One-Day Conference on "Semantic Technologies for eGov"
White House Conference Center, Monday, September 8th, 2003

We would like to invite you to a one-day conference on Semantic Technologies for E-Gov at the White House Conference Center on Monday, September 8th. This is a conference organized by Brand Niemann, the chair of the CIO Council's XML Web Services Working Group, and TopQuadrant, the lead for the Semantic Technologies Pilots in e-Gov. The conference is by invitation only and no fees are involved.

Registration for the event is at http://www.topquadrant.com/conferences/stgov03.html. If you have any questions or problems with registration contact Brand Niemann at niemann.brand@epa.gov or Ralph Hodgson at ralph@topquadrant.com.

Semantic Technologies are driving the next generation of the Web, the Semantic Web, "a machine-readable web of "smart data" and automated services that amplify the Web far beyond current capabilities". We define semantic technology as a software technology that allows the meaning of and associations among information to be known and processed at execution time. Semantic technologies help solve the problems of application and data interoperability, improved search, discovery and content provisioning in knowledge-centric systems and dynamic integration across distributed systems. A definition of semantic technologies is provided at the end.

The conference will be a combination of talks, a poster session of solution stories from both vendors of semantic technologies and agencies that are already using them in applications. There will be a moderated panel discussion with agency representatives and product/technology vendors. There will be a number of distinguished speakers, people with hands-on experience and vendors who have made important progress in addressing business problems with innovative products based on semantic technologies. The participating vendors are listed in the agenda.

The event continues a series of meetings that the XML Web Services Working Group has organized to foster innovation and collaboration across federal agencies (see additional background at the end of this note).

How the Day is Organized

Brand Niemann, Leader XML Web Services Working Group and conference sponsor, will open the day.

The morning key note will be given by Eric Miller, who, as activity lead for the W3C World Wide Web Consortium's Semantic Web Initiative, will describe his work with Tim Berners-Lee. Following this, Michael Daconta, Chief Scientist, APG, McDonald Bradley, Inc. will give a talk on "Data Independence and the Roadmap to the Semantic Web", based on the book he co-authored with Dr. Leo Obrst, recently published by Wiley in June 2003.

Irene Polikoff, Executive Partner, TopQuadrant will give a presentation on "Positioning Semantic Technologies: The Emerging Vendor Landscape". Then, Dr. Dean Allemang of TopQuadrant will describe the lunch-time gallery of examples of semantic applications and facilitate brief introductory statements from the participating vendors.

During lunch, the "gallery" of semantic application “cases” featuring solution stories about what Semantic Technologies have, or could, realize, will be open for touring. The gallery will be arranged as poster sessions arranged on perimeter tables. There will be a number of software vendors attending and some representation of projects in government that are already using semantic technologies.

After lunch, Ralph Hodgson, CEO, TopQuadrant Inc., will kick off the afternoon session. He will provide an overview of the integrated Semantic Technology Pilots identified by the CIO Council's Emerging Technology Subcommittee. He will illustrate his presentation with the targeted focus of utilizing semantic technologies for inter-agency partnering in building business cases for developing IT capabilities.

Following this, the "Semantic Technology Panels – Interactive Discussion Session" will begin. The panels will focus on the critical issues and opportunities involved in "Applying Semantic Technologies in Government". Organized as two groups of panelists: "Problem Owners" and "Solution Providers", the session is designed to support a dialog about business problems and technology solutions.

In, Part I, the questions and problems to be addressed will be framed. First, the "Problem Owners" will each have a short time to state their challenges and/or questions for the response and discussion. The audience will also be invited to submit questions.

Part II will feature responses from the “Solution Providers” panel. In addition to selected vendors, rounding out the perspectives and experience on the Solution Provider’s panel will be Dr. Leo Obrst, Information Semantics, Center for Innovative Computing & Informatics, The MITRE Corporation and York Sure from OntoWeb, University of Karlsruhe. They will represent the strategic vision and evolutionary perspective of thought leaders and researchers (including the vigorous European perspective on Semantic Technologies) to complement the vendors’ responses to adoption and application issues based on what is real and possible today.

Capping the day’s full agenda, Professor Jim Hendler, Director, Semantic Web and Agents Research at the University of Maryland Information and Network Dynamics Lab, and co-chair of the W3C Web Ontology Working Group, will give a closing keynote talk on "Semantic Web Services".

AGENDA

"Semantic Technologies for e-Gov ", Monday, September 8th, 2003

White House Conference Center, Truman Room (3rd floor), 8:30 a.m. - 5:00 p.m.

8:30 a.m. - Welcome and Logistics - Brand Niemann, Chair, XML Web Services Working Group;

8:45 a.m. - Keynote: "The Semantic Web" - Eric Miller, Activity Lead for W3C's Semantic Web Initiative.

9:30 a.m. - Invited Talk: "Data Independence and the Roadmap to the Semantic Web", Michael Daconta, Chief Scientist, McDonald Bradley

10:00 a.m. Morning Break.

10:15 a.m. "Positioning Semantic Technologies: The Emerging Vendor Landscape", Irene Polikoff, Executive Partner, TopQuadrant.

10:45 a.m. Introducing the Semantic Application Gallery and Vendors, Dr. Dean Allemang, TopQuadrant, with short statements from participating Vendors.

11:15a.m. Audience Participation, Susan Turnbull, CIO Council's Emerging Technology Subcommittee

12:00 p.m. - Gallery Lunch (for those who order catered in) and **tour** of the exhibits in the "Semantic Applications Gallery" in the Lincoln Room (2nd floor).

Exhibitors: Ecosystems, Coolheads Consulting, Linkspace, McDonald Bradley, Modulant, Ontoprise, Semagix, Software AG, TopQuadrant, Unicorn Solutions.

1:30 p.m. Ralph Hodgson, CEO TopQuadrant and Pilot Lead: Welcome back- Overview of Part 2 and Semantic Technology Pilots "Potential of Semantic Technologies for eGovernment"

2:00 p.m. Interactive Panels discuss "Applying Semantic Technologies in Government":

Part 1: "Framing the Questions"

- The Problem Owners Panelists introduce themselves and state their problems, challenges and questions
- (Panel will consist of: George Strawn, CIO of NSF, William Sonntag, Chief of Staff at EPA, Lillian W. Gassie, Senior Systems Librarian of Homeland Security Digital Library, Con Kenney, Chief Enterprise Architect at FAA, Luis G. Kun, Ph.D. Professor of Systems Management at National Defense University/ DOD, Tom West, Chair, IC Metadata Working Group at DIA)
- The Solution Provider Panelists introduce themselves
- (The Panel will comprise Leo Obrst (Mitre), York Sure (University of Karlsruhe/Ontoprise), Zvi Schreiber (Unicorn), Clemens Bertram (Semagix), Dwight Lodge (enLeague) JP Morgenthal (Software AG))

Part II: "Exploring Answers"

- Panel Discussion Moderators, Ralph Hodgson and Dr. Robert Coyne, TopQuadrant, Inc. briefly summarize what has been heard to initiate and facilitate the discussion
- **Discussion between Problem Owners, Solution Providers and the audience**

3:45 p.m. Closing Keynote: "Semantic Web Services", Professor Jim Hendler, Director, Semantic Web and Agent Technologies, Maryland Information and Network Dynamics Lab, University of Maryland and co-chair of the W3C Web Ontology Working Group.

4:30 p.m. - Closing Remarks and Some Next Steps - Brand Niemann and Susan Turnbull, CIO Council's Emerging Technology Subcommittee

4:45 p.m. - **Adjourn**

5:00 p.m. - Conference Center closes

XML Web Services Working Group Overview

The CIO Council's Architecture and Infrastructure Committee in cooperation with the XML Working Group and the Universal Access Collaboration Workshops has the mission to achieve interoperability and reuse of IT capabilities across agencies by exploring the value of emerging technologies and standards through incubator projects. The XML Web Services Working Group maintains a log of its activities at http://web-services.gov/ (courtesy of USGS Geology Division CIO's Office), a GSA ListServ (courtesy of GSA CIO's Office): http://listserv.gsa.gov/archives/cioc-web-services.html, and a Collaboration Place (courtesy of GSA's CIO's Office of Government-wide Policy: http://ioa-qpnet-co.gsa.gov/WebServices. The purpose of the XML Web Services Working Group is to support the Emerging Technology Subcommittee of the CIOC/AIC in its work with the other two subcommittees (Enterprise Architecture Governance and Components) and to produce incubator pilot projects in support of the e-Gov Initiatives that use XML Web Services to demonstrate increased accessibility and interoperability.

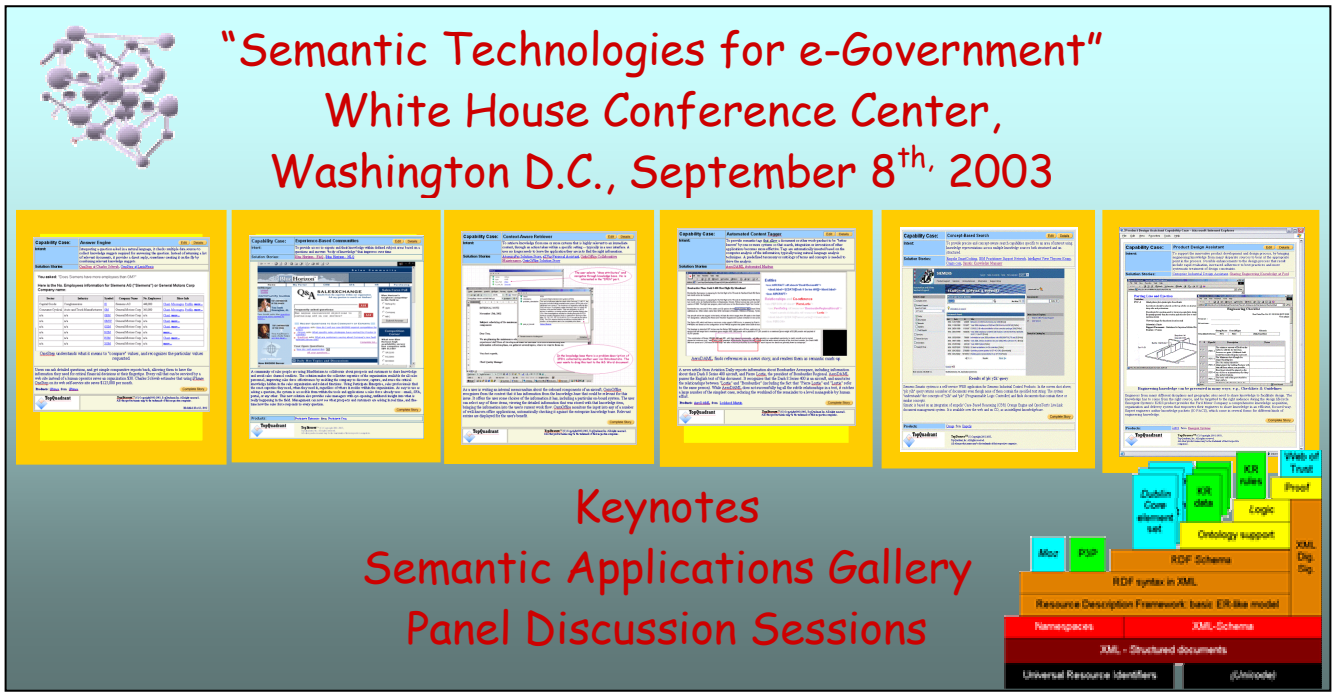
Semantic Technologies in a Nutshell

We define semantic technology as a software technology that allows the meaning of and associations between information to be known and processed at execution time. For a semantic technology to be truly at work within a system there is a knowledge model of some part of the world (an active ontology) that is used by one or more applications at execution time. Semantic technologies use knowledge representation languages (for example KIF, LOOM, and KL-ONE), knowledge markup languages (for example, Topic Maps, RDF[S], RuleML and OWL), and model-based design approaches (example, MOF, MDA and OCL). Semantic technologies help solve the problems of application and data interoperability, improved search, discovery and content provisioning in knowledge-centric systems and dynamic integration across distributed systems. Semantic Technologies are driving the next generation of the Web, the Semantic Web, "a machine-readable web of "smart data" and automated services that amplify the Web far beyond current capabilities".

Problem Statements for

"Semantic Technology Panels – Interactive Discussion Session"

at the event



This one day conference-workshop on the use of semantic technologies for e-Government is jointly organized by Brand Niemann, CIO Council AIC XML Web-Services Working Group <http://www.web-services.gov/> and TopQuadrant, Inc. http://www.topquadrant.com, lead for the Semantic Technologies in e-Gov Pilot.

One-Day Conference-Workshop on "Semantic Technologies for e-Government" Problem Statements

This document contains the collected 'Problem Statements' submitted by government agency personnel and representatives for the "Semantic Technology Panels – Interactive Discussion Session". By the term *Problem Statement*, we mean simple statements of business needs, problems or challenges related to government agencies or contexts that may benefit from semantic technology-enabled applications and solutions. These statements were solicited and submitted in response to the note below.

Problem Statements selected for discussion on the Panel at the event are indicated with a '*'.

The following is the text of the problem statement solicitation sent out by the conference organizers.

Thank you for your interest in the upcoming event: **Semantic Technologies for E-Government on September 8th at White House Conference Center** (see enclosed for a complete description).

A panel discussion between "Problem Owners" and potential Solution Providers is a central feature of the day's activities. By "problem" we simply mean a business situation that requires an IT solution to be realized.

For this purpose, we are soliciting simple statements of business needs, problems or challenges related to government agencies or contexts. We would like to collect as many meaningful problem statements as we can. Some of the statements will be selected for representation on the panel. All of the statements will be conveyed to the Solution Providers panel for their possible response. With the author's permission, the statements will also be published on the web site for the event.

If you wish to submit a problem statement or a question, please use the enclosed, simple one-page 'Problem Owners Submission Template'. It is pretty much self-explanatory. We expect to shortly post one or more examples of this already filled in by other problem owners and will let you know where you can look at those as soon as they are available.

Conference organizers are planning to select ~ 4-5 of the submissions to be represented by panelists. Each of the selected statements will be presented by its representatives in a short overview (maximum of 8 minutes for each).

The final selection of panelists will be based on the description and content of problems statements most meaningful to the audience -- cutting across a broad range of government agencies and services. To optimize the presentation of problem areas and possible solutions that will contribute most to the success of the event for all participants, we are deferring the final selection until 10 days prior to the event.

We greatly appreciate your interest in participating in the event, and invite you to consider making a submission. If you do so, please indicate on the template whether you would consider being on the panel if selected.

* **Problem Statement Submitted by:** United States Environmental Protection Agency, Office of Environmental Information, William Sonntag, Chief of Staff, Sonntag.william@epa.gov Phone: 202 564-3871

1) Please describe in (one paragraph) the nature of the problem or challenge you are facing?

Pound for pound, children are more susceptible to environmental contaminants than adults. Given the cumulative and multiple impacts a child can face from air pollution, lead paint, contaminated soil, it is almost impossible to scientifically predict their "risk" in such a way to support a regulatory solution. The answer seems to lie in providing both public health data (how many chemicals are in a child's blood stream, and health effects) and environmental data (what chemicals/toxins is a child exposed to in their environment) to a wide variety of decision makers at all levels of government so that risk can be determined at the local level. Almost all of this information now resides in different data systems and is often expressed in terms that are not usable for either health or environmental decision making. While EPA and Center for Disease Control and Prevention (CDC) and their state partners are addressing this issue in developing internet based networks of information, these networks are not yet funded to include the wide scope of individuals who could use the data-nor are they envisioned to include linkages to numerous studies, reports and systems which could illuminate the answer to the public's question: Is my child safe from environmental toxins?

2) What business forces are making this problem especially critical or important? (For example, new regulations? Changes in expectations for services provided?)

The expectation of the public is that land-use decisions and the regulations of environmental contamination should ensure that children are protected. The President, through Executive Order has set up a Cabinet level Children's health Taskforce. EPA created a children's health office reporting directly to its Administrator. CDC recognizes the frustration and the federal government's limitations in the explaining of cancers, respiratory problems, increasing asthma rates in environmental hot spots. It would be fair to say that public's expectations are rising and that Congressional interest is high. EPA recognizes that the next round of environmental improvements including those that impact children's health will often come at a high cost, either in terms of changes in social behavior or major infrastructure and technology changes. These will need the most robust of health and ecological-based benefit determinations for success.

3) Describe what you are currently doing to address the problem/challenge (2-5 bullets)

- ✓ EPA and CDC are coordinating their developing of national internet based data networks and determining which data sets and questions should be given priority ie childhood asthma, drinking water contaminants, indoor air quality and pesticide exposure
- ✓ Convening CDC and EPA technical and terminology workgroups to examine convergence and identifying divergence
- ✓ Federal wide research agenda into the science of determining the risk to children
- ✓ Strong outreach effort to educate and involve key stakeholder groups
- ✓ International collaboration within the UN WHO, European Union and others on all of the above

4) In your ideal world, how could the advances in information management and technology help solve your problem?

In the ideal world, within the next 5 years, public servants should be able to query over the Internet all sources of data that could help them diagnose and mitigate environmental threats to children. For example, a public servant should be able to access arrays of data from heterogeneous sources and obtain answers to the following questions "What environmental containments are in my county? Where are the children living? What is the risk of the chemicals present? What does the research say on these chemicals? what is the level of exposure to children? What diseases do these children already have and which of them could be attributed to or aggravated by environmental contaminants?

5) Please list any websites that provide additional information or context for question #1.

USEPA Environmental Data Registry - http://www.epa.gov/edr/

Environmental Information Exchange Network - http://www.exchangenetwork.net/

Children's Environmental Health and Safety Inventory of Research - http://www.epa.gov.chehsir

USEPA Children's Health Protection Office, America's Children and the Environment: Measures of Contaminants, Body Burdens, and Illnesses (Second Edition) - http://yosemite.epa.gov/ochp/ochpweb.nsf/content/publications.htm

CDC National Health and Nutrition Examination Survey (NHANES) Second Report – www.cdc.gov/exposurereport

CDC National Environmental Health Tracking Program – http://www.cdc.gov/nceh/tracking

* **Problem Statement Submitted by:** Dr. George Strawn, CIO, National Science Foundation, Strawn, gstrawn@nsf.gov

1) Please describe in (one paragraph) the nature of the problem or challenge you are facing?

NSF is considering a Knowledge Management (KM) initiative. NSF's business could be described as a "service-based, low-volatility" set of activities (centered around merit review of research proposals and award management for grants to scientists and engineers to conduct that research). Some proponents suggest that KM for an organization like ours should consist of "high quality knowledge repositories with effective search capabilities". Thus we would like to create such repositories with "more than google" to support effective search.

2) What business forces are making this problem especially critical or important? (for example, new regulations? Changes in expectations for services provided?)

We utilize many temporary employees (eg, scientists on 2-year loan from universities) who need to be "jump started" quickly into the NSF culture.

3) Describe what you are currently doing to address the problem/challenge (2-5 bullets)

We currently produce conventional manuals and put them on the web.

4) In your ideal world, how could the advances in information management and technology help solve your problem?

Semantic meta-information seemly could give us more effective search capabilities.

5) Please list any websites that provide additional information or context for question #1.

* **Problem Statement Submitted by:** Lillian W. Gassie, Senior Systems Librarian, Project Manager, Homeland Security Digital Library, Naval Postgraduate School, 411 Dyer Rd., Monterey, CA 93943
Phone: (831) 656-3342, Email: lgassie@nps.navy.mil

1) Please describe in (one paragraph) the nature of the problem or challenge you are facing?

We are collecting a lot of content related to homeland security into a digital library. Although the content is mainly open source, because of its aggregation and availability in one central site, the aggregated content has intelligence value to those who may wish to do harm to the United States, and as such access should be restricted. On the other hand we would like to place as much information as possible in the hands of the public especially if those documents that are of public interest, and would be extremely useful for scholarship and research. The question is: how does one determine at what level is aggregation of open information a security risk, and how does one design intelligent tools/processes to identify content that may be put on a public site vs. one that would be put within a restricted site.

2) What business forces are making this problem especially critical or important? (for example, new regulations? Changes in expectations for services provided?)

Since September 11, 2001 a lot of government agencies have pulled data from their public website; there have been discussions on how one determines what is considered sensitive information in terms of national security vs. the loss of information useful for public interest, research and education.

3) Describe what you are currently doing to address the problem/challenge (2-5 bullets)

Discussed guidelines and policies on what should be considered public access, and what should be restricted. Have these guidelines reviewed by the appropriate security personnel to determine what is appropriate for posting on a public site. Problem is in getting access to senior folks who has the experience to provide this guidance, and that this issue is recurring.

4) In your ideal world, how could the advances in information management and technology help solve your problem?

Expert systems, using artificial intelligence, rules, ontologies, etc. that can map the knowledge in the heads of these senior staff so that we can re-apply these rules instead of asking them to make decisions repeatedly on a case-by-case basis.

5) Please list any websites that provide additional information or context for question #1.

http://library.nps.navy.mil/home/staff/lgassie/SLA03_files/frame.htm (check out the notes that go with each slide for context)

* **Problem Statement Submitted by:** FAA/DOT, Con Kenney, Chief Enterprise Architect, 202.267.3882, con.kenney@faa.gov

1) Please describe in (one paragraph) the nature of the problem or challenge you are facing?

An enterprise architecture is a set of information models of information systems and business processes. These models must answer the questions of many different stakeholders. The models must also lend themselves to being repurposed for questions beyond what the enterprise architects anticipated. As we learn how to make architecture more useful, new types of models will emerge that we need to connect with the old. To meet these requirements, I believe it is essential to be explicit about our assumptions about the purposes and parts of these models and their relationships to each other. The problem of assuring consistency across models is particularly important and difficult. Otherwise I worry that our stakeholders won't use or trust the enterprise architecture.

2) What business forces are making this problem especially critical or important? (for example, New regulations? Changes in expectations for services provided?)

The Clinger-Cohen act of 1996 provides the initial impetus for enterprise architecture, and the President's Management Agenda and the Federal Enterprise Architecture Program Management Office at OMB are bringing a lot of focus and energy in the executive branch. By tying agency budgets to enterprise architecture, OMB has signaled the importance of modeling information systems and business processes in eliminating redundant or unnecessary information technology investment. Additional drivers include reducing turnaround time for government-to-citizen and government-to-business transactions and improving information security.

3) Describe what you are currently doing to address the problem/challenge (2-5 bullets)

a. We're comparing applications and data base schemas for our National Airspace System (NAS) architecture, DOT EA, legacy EA, and the Federal Enterprise Architecture Management System used by OMB.
b. We've developed a set of use cases and plan to keep adding use cases as we discover new requirements.
c. We're sharing best practices in enterprise architecture from other Federal agencies such as the IRS and EPA.
d. We're researching ontologies developed elsewhere such as TopQuadrant's or the Business Process Ontology (Dieter Jenz).
e. We're urging researchers at CMU and USC to look at ontology for enterprise architecture.

4) In your ideal world, how could the advances in information management and technology help solve your problem?
The development and management of information systems requires mastery of a huge amount and range of information, far beyond any human being's capacity. I'm looking for an integrated toolset for business process modeling, architecture, and software development so we can define our requirements in terms of process and trace those requirements throughout an electronic lifecycle. The tools would support the encoding of business rules and policy that project teams often violate because of ignorance. Combined with the appropriate methodology and management controls these tools could enable us to reduce the cost

of complying with laws, regulations, and policies. We could also reduce the incidence of non-compliance by detecting and preventing deviations from policies and standards. Reuse of artifacts across the length of the software development lifecycle is another gain that could be achieved by the right set of models, tools, processes, and controls.

5) Please list any websites that provide additional information or context for question #1.

* **Problem Statement Submitted by:** Timothy N. West, Chair, Intelligence Community Metadata Working Group, Defense Intelligence Agency (DIA/DS), (202)231.2256, Timothy.West@dia.mil

1) Please describe in (one paragraph) the nature of the problem or challenge you are facing?

A vital component of homeland security and overall defense of the Nation is the ability of the US Government intelligence community to share information among dozens of Agencies, Military Services, and Combatant Commands (e.g., Central Command (USCENTCOM)). This shared information must be able to present a common perspective of the new asymmetric threat to our country and provide decisionmakers the ability to quickly find actionable information. Seven key characteristics describe the intelligence community's objectives for bringing semantic richness to actionable information:

1. data interoperability – speak a common language
2. data precision – expressly focused search or discovery
3. data recall – specific relevance
4. data consistency – repeatable results
5. data confidence – completeness and trust
6. data reuse – focused dissemination and constructive repurposing
7. data productivity – reduced fusion effort

2) What business forces are making this problem especially critical or important? (for example, New regulations? Changes in expectations for services provided?)

East Timor, New York City, the Pentagon, Afghanistan, Iraq, Liberia, . . .and most importantly, the next crisis event that we may not have yet even considered!

3) Describe what you are currently doing to address the problem/challenge (2-5 bullets)

- Defining and prescribing intelligence community-wide metadata and XML standards
- Insuring intelligence community standards are compatible with DOD, USG, and international standards
- Documenting intelligence community implementation and resource plans for applying these standards
- Understanding the role of XML in search, discovery, and exploitation activites
- Understanding the impact of XML on search, discovery, and exploitation tools

4) Please list any websites that provide additional information or context for question #1.

http://www.xml.saic.com/icml/

* **Problem Statement Submitted by:** National Defense University/ DOD: Luis G. Kun, Ph.D. Professor of Systems Management, IRM College, kunl@ndu.edu, 202-685-2786. (Course manager for Homeland Security & course manager for Intelligent Agents).

1) Please describe in (one paragraph) the nature of the problem or challenge you are facing?

Background: As presented last October [1], different terrorist threats to the public health are being addressed by different agencies, i.e. CDC/NIH (biological, chemical, etc.), EPA (water, air, etc.), DOE (nuclear, radiological, etc.), USDA (livestock, plants/food, etc.), DOC/HLS (cyber security), etc. Not only each agency has its own semantic/vocabulary but when trying to create a unified Public Health Information Infrastructure (multiple lanes in the same highway where they all address Public Health from their respective area of expertise), these vocabularies need to be "adjusted and/or translated" depending if the reader is for example someone from the Department of Health and Human Services or the Department of Homeland Security.-

2) What business forces are making this problem especially critical or important? (for example, New regulations? Changes in expectations for services provided?)

Question: Homeland Security and/or Public Health require the integration of information from multiple stakeholders. How can semantic technologies help resolve the issue of interpretation of the integrated information, depending on who the stakeholder analyzing it may be? [For example, people doing intelligence may not be from the same department that the one that collected the data therefore their interpretation may be different.]

3) Describe what you are currently doing to address the problem/challenge (2-5 bullets)

Personally interested in addressing this issue through the use of Intelligent Agents for the classes I teach
.

4) In your ideal world, how could the advances in information management and technology help solve your problem?

Intelligent Agents could be easily taught to differentiate who the audience reading the information is and therefore apply rules that would predetermine the correct semantic interpretation.

5) Please list any websites that provide additional information or context for question #1.

Reference: [1] Kun, L.G.: Editorial: "Homeland Security: The possible, the probable and the perils of Information Technology" IEEE Engineering in Medicine and Biology Magazine, Volume 21, No 5. Special Issue on Bioterrorism, Co Guest Editors: Kun & Laxminarayan, September-October, 2002.

Problem Statement Submitted by: Phillip B. Grove, USDA, Manual's Unit, Phone: (240) 629-1936, Philip.B.Grove@aphis.usda.gov

1) Please describe in (one paragraph) the nature of the problem or challenge you are facing?

Locating people and other resources to solve business problems in a timely manner is a primary goal of any organization. Our challenge is achieving this goal in a chaotic environment where responsibilities, knowledge, and resources are constantly shifting and difficult to identify.

2) What business forces are making this problem especially critical or important? (For example, new regulations? Changes in expectations for services provided?)

The process of regulating imports and exports requires quick decision making; however, the consequence of making a decision too quickly can have a devastating effect on agricultural commodities by letting a disease enter the United States and destroy a market.

3) Describe what you are currently doing to address the problem/challenge (2-5 bullets)

- An early effort to develop a knowledge and skills database failed because it could not be maintained in the current business infrastructure.
- Subject-Matter-Experts are identified to help in the development of operational manuals; however, their contact information is not maintained after the publication of the manual.
- We are currently looking at the "category" function of Lotus Notes to develop a database to identify decision makers, Subject-Matter-Experts, and program support personnel.
-

4) In your ideal world, how could the advances in information management and technology help solve your problem?

Individuals would be responsible for updating a skills and knowledge database with their personnel information. They would be automatically notified to update the database and fill in the required fields that would be validated on the fly. Some fields would have entries restricted to a pick list. The database would be fully searchable to identify decision makers for operations, or policy; Subject-matter-experts for general knowledge or as accomplished performers, and to identify personnel who are responsible for specific tasks in activity areas of the organization.

5) Please list any websites that provide additional information or context for question #1.

None

Problem Statement Submitted by: Moonja Kim, Ph.D., Assistant to Chief Knowledge Officer, Knowledge Management Center Defense Contract Management Agency, 6350 Walker Lane, Alexandria, VA 22310, (703) 428-1483, Moonja.Kim@dcma.mil

1) Please describe in (one paragraph) the nature of the problem or challenge you are facing?

Our agency is starting an effort for Enterprise Content Management, and initial effort is generating all words and phrases with their definitions so that various groups of people can review these if they are acceptable words to be included in the taxonomy. We are going to use some tools to develop enterprise taxonomy to be used for content management with a good search engine. The challenge we have is that there are too many tools and it's difficult to know which one will do the best job for our situation.

2) What business forces are making this problem especially critical or important? (for example, new regulations? Changes in expectations for services provided?)

Our agency is trying to serve Defense that is transforming, so transformation is a big initiative we have. The transformation is to move toward Customer-Centered Culture, and a good content management will provide better services to our customers. So the taxonomy development is the initial step for the transformation.

3) Describe what you are currently doing to address the problem/challenge (2-5 bullets)

- Searching a software tool to develop glossary from many hyperlinked documents that describes all business processes we have at our agency.
- Planning to try "Concodance" to develop Glossary
- Planning to interview various functional groups about words and phrases for specific functional areas for their approval of usage and definitions.
- Then use Masterfile by ConcordUSA to import the taxonomy to be used for search within Blackboard supported communities of practice
- We may be switching Blackboard to Verity or some other software later

4) In your ideal world, how could the advances in information management and technology help solve your problem?

If there is one tool that will perform text analysis and develop taxonomy that will be acceptable to the users and also incorporate a good search engine to help to locate only the relevant information when needed. I am not sure if there exist such tool yet.

5) Please list any websites that provide additional information or context for question #1.

Our home page is http://www.dcma.mil The hyperlinked documents we want to analyze is under policy, it's called "Policy:DCMA One Book"

Problem Statement Submitted by: Owen Ambur, Co-Chair, XML Working Group, http://xml.gov, Owen_Ambur@fws.gov

1) Please describe in (one paragraph) the nature of the problem or challenge you are facing?

The Emerging Technologies Subcommittee (ET S/C) has been charged with developing a process whereby the information technology innovation life-cycle can be managed on a Governmentwide basis. The driving force is the inability of agency chief information officers (CIOs) to respond effectively to myriad vendors and other proponents of technology components, particularly those that are new, innovative, and perhaps untested and unproven in practical application. The expectation is that the process will help to structure such input for better coordinated and more productive response, in support of the eGov initiatives and within the framework of the Federal Enterprise Architecture (FEA). The desired outcome is the well-coordinated acquisition of logically separable technology components for potential Governmentwise usage. However, the initial challenge is to clearly ***identify*** proposed emerging technology components in terms that are unambiguous and meaningful to government decision-makers who may have limited technical expertise. Thus, it will be interesting to see whether this conference and "semantic technologies" in general can make a siginificant contribution to meeting this challenge.

2) What business forces are making this problem especially critical or important? (for example, new regulations? Changes in expectations for services provided?)

The government is spending billions of dollars on information technology systems that do not efficiently and effectively manage and share records having the attributes outlined in ISO 15489. Moreover, many of those systems are: a) redundant as well as failing to b) comply with applicable voluntary consensus standards for interoperability, and c) take advantage of the potentials of newer, more innovative technologies.

3) Describe what you are currently doing to address the problem/challenge (2-5 bullets)

I have drafted a proposal for more effectively managing the information technology innovation life-cycle and will be pursuing its implementation through the Emerging Technologies Subcommittee of the CIO Council's Architecture and Infrastructure Committee.

4) In your ideal world, how could the advances in information management and technology help solve your problem?

Proposed emerging technology components could clearly be understood and distinguished from each other by the average person, as well as by others who think in different terms. Moreover, individuals as well as .gov agencies could easily associate themselves with and thereby form communities of interest around emerging technology components that may better serve their needs and wishes.

5) Please list any websites that provide additional information or context for question #1.

My draft proposal is available at http://xml.gov/draft/etLifeCycle.htm and my strawman draft for the elements of the first stage of the process is at http://xml.gov/draft/etsc300form1stStage.pdf

Problem Statement Submitted by: Lillian W. Gassie, Senior Systems Librarian, Project Manager, Homeland Security Digital Library, Naval Postgraduate School, 411 Dyer Rd., Monterey, CA 93943
Phone: (831) 656-3342, Email: lgassie@nps.navy.mil

1) Please describe in (one paragraph) the nature of the problem or challenge you are facing?

We are using a metasearch tool to perform simultaneous searching of multiple databases in a digital library. Because metasearching tend to be more successful with broad keyword searches, users are faced with high recall. We want a search query to expand to synonyms identified as pertinent to a search term so that users do not need to know all the possible ways to enter a search term. In return, we want to enable intelligent filtering of search results using categories that are meaningful to our users, in our case, in the domain of homeland security. Going one step further, we want these categories to display in the terms that are familiar to different groups of users, e.g. users from the Coast Guard will see the same results displayed in categories that are meaningful to the Coast Guard, while those at FEMA would see the same results, but filtered using FEMA terms.

2) What business forces are making this problem especially critical or important? (for example, new regulations? Changes in expectations for services provided?)

Increasing number of databases, availability of metasearch tools, which alleviates the user having to repeat a search query in different databases. The problem of keyword searching is well-known in that most database search engines are not intelligent nor intuitive enough to recognize variant word forms, nor synonyms.

3) Describe what you are currently doing to address the problem/challenge (2-5 bullets)

Creating a taxonomy on homeland security that ideally would cut across agencies.

4) In your ideal world, how could the advances in information management and technology help solve your problem?

Technology can assist in suggesting new terms, automatically link or invoke like/similar terms during search and retrieval, and based on knowledge of the user (e.g. user's background, agency affiliation, or domain expertise), will display results in the language that the users understand.

5) Please list any websites that provide additional information or context for question #1.

http://library.nps.navy.mil/home/staff/lgassie/SLA03_files/frame.htm (check out the notes that go with each slide for context)

Problem Statement Submitted by: Duane Degler, consultant to SSA's Usability Center. Duane.Degler@ssa.gov. Telephone: (work) 410-966-8930; (cell) 301-523-8819.

Formal taxonomies go out of date, drift from the terms/jargon/acronyms that people use, and eventually become a barrier to understanding and content access. The problem multiplies with the difference in the terms used by "average citizen," those used by internal staff, and those found in documents. Language – and its semantic representation – is highly dynamic. Taxonomies and ontologies become institutionalized and very resistant to change – yet if they change too erratically, they risk confusing the user. Proponents of the semantic web say technology can address these issues, but it is both a human and a technological challenge. Technology can't maintain ontologies without human interpretation and judgment, yet there is rarely a budget or mandate for knowledgeable, dedicated staff to maintain and refresh ontologies, to update/retag content, and to communicate changes across the organization. So who should do this? How can it be made easier, so more people can do this? Can technology help, or is it trapped in "document speak" rather than "human speak"? Is automating the ontology refreshing process too risky?

The loss of knowledgeable staff to the "retirement wave" is bordering on a crisis for knowledge retention in many agencies. Increasing "self service" via the web changes staff contact with the public, reducing the ability to stay "in sync" semantically. The amount of content is increasing rapidly as more people are able to author/publish formal content, and generate informal content electronically. Automated content management and semantic tools are still too rudimentary to support the few people who need to address the dynamic "semantic drift." IT priorities are focused on service delivery and operations issues, not maintaining semantic relevance. Their answer is often "it's a training issue."

What small steps are we taking toward addressing these issues?

- Assessing user needs (both citizen and staff) and their information expectations using formal user-centered design methods within the software development lifecycle (to identify problems)
- Building an SME/author-maintained topic map application that supports non-technologists creating and updating semantic maps and relationships to content
- Raising awareness with agency executives and managers, and opening informal conversations with other government problem owners and with vendors

What is the role of technology products and tools in a blended approach to the problem? A "blended" approach would need to incorporate things like: simplifying users' interactions with government and its content; ongoing dynamic analysis of a wide range of content; pattern identification in structured and unstructured repositories; exposing/representing the drifts to human problem owners for subjective judgment and management; facilitated collaboration between people to support keeping semantic representations/maps fresh; allowing easy/dynamic feedback and dialog with users whose needs are not met; and particularly tools that keep users and subject experts – and their knowledgeable, subjective insights – in control of the process.

TopQuadrant Technology Briefing

Semantic Technology

Version 1.1

September 2003

Table of Contents

Tables

1. Semantic Technology

1.1 What is Semantic Technology?

We define semantic technology as a software technology that allows the meaning of and associations between information to be known and processed at execution time. For a semantic technology to be truly at work within a system there must be a knowledge model of some part of the world that is used by one or more applications at execution time.

How is it distinguished from more conventional applications?

- Semantic technologies represent meaning through connectivity. The meaning of terms, or concepts, in the model is established by the way they connect to each other.
- A semantic model expresses multiple viewpoints.
- Semantic models represent knowledge that is in the world in which the system operates. Several interconnected models could be used to represent different aspects of the knowledge. The models are consultable (accessible) by applications at run time.
- A semantic application uses knowledge models in an essential way as part of its operation. Use of a model is often referred to as "reasoning over the model". Reasoning can range from a very simple process of graph search to intricate inferencing over the model.
- Semantic applications are thin because they work with “smart” data. All the business rules logic is held in the models shared across applications.

Figure 1 shows a simplest form of a semantic model, a taxonomy. The model describes government concepts that are part of Federal Enterprise Architecture (FEA). In a taxonomy connections between terms exist, but are not named. Therefore, the structure itself becomes a way to identify the nature of relationships. Taxonomies are hierarchies that establish “parent-child” relationship between its concepts.

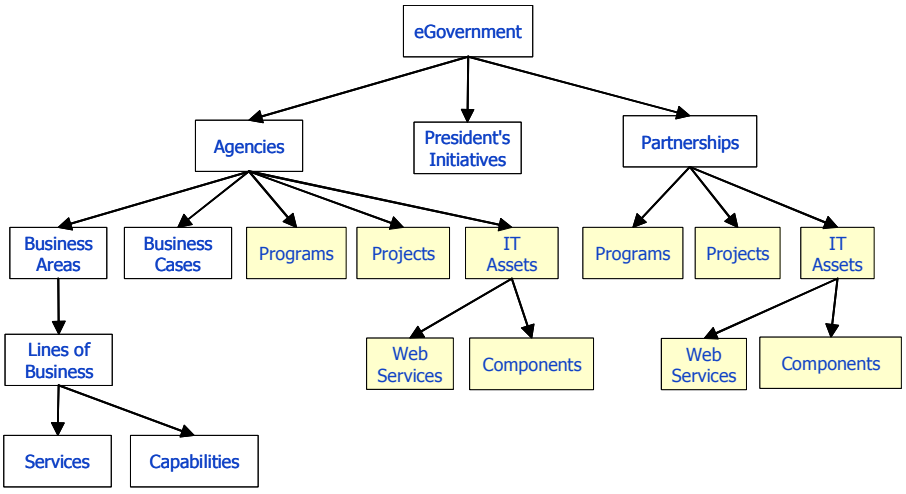


Figure 1: Example of a Taxonomy for e-Government

Because of the hierarchical nature of a taxonomy, some concepts have to be grouped under more then one category.

Figure 2shows a richer model where relationships are explicitly named and differentiated. This model is called an ontology. Because the relationships are specified there is no longer a need for a strict structure. The model becomes a network of connections. New knowledge could be inferred by examining the connections between concepts. For example, the model below could be used to infer that a specific IT components is a way to deliver support for a given president's initiative.

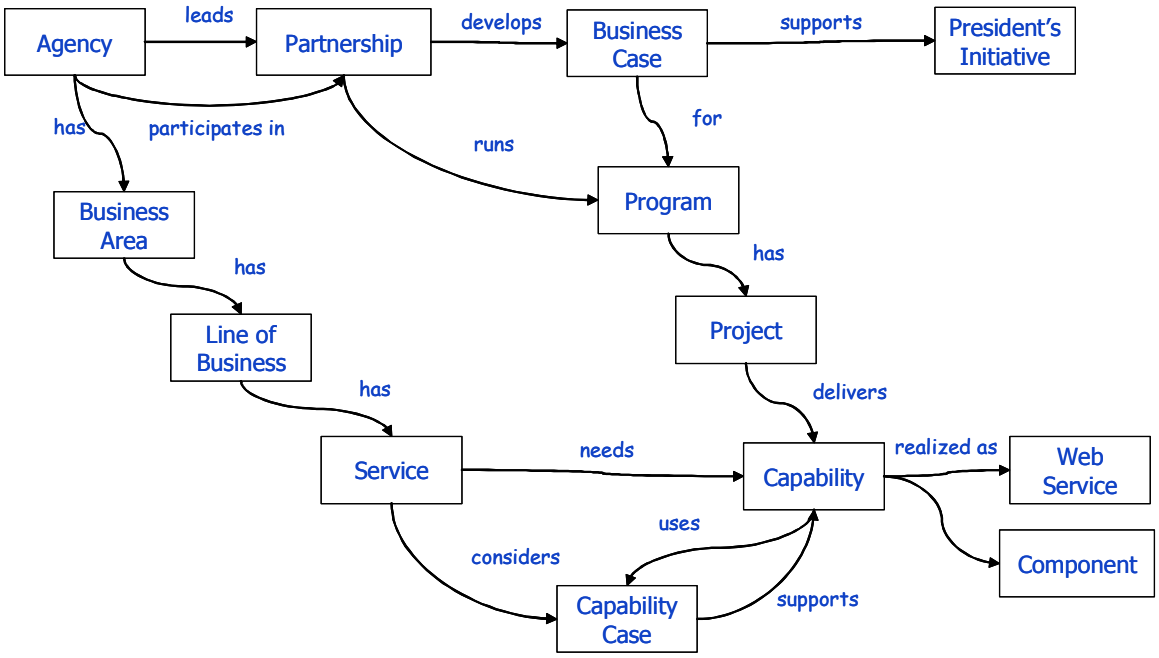


Figure 2: Part of the FEA Capabilities Manager Ontology Model

Simpler ontologies are just networks of connections; richer ontologies include rules and constraints governing these connections as illustrated in Figure 3. The model shows how business cases have to be constructed with compliance to the FEA models. A simple rule for checking baseline values of measurement indicators is illustrated. What the rule says is that the baseline values of all measures must be greater than or equal to the baseline values of their respective indicators.

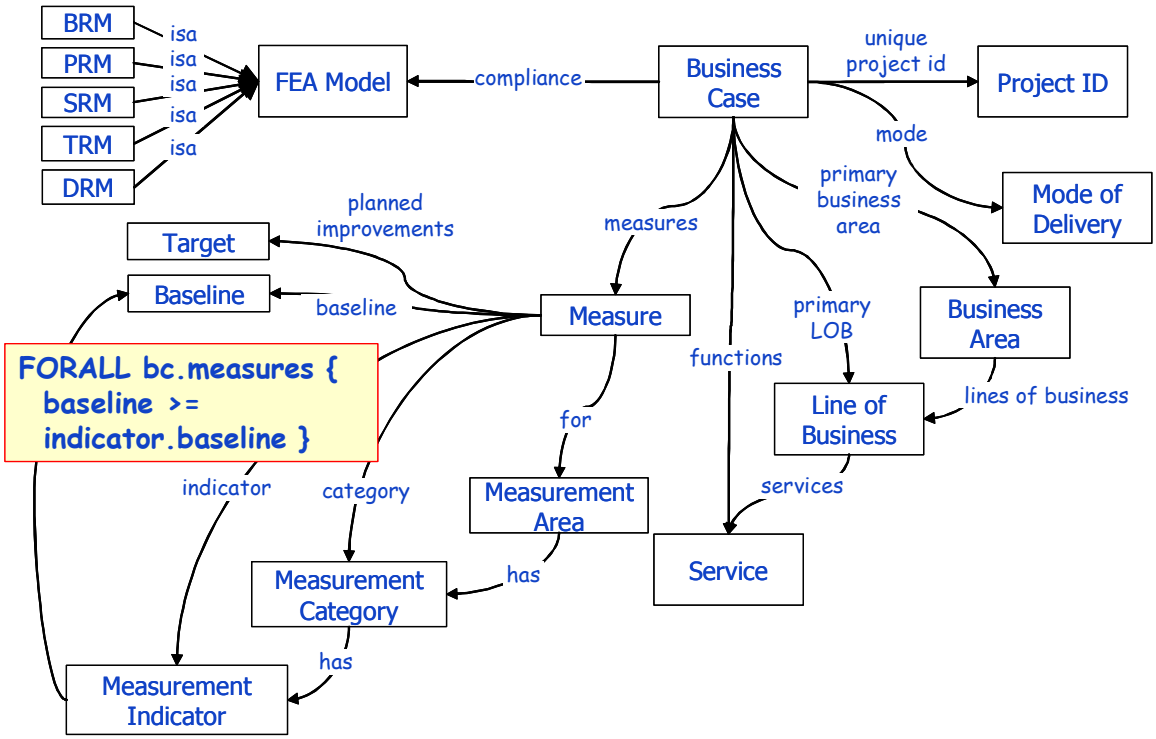


Figure 3: Example of a Rule for Exhibit 300 Measures

An example, from the FEA Project Management Office "Additional Guidance On The FE-Related Requirements in OMB Circular A-11" document, is shown in the table below.

Table 1: Example of a Business Measure Baseline

Fiscal Year	Measurement Area	Measurement Category	Measurement Indicator	Baseline	Planned Improvements	Actual Results
2005	Mission and Business Results	International Affairs and Commerce	# of US Exporters entering new market	5,386		

1.2 How Knowledge Models are different from other Software Models

A model describes how concepts and phenomena are similar, and how they differ, what is called the commonality and variability of concepts in a chosen area of interest, sometimes also referred to as a domain of discourse. The most commonly used models in software engineering are object and data models:

The Object Model in an object-oriented program is a networked data set that describes the system itself. In an object model, classes high in the hierarchy express properties that are shared by many system elements; classes low in the hierarchy describe properties that are specific to small sets of elements. Therefore, it is a model that reflects and describes properties and functions of a specific system.

The Data Model describes the world outside of the system. Many applications can share the same database, but in reality the schema of the data is typically fine-tuned to the needs of specific application. In a data model, each table in the schema dictates what this collection of records has in common; another schema denotes this for other records. Differences are represented both by individual records, as well as record types. The relationships are held in special index tables and are not explicitly defined.

Semantic models are intended as a way for different agents (applications and/or people) to interoperate and to share meaning. Unlike object models they describe the world that is outside of any of the application that uses the model. Furthermore, the variations and commonalities semantic models represent are not of a single entity or stakeholder. By definition semantic models support multiple viewpoints. This makes them especially suitable for solving interoperability problems.

1.3 Standard Languages for Knowledge Modeling

What languages can be used for knowledge or semantic modeling? By now, we all have heard of HTML and XML. A few important developments preceded HTML, but many have occurred since XML became popular. What we are witnessing today is the emergence of standards for the semantic WEB. These and other important influences from AI, Software Engineering and Process Modeling make up what we are illustrating in Figure 4 as "The Tree of Knowledge Technologies"

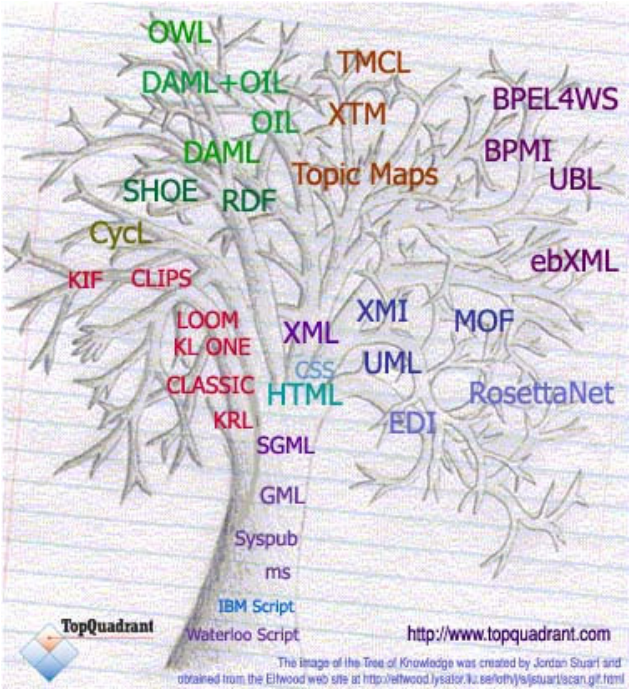


Figure 4: Tree of Knowledge Technologies

1.3.1 The History and the Current State

The current state of the art on representing and using ontologies has grown out of several efforts that started in the 1980s. Back then, KL-ONE was the most influential of the frame-based representation languages; it allowed for the representation of categories and instances, with inheritance of category properties, and a formal logic for expressing the meaning of properties and categories. At about the same time, rule-based systems were a promising technology. The NASA-sponsored C-Language Integrated Production System (CLIPS) became a de-facto standard for building and deploying rule-based systems.

The Knowledge Interchange Format (KIF), and it accompanying translation tool Ontolingua, were developed to allow knowledge to be shared among these different efforts, and provided the capability to translate knowledge bases in one representation language to another. These languages were ahead of their time. As a result, they have remained largely within the purvey of academia, gaining little commercial support.

With the advent of the World Wide Web, and the acceptance of XML as a de-facto standard for representation of information on the web, ontology efforts joined in. An early project at the University of Maryland produced SHOE, a system for expressing ontologies in XML, and marking up web pages with ontology-based annotations. Many of the ideas from this work made it into the World Wide Web Consortium (W3C) proposal for the Resource Description Framework (RDF) Language.

The DARPA Agent Markup Language (DAML) is built on RDF providing particular logical relationships that standardize the semantics of inferences that can be made over the information in a resource description. The DAML effort drew much of the formal semantics for its logical approach from a parallel effort called OIL (Ontology Inference Layer), which encoded the semantics of Description Logic into an XML-based language. The joining of the two efforts resulted in DAML+OIL language. It allows for a strict interpretation of the statements, so that reasoning agents can collaborate in their use of ontologies. DAML+OIL became a foundation for W3C Web Ontology Language (OWL).

While we have seen some use of UML as a knowledge language and a few MOF (Meta Object Framework) based integration solutions, RDF-based languages have the most potential for success. Table 2: provides a high level view of standards and an indication of the marketplace adoption.

Table 2: View of Knowledge Modeling Standards and Marketplace Adoption

	KIF/OKBC/ CG/CycL	UML	Topic Maps/XTM	RDF(S)	DAML+OIL	OWL
Description	*Legacy KR Languages*	*Universal Modeling Language*	*Topic Maps/XML Topic Maps*	*Resource Description Framework*	*DARPA ML + Ontology Inference*	*Web Ontology Language*
Governance	ANSI American Natio and others	OMG	ISO Internat Organi Standar	W3C	DARPA	W3C
Years since proposed	*>10*	*>5*	*>5*	*>3*	*>2*	*>1*
Commercial Support (KL*)	□	□	□□	□□□□	□□□	□
Open Source Support	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	

□ 2 or less vendors
□□ 5 or less vendors
□□□ 10 or less vendors
□□□□ > 10 vendors
* - knowledge language

1.3.2 XML-based Knowledge (Ontology) Modeling Languages

XML is being used to represent hierarchies of data. To go beyond hierarchies and simple taxonomies requires different kind of standards. The standards below represent convergence of conceptual modeling (AI heritage) and mark up languages (HTML and XML heritage):

ISO/IEC 13250 Topic Maps

Topic Maps defines a method of using SGML to represent networks of concepts to be superimposed on content resources (documents of various types), providing a means to represent, navigate, and query the network itself, rather than the full text of a document collection. ISO Topic Maps is an approach for representing topics, their occurrences in documents, and the associations between topics.

XTM is an XML representation of Topic Maps.

Standard Status = Released

There are 3 commercial vendors that offer Topic Maps tools. The Topic Maps standard has been developed in an effort parallel to RDF-based ontology languages. Convergence is not likely, but interoperability is possible. Several approaches for mapping between Topic Maps and RDF have been published. Topic Maps are applicable for building indices over *information objects* that represent unstructured information. We do not recommend Topic Maps for semantic integration in enterprise systems and across decentralized knowledge spaces.

RDF/S

The **Resource Description Framework** [W3C-RDF] defines a model and XML syntax to represent and transport metadata. RDF integrates a variety of applications from library catalogs and world-wide directories to syndication and aggregation of news, software, and content to personal collections of music, photos, and events using XML as interchange syntax. The RDF specifications provide a lightweight ontology system to support the exchange of knowledge on the Web.

Standard Status = Released

The Resource Description Framework (RDF) is a foundation for representing and processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web.

RDF Schema, RDF's vocabulary description language, is an extension of RDF. It provides mechanisms for describing groups of related resources and the relationships between these resources. RDF Schema does the same thing for RDF that DTD and XML Schema do for XML.

Standard Status = Draft

RDF is making good inroads in terms of vendor support. Commercially available tools range from development environments to RDF databases to semantic integration and search/categorization solutions.

DAML+OIL and OWL

DAML + OIL is a semantic markup language for Web resources. It builds on earlier W3C standards such as RDF and RDF Schema, and extends these languages with richer modeling primitives. DAML+OIL was built from the original DAML ontology language DAML-ONT (October 2000) in an effort to combine many of the language components of OIL.

A DAML+OIL knowledge base is a collection of RDF triples. DAML+OIL prescribes a specific meaning for triples that use the DAML+OIL vocabulary.

The W3C Web Ontology Working Group (WebOnt) has been tasked with producing a web ontology language extending the reach of XML, RDF, and RDF Schema. This language, called OWL, is based on the DAML+OIL web ontology language. The only substantive changes from DAML+OIL are the removal of qualified number restrictions, the ability to directly state that properties can be symmetric; and the removal of some unusual DAML+OIL constructs, particularly restrictions with extra components. There are also a number of minor differences, including a number of changes to the names of the various constructs.

There are three levels of OWL defined (OWL Lite, OWL DL and OWL Full) with progressively more expressiveness and inferencing power. These levels were created to make it easier for tool vendors to support a specified level of OWL.

Standard Status = Recommendation released in August, 2003.

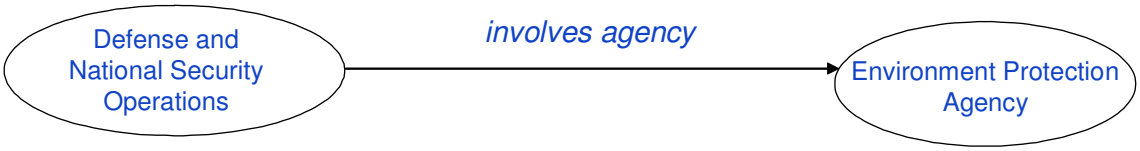
DAML+OIL and OWL both depend on RDF/S semantics. Thus, the development of these standards is presently a fairly interlocking sequence. Today a number of vendors offer

DAML+OIL support. As OWL matures we expect to see them moving from DAML+OIL to OWL.

Different approaches to semantic technology are distinguished by the different ways knowledge representation languages express the connections between concepts:

- Taxonomies and Thesauri have very simply connection
- RDF and Topic Maps have somewhat more complex ones:
  - RDF has very formal connections,
  - while Topic Maps have intuitive ones
- DAML and OWL have very powerful logical connections

A small example below shows a part of the FEA knowledge model, an RDF statement describing one of the relationships between Defense and National Security Operations and Environmental Protection Agency.




```
<FEA:Service rdf:about="&FEA;Anti-Terrorism"
             rdfs:label="Anti-Terrorism">
             <FEA:service_of rdf:resource="&FEA;Defense and National Security
Operations"/>
             <FEA:involves_agency rdf:resource="&FEA;Department of Commerce"/>
             <FEA:involves_agency rdf:resource="&FEA;DoJ"/>
             <FEA:involves_agency rdf:resource="&FEA;Environment Protection Agency"/>
             <FEA:involves_agency rdf:resource="&FEA;FEMA"/>
             <FEA:involves_agency rdf:resource="&FEA;General Services Administration"/>
             <FEA:involves_agency rdf:resource="&FEA;State"/>
             <FEA:involves_agency rdf:resource="&FEA;Transportation"/>
             <FEA:involves_agency rdf:resource="&FEA;Treasury"/>
</FEA:Service>
```

Figure 5: RDF Example

Applications of Semantic Technology

Semantic technology can be applied in a number of different situations. The key to getting value out of it is picking the most appropriate application area. The table below lists a number of capabilities known to be successfully delivered by semantic technology. For each, we identify the reason why semantic technology is a good fit for implementing the capability. Alternative technical approaches are also described. The common downside many of the alternative approaches share is lack of scalability and flexibility needed to support the solution as the new information sources, new users and new applications are added or new requirements become important. Another words, they are simple to implement and work well in well bounded situations, but do not grow well. One exception is neural networks and other machine learning approaches. In many cases this technology is complementary to semantic technology, a knowledge representation approach, and could be used together very successfully.

Therefore, one of the key success criteria for implementing semantic technology is picking an area where the situation is fairly complex and/or extensibility of the solution is important. On the other hand, such situations are often perceived by companies as mission-critical. The tolerance to risk associated with new technology is low. A number of success stories are becoming available from early adoption are paving the road to broader adoption.

Table 3: Semantic Capabilities

Capability Intent	Semantic Technology Fit	Other Approaches
Answer Engine		
To provide a direct reply to a search questions as opposed to returning a list of relevant documents. It interprets a question asked in a natural language, checks multiple data sources to collect knowledge nuggets required for answering the question and may even create an answer on the fly by combining relevant knowledge nuggets.	Interpretation of questions using domain knowledge. Aggregation and composition of the answer. Also see Generative Documentation below.	Identifying frequently asked questions and posting answers to them.
Automated Content Tagger		
To provide semantic tags that allows a document or other work-product to be "better known" by one or more systems so that search, integration or invocation of other applications becomes more effective.	Tags are automatically inserted based on the computer analysis of the information, typically using natural language analysis techniques. A predefined taxonomy or ontology of terms and concepts is used to drive the analysis.	Machine learning approaches based on statistical algorithms such as Bayesian networks.
Concept-based Search		
To provide precise and concept-aware search capabilities specific to an area of interest using knowledge representations across multiple knowledge sources both structured and un-structured.	Knowledge model provides a way to map translation of queries to knowledge resources.	Dictionary of synonyms and domain specific jargon could provide an approximation to concept-based search.

Capability Intent	Semantic Technology Fit	Other Approaches
Connection and Pattern Explorer		
Discover relevant information in disparate but related sources of knowledge, by filtering on different combinations of connections or by exploring patterns in the types of connections present in the data.	Inferences over models to identify patterns using the principles of semantic distance.	Statistical algorithms such as Bayesian networks. Technologies could create visualization of complex data, thereby facilitating pattern discovery by humans or potentially by machine vision algorithms.
Content Annotator		
Provide a way for people to add annotations to electronic content. By annotations we mean comments, notes, explanations and semantic tags.	Knowledge model is used to assist people in providing consistent attribution of artifacts.	Using fix templates for each type of artifact.
Context-Aware Retriever		
To retrieve knowledge from one or more systems that is highly relevant to an immediate context, through an action taken within a specific setting -- typically in a user interface. A user no longer needs to leave the application they are in to find the right information.	Knowledge model is used to represent context. This “profile” is then used to constrain a concept-based search.	Machine learning techniques based on statistical algorithms could be used to “understand” the context.
Dynamic User Interface		
To dynamically determine and present information on the web page according to user's context. This may include related links, available resources, advertisements and announcements. Context is determined based on user's search queries, web page navigation or other interactions she has been having with the system.	A model of context and a memory of activities are used to control UI generation.	Using XML interaction mark up languages and XSLT against a set of predetermined dialog choices.
Enhanced Search Query		
To enhance, extend and disambiguate user submitted key word searches by adding domain and context specific information. For example, depending on the context a search query "jaguar" could be enhanced to become "jaguar, car, automobile", "jaguar, USS, Star Trek", "jaguar, cat, animal" or "jaguar, software, Schrödinger".	Knowledge models are used to express the vocabulary of a domain.	A dictionary of synonyms and domain specific jargon can be used.

Capability Intent	Semantic Technology Fit	Other Approaches
Expert Locator		
To provide users with convenient access to experts in a given area who can help with problems, answer questions, locate and interpret specific documents, and collaborate on specific tasks. Knowing who is an expert in what can be difficult in an organization with a large workforce of experts. Expert Locator could also identify experts across organizational barriers.	The profiles of experts are expressed in a knowledge model. This can then be used to match concepts in queries to locate experts.	Simple profile-based approaches using fixed templates. Alternatives usually give poor results because of the lack of support for determining semantic distance and semantic similarity.
Generative Documentation		
Maintain a single source point for information about a system, process, product, etc., but deliver that content in a variety of forms, each tailored to a specific use. The format of the document, and the information it contains, is automatically presented as required by each particular audience.	Knowledge model is used to represent formatting and layout. Semantic matching is a key component of the solution.	Manual repurposing of the information. Creation of special one-to-one repurposing programs.
Interest-based Information Delivery		
Filter information for people needing to monitor and assess large volumes of data for relevance, volatility or required response. The volume of targeted information is reduced based on its relevance according to a role or interest of the end user. Sensitive information is filtered according to the "need to know".	A profile of each user's interests is expressed in a knowledge model. This is then be used to provide "smart" filtering of information that is either attributed with meta-data or has knowledge surrogates.	Rules and collaborative filtering could be used for personalization.
Navigational Search		
Use topical directories, or taxonomies, to help people narrow in on the general neighborhood of the information they seek.	A Taxonomy that takes into account user profiles, user goals and typical tasks performed is used to drive a search engine. To optimize information access by different stakeholders, multiple inter-related taxonomies are needed. Taxonomies and ontologies are used to suggest related subjects.	
Product Design Assistant		
To support the innovative product development and design process, by bringing engineering knowledge from many disparate sources to bear at the appropriate point in the process. Possible enhancements to the design process that result include rapid evaluation, increased adherence to best practices and more systematic treatment of design constraints.	Knowledge models are used to express design constraints and best practices.	Expert systems.

Capability Intent	Semantic Technology Fit	Other Approaches
Semantic Data Integrator		
Systems developed in different work practice settings have different semantic structures for their data. Time-critical access to data is made difficult by these differences. Semantic Data Integration allows data to be shared and understood across a variety of settings.	A common knowledge model is used to provide one or more unified views of enterprise data. Typically this is done by using mapping. Rules are executed to resolve conflicts, provide transformations and build new objects from data elements.	One to one mappings and transformation of data sources.
Semantic Form Generator and Results Classifier		
To improve the data collection process and data input analysis by providing knowledge-driven dynamic forms.	A knowledge model is used to intelligently guide the user through data capture. The results are automatically classified and analyzed according to the model	Pre-defined forms.
Semantic Service Discovery and Choreography		
Service Oriented Architectures enable increased reuse of existing services and the dynamic automation of processes through service composition and choreography.	Knowledge models are used to enhance the functionality of service directories. Invocation methods, terminology and semantic description allow the dynamic discovery of services by machines.	
Virtual Consultant		
Offer a way for customers to define their individual goals and objectives, and then show them what products and services can help them meet those goals. Understanding customer's goals and requirements through a questionnaire or dialog establishes a profile that helps you communicate effectively with them now and in the future.	A knowledge model of users and their work within a domain is used to provide intelligent guided support of interactive sessions.	Canned dialogs and responses.

1.4 Application Architecture

How does semantic technology fit into overall architecture of business applications? Figure 6 depicts typical application architecture.

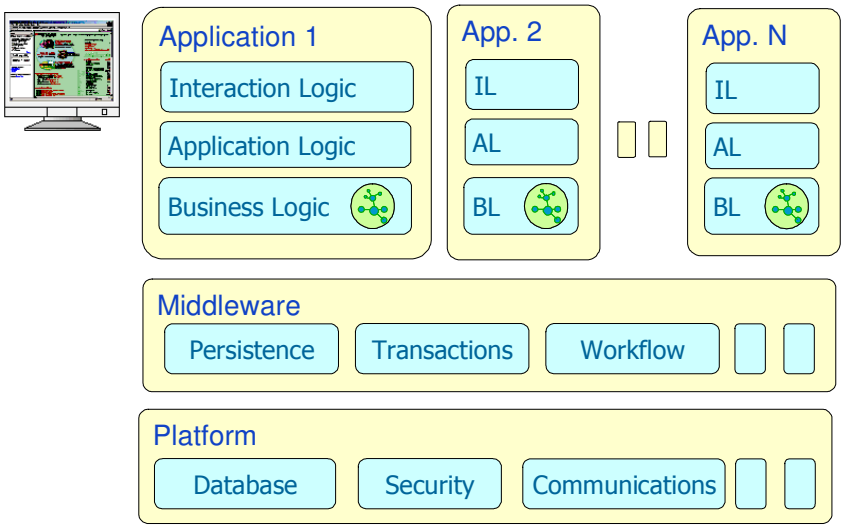


Figure 6: Typical Application Architecture

Semantic technology could be used to encapsulate business domain knowledge used by many applications. This means that the applications would become thinner as they no longer need to have their own representation of business logic. Instead they would need to have a way to consult a knowledge model. Such access is made possible through the use of semantic engines. Figure 7 shows a modified architectural view with each application having semantic interface (SI).

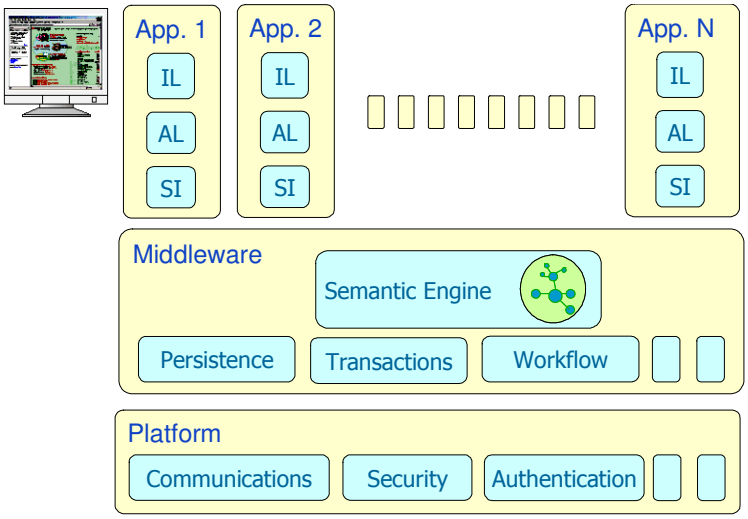


Figure 7: The role of Semantic Engine in the Application Architecture

This architectural approach ensures interoperability between diverse set of applications that operate in the same or related business domains. The interoperability is achieved by using a common set of models describing business concepts and their relationships as illustrated in Figure 8.

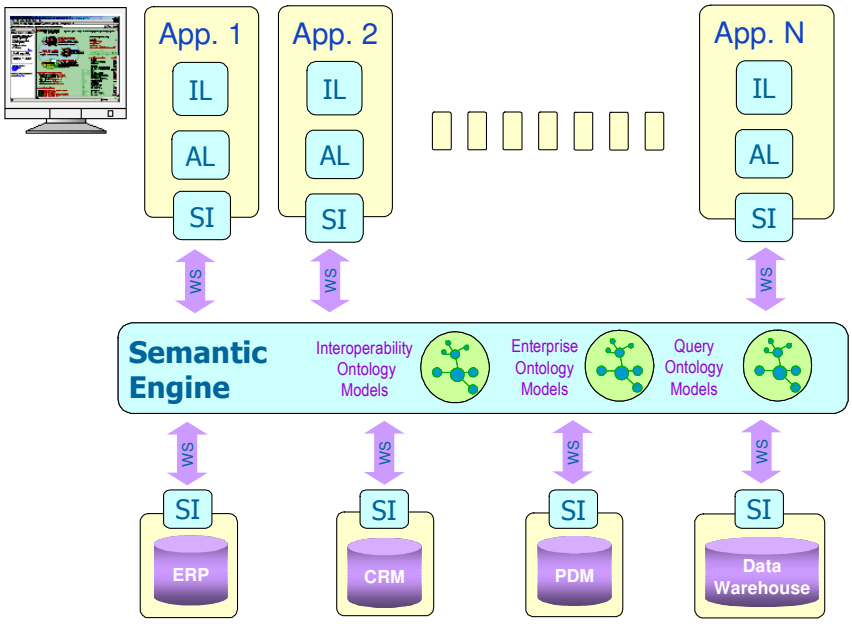


Figure 8: Architecture for Semantic Interoperability

2. Ontology Development Tool Support

2.1 Executive Summary

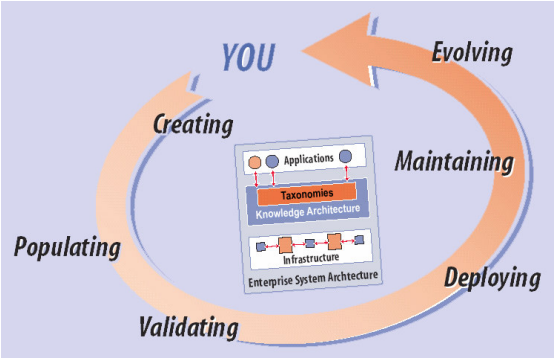
Knowledge models, or ontologies, are a necessary pre-condition to any semantic application. Therefore, the state of ontology development tooling is a key factor in adoption of semantic technology. Today some tool support is available for all stages of the ontology lifecycle. Many tools are still offered as research prototypes, but many others have begun to be commercialized (they are often commercial versions of their direct research counterparts). Standard compliance and support for RDF(S) and OWL is growing. However, due to the different forms of RDF(S), ontology tools still do not interoperate well.

Currently, ontology creation tools require their users to be trained in knowledge representation and predicate logic. The user interface and development paradigms of these tools are different from the standard application development tools. Support for semi-automation (for example, term extraction) is maturing. Support for collaborative authoring is still weak. To scale ontology-based applications from the pilot/prototype stage to enterprise-level implementations a new generation of tools is needed that will:

- Improve user interface for ontology building by leveraging familiar interfaces of widely used application development tools or MS Office applications
- Offer a server-based environment with support for consistency checking of interconnected ontologies
- Offer a collaborative environment for model review and refinement that does not require reviewers to be expert modelers
- Feature SOAP interfaces for ease of integration

2.2 Ontology Lifecycle and Tools

Ontology lifecycle spans from creation to evolution as shown in the picture below:



Tool support is available for all stages of the lifecycle:

Creating **-** This can be done from scratch, using a tool for editing and creating class structures (usually with an interface that is similar to a file system directory structure or bookmark folder interface). However, there is also a good deal of assistance available at this stage:

- Text mining can be used to extract terminology from texts, providing a starting point for ontology creation.
- Often, ontology information is available in legacy forms, such as database schemas, product catalogues, and yellow pages listings. Many of the recently released ontology editors import database schemas and other legacy formats i.e., Cobol copybooks.
- It is also possible to re-use, in whole or in part, ontologies that have already been developed in the creation of a new ontology. This brings the advantage of being able to leverage detailed work that has already been done by another ontology engineer.

Populating - This refers to the process of creating instances of the concepts in an ontology, and linking them to external sources:

- Ordinary web pages are a good source of instance information; so many tools for populating ontologies are based on annotation of web pages.
- Legacy sources of instances are also often available; product catalogues, parts lists, white pages, database tables, etc. can all be mined while populating an ontology.

Population can be done manually or be semi-automated. Semi-automation is highly recommended when a large number of knowledge sources exist.

Deploying - There are many ways to deploy an ontology once it has been created and populated:

- The ontology provides a natural index of the instances described in it, and hence can be used as a navigational aid while browsing those instances.
- More sophisticated methods, such as case-based reasoning, can use the ontology to drive similarity measures for case-based retrieval.
- DAML+OIL and OWL have capabilities for expressing axioms and constraints on the concepts in the ontology; hence powerful logical reasoning engines can be used to draw conclusions about the instances in an ontology.
- Semantic integration across all of the various applications is probably the fastest growing area of development for ontology-based systems.

Validating, Evolving and Maintaining - Ontologies, like any other component of a complex system, will need to change as their environment changes. Some changes might be simple responses to errors or omissions in the original ontology; others might be in response to a change in the environment. There are many ways in which an ontology can be validated in order to improve and evolve it; the most effective critiques are based on strict formal semantics of what the class structure means:

- Extensive logical frameworks that support this sort of reasoning have been developed, and are called *Description Logics*.
- A few advanced tools use automated description logic engines to determine when an ontology has contradictions, or when a particular concept in an ontology can be classified differently, according to its description and that of other concepts.
- These critiques can be used to identify gaps in the knowledge represented in the ontology, or they can be used to automatically modify the ontology, consolidating the information contained within it.

The task of ontology maintenance may require merging ontologies from diverse provenance. When this is the case tool support is important:

- Some tools provide human-centered capabilities for searching through ontologies for similar concepts (usually by name), and provisions for merging the concepts.
- Others perform more elaborate matching, based on common instances or patterns of related concepts.

2.3 Ontology Tools Survey

Many tools are still offered as research prototypes, but many others have begun to be commercialized (they are often commercial versions of their direct research counterparts). Current ontology tools require their users to be trained in knowledge representation and predicate logic.

Typically ontologies are built by highly trained knowledge engineers working with domain specialists or subject matter experts. In order to scale this approach in a large enterprise, a wider group of people must be able to independently perform some of the ontology lifecycle activities. They need to be able to create and modify knowledge models directly and easily, without the requirement for specialized training in knowledge representation, acquisition, or manipulation. A new generation of tools is needed that will:

- Improve user interface for ontology building by leveraging familiar interfaces of widely used application development tools or MS Office applications
- Offer a server-based environment with support for consistency checking of interconnected ontologies
- Offer a collaborative environment for model review and refinement that does not require reviewers to be expert modelers
- Feature SOAP interfaces for ease of integration

Table 1 below gives a list of representative tools available today, and a brief description of their capabilities. A number of commercial tools addresses multiple, sometimes all, stages of ontology lifecycle. When this is the case, we have placed the tools in the category(s) where their capabilities are the strongest.

There has been a significant growth in the number of ontology technology products; this report doesn't cover all the available tools. In composing this list we have selected the tools that:

- Support all or some of RDF(S)-DAML+OIL/OWL standard (or have committed to support in the very near future – by mid 2003).
- Have strong technical vision for ontology-based solutions
- Are robust and ready to be used

In our research we have identified several powerful and mature products that have strong value proposition, but currently do not offer standard compliance. We plan to publish results of this research in one of the upcoming issues.

Table 4: Ontology Tool Support

Tool	Capabilities	Vendor	Standards Compliance and General Comments
Lifecycle Phase: Creating			
Protégé-2000	Create concept hierarchies, create instances, and view in several formats. Typically used as a single user tool. Multi-user support is becoming available.	Stanford KSL	Open Source (Mozilla); plug-in architecture. Supports RDF, DAML+OIL. OWL support is currently at an alpha status.
OntoEdit	Create concept hierarchies, create instances. Integrate with common databases. Single user tool.	Ontoprise	Claims to be RDF, DAML+OIL compliant, plug-in architecture. Our tests were not able to confirm the compatibility.
OilEd	Create concept hierarchies, create instances, and analyze semantic consistency (according to DL). Single user tool.	U of Manchester	RDF, DAML+OIL support. From the creators of OIL. Free download, integrates with DL reasoner.
Medius Visual Ontology Modeler	Ontology creation. Some support for collaboration.	Sandpiper Software	A limited beta started in January 2002. Extends UML, requires Enterprise Edition of Rational Rose. Supports RDF and DAML + OIL. Includes a library of ontologies that represent the IEEE Standard Upper Ontology (SUO).
Cerebra Construct	An advanced ontology construction toolset (with semi-real time reasoning support). Enables ontology seeding - the absorption of existing ontologies, taxonomies, database schemas, and wrapping with ontology.	Network Inference	Commercial version of OilED extended to integrate with a commercially available graphical editor and enable collaborative authoring. First release in March, 2003. New releases are expected in 3Q03. Supports RDF, DAML+OIL, OWL, SOAP interfaces.
LinKFactory Workbench	Collaborative authoring environment. Originally designed for very large medical ontologies. Has a Java beans API and optional Application Generators for semantic indexing, automatic coding, and information extraction. Compares and links ontologies via a core ontology; related concepts matched on formal relationships and lexical information.	Language and Computing	Supports RDF(S); DAML+OIL/OWL. Some support for population and maintenance.
K-Infinity	Collaborative authoring environment.	Intelligent Views	Modularized tools supporting all stages of lifecycle. Supports RDF and Topic Maps.
Lifecycle Phase: Populating			
OntoAnnotate	Copy items from web pages, Create mark-up	Ontoprise	Integrated with MS IE RDF, DAML+OIL compliant

Tool	Capabilities	Vendor	Standards Compliance and General Comments
OntoMat	Copy items from web pages, create mark-up	AIFB Uni Karlsruhe	Free download. Adopted by DARPA On-To-Agents project. JAVA-based and provide a plug-in interface for extensions. DAML+OIL compliant.
AeroDAML	Natural language parse of documents to create mark-up	Lockheed-Martin	Demo available as web service. RDF, DAML+OIL compliant.
CORPORUM OntoBuilder	Basic concepts and relations are represented with single inheritance. Representation of concepts and relations extracted from content may be extended with WordNet information.	CognIT AS	Spun off from OnToKnowledge. Supports RDF, DAML+OIL support. Requires Sesame RDF repository. Focuses on generating editable ontologies automatically from natural language documents. Also supports creation.
Freedom Enterprise Semantic Platform	Semantically enhancing the metadata with associations and concepts unique to the language, structure and needs of an industry. Automatic ontology directed classification and semantic annotation of heterogeneous content.	Semagix	Supports XML with RDF planned for 2003. Also supports deployment and creation. It includes the Knowledge Toolkit for building ontologies.
Lifecycle Phase: Deploying			
Orenge	Performs natural language search using ontologies. Supported Capability: Concept-based Search	Empolis	Also supports ontology creation. XML and Topic Map compliant.
Freedom Enterprise Semantic Platform	Categorization and search using ontologies. Aggregating and normalizing content from a wide variety of content sources. ESP is an application platform for semantic integration of heterogeneous content including media and enterprise databases. Supported Capability: Automated Concept Tagger, Navigational Search, and Concept-based Search.	Semagix	Also supports ontology creation and population. XML-based, RDF support planned for 2003.
OntoBroker	Provides framework for processing rules organized by an ontology. Supported Capability: Product Design Assistant	Ontoprise	RDF, DAML+OIL compliant.
Semantic Miner	Constructs semantically meaningful queries from natural language queries.	Ontoprise	RDF, DAML+OIL compliant.
OntoOffice	Just-in-time content delivery based on ontologies. Supported Capability: Context-aware Retriever.	Ontoprise	Integrates with MS Office. RDF, DAML+OIL compliant.

Tool	Capabilities	Vendor	Standards Compliance and General Comments
Coherence	Ontology-based Enterprise Data Management. Focused on structured data. Supported Capability: Enterprise Data Model Manager	Unicorn Solutions	Download with registration. RDFS/DAML+OIL support. OWL support in 2003. Also supports ontology creation.
Contextia	Enterprise Data and Application Integration. Accepts a variety of inputs for mapping and modeling, including XML schemas, native schemas, database tables, and delimited files. Can work stand alone or in conjunction with existing IT infrastructures — such as EAI, B2Bi, business process management, message brokers and off-the-shelf connectors. Can handle information transformations involving complex data elements, nested structures and incompatible or conflicting semantics. Supported Capability: Semantic Data Integrator, Semantic Application Integration.	Modulant	Ontology creation supported by FirstStep XG included with Contextia. Some support for critiquing: Express model (ISO 10303) is used for validation; cross-ontology consistencies. Supports XML, Web Services ready and supports SOAP.
Cerebra Server	Semantic integration of Enterprise Data and Applications leveraging Cerebra's inference engine. This is a commercial version of OilED semantic engine. Entire platform of tools based on Cerebra's engine is planned. Supported Capability: Semantic Data Integrator, Semantic Form Generator and Results Classifier, Knowledge Pulse	Network Inference	Also supports population. Supports maintenance and evolving of ontologies with live addition of new axioms/relationships without system downtime. Supports RDF, DAML+OIL, OWL, SOAP interfaces.
Tucana KnowledgeStore	Distributed database designed especially for metadata and metadata management. The database has been architected to persist and retrieve metadata with extremely fast performance levels while maintaining permanent integrity and secure access.	Plugged In Software	Also supports creation and population with Tucana Metadata Extractor™. Supports RDF, has SOAP, COM and Java Interfaces.
Lifecycle Phase: Critiquing activity of Maintaining and Evolving			
ConVisOr	Checks conformance of ontology to description logic	Lockheed Martin	
OntoClean (ODE)	Checks for consistency of ontologies	U of Madrid	Research prototype, DAML+OIL compliant.

Tool	Capabilities	Vendor	Standards Compliance and General Comments
OilED	Analyze consistency of ontologies according to DL	U of Manchester	Free download. Integrates with MS Office. RDF, DAML+OIL compliant.
Cerebra Inference Engine	Analyze consistency and draw conclusions based on DL	Network Inference	Commercial version of OilED. Supports deployment. Also supports maintenance and evolving of ontologies with live addition of new axioms/relationships without system downtime. Supports RDF, DAML+OIL, OWL, SOAP interfaces.
Lifecycle Phase: Merging activity of Maintaining and Evolving			
PROMPT	Supports merging two or more ontologies	Stanford KSL	Plug-in for Protege2000.
Chimera	Allows multiple ontologies to be processed together, provides analysis to find merges	Stanford KSL	Planned support for RDF and OWL.
FCA-Merge	Merges ontologies bottom-up based on common instances	AIFB, Karlsruhe	Research prototype.

2.4 Tool Interoperability

We have performed Protégé-2000, OilEd and OntoEdit exports and imports and concluded that at this point currently ontology creation tools do not interoperate well.

RDF(S) has several entries, because it can be used in different ways, and the behavior of the systems depends not only on the usage of RDF, but also the provenance. We have identified basically three different RDF(S) forms:

- Two of them are "plain" RDFS, but differ as export from OilED and Protégé-2000
- The third one is "Standard Oil RDFS", as exported by Protégé-2000

While Protégé-2000 and OilED respectively can produce Oil (RDFS) and OWL outputs respectively, no tool can read these, even the tools that produced them.

The following table shows the conversion capabilities of the major ontology editing tools with respect to the main representation languages.

Table 5: Conversion Capability of major Ontology Creation tools

Read by:	OilEd	Protégé-2000	OntoEdit
Plain RDFS from OilED (lost sameClassAs,axiom,restriction)	Yes	Yes	Remove spaces
Plain RDFS from Protégé-2000	Version[1]	Yes	No
Standard Oil RDFS from Protégé-2000 No(kb:oil:)	No	No	No

[1] The version issue from Protégé-2000 to OilED has to do with the fact that Protégé-2000 outputs RDF using the version marked "19990303", while OilED uses an unmarked version. Unfortunately, some things that Protégé-2000 uses in RDFS are only supported in version 19990303; thus OilED gets a (recoverable) error when it tries to process them.

Read by:	OilEd	Protégé-2000	OntoEdit
DAML+OIL from OilEd	Yes	No	Minor problems
OWL from OilEd	No	No	No
DAML+OIL,RDFS from OntoEdit is empty[2]	N/A	N/A	N/A

[2] Notice that while OntoEdit has capabilities on its menu to produce RDFS and DAML+OIL formats, neither of these produces files longer than 0 length. We have experimented with the public domain version, not the commercial version of the tool. Ontoprise acknowledged the problem and expected it to be fixed in the next release, Q1 2003.

3. Semantic Integration, Strategies and Tools

3.1 Executive Summary

A growing number of semantic technology vendors are responding to the critical need to manage and integrate large numbers of disparate applications and data sources present in today's enterprise. This briefing is focused on the use of semantic technology to integrate structured data and applications and includes analyzes of offerings from 9 leading vendors.

The most common current solution to integration and translation is field to field mapping. Schemas from two data sources are imported and fields are mapped to each other. This approach doesn't scale well as the number of maps grows exponentially with each new data source. Enterprises working with this technology often discover that creating correct maps is a challenge because it requires that the person doing each mapping has an in depth knowledge of both data sources, which is rarely possible.

Semantic technologies offer a new way to integrate data and applications. Before making mappings, a model (or an ontology) of a given business domain is defined. The model is expressed in a knowledge representation language and it contains business concepts, relationships between them and a set of rules. By organizing knowledge in a discrete layer for use by information systems, ontologies enable communication between computer systems in a way that is independent of the individual system technologies, information architectures and applications.

Compared to one-to-one mappings, mapping data sources to a common semantic model offer a much more scaleable and maintainable way to manage and integrate enterprise data. The "common business model" terminology used here may remind readers of the enterprise data and process modeling initiatives. These initiatives have proven to be long on cost and resources and short on ROI. Does the use of semantic integration solutions depend on an enterprise-wide modeling effort? We don't believe so. In fact, we recommend a targeted way to start by situating your first semantic integration solution within a specific project, as opposed to having it as a separate initiative. The model has to be large enough to provide value – sufficient to integrate specific data or applications. It doesn't need to be enterprise-wide. Using knowledge representation approaches based on W3C standards ensures open, future proof implementations where models can be expanded, interlinked, merged and federated.

Semantic technologies are proving to offer enterprises competitive advantage. With the growing adoption of XML and the attendant need to reconcile meanings across different vocabularies, these technologies are becoming increasingly important. Beyond managing and connecting disparate enterprise data, key future capabilities include intelligent web services discovery and orchestration.

Now is the right time to begin developing the expertise in modeling and learning more about semantic technologies. As forecast by Gartner: "By 2005, lightweight ontologies will be part of 75 percent of application integration projects. The relative scarcity of skills in semantic modeling and the unification of information models may be the greatest challenge. Beyond initial development, the need for ongoing information-management processes at the enterprise level will severely tax most enterprises"[3].

[3] Gartner, "Semantic Web Technologies Take Middleware to the Next Level", 8/2002

To begin understanding and responding to these challengers, learning more about RDF/S and OWL is an important suggested step. Likewise, acquiring methodologies for modeling and information management is recommended.

3.2 A Need to Integrate and a Need to Manage

Integration is arguably the most pressing and expensive IT problem faced by companies today. A typical enterprise has a multitude of legacy databases and corresponding applications. The disconnected systems problem is the result of mergers, acquisitions, abundance of "departmental" solutions and simply implementation of many silo applications created for a specific purpose.

We know of a bank with over 40 different call center systems, a financial services company with more than 1,000 databases and a manufacturing company with over 2,000 CAD/CAM systems. These systems contain valuable information and often are still good for supporting specific tasks they were intended for. Unfortunately, the information they contain can not be leveraged by other systems without a considerable effort. When the changes in business needs or available technology require modifications to these applications to provide additional capabilities and to streamline workflows, integration and extension become a very expensive undertaking. Simply tracking all the enterprise data sources and their relationship to each other is proving to be a challenge. In fact, many IT organizations spend up to 80% of their budgets maintaining the legacy systems leaving limited funds to support new business opportunities or to satisfy new regulatory requirements.

Many companies have been moving to XML to take advantage of standards based integration. However, XML doesn't capture the contextual meaning (or semantics) of the data. And a growing number of "standard" XML dialects (currently over 400) intended to standardize business vocabularies make the need for a semantic translation layer even more apparent.

3.2.1 The Most Common Solution Strategy

The most common solution to data integration and translation is field to field mapping. Schemas from two data sources are imported and fields are mapped to each other. Rules can be defined to split or concatenate fields or to perform other simple transformations. Once this is done the tool can do data translations either directly at run time or by generating code that will perform the transformations. There are a number of tools on the market that support this approach. Vendors include IBM and Microsoft. Some of the tools have been available for nearly a decade, but the adoption has been slow for a number of reasons:

- Field to field mapping works on a small scale. However, the number of maps grows exponentially with each new data source. Maintenance and evolution become a problem since any change in the schema of one data source will require you to redo multiple maps.
- Enterprises working with this technology often discover that creating correct maps is a challenge. It requires that the person responsible for each mapping has an in depth knowledge of both data sources, which is rarely possible. As a consequence, mapping mistakes are quite common.
- Mapping and translating between two schemas that are using a different design paradigm (i.e., different degree of normalization or nesting) can be very difficult. There is more then one way to design a schema. Performance considerations may result in de-normalized database schemas. When schemas are expected to change, designer may opt for a reflective design. Some XML schemas are

deeply nested, others are shallow. Mapping between relational (RDBMS) and hierarchical (XML) stores can suffer from significant impedance mismatch of the models.

- Direct mapping may fail in the situations requiring more conceptual and conditional transformations.

Is there a better solution?

3.2.2 Semantic Solutions

Semantic technologies offer a new way to integrate data and applications. Before making mappings, an ontology (or a model) of a given business domain is defined. It can be "jump started" by importing data schemas. The model is expressed in a knowledge representation language and it contains business concepts, relationships between them and a set of rules. This is the knowledge that the users of the systems want to store and access, rather than the data that implements that knowledge. The knowledge model is then mapped to fields in databases, XML Schema elements, or operations, such as SQL queries or sets of screen interactions. This approach solves many maintenance, evolution and schema compatibility problems.

The key ingredients that make up an ontology are a vocabulary of basic terms, a precise specification of what those terms mean and how they relate to each other. The term 'ontology' has been used in this way for a number of years by the artificial intelligence and knowledge representation community, but is now becoming part of the standard terminology of a much broader community including object modelers and XML users. By organizing knowledge in a discrete layer for use by information systems, ontologies enable communication between computer systems in a way that is independent of the individual system technologies, information architectures and applications. As a common model an ontology helps in the management of enterprise data sources.

Once the data sources are mapped to the model it can be used as an enterprise data management tool and to transform and validate data at design or run time. We can also envision future applications composed of very thin components that dynamically change their behavior based on the interactions with the business knowledge embedded in the model.

The distinct advantage of knowledge representation languages as ways to express the model is that they are optimized for capturing relationships between concepts and for defining generic and specific rules (assertions) that logical reasoning can be based on. Some examples of such rules are:

- If A is a part of B and B is a part of C then A is a part of C
- If a person has blood-contact with someone at risk of an HIV infection risk, then they are potentially at risk of an HIV infection
- If John wrote a paper on semantic integration, he knows about semantic integration

The attraction of logic as a technology for supporting semantic integration stems from the capability of logical languages to express relationships in generic ways, and the availability of sophisticated automated systems for finding combinations of related items that satisfy certain constraints. The variants of logic used for semantic integration (including Horn logic (prolog), frame logic, and description logic) differ primarily in the expressiveness of the logic and the tractability of the reasoning system. Another technology that provides similar capabilities is "means-ends analysis", which grew out of a different research background. Some vendors (Celcorp) base their integration products on this

technology. Using models of knowledge, semantic engines can make inferences and create dynamic (on the fly) relationships between different concepts.

The model in the Figure 9 shows a unified view of billing and contractual databases. The blue arrows indicate explicitly defined relationships, while yellow arrows indicate derived ones. The derived relationships were established by the system based on the defined rules some of which are also shown in the figure below. For example:

- The rule "If customer is subject to a contract and invoice is billed to the customer then invoice is subject to a contract" has resulted in establishing a dynamic runtime connection between an invoice and a customer
- The rule "If contract has terms and invoice is subject to the contract then invoice is subject to each term" has built on the connection inferred by applying the previous rule and established connections between an invoice and the specific terms of the contract.

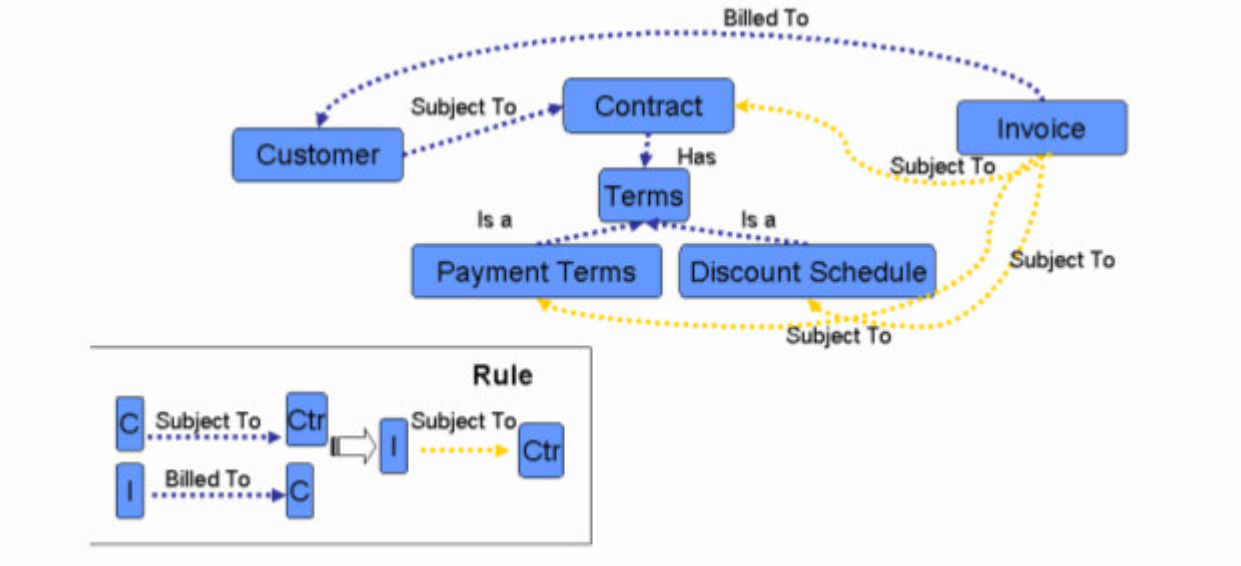


Figure 9: Illustration of a Unified View of Billing and Contractual Databases

Some ideas behind semantic models or ontologies for integration may remind you of metadata management. It is, in fact, based on the similar concepts. However, proponents of semantic integration argue that the use of W3C standard knowledge representation languages gives them distinct advantages:

- ***Open Standards.*** Using knowledge representation approaches based on W3C standards ensures open, future proof implementations where models can be expanded, interlinked, merged and federated.
- ***Rich Semantics.*** Knowledge representation languages offer support for richer and more precise semantics then UML, a standard language behind meta-data repositories. W3C languages like RDF (resource description framework), RDF Schema and the new Web Ontology Language (OWL) have been specifically designed to capture relationships between concepts and to define generic and specific rules (assertions) with the precision that logical reasoning needs.

- ***Native to the Web.*** RDF and OWL are serialized in XML and are, therefore, native to the Web. W3C sees semantic standards as a fundamental enabler for the next phase of web solutions.

3.3 Semantic Integration Vendors

Table 1 lists companies offering semantic integration solutions. Most of the vendors in this emerging technology field are relatively young (less then 5 years old), privately held companies. Many are capitalizing on the research work that started in early 1990s.

Vendor Name	Product Name	Description	Year Founded	Company
Celcorp	Celware	**Engine**: Server and Real-time Planner integrate applications streamlining users' workflow where multiple systems must be accessed in order to perform a task. The software uses intelligent agent technology based on proprietary extensions to the "Plan Domain Model and the Graph Plan Algorithm." **Modeling**: Models are automatically generated by running Celware Recorder, a design time tool.	1990	Celcorp is privately held and based in Santa Monica, California. The company was originally established in Canada and has been offering business integration software for sometime. It has a number of reference clients.
Contivo	Enterprise Integration Modeling (EIM) Server	**Engine**: Server includes a Semantic Dictionary containing enterprise vocabularies, such as various XML, EDI, and ERP standards; a Thesaurus with synonyms that match business concepts; and a Rules Dictionary that governs the field level data transformation. **Modeling**: Modeling (mapping) is done using Contivo Analyst tool. Some pre-built maps are available.	1998	Contivo is a privately held company with offices in Palo Alto, California. Contivo's corporate investors include industry leaders BEA Systems, TIBCO Software and webMethods. Venture capital investors include BA Venture Partners, Voyager Capital and MSD Capital LP. It has received a 3rd round of funding in January 2003.
enLeague	Semantic Broker	The Vitirus V3 Integration Platform: **Engine:** The base product, **Vitirus Envoy**, provides data connectivity for a peer-to-peer or hub and spoke interoperability solution. It allows the deployment of XML and SOAP messaging. The Services Manager acts as the "traffic cop" that monitors the interaction of services based on the rules and flows defined with the Modeler and Services Flow Manager. Product focus is on creating a scaleable run time environment integrated with popular web application servers. **Vitirus Mediator** addresses the problem of scalability. As the	2000	Partially owned by Coca-Cola and located in Atlanta, Georgia (on Coca-Cola campus) enLeague was formed in September 2000. The company has recently acquired Killdara's XML Integration Platform bringing a total number of employees to 16.

Vendor Name	Product Name	Description	Year Founded	Company
		number of Envoy connections expand and the complexity of the data exchanges grow, Mediator provides for standards based, real-time integration of XML and Web services. It manages, mediates and coordinates the requests for and delivery of information throughout the enterprise. **Modeling**: Business and Service Flow Modeler uses ontologies to rapidly describe and model critical business processes, goals, and objectives. The Modeler also enables companies to use existing business models (e.g. database schemas), industry standards, and information from legacy systems by importing and integrating them. There is also a Business Activity Monitor called **Vitirus Insight**.		
Modulant	Contextia Product Suite	**Engine**: Contextia Dynamic Mediation uses a central description of enterprise data called Abstract Conceptual Model (ACM) to enable disparate applications exchange information by transforming messages at runtime. It reconciles semantic conflicts among disparate applications and data sources. **Modeling**: Modeling is done using Contextia™ Interoperability Workbench capturing the meaning, relationships, and context of data elements of all source and target applications, and mapping them to ACM. The mapping specifications and ACM are then used by the Modulant Contextia Dynamic Mediation to transform data from source to target at runtime. The Interoperability Workbench accepts a variety of inputs for mapping and modeling, including XML schemas, native schemas, database tables, and delimited files.	2000	Modulant was founded in 2000, and subsequently merged with Product Data Integration Technologies (founded in 1989) in order to develop commercially-deployable software based on PDIT's proprietary technology and patent-pending methodology. Modulant is a private, venture-backed company whose existing investors include Sandler Capital Management, Guardian Partners and First Lexington Capital. Modulant's world-wide headquarters is in Charleston, SC, with additional offices in Long Beach and San Francisco, CA, Chicago, IL, Dallas, TX, Washington, DC, London, England and Stockholm, Sweden.
Network Inference	Cerebra Platform	**Engine**: Cerebra Inference Engine creates dynamic connections between different ontologies using reasoning based on description logic. While Cerebra can work with the central model its value proposition is based on the assertion that only a few key	2000	Founded in late 2000 to commercialize a description logic reasoner from the University of Manchester. The company is headquartered in London, UK with plans to open US offices. Network Inference is backed by Nokia Ventures.

Vendor Name	Product Name	Description	Year Founded	Company
		connections between disparate schemas are needed. Cerebra can dynamically infer the rest of the connections thereby minimizing mapping efforts. **Modeling**: Modeling is done using Cerebra Construct, a MS Visio based graphical modeling tool.		
Ontology Works	IODE	**Engine**: IODE utilizes a central description of enterprise data to determine answers to complex queries. Each link in the enterprise ontology is mapped to a query in the "ontology database"; this can either be a warehoused database created as part of the ontology engineering process, or a mediated connection to a legacy database. Solutions to queries in the ontology are build using the rules and relations in the ontology, so that the "proof" of the result can be translated in a simple fashion into a program that runs over the databases, to determine the correct answer. **Modeling**: Modeling can be done using UML tools, translated into a proprietary Ontology Works Language.	1998	The company is privately held and has offices in Maryland and Arkansas. In the first quarter of 2000, it completed development of an initial version (V 1.0) of its tool set and secured its first customer.
Ontoprise	OntoBroker	**Engine**: Data integration is done via a several step process that includes importing data schemas from existing databases, and using OntoMap to map concepts and relations from one ontology to the next. These mappings are translated into F-Logic statements, so that Ontobroker can reason over the combined ontology results in data references in the original data sources. **Modeling**: Modeling is done using OntoEdit and OntoMap. Two more tools are needed to complete this picture, which are a rule editor and a rule debugger, both of which are currently in the proposal stage. The rules state the actual connections between the newly merged concepts, and are susceptible to bugs; hence they must be viewable and debuggable.	1999	The Ontoprise® GmbH is venture capital backed; it achieved a break even point in 2002. The company is headquartered in Germany. Ontoprise was founded as a spin off of the University of Karlsruhe which implemented the first version of technology in 1992.
SchemaLogic	SchemaServer	**Modeling**: SchemaServer captures	2001	Privately held company founded by ex-Microsoft employees. Located in

Vendor Name	Product Name	Description	Year Founded	Company
		and communicates data definitions (enterprise schema) used across all applications and languages. To help create the active repository of schema and metadata, SchemaServer imports existing schema, taxonomy and classification criteria from databases, applications or content management systems. It supports distributed, collaborative management of enterprise taxonomy. SchemaServer manages the associations and links among the separate schemas by providing the tools necessary to model, map, and describe the multiple relationships.		Redmond, WA.
Unicorn Solutions	Unicorn System	**Engine**: The Unicorn is a design time tool and a script generator for integration with third party engine, such as WebMethods. **Modeling**: The Unicorn too imports schemas from multiple data sources including XML, RDBMS, COBOL, IMS, and EDI. They are then mapped to a central enterprise model (ontology). Mapping supports creation of data transformation rules. Unicorn can generate transformation scripts as executable SQL, XSLT, and Java Bean code.	2001	The company is privately held. It is headquartered in New York City with R&D in Israel. Unicorn's investors include: Jerusalem Global Ventures, Bank of America Equity Partners, Intel Capital, Israel Seed Partners, Tecc-IS and Apropos.

Table 1: Overview of Semantic Integration Vendors

Other companies worth mentioning in this category include IGS (www.igs.com) and MetaMatrix (www.metamatrix.com) that have UML and MOF based approaches to integration, Miosoft (www.miosoft.com) that offers a highly scaleable run time data validation and consolidation platform based on a central model with a rich set of rules, as well as Vitria (www.vitria.com), an EAI vendor that incorporates business vocabularies.

The market for semantic integration is expected to grow fairly quickly fueled by the needs of enterprises and by the growing maturity of the AI (Artificial Intelligent) technologies that underlie many of these solutions. According to a report released by Business Communications Company, Inc., (*www.bccresearch.com*) they expect AI technologies that assist existing applications handle more complex data analysis, addressing the potential variability in a situation via a set of rules, will see strong growth and implementation across sectors. Their estimate is that this technology will reach $4.8 billion in sales and an AAGR of 14.5% through 2007.

3.4 Capabilities of Semantic Integration Platforms:

We have identified the following as key capabilities offered by semantic integration solutions:

Enterprise Data Management

- Creating and publishing shared vocabularies of business concepts
- Cataloging data assets, including their schemas and other metadata.
- Formally capturing the semantics of corporate data by mapping database and message schemas to the ontology
- Importing a variety of standard data definition formats
- Supporting model management and evolution

Data Transformation

- Generating scripts and transformations to copy or move the data from one data source to another

Dynamic Code Generation

- Generating executable code such as SQL, XSLT and Java
- Generating "wrappers" for data sources
- Embedding of business rules in models
- Automatic updates after change in the model and schemas

Semantic Data Validation

- Using inference rules to validate integrity of the data based on a set of restrictions. The inference rules will automatically identify inconsistencies when querying for information.

Run-time Support

- Scaleable semantic engine that supports high volume of real time queries

Orchestration of Web Services

- Integration broker
- Intelligent discovery and orchestration (composition and chaining) of web services

Table 2 compares capabilities currently offered by each of the vendors.

Table 2: Comparison of Capabilities Offered by Vendors

	Enterprise Data Management	Data Transformation	Dynamic Code Generation	Semantic Data Validation	Run-time Support	Web Services Orchestration
Celcorp Celware	-	-	Yes	-	Yes	-
Contivo EIM Server	Yes	Yes	Yes	-	-	-
enLeague Semantic Broker	Yes	-	-	-	Yes	Yes
Modulant Contextia	-	Yes	-	-	Yes	-

	Enterprise Data Management	Data Transformation	Dynamic Code Generation	Semantic Data Validation	Run-time Support	Web Services Orchestration
Product Suite						
Network Inference Cerebra Platform	-	-	-	Yes	Yes	-
Ontology Works IODE	-	Yes	Yes	Yes	Yes	-
Ontoprise Ontobroker	-	-	-	Yes	Yes	-
SchemaLogic SchemaServer	Yes	Yes	-	-	-	-
Unicorn System	Yes	Yes	Yes	-	-	-

Table 3 provides a detailed look at each product and its support for open standards.

Table 3: Support for Open Standards of Semantic Integration Products

Product	Product Adoption and Usage	Knowledge Representation	Reasoning Capabilities	Interfaces	Support for Web Services Standards
Celcorp Celware	Mature product, offers a unique approach to application integration. Have a number of reference customers in the financial services industry.	Proprietary, planning to go to RDF in 2003.	Based on proprietary extensions to the "Plan Domain Model and the Graph Plan Algorithm."	Import: Screen scraping, SQL statements	
Contivo Enterprise Integration Modeling (EIM) Server	Relatively mature, has a number of reference customers. Focused on complementing webMethods and Tibco.	Proprietary on top of relational database, evaluating RDF	None evident, integration with a reasoning engine would be hard to implement until support for RDF is offered	Import: XML Schema, RDB (Oracle only), flat files Export: XML Schema (XSLT), EAI (WebMethods, TIBCO), Java	XML, SOAP, WSDL
enLeague Semantic Broker	New, currently in beta	RDF, DAML+OIL, OWL intentions	Is designed to Incorporate 3rd party inference engines	Import: XML Schema, RDB (JDBC), RDF/S, DAML+OIL Export: XML Schema (XSLT), RDF/S,	XML, SOAP, WSDL, UDDI

Product	Product Adoption and Usage	Knowledge Representation	Reasoning Capabilities	Interfaces	Support for Web Services Standards
				DAML+OIL	
Modulant Contextia Product Suite	Relatively mature, has a number of reference customers. Focused on government, STEP customers.	XML, proprietary, evaluating RDF	None evident	Import: XML, RDB, flat files, STEP 21 files Export: XML	XML, SOAP
Network Inference Cerebra Platform	New, currently in beta. Initial focus on biotechnology.	RDF, DAML+OIL, OWL	Description Logic	Import: XML Schema, RDB (JDBC), RDF/S, DAML+OIL Export: XML Schema (XSLT), RDF/S, DAML+OIL	XML, SOAP, WSDL
Ontology Works IODE	Relatively mature, has a number of reference customers in government.	Proprietary	Robust, based on a proprietary Ontology language OWL (a variant of KIF, not related to w3c standard by the same name)	Import: UML, RDF/S Export: RDB (Oracle, DB2), DDB, RDF/S, XML	XML
Ontoprise Ontobroker	Relatively mature semantic engine has a number of reference customers. New to the integration market.	RDF, DAML+OIL, OWL support planned	F-Logic	Import: RDB, RDF/S, DAML+OIL, XML Schema Export: RDF/S, DAML+OIL	XML
SchemaLogic SchemaServer	New. The product can unify structured and unstructured data management. Focuses on helping existing customers of Portal and Content Management products.	XML, Proprietary	No	Import: RDF, XML Schema Export: ?	XML, SOAP
Unicorn System	Relatively new, focused on enterprise data management. First customer implementations	RDF, DAML+OIL, OWL support planned	A third party reasoning engine could be integrated with this standards-	Import: RDB (Oracle 7i/8i/9i, MS SQL Server 7/2000, DB2), XML Schema, UML (via	XML

Product	Product Adoption and Usage	Knowledge Representation	Reasoning Capabilities	Interfaces	Support for Web Services Standards
	are in progress.		based tool	adopter), ERWin, RDF/S, DAML+OIL Export: RDF/S, DAML+OIL, SQL Transformation Scripts, XSLT	

Figure 10 compares how these solutions are positioned within the semantic integration space. The vertical axis represents a vendor's ability to integrate disparate information based on semantics. The horizontal positioning represents a vendor's solution focus. The vertical axis represents a progression – the higher positioning indicates more powerful semantic capabilities. The horizontal line doesn't end with an arrow because, unlike the vertical axis, it is not intended to represent a progression of capabilities. The right most position of a vendor indicates that its major strength is in "Integration and Orchestration". The vendor may also offer some support, but not the full functionality, in the areas of "Data and Schema Management", "Validation" or "Run-time".

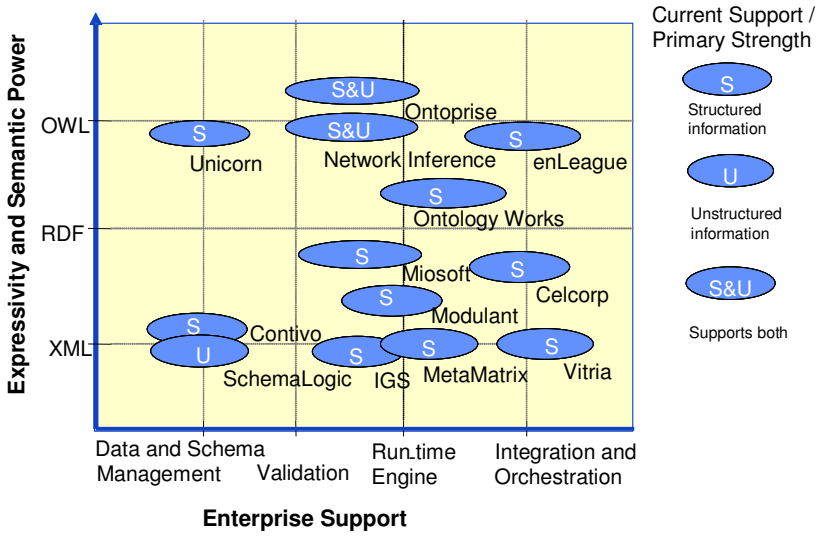


Figure 10: Positioning of Vendor's Solutions within the Semantic Integration Space

3.5 Recommendations for Getting Started:

The "common business model" terminology used by some vendors may remind readers of this report of the enterprise data and process modeling initiatives. These initiatives have proven to be long on cost and resources and short on ROI. Does the use of semantic integration solutions depend on an enterprise-wide modeling effort? We don't believe so. In fact, we recommend a targeted start by situating your first semantic integration solution within a specific project, as opposed to having it as a separate initiative. The model has to be large enough to provide value – sufficient to integrate specific

data or applications. It doesn't need to be enterprise-wide. Using knowledge representation approaches based on W3C standards ensures open, future proof implementations where models can be expanded, interlinked, merged and federated.

You may be implementing or enhancing a CRM, portal or a supply chain solution. Any of these projects can be a good starting ground for the semantic integration. It could be used to help you with the data migration or to actually serve as an integration broker. Start with a limited model necessary to support your project. Grow it as needed. Using open standards based technology will enable you to leverage this model with other tools and projects.

Now is the right time to begin developing the expertise in modeling and learning more about semantic technologies. As forecast by Gartner: "By 2005, lightweight ontologies will be part of 75 percent of application integration projects. The relative scarcity of skills in semantic modeling and the unification of information models may be the greatest challenge. Beyond initial development, the need for ongoing information-management processes at the enterprise level will severely tax most enterprises"[4].

To begin understanding and responding to these challengers, learning more about RDF/S and OWL is an important suggested step. Likewise, acquiring methodologies for modeling and information management is recommended.

3.5.1 About Vendor Selection

Vendors covered in this issue have different strengths as well as different industry and problem focus areas. Choosing the right product will depend on:

- How well it integrates with your data and content sources, infrastructure and applications
- The degree to which you need run time support
- Product's support for the industry specific XML schemas and vocabularies
- Vendor's flexibility and interest in evolving the product to support your requirements

[4] Gartner, "Semantic Web Technologies Take Middleware to the Next Level", 8/2002

About the Authors

Irene Polikoff is an Executive Partner with TopQuadrant. She is an editor and the main writer of TopQuadrant's quarterly Technology Briefings.

Irene has over 15 years of experience in business application development and deployment, consulting, software development and strategic planning. Irene has held a number of executive positions at IBM. She was Senior Development Manager and Project Executive for worldwide consultant's tooling and methods. Most recently she was a Principal in the national Knowledge, Content Management and Portals Practice in IBM Global Services. Ms. Polikoff was part of the team that developed and deployed a world-wide project management method for IBM Global Services.

Prior to IBM, Ms. Polikoff held IT management positions at Fortune 500 companies where she was responsible for development and deployment of enterprise-wide mission critical information systems. Irene has a background in Operations Research and a strong interest in technologies for software innovation.

Dean Allemang is a Senior Consultant with TopQuadrant who has contributed content to sections of this report.

Dr. Allemang specializes in innovative applications of knowledge technology, and brings to TopQuadrant over 15 years experience in research, deployment and development of knowledge-based systems. Prior to joining Top Quadrant, Dr. Allemang was the Vice-President of Customer Applications at Synquiry Technologies, where he helped Synquiry's customers to understand how the use of semantic technologies could provide measurable benefit in their business processes.

Dr. Allemang has filed two patents on the application of graph matching algorithms to the problems of semantic information interchange. In the Technology Transfer group at Swiss Telecom, he co-invented patented technology for high-level analysis of network switching failures. He is a co-author of the Organization Domain Modeling method, which addresses cultural and social obstacles to semantic modeling, as well as technological ones.

Dr. Allemang combines a strong formal background (M.S. in Mathematics, University of Cambridge, PhD in Computer Science, Ohio State University) with years of experience applying knowledge-based technologies to real business problems. Dr. Allemang is a lecturer in the Computer Science Department of Boston University Metropolitan College.

Companies interviewed for this report:

Celcorp - www.celcorp.com

Contivo - www.contivo.com

enLeague Systems – www.enleague.com

Network Inference – www.networkinference.com

MetaMatrix - www.metamatrix.com

Miosoft – www.miosoft.com

Ontology Works – www.ontologyworks.com

Ontoprise – www.ontoprise.com

Unicorn – www.unicorn.com

Additional TopQuadrant Technology Briefings are Available

- Dictionary of Search
- Modeling Techniques
- Semantic Solutions for Search and Self Service
- Interoperability

To access these papers, please visit our web site at www.topquadrant.com

About TopQuadrant

TopQuadrant is a trusted intermediary for the intelligent application of knowledge technologies. As knowledge system architects, we are assisting leading enterprises to envision, architect, plan and realize knowledge-based solutions. Our consultants have many years of experience in large consulting organizations, for example IBM Global Services, and have a background in AI, Object Technology, Knowledge Management and Methodologies for Knowledge, Software and Systems Engineering.

Using the following unique tools, we address major obstacles to success in building knowledge solutions:

- ***Solution Envisioning***, a scenario-driven approach to experiencing a future system through analogies and examples using a Database of Capability Cases.
- ***Capability Cases***, application solution patterns (e.g., for ontology-based knowledge applications) expressed in a business context with examples of known uses, applicable technologies and leading practices.
- ***TopDrawer***™, a comprehensive knowledge base for storing and dynamically working with Capability Cases.

With a proven track record in the practical application of knowledge technologies, **TopQuadrant** helps clients transition to next generation, semantically integrated systems, while sustaining and optimizing their investments in current systems.