



Rethinking the Semantic Web, Part I

Rob McCool • Yahoo

The Semantic Web is a compelling vision, laid out in 2001 by Tim Berners-Lee and others,¹ in which the World Wide Web will include a notion of meaning in data and services. Intelligent agents will exchange information and rules for how to interact with that information, with or without human intervention; appointments will be automatically scheduled; and automated agents will select and invoke services. Information will be easy to find without depending solely on keywords.

In part one of this column, I propose several reasons that this vision hasn't yet been adopted despite substantial research funding in the US and European Union (EU). These reasons will provide the foundation for a new approach, which I'll propose in part two.

Fundamentals

The Semantic Web has three fundamental parts. The foundation is the set of data models and formats that provide semantics to applications that use them. The second layer is composed of services – purely machine-accessible programs that answer Web requests and perform actions in response. At the top are the intelligent agents, or applications.

Data Models and Formats

There are three major information formats on the Semantic Web:

- the Resource Description Format (RDF),²
- RDF Schema,³ and
- the Web Ontology Language (OWL).⁴

OWL is a superset of RDF Schema, which is a superset of RDF. RDF provides a format, and both RDF Schema and OWL provide ontological data models.

The fundamental unit of representation in RDF is the *triple*. RDF forms a directed graph, in which each triple consists of a *subject* node, a *predicate* name, and a *target* node or literal. Anything that can't be represented as a triple must use an excessively specific predicate name or be *reified*. Reification is the process of taking a triple and stating sentences about it using another set of triples. If reification isn't used, we must typically embed contextual information into the predicate name. When we use RDF Schema or OWL for vocabulary, these triple sets form ontologies. Ontological vocabularies define basic structure for triples, including class hierarchies and conventions for human-readable node names, as well as define behaviors for certain types of properties.

RDF triples must be maintained, either as separate files or within separate blocks inside HTML files. They are separate from any natural language representations of their contents, and they can't include HTML markup. The formats are for machines and have nothing to do with language or markup. The ontological data model makes representation of any nontrivial factual information difficult because it can't represent context of any kind. The Semantic Web thus represents a "shadow web" that's entirely separate from today's Web.

Each of these formats has seen

phenomenally low adoption rates. A quick visit to the Swoogle Semantic Web Search and Metadata Engine (<http://swoogle.umbc.edu>) reveals problems in both coverage and quality. The Alpiri Project (TAP; <http://tap.stanford.edu/TAP/>), a large-scale RDF project that I cofounded with Ramnathan V. Guha, suffered from similar spotty coverage and inconsistent quality. Other examples include the Semantic Interoperability of Metadata and Information in unlike Environments (Simile) project at MIT (<http://simile.mit.edu>) and the KnowItAll project at the University of Washington (www.cs.washington.edu/research/knowitall/). There is very little breadth or depth to any of this information. Further, automated extractors introduce numerous errors.

Services

Most consultants say that the Web Services Definition Language (WSDL)⁵ is seeing increasingly wide deployment inside enterprises. Yet, such services contribute nothing to the public Web. Major search engines and e-commerce sites have deployed the only public services to date. Amazon and eBay use their Web services to attract new buyers and sellers and to allow external developers to create new functionality. Google and Yahoo provide Web services simply because their success encourages "screen scraping." Both companies found that developers were using programs to parse their HTML and extract search results, causing errors, and decided it was in their bet-

continued on p. 86

continued from p. 88

ter interests to support and learn from these developers rather than hamper them with the error rates that result from programmatic HTML analysis.

Beyond these efforts, most public services are trivial. A quick visit to the XMethods Web services directory (www.xmethods.net) or any UDDI node (www.uddi.org) confirms this fact. The Semantic Web vision argues that adding semantics to Web services will allow intelligent agents to discover

lus led to the relational database revolution in the 1980s. Before his work, information representation in computers typically used ad hoc, hierarchical, inflexible techniques.

Knowledge representation uses the fundamental mathematics of Codd's theory to translate information, which humans represent with natural language, into sets of tables that use well-defined schema to define what can be entered in the rows and columns. The technique is very similar to a database

tially. New approaches such as Web Service Modeling Ontology (www.wsmo.org) are emerging, but they suffer from the same fundamental flaws as knowledge representation.

Consider the Semantic Web vision paper. It describes how a brother and sister might use "a trusted rating service" to find a doctor "with a rating of excellent" for their mother. Yet, as anyone who has used contractor-referral services for home remodeling knows, these ratings simply don't work, no matter how many axes they're applied across. Speaking to a friend or participating in a trusted community is nearly always a more informed approach for locating referrals because language representation fundamentally invokes catastrophic translation and maintenance costs. Complexities should not be distilled to simplicities.

Because of these high costs and a desire not to assist their competitors, corporations typically don't share databases unless they have to. Even hobbyists, such as those who started the Internet Movie Database (www.imdb.com) or the Roller Coaster Database (www.rcdb.com/IMDB/), typically want to ensure that they have a way to recover labor and hosting costs. On the Web, advertising provides this revenue. Really Simple Syndication (RSS),⁹ another "shadow web" format that has nothing to do with markup or language, has also enjoyed immense success because RSS feed views nearly always result in page views, which can be monetized. The Semantic Web offers no equivalent.

Because it's a complex format and requires users to sacrifice expressivity and pay enormous costs in translation and maintenance, the Semantic Web will never achieve widespread public adoption. Some say that the answer to this is more funding — a Manhattan Project (the US effort to build an atomic weapon in World War II) or Apollo Project (the US effort to send astronauts to the moon) level of effort. Cycorp

Logic, which forms the basis of OWL, suffers from an inability to represent exceptions to rules and the contexts in which they're valid.

and compose sets of useful services, enabling a revolution in Internet applications.⁶ But there are no services and, thus, no semantics.

What About the Applications?

Given the lack of deployment of both information and services, it's not surprising that the third fundamental part of the Semantic Web, the applications, hasn't arrived. The reason for this is the fundamental problem in the approach: the Semantic Web's basis in *knowledge representation*. (See www.aaai.org/AITopics/html/repr.html for a good overview.)

The Semantic Web has enjoyed neither widespread deployment nor the formation of scalable, simple systems. It hasn't developed user communities that perceive themselves as contributing to a large, societal effort.⁷ To understand why, we must visit its roots. Knowledge representation is a technique with mathematical roots in the work of Edgar Codd,⁸ widely known as the one whose original paper using set theory and predicate calcu-

but with a large number of columns and a relatively sparse set of non-empty cells. The problem is that this creates a fundamental barrier, in terms of richness of representation as well as creation and maintenance, compared to the written language that people use. Logic, which forms the basis of OWL, suffers from an inability to represent exceptions to rules and the contexts in which they're valid.

Despite decades of effort, database tools remain extremely difficult to create. Typically, databases are deployed only by corporations whose information-management needs require them or by hobbyists who believe they can make some money from creating and sharing their databases. Because information theory removes nearly all context from information, both knowledge representation and relational databases represent only facts. Complex relationships, exceptions to rules, and ideas that resist simplistic classifications pose significant design challenges to information bases. Adding semantics, including classifications, properties, and rules, increases the burden exponen-

(www.cyc.com) is notable in this area. In the 1980s, Doug Lenat decided that the only way to create the world's first true artificial intelligence would be to simply do it. In creating Cyc, he created one of the largest knowledge-representation projects in history.

Large-scale efforts like these will fail. Finding tangibly useful results from the Cyc effort, 20 years later, is difficult even for its proponents. The US and EU have already funded Semantic Web efforts considerably. Funding must instead be directed toward finding this generation's Edgar Codd to solve the representation problem. New representations must be easy to translate to and from natural language. Any other approach ignores the representation problem, assumes that context-free facts and logical rules are sufficient, and will fail. The Semantic Web will fail because it inherits these problems and then couples them to the Web, which represents the breadth of human knowledge. It will fail.

In part two of this article, I'll propose our solution. □

References

1. T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web: A New Form of Web Content that Is Meaningful to Computers Will Unleash a Revolution of New Possibilities," *Scientific Am.*, May 2001, pp. 28-37; www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.
2. *RDF/XML Syntax Specification*, D. Beckett, ed., W3C recommendation, 10 Feb. 2004; www.w3.org/RDF/.
3. *RDF Vocabulary Description Language 1.0: RDF Schema*, Dan Brickley and R.V. Guha, eds., W3C recommendation, 10 Feb. 2004; www.w3.org/TR/rdf-schema/.
4. *OWL Web Ontology Language Overview*, D. McGuinness and F. van Harmelen, eds., W3C recommendation, 10 Feb. 2004; www.w3.org/2004/OWL/.
5. E. Christensen et al., *Web Services Description Language 1.1*, W3C note, 15 Mar. 2001; www.w3.org/TR/wsdl/.
6. C. Petrie, "World Wide Wizard: Why Web

Services Might Matter," presented at the Montreal Conf. eTechnologies, Jan. 2005; available at <http://logic.stanford.edu/talks/Wizard/>.

7. C. Petrie, "Emergent Collectives for Work and Play," *Société de Stratégie AGIR Revue Generale de Stratégié*, Jan. 2005; available at <http://www-cdr.stanford.edu/~petrie/revue/>.
8. E.F. Codd, "A Relational Model of Data for Large Shared Data Banks," *Comm. ACM*, vol. 13, no. 6, June 1970, pp. 377-387.
9. M. Pilgrim, "What Is RSS?" XML.com, 18 Dec. 2002; www.xml.com/pub/a/2002/12/18/dive-into-xml.html.

Rob McCool is a software developer and architect at Yahoo. He worked on the Alpiri Project (TAP) on Semantic Web technologies at Stanford University from 2002 until earlier in 2005. He authored the National Center for Supercomputing Applications (NCSA) httpd Web server, which became the Apache Web server. McCool has a Bachelor's degree in computer science from the University of Illinois at Urbana-Champaign. Contact him at robm@robm.com.

ADVERTISER INDEX NOVEMBER / DECEMBER 2005

Advertising Personnel

Marion Delaney

IEEE Media, Advertising Director
Phone: +1 212 419 7766
Fax: +1 212 419 7589
Email: md.ieeemedia@ieee.org

Marian Anderson

Advertising Coordinator
Phone: +1 714 821 8380
Fax: +1 714 821 4010
Email: manderson@computer.org

Sandy Brown

IEEE Computer Society,
Business Development Manager
Phone: +1 714 821 8380
Fax: +1 714 821 4010
Email: sb.ieeemedia@ieee.org

Advertising Sales Representatives

Mid Atlantic (product/recruitment)

Dawn Becker
Phone: +1 732 772 0160
Fax: +1 732 772 0161
Email: db.ieeemedia@ieee.org

New England (product)

Jody Estabrook
Phone: +1 978 244 0192
Fax: +1 978 244 0103
Email: je.ieeemedia@ieee.org

New England (recruitment)

John Restchack
Phone: +1 212 419 7578
Fax: +1 212 419 7589
Email: j.restchack@ieee.org

Connecticut (product)

Stan Greenfield
Phone: +1 203 938 2418
Fax: +1 203 938 3211
Email: greeneco@optonline.net

Midwest (product)

Dave Jones
Phone: +1 708 442 5633
Fax: +1 708 442 7620
Email: dj.ieeemedia@ieee.org
Will Hamilton

Phone: +1 269 381 2156
Fax: +1 269 381 2556
Email: wh.ieeemedia@ieee.org
Joe DiNardo

Phone: +1 440 248 2456
Fax: +1 440 248 2594
Email: jd.ieeemedia@ieee.org

Southeast (recruitment)

Thomas M. Flynn
Phone: +1 770 645 2944
Fax: +1 770 993 4423
Email: flyntom@mindspring.com

Southeast (product)

Bill Holland
Phone: +1 770 435 6549
Fax: +1 770 435 0243
Email: hollandwfh@yahoo.com

Midwest/Southwest (recruitment)

Darcy Giovingo
Phone: +1 847 498-4520
Fax: +1 847 498-5911
Email: dg.ieeemedia@ieee.org

Southwest (product)

Josh Mayer
Phone: +1 972 423 5507
Fax: +1 972 423 6858
Email: jm.ieeemedia@ieee.org

Northwest (product)

Peter D. Scott
Phone: +1 415 421-7950
Fax: +1 415 398-4156
Email: peterd@pscottassoc.com

Southern CA (product)

Marshall Rubin
Phone: +1 818 888 2407
Fax: +1 818 888 4907
Email: mr.ieeemedia@ieee.org

Northwest/Southern CA (recruitment)

Tim Matteson
Phone: +1 310 836 4064
Fax: +1 310 836 4067
Email: tm.ieeemedia@ieee.org

Japan

Tim Matteson
Phone: +1 310 836 4064
Fax: +1 310 836 4067
Email: tm.ieeemedia@ieee.org

Europe (product)

Hilary Turnbull
Phone: +44 1875 825700
Fax: +44 1875 825701
Email: impress@impressmedia.com