Editor: Charles Petrie • *petrie@nrc.stanford.edu*

Possible Ontologies

How Reality Constrains the Development of Relevant Ontologies

Martin Hepp • *Digital Enterprise Research Institute, University of Innsbruck*

For about a decade, ontologies have been known in computer science as consensual models of domains of discourse, usually implemented as formal definitions of the relevant conceptual entities.[1] Researchers have written much about the potential benefits of using them, and most of us regard ontologies as central building blocks of the Semantic Web and other semantic systems. Unfortunately, the number and quality of actual, "non-toy" ontologies available on the Web today is remarkably low. This implies that the Semantic Web community has yet to build practically useful ontologies for a lot of relevant domains in order to make the Semantic Web a reality.

Theoretically minded advocates often assume that the lack of ontologies is because the "stupid business people haven't realized ontologies' enormous benefits." As a liberal market economist, I assume that humans can generally figure out what's best for their well-being, at least in the long run, and that they act accordingly. In other words, the fact that people haven't yet created as many useful ontologies as the ontology research community would like might indicate either unresolved technical limitations or the existence of sound rationales for why individuals refrain from building them — or both. Indeed, several social and technical difficulties exist that put a brake on developing and eventually constrain the space of possible ontologies.

Four Bottlenecks

We can classify ontology-related tasks into two main groups: *building* or contributing to the development of ontologies and *committing* to a particular ontology. Committing to a given ontology, explicitly or implicitly, means agreeing that it properly represents the domain's conceptual elements. An example of implicitly committing is to use an ontology to annotate your own data or to express a query using elements from it. Such ontological commitment can be based either on checking the specification (that is, verifying that the formal part and the documentation specify concepts in a way that's compatible with your subjective view of the domain, at least for the respective task) or trusting the creators (assuming, for example, that an ontology of countries as provided by the United Nations is politically correct). Charles Petrie goes even further in his opinion that ontology commitment can be achieved only by successful joint action — that is, successful usage of the ontology.[2]

Although numerous fine-grained methodologies exist for building ontologies,[3] most reflect best practices for settings in which the individuals have agreed to build a particular ontology (as part of an academic research project, for instance). They address only lightly issues such as legal constraints and future usage by individuals that weren't involved in building the ontology.

In a nutshell, current ontology-engineering practices insufficiently address at least five fundamental aspects of building and committing to ontologies:

1. *Ontology engineering lag versus conceptual dynamics.* Can we build ontologies fast enough to reflect quickly evolving domains?
2. *Resource consumption.* Does the gain in automation that the ontology provides justify the resources needed to develop it? From another perspective, do the technical problems that the ontology can help us solve outweigh the problems we must master to create it?
3. *Communication between creators and users.* Can the individuals who consider using an ontology to annotate data or express queries easily grasp the meaning of all the elements as intended by the ontology creators?

1089-7801/07/$25.00 © 2007 IEEE

4. *Incentive conflicts and network externalities.* Is the incentive structure for relevant actors in the process compatible with the required contributions? For example, are those who must dedicate time and resources benefiting from the ontologies? Moreover, ontologies exhibit *positive network effects*,[4] such that their perceived utility increases with the number of people who commit to them. This implies that convincing individuals to invest effort into building or using ontologies is particularly difficult while the user base associated with it is small or nonexistent.
5. *Intellectual property rights.* For many applications, we need ontologies that represent existing standards. However, standards are often subject to intellectual property rights.[5] Establishing the legal framework for deriving ontologies from relevant standards is thus nontrivial.

The fact that resource consumption and incentives for building ontologies are two sides of the same coin leads to four major bottlenecks that might explain today's shortage of actual ontologies on the Web.

Obstacle 1: Conceptual Dynamics

Most practically relevant domains include some degree of conceptual dynamics — new elements arise as some old ones become irrelevant. In the products and services domain, for instance, manufacturers are continually inventing new types of goods; in physics, scientists can discover new types of particles or relations among them; and in the geopolitical domain, new states form and political borders change. Philosophically, we could argue over whether all abstract concepts exist independently of time, but practically, dynamics among conceptual elements is relevant when building an ontology for a particular domain.

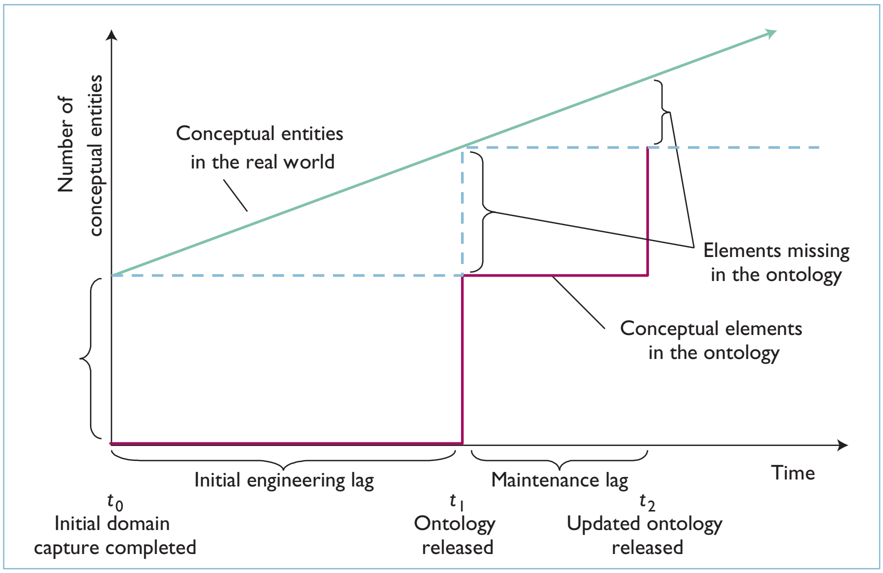

Figure 1. Conceptual dynamics and ontology content and coverage. During the ontology-building process, new conceptual elements become relevant in the domain of discourse, which aren't included in the initial domain capture.

To date, this is a widely ignored fact. Indeed, Helena Pinto and João Martins completed the only work that I know of on the matter,[6] identifying dynamism as a relevant dimension of the ontology-engineering process. This lack of interest is likely because conceptual dynamics is less obvious when dealing with abstract concepts such as "physical matter," "agent," or "intangible." Because finding ontological truth has historically been a major guideline of building ontologies in computer science, we often falsely assume that creating lasting ontologies is just a matter of proper conceptual modeling. That is, once we've discovered the correct model for a domain of discourse, the conceptualization will be stable for ages.

It is a triviality that a domain's conceptual dynamics increases with the specificity of modeling, (a class hierarchy's granularity, for example). It's also obvious that materializing most of the promises about ontologies will require very detailed domain ontologies rather than just philosophical abstractions.

The main problem is that creating an ontology or updating an existing one takes time. If we assume that some conceptual dynamics always exists in a domain of discourse, we face the fundamental problem illustrated in Figure 1. That is, once the domain capture is completed (t_0), it takes some time to formalize and release the ontology. Thus, the first version will become available only by t_1. This version contains all the elements (classes, instances, attributes, relations, and axioms) that were included in the initial domain capture in t_0. In the meantime, however, additional conceptual elements will have become relevant in the real world (depicted by the green line), which are again unavailable for annotating data or for expressing queries. This is particularly unfortunate because the novel concepts in a domain are often the most interesting when applying semantic technology.

In actively maintaining the ontology, we might carry out an updated domain capture at t_1, but producing the updated ontology and documentation again takes time, which means the new version is available at t_2 and again lacks newly relevant concepts.

As long as the ontology-engineer-

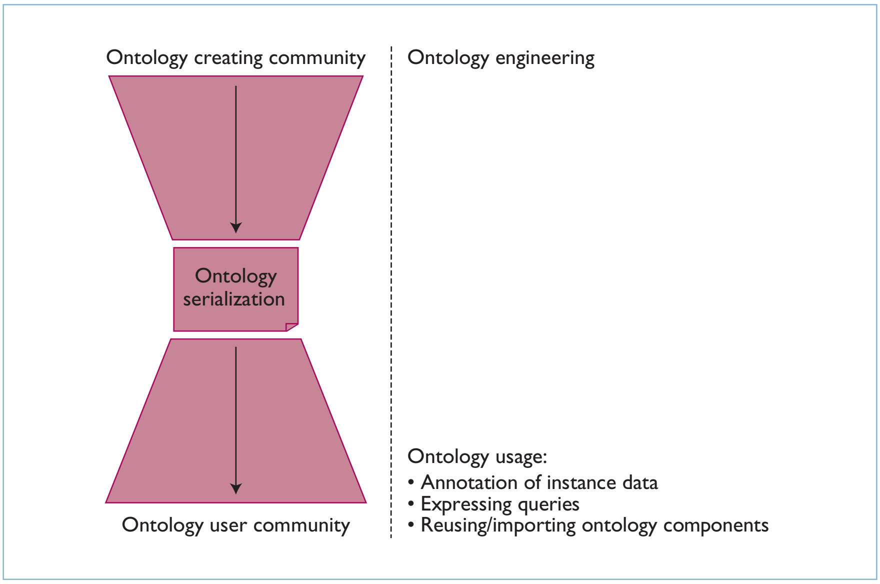

Figure 2. The ontology-perspicuity bottleneck. Users can communicate with the community that built the ontology only through the narrow channel of the ontology specification.

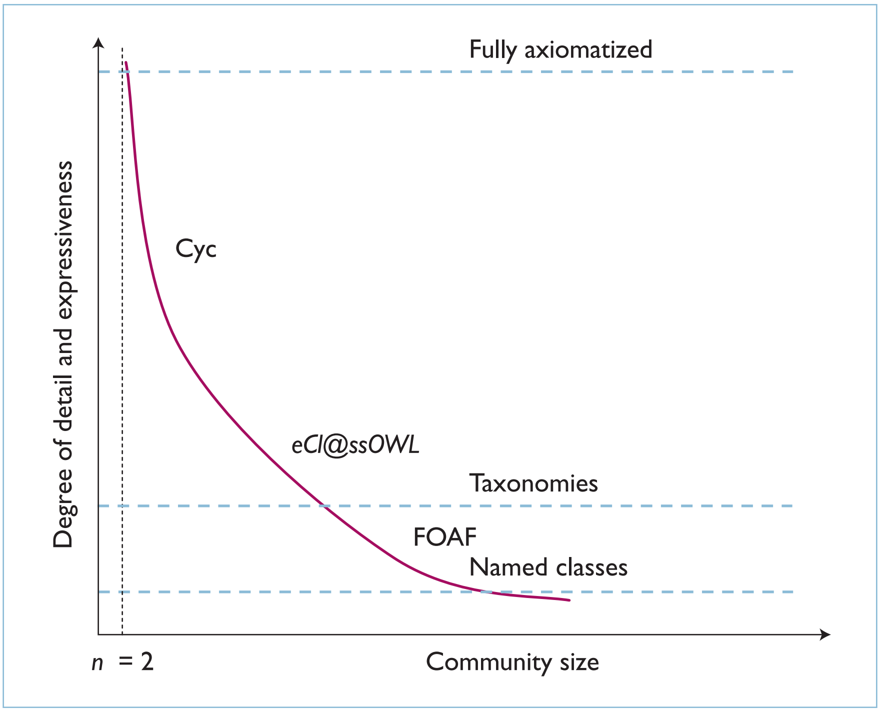

Figure 3. The expressivity–community-size frontier. A trade-off exists between an ontology's degree of detail and expressiveness and the achievable community size because the more detailed the ontology the fewer people will be willing to dedicate the resources for reviewing it prior to adopting it.

ing and ontology maintenance lags are small and the conceptual dynamics is limited, this problem might not be too substantial. Assuming ontologies follow the principle of minimal ontological commitment, thus making them more abstract, this knowledge-acquisition and maintenance bottleneck should be less problematic than with detailed knowledge bases. However, as part of my PhD,[7] I analyzed several relevant domains and found substantial conceptual dynamics in typical domains such as IT components, pharmaceuticals, and chemical substances. We should also consider that release cycles for ontologies are typically six to 12 months at best.

Obstacle 2: Economic Incentive

The second major bottleneck for building ontologies is in the economic dimension. Quite obviously, creating an ontology consumes resources (such as human labor) provided by the contributing individuals. Unfortunately, for many domains, whether the gain in automation enabled by an ontology outweighs the resources necessary for creating it remains a completely open research question. Worse yet, even if an ontology's overall benefit over its entire lifetime exceeds the creation costs, creating it must still be economically feasible for each individual required to contribute.

As with standards, ontologies exhibit positive network externalities (that is, their utility for an individual user increases with the absolute number of users). This implies that reaching a critical mass of users is difficult in the beginning because the utility for early adopters is very low, whereas the effort of adopting the ontology might be higher than at a later stage of diffusion (because less expertise and support is available, for instance).

Decisions about adopting standards also tend to be based on higher-order expectations – those about the expectations of other relevant individuals or groups. For standards and standardization, researchers in economic theory have long been analyzing such problems (namely, Michael L. Katz and Carl Shapiro[4]), but I have yet to see any discussion of network externalities in the context of ontology engineering. Some novel work exists on cost-estimation models for ontologies,[8] but this addresses only the resource-consumption aspects rather than all relevant economic driving forces.

Diffusion is even more complex with ontologies than with traditional standards because, to attract individual adopters, ontologies must properly cover their potential users' domain-representation needs as well as reach a critical mass of users. With traditional standards, such as thread diameters or track gages for railways, it's often sufficient to agree on one of many possible specifications. For example, it doesn't matter technically whether the agreed track gauge is 1,000 or 1,200 mm, as long as railway technology manufacturers and railway lines use the same value.

The incentive bottleneck is relevant not only for those creating or contributing to a given ontology but also for all pure users of it. The latter face familiarization and commitment costs, which is why good documentation helps increase the likelihood of adoption. That leads to the third bottleneck.

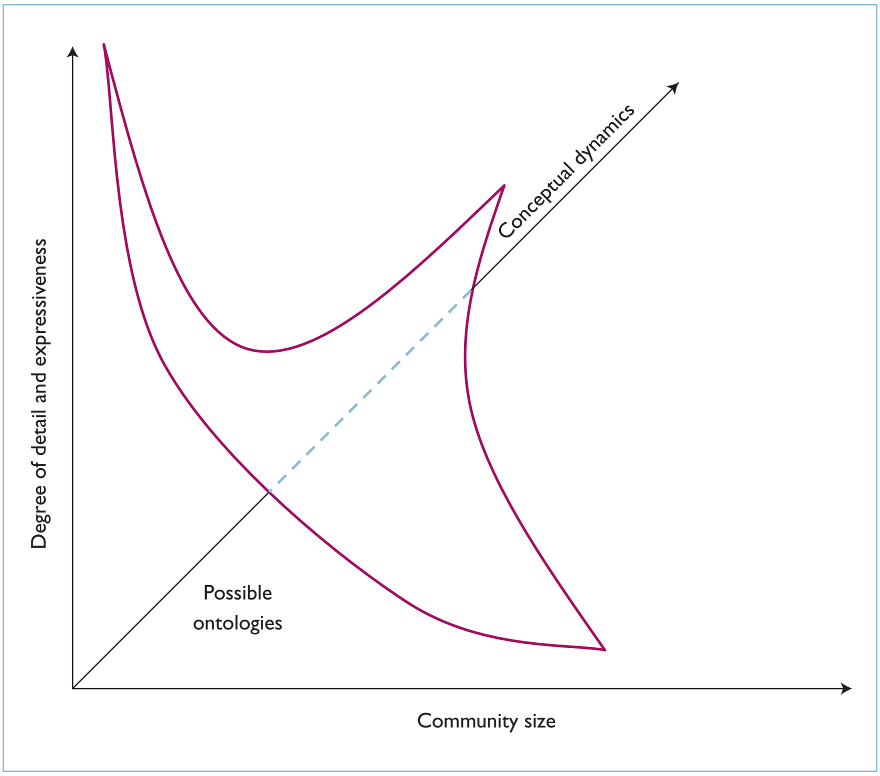

Figure 4. Possible ontologies. Building ontologies is constrained by trade-offs between conceptual dynamics, expressivity, and the number of users.

Obstacle 3: Ontology Perspicuity

As I mentioned, committing to an ontology can mean trusting its creator or verifying the specification. An example of the former is to say, "I believe the W3C that their definition of `foaf:knows` in the Friend-of-a-Friend vocabulary specification is compatible with my definition; if there are discrepancies, I'm willing to take the consequences."

Back in 1997, Marc S. Fox and Michael Gruninger identified "perspicuity" as a requirement, arguing that a good ontology should be "easily understood by the users, so that it can be consistently applied and interpreted."[9]

Ontologies are usually created by small communities but intended for much wider use. Those not involved in creating an ontology have nothing but the specification and documentation at hand to understand the semantics of all the elements. In other words, the larger group is communicating with the initial community only through the ontology specification. Figure 2 illustrates this bottleneck.

This situation presents several practical problems. First, a large share of the intended semantics exists only in the informal part of the ontology — in human-readable names for ontology elements or in `rdfs:label` and `rdfs:comment` properties, for example — and ontology creators (particularly those with strong backgrounds in logic) often dedicate little effort to creating good labels and natural-language definitions. Also, in terms of media richness, communication through an RDF/XML file about an ontology element's meaning isn't ideal. To help address these issues, my colleagues and I (and perhaps others that we don't know of) have proposed using multimedia elements to enrich the informal parts of ontology specifications.[10]

Another problem is that only a few domain experts can successfully interpret formal ontology specifications. Description-logics modeling is often particularly unintuitive — teaching university classes, I've found that some students even have great difficulty in grasping the exact semantics of `rdfs:subClassOf`.

Now, which individual or organization will authorize all the inferences that are to be drawn from a particular ontology if they can't understand them up front?

Obstacle 4: Intellectual Property Rights

Many of the interoperability problems that ontologies could help overcome deal with competing, incompatible standards, such as the automated mediation between two electronic data interchange (EDI) message formats. This requires that we first lift all relevant input standards to an ontological level; otherwise, systems can't process respective data on a semantic level.

Again, standards specifications, controlled vocabularies, and existing taxonomies are often subject to intellectual property rights.[5] This constrains creating and republishing ontologies

Table 1. Snapshot of popular ontologies.

Swoogle rank as of 11 Nov. 2005	Name	Namespace URI	File size (Kbytes)	Semantic Web documents referring to this ontology
1	RDF	http://www.w3.org/1999/02/22-rdf-syntax-ns#	7	321.108
2	DC-E	http://purl.org/dc/elements/1.1/	15	238.346
3	RSS	http://purl.org/rss/1.0/	4	195.018
4	FOAF	http://xmlns.com/foaf/0.1/	39	79.226
5	RDF-S	http://www.w3.org/2000/01/rdf-schema#	7	65.486
6	BIO	http://purl.org/vocab/bio/0.1/	7	16.588
7	DC-T	http://purl.org/dc/terms/	48	12.738
8	WGS84	http://www.w3.org/2003/01/geo/wgs84_pos#	6	11.570
9	VCARD	http://www.w3.org/2001/vcard-rdf/3.0#	10	11.185
10	CC	http://web.resource.org/cc/	6	11.023

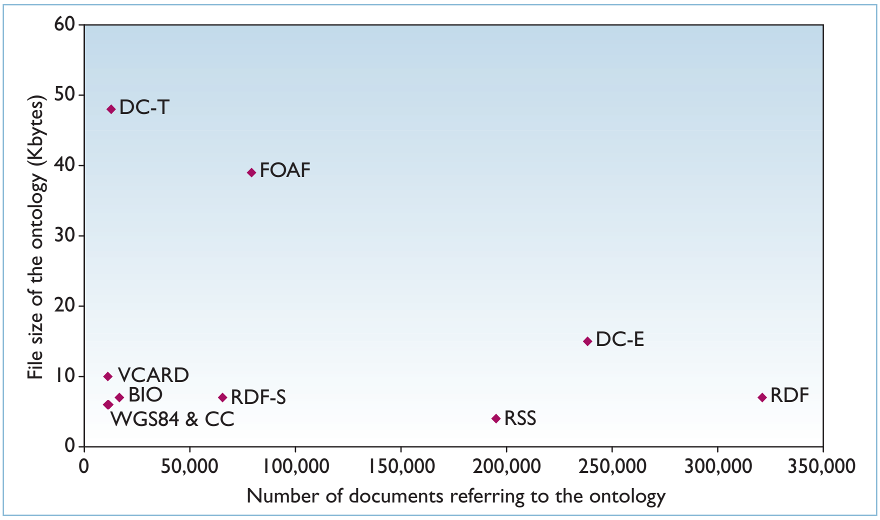

Figure 5. Ontology size and popularity. A negative correlation exists in the relationship between ontology file size (given in Kbytes) and the number of Semantic Web documents that refer to them for the 10 most popular ontologies (based on Swoogle data, retrieved 11 Nov. 2005).

as derived works, even if the original standard can be used rather freely. In other words, ontologizing industrial standards will often require explicit legal agreements with the owners of those standards. Some academic prototypes of ontologies are currently derived from common standards on the Web. In many cases, however, the general licenses offered by the respective standards owners disallow the creation and publication of derived works. The ontology creators must therefore have established additional legal agreements, or else they're playing a risky game.

A Prediction

The four obstacles I've identified constrain the space of possible ontologies — those that are both technically feasible and adoptable by rationally acting, free individuals. Figure 3 shows the "expressivity–community-size frontier" we can predict as a result. Basically, the more detailed and expressive the ontology, the smaller the actual user community will be because it increases the resources necessary for reviewing and understanding the specification and associated documentation, which makes committing to the ontology reasonable only for a smaller number of individuals. The extreme would be an ontology that is so detailed that only two entities shared it.

In practice, useful ontologies must be small enough to have reasonable familiarization and commitment costs and big enough to provide substantial added value for using them. This is a proper extension of the classical idea of minimal ontological commitment.[3] FOAF shows that such shallow, small ontologies have driven the Semantic Web's development so far.

If we combine this constraint with the conceptual dynamics dimension, we see that the space of possible ontologies is limited in at least three dimensions, as Figure 4 shows. A high degree of conceptual dynamics further constrains the possible degree of detail and expressiveness because it takes more resources and time to update concepts in a highly axiomatized ontology. Also, frequent updates increase the resources that ontology users must invest to verify and renew their commitment.

Reality Check

To support my predictions regarding possible ontologies, I gathered data about the popularity of individual ontologies from statistics at the Swoogle Semantic Web search engine (http://swoogle.umbc.edu).

Table 1 shows the resulting data. I first retrieved a list of the most popu-

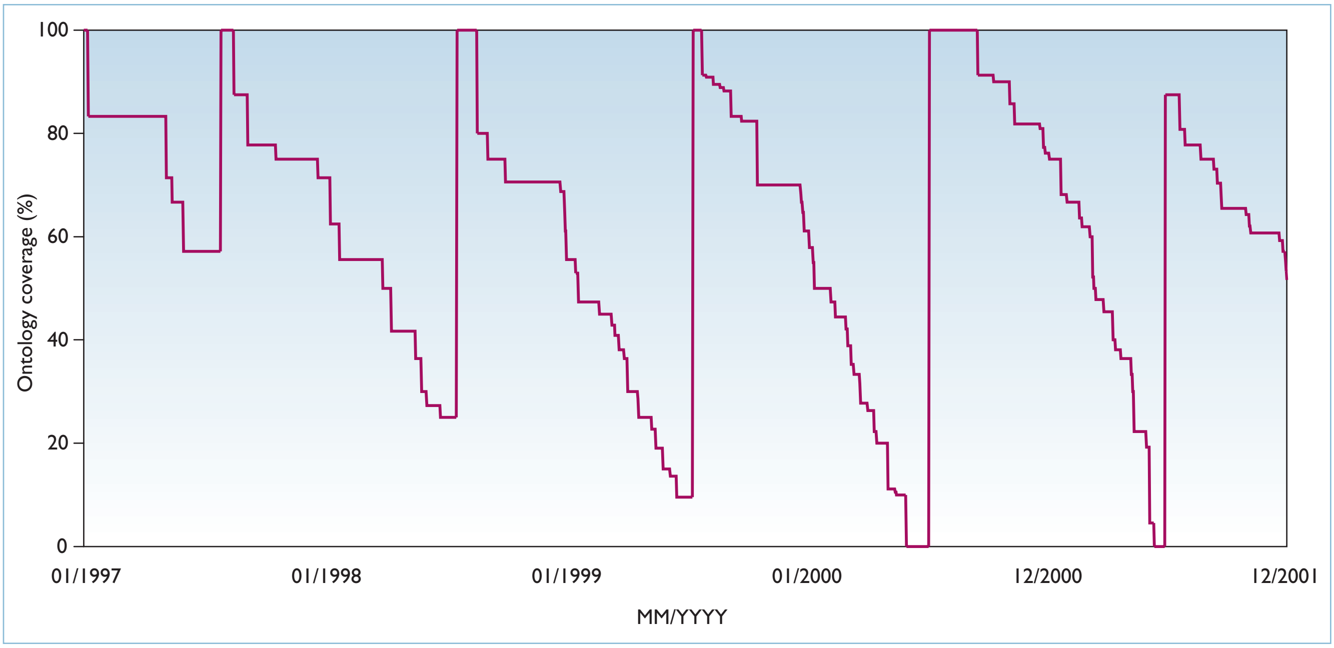

Figure 6. Conceptual coverage of an Intel CPU ontology. The coverage rate of current concepts in this ontology continuously decreases because a subset of the elements becomes obsolete and new ones are included only after the next release cycle.

lar ontologies — those referenced in the highest number of Semantic Web documents (SWDs) indexed in Swoogle. I then eliminated those for which Swoogle didn't find at least one specification document (called a Semantic Web Ontology [SWO] in Swoogle terminology), which usually indicates that they are mere namespaces of XML schemas or other informal specifications.

I then retrieved the ontology specifications for the 10 most popular ontologies and computed the ratio between their specification file sizes and the number of Semantic Web documents referring to each. In other words, I used the ontology-specification file size as an approximation for the level of detail and expressiveness and the number of Semantic Web documents as an approximation for the size of the community using the ontology.

Figure 5 visualizes as an XY scatter plot the negative correlation in the relationship between an ontology's specification file size and the number of SWDs referring to it (the correlation coefficient is –0.20). It's striking that all 10 of the most popular ontologies are less than 50 Kbytes in size — indeed, seven of them are 10 Kbytes or less.

I also did a Petri-net-based simulation on how complete the coverage of current concepts an ontology of Intel CPUs could be, given realistic assumptions. CPUs are particularly interesting because, although instances of the concept "CPU," CPU models are likely part of a PC-components ontology rather than just data, given that we use them to classify PC categories in terms of configuration and performance.

I used the "birth dates" that were available for each individual CPU model for a five-year timeframe in order to determine when each conceptual entity became relevant. I then approximated these concepts' life spans — that is, how long they really belonged to the current vocabulary. This is necessary because, when measuring the coverage as a percentage, we should exclude old concepts that are no longer part of the active vocabulary; otherwise, the vast amount of outdated (but still specified) concepts will obscure the amount of current concepts missing in the ontology.

I describe the experiment details elsewhere,[6] but my basic assumptions for the simulation were that:

- CPUs released through 1997 would belong to the relevant concepts for 720 days after their introduction, and those released starting in 1998 would be relevant for 360 days, and
- the body managing the ontology would update the ontology every 360 days to include all new elements introduced at least seven days before the ontology update.

Figure 6 shows how the coverage rate of current concepts in this ontology continuously decreases because a subset of the elements becomes obsolete and new ones are included only after the next release cycle.

Making the Semantic Web a reality demands more and better ontologies. Yet, building ontologies is inherently a social process constrained by technical, social, economic, and legal bottlenecks. That means that researchers must bring the same interest they do to purely technical issues to addressing the other challenges reality imposes on ontology projects.

Acknowledgments

The European Commission supports this work under the projects DIP (FP6-507483), SUPER (FP6-026850), and MUSING (FP6-027097), and the Austrian BMVIT/FFG supports it under the FIT-IT project myOntology (grant no. 812515/9284).

References

1. M. Uschold and M. Grüninger, "Ontologies: Principles, Methods, and Applications," *Knowledge Eng. Rev.*, vol. 11, 1996, pp. 93–155.
2. C. Petrie, "Pragmatic Semantic Unification," *IEEE Internet Computing*, vol. 9, no. 5, 2005, pp. 96, C3.
3. A. Gómez-Pérez, M. Fernández-López, and O. Corcho, *Ontological Eng.*, Springer, 2004.
4. M.L. Katz and C. Shapiro, "Network Externalities, Competition, and Compatibility," *Am. Economic Rev.*, vol. 75, 1985, pp. 424–440.
5. P. Samuelson, "Copyrighting Standards," *Comm. ACM*, vol. 49, 2006, pp. 27–31.
6. H.S. Pinto and J.P. Martins, "Ontologies: How Can They Be Built?" *Knowledge and Information Systems*, vol. 6, 2004, pp. 441–464.
7. M. Hepp, *Güterklassifikation als semantisches Standardisierungsproblem* (*Classification of Goods as a Problem of Semantic Standardization*), Deutscher Universitäts-Verlag, 2003.
8. E. Paslaru Bontas Simperl, C. Tempich, and Y. Sure, "ONTOCOM: A Cost Estimation Model for Ontology Engineering," *Proc. 5th Int'l Semantic Web Conf.* (ISWC 2006), LNCS 4273, Springer, 2006, pp. 625–639.
9. M.S. Fox and M. Gruninger, "On Ontologies and Enterprise Modeling," *Proc. Int'l Conf. Enterprise Integration Modeling Technology*, 1997; www.eil.utoronto.ca/enterprise-modelling/papers/fox-eimt97.pdf.
10. M. Hepp, D. Bachlechner, and K. Siorpaes, "OntoWiki: Community-Driven Ontology Engineering and Ontology Usage based on Wikis," *Proc. 2006 Int'l Symp. on Wikis* (WikiSym 2005), ACM Press, 2006, pp. 143–144.

Martin Hepp is a senior researcher at the Digital Enterprise Research Institute (DERI) at the University of Innsbruck, Austria, where he leads the Semantics in Business Information Systems research cluster. He created eClassOWL, the first industry-strength ontology for products and services, and is currently working on using Semantic Web services technology for business process management. Hepp has a master's degree in business management and business information systems and a PhD in business information systems from the University of Würzburg, Germany. Contact him at mhepp@computer.org; www.heppnetz.de.

IEEE computer society

PURPOSE: The IEEE Computer Society is the world's largest association of computing professionals and is the leading provider of technical information in the field.
MEMBERSHIP: Members receive the monthly magazine *Computer*, discounts, and opportunities to serve (all activities are led by volunteer members). Membership is open to all IEEE members, affiliate society members, and others interested in the computer field.
COMPUTER SOCIETY WEB SITE: www.computer.org
OMBUDSMAN: Email help@computer.org.

Next Board Meeting: 9 Feb. 2007, Vancouver, B.C.

EXECUTIVE COMMITTEE

President: Michael R. Williams*
President-Elect: Rangachar Kasturi;* **Past President:** Deborah M. Cooper;* **VP, Conferences and Tutorials:** Susan K. (Kathy) Land (1ST VP);* **VP, Electronic Products and Services:** Sorel Reisman (2ND VP);* **VP, Chapters Activities:** Antonio Doria;* **VP, Educational Activities:** Stephen B. Seidman;† **VP, Publications:** Jon G. Rokne;† **VP, Standards Activities:** John Walz;† **VP, Technical Activities:** Stephanie M. White;* **Secretary:** Christina M. Schober;* **Treasurer:** Michel Israel;† **2006–2007 IEEE Division V Director:** Oscar N. Garcia;† **2007–2008 IEEE Division VIII Director:** Thomas W. Williams;† **2007 IEEE Division V Director-Elect:** Deborah M. Cooper;* ***Computer* Editor in Chief:** Carl K. Chang;†
** voting member of the Board of Governors* *† nonvoting member of the Board of Governors*

BOARD OF GOVERNORS

Term Expiring 2007: Jean M. Bacon, George V. Cybenko, Antonio Doria, Richard A. Kemmerer, Itaru Mimura, Brian M. O'Connell, Christina M. Schober
Term Expiring 2008: Richard H. Eckhouse, James D. Isaak, James W. Moore, Gary McGraw, Robert H. Sloan, Makoto Takizawa, Stephanie M. White
Term Expiring 2009: Van L. Eden, Robert Dupuis, Frank E. Ferrante, Roger U. Fujii, Anne Quiroz Gates, Juan E. Gilbert, Don F. Shafer

EXECUTIVE STAFF

Assoc. Executive Director: Anne Marie Kelly; **Publisher:** Angela R. Burgess; **Associate Publisher:** Dick J. Price; **Director, Administration:** Violet S. Doan; **Director, Finance and Accounting:** John Miller

COMPUTER SOCIETY OFFICES

Washington Office. 1730 Massachusetts Ave. NW, Washington, DC 20036-1992
Phone: +1 202 371 0101 • Fax: +1 202 728 9614 • Email: hq.ofc@computer.org
Los Alamitos Office. 10662 Los Vaqueros Circle, Los Alamitos, CA 90720-1314
Phone: +1 714 821 8380 • **Email: help@computer.org**
Membership and Publication Orders:
Phone: +1 800 272 6657 • Fax: +1 714 821 4641 • **Email: help@computer.org**
Asia/Pacific Office. Watanabe Building, 1-4-2 Minami-Aoyama, Minato-ku, Tokyo 107-0062, Japan
Phone: +81 3 3408 3118 • Fax: +81 3 3408 3553
Email: tokyo.ofc@computer.org

IEEE OFFICERS

President: Leah H. Jamieson; **President-Elect:** Lewis Termin; **Past President:** Michael R. Lightner; **Executive Director & COO:** Jeffry W. Raynes; **Secretary:** Celia Desmond; **Treasurer:** David Green; **VP, Educational Activities:** Moshe Kam; **VP, Publication Services and Products:** John Baillieul; **VP, Regional Activities:** Pedro Ray; **President, Standards Association:** George W. Arnold; **VP, Technical Activities:** Peter Staecker; **IEEE Division V Director:** Oscar N. Garcia; **IEEE Division VIII Director:** Thomas W. Williams; **President, IEEE-USA:** John W. Meredith, P.E.

revised 11 Dec. 2006

IEEE