- Is the group that develops an ontology open for participation?
- Is a good versioning system in place?

Final Outcome and Conclusions

The NIST Workshop on Ontology Evaluation highlighted both the importance that workers in ontology engineering place on the problem of evaluation and the extreme lack of consensus in how to approach it. It was clear that, apart from static inspection of ontologies, there is no shared procedure in place that has been adopted by the ontology-development community broadly. This is not to say any particular approach is right or wrong, but there is no overarching framework into which the approaches might relate to each other. At the same time, the workshop participants acknowledged that static inspection does not allow one to quantify ontology quality, is rather subjective, and lacks formal study of inter-rater reliability.

The presentations at the workshop clarified a number of important dichotomies in ontology evaluation: human review *versus* the use of computational metrics; central editorial control *versus* open peer review; viewing ontologies with philosophical rigor *versus* viewing ontologies as software artifacts. The workshop participants agreed that the manner in which ontology developers construe their work necessarily determines how ontologies should be evaluated, and yet the absence of formal studies of alternative evaluation methods prevents the community from moving toward established methods for assuring ontology quality.

A recurring theme at the workshop was that ontologies are constantly evolving and that any evaluation performed at a particular point in time is not going to carry much meaning for future versions of an ontology as developers refine their work. All evaluation methods, including central editorial control, community-based peer review, and inspection of ontology content by either people or machines, must be part of an ongoing process of quality assurance to remain relevant as ontologies develop over time. This indicates a need for extensibility measures and metrics.

The breakout sessions at the workshop were helpful in laying out in broad strokes ontology-evaluation strategies that can be considered by the ontology community. At the same time, the sketchiness of these strategies and the lack of firm, empirical data on which to base them was troubling to the workshop attendees. Ontologies are playing a central role in e-commerce, industrial manufacturing, e-science, and virtually every other knowledge-intensive human endeavor, and yet we do not have strong theories to help us know whether our ontologies are any good for the purposes to which we apply them.

The primary conclusion of the workshop is that evaluation of ontologies needs to be viewed as a first-class discipline in its own right. There is an urgent need for research that can lead to empirically validated methods for measuring ontology quality and in particular, fitness for purpose for use and reuse. Such methods offer the possibility to inform the evolution of ontologies as both our knowledge of the world and the tasks to which we apply our ontologies change over time.

Acknowledgments & Disclaimer

Meredith Keybl at MITRE provided assistance with editing the manuscript. Financial support was provided by the NIST's "Standards for Bioimaging" project.

Certain commercial software systems are identified in this paper. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology (NIST); nor does it imply that the products identified are necessarily the best available for the purpose. Further, any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NIST or any other supporting US government or corporate organizations

Appendix: Workshop Participant List

Organizers:
Mark Musen, Stanford University, Ram D. Sriram, National Institute of Standards and Technology, and Carol Bean, National Institutes of Health (currently with the Office of the National Coordinator for Health Information Technology)

Scribes:
Fabian Neuhaus, Conrad Bock, Evan Wallace

Participants:
Mary Brady, National Institute of Standards and Technology
James Brinkley III, University of Washington, Seattle
Arthur Castle, NIDDK, National Institutes of Health
German Cavelier, NIMH, National Institutes of Health
Kevin Cohen, University of Colorado Health Sciences Center
Vinay Chaudhri, SRI, International, Inc.
Peter Good, National Institutes of Health
Michael Gruninger, University of Toronto
Larry Hunter, University of Colorado Health Sciences Center
Cliff Joslyn, Pacific Northwest National Laboratory
Vipul Kashyap, Partners Healthcare (currently with CIGNA)
Joanne Luciano, The MITRE Corporation (currently with Rensselaer Polytechnic Institute)
Onard (Jose) Mejino, University of Washington, Seattle
Natasha Noy, Stanford University
Leo Obrst, The MITRE Corporation
Steve Ray, MSID, National Institute of Standards and Technology
(currently an independent consultant)
Karen Skinner, National Institutes of Health
Barry Smith, State University of New York, Buffalo
Robert Stevens, University of Manchester, UK
Michael Uschold, The Boeing Company
Chris Welty, IBM Research

References

[1] Berners-Lee, T. and Hendler, J., "Scientific publishing on the semantic web," *Nature*, 410:1023-1024, 2001. World Wide Web Consortium, "Semantic Web," http://www.w3.org/standards/semanticweb, 2010.

[2] World Wide Web Consortium, "Semantic Web," http://www.w3.org/standards/semanticweb, 2010.

[3] Health Level 7 International, "HL7 Reference Information Model," http://www.hl7.org/implement/standards/rim.cfm, 2010.

[4] The Gene Ontology, "Gene Ontology Website," http://www.geneontology.org, 2010.

[5] Object Management Group, "Model-driven Architecture," http://www.omg.org/mda, 2010.

[6] OpenGALEN, "OpenGALEN Mission Statement," http://www.opengalen.org, 2010.

[7] University of Washington, "Foundational Model of Anatomy," http://sig.biostr.washington.edu/projects/fm, 2010.

[8] The Cyc Foundation, "Computable Common Sense," http://www.cycfoundation.org, 2010.

[9] World Wide Web Consortium, "OWL 2 Web Ontology Language Document Overview," http://www.w3.org/TR/owl2-overview, 2009.

[10] Laboratory for Applied Ontology, "OntoClean Central," http://www.ontoclean.org, 2010.

[11] Stanford University, "CALO Explanation Project," http://www.ksl.stanford.edu/projects/CALO, 2009.

[12] University of Texas, "KM: The Knowledge Machine," http://www.cs.utexas.edu/~mfkb/km, 2006.

[13] Morley, D., Myers, K., "The SPARK Agent Framework," in Proceedings of the International Conference on Autonomous Agents and Multi-agent Systems, 2004.

[14] Heras, F., Larrosa, J., Oliveras, A., "MINIMAXSAT: An Efficient Weighted Max-SAT Solver," Journal of Artificial Intelligence Research, 31:1-32, http://www.aaai.org/Papers/JAIR/Vol31/JAIR-3101.pdf, 2008.

[15] W3C, "Resource Description Framework (RDF): Concepts and Abstract Syntax," http://www.w3.org/TR/2004/REC-rdf-concepts-20040210, 2004.

[16] New York University at Buffalo, "The Open Biological and Biomedical Ontologies," http://www.obofoundry.org, 2010.

[17] Smith, B., et al, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," Nature Biotechnology, 25, pp. 1251 – 1255, http://www.nature.com/nbt/journal/v25/n11/full/nbt1346.html, 2007.

[18] Netscape, "Open Directory Project," http://www.dmoz.org, 2010.

[19] National Cancer Institute, "Terminology Resources", http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources, 2010

[20] International Health Terminology Standards Organization, "Systematized Nomenclature of Medicine-Clinical Terms," http://www.ihtsdo.org/snomed-ct, 2010.

[21] National Library of Medicine, "Medical Subject Headings," http://www.nlm.nih.gov/mesh, 2010.