Semantic Interoperability Community of Practice (SICoP)

Introducing Semantic Technologies and the Vision of the Semantic Web

White Paper Series Module 1

Updated on 02/16/05

Version 5.4

SICoP White Paper Series Module 1

Introducing Semantic Technologies and the Vision of the Semantic Web

Executive Editors and Co-Chairs
Dr. Brand Niemann, *U.S. EPA, Office of the CIO (SICoP Co-Chair)*
Dr. Rick (Rodler F.) Morris, *U.S. Army, Office of the CIO (SICoP Co-Chair)*
Harriet J. Riofrio, *Senior Staff Officer for Knowledge Management, Office of Assistant Secretary of Defense for Networks and Information Management, Deputy Chief Information Officer, Information Management (OASD NII DCIOIM), U.S. Department of Defense (KM.Gov Co-Chair)*
Earl Carnes, *Nuclear Industry Liaison, Environment, Safety & Health, Office of Regulatory Liaison, U.S. Department of Energy (KM.Gov Co-Chair)*

Managing Editor
Jie-hong Morrison, *Computer Technologies Consultants, Inc.*

Editor
Kenneth R. Fromm, *Loomia Inc.*

Copy Editor
Michael J. Novak, *Senior Analyst, Headquarters Office of Research, Internal Revenue Service*

Primary Contributors
Kenneth R. Fromm, *Loomia Inc.*
Irene Polikoff, *TopQuadrant, Inc.*
Dr. Leo Obrst, *The MITRE Corporation*
Michael C. Daconta, *Metadata Program Manager, Department of Homeland Security*
Richard Murphy, *U.S. General Services Administration*
Jie-hong Morrison, *Computer Technologies Consultants, Inc.*

Contributors
Jeffrey T. Pollock, *Network Inference Inc.*
Ralph Hodgson, *TopQuadrant, Inc.*
Joram Borenstein, *Unicorn Solutions, Inc.*
Norma Draper, *Northrop Grumman Mission Systems*
Loren Osborn, *Unicorn Solutions, Inc.*
Adam Pease, *Articulate Software Inc*.

Reviewers
Irene Polikoff, *TopQuadrant, Inc.*
Jeffrey T. Pollock, *Network Inference, Inc.*
Adam Pease, *Articulate Software, Inc.*
Dr. Yaser Bishr, *ImageMatters LLC*
Kathy M. Romero, *U.S. Army Training and Doctrine Command, Futures Center*
David Wood, *Tucana Technologies, Inc.*

NOTE: The views expressed herein are those of the contributors alone and do not necessarily reflect the official policy or position of the contributors' affiliated organizations.

TABLE OF CONTENTS

TABLE OF FIGURES

Introduction to the White Paper Series

This set of white papers is the combined effort of KM.Gov (http://km.gov) and the Semantics Interoperability Community of Practice (SICoP), two working groups of the Federal CIO Council. The purpose of the white papers is to introduce semantic technologies and the vision of the Semantic Web. They will make the case that these technologies are substantial progressions in information theory and not yet-another-silver-bullet technology promising to cure all IT ills.

The papers are written for agency executives, CIOs, enterprise architects, IT professionals, program managers, and others within federal, state, and local agencies with responsibilities for data management, information management, and knowledge management.

Module 1:
Introducing Semantic Technologies and the Vision of the Semantic Web

This white paper is intended to inform readers about the principles and capabilities of semantic technologies and the goals of the Semantic Web. It provides a primer for the field of semantics along with information on the emerging standards, schemas, and tools that are moving semantic concepts out of the labs and into real-world use. It also explains how describing data in richer terms, independent of particular systems or applications, can allow for greater machine processing and, ultimately, many new and powerful autonomic computing capabilities.

This white paper focuses upon applications of semantic technologies believed to have the greatest near-term benefits for agencies and government partners alike. These include semantic web services, information interoperability, and intelligent search. It also discusses the state and current use of protocols, schemas, and tools that will pave the road toward the Semantic Web.

Takeaways: We want readers to gain a better understanding of semantic technologies, to appreciate the promises of the next generation of the World Wide Web, and to see how these new approaches to dealing with digital information can be used to solve difficult information-sharing problems.

1.0 Executive Summary

> "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."
>
> *Semantic Technologies for eGov Conference (Sept. 8th, 2003)*

Semantic technologies hold great promise for addressing many of the federal government's more difficult information technology challenges. One example is the Environmental Protection Agency's preliminary efforts to reconcile public health data with environment data in order to improve the well being of children. Children are extremely susceptible to environmental contaminants, much more so than adults, and so the public is rightly concerned about the quality of their environment and its effects on our children. The increased public awareness of environmental dangers, in combination with the accessibility of the Internet and other information technologies, have conditioned both the public and various government officials to expect up-to-date information regarding public health and the environment. Unfortunately, these expectations are not adequately being met using the federal government's existing information technology tools and architectures.

The problem is not one of resources. Significant resources are being spent on data gathering and analysis to assess the health risks that environmental contaminants pose to our children. Unfortunately, the current state of the information sharing between agencies, institutions, and other third parties as well as the level of tools to intelligently query, infer, and reason over the amassed data do not adequately meet these expectations.

Public health and environmental data sets come from many sources, many of which are not linked together. Vocabularies and data formats are unfamiliar and inconsistent, especially when crossing organizational boundaries (public health vs. environmental bodies). Data structures and the relationships between data values are difficult to reconcile from data set to data set. Finding, assembling, and normalizing these data sets is time consuming and prone to errors and, currently, no tools exist to make intelligent queries or reasonable inferences across this data.

In fairness, tremendous strides have been made in physically connecting computers and exchanging large amounts of data in highly reliable and highly secure manners. A number of reputable vendors offer proven middleware solutions that can connect a wide variety of databases, applications, networks, and computers. But while these technologies can connect applications and various silos of information and enable them to move data around, they do not address the real challenge in connecting information systems – that of enabling one system to make transparent, timely, and independent use of information resident in another system, without having to overhaul IT systems or fundamentally change the way organizations operate.

It is this logical transformation of information – understanding what the information means and how it is used in one system versus what it means and how it is used in another – that is one of the larger impediments to making rational use of the available data on public health and the environment. The goal is not just to connect systems, but also to make the data and information resident within these systems interoperable and accessible for both machine processing and human understanding.

In an attempt to provide solutions to redress these issues, a pilot program is underway in the Environmental Protection Agency (EPA) to make use of semantic technologies to connect information from the Centers for Disease Control and Prevention (CDC) and the EPA, as well as from their state partners, in ways that can move the EPA farther down the path to answering the public's question: Is my child safe from environmental toxins? (Sonntag, 2003) While the focus of this pilot is primarily technical in nature, the successful deployment of more expansive capabilities holds enormous human considerations, offering great potential for improving the health and livelihood of millions of children across the country. Quickly identifying potential toxic exposures, knowing the location and severity of infected sites, and effectively prioritizing environmental cleanups are just three of the most basic priorities for agencies and industry and for the benefactors of these efforts – children, their parents, and all other members of society.

This story is one illustration of the tremendous IT challenges that the federal government faces. The complexity of the federal government, the size of its data stores, and its interconnected nature to state, local, and tribal government agencies as well as, increasingly, to private enterprise and Nongovernmental Organizations (NGOs) has placed increasing pressure on finding faster, cheaper, and more reliable methods of connecting systems, applications, and data. Connecting these islands of information within and between government agencies and third parties is seen as a key step to improving government services, streamlining finances and logistics, increasing the reliable operation of complex machinery, advancing people's health and welfare, enabling net-centric defense capabilities, and ensuring the safety of our nation.

Widespread information interoperability is one of the benefits that many researchers, thought-leaders, and practitioners see for semantic technologies. But by no means is it the only benefit. Building on top of this notion of richer, more accessible and autonomic information, far greater capabilities such as intelligent search, intelligent reasoning, and truly adaptive computing are seen as coming ever closer to reaching reality.

Although pioneers in the field of semantic computing have been at work for years, the approval of two new protocols by the World Wide Web Consortium (W3C) early in 2004 marked an important milestone in the commercialization of semantic technologies, also spurring development toward the goal of the Semantic Web. In the words of the W3C, "The goal of the Semantic Web initiative is as broad as that of the Web: to create a universal medium for the exchange of data."[1] "The Semantic Web is a vision: the idea of having data on the web defined and linked in ways so that it can be used by machines – not just for display purposes – but for automation, integration and reuse of data across various applications, and thus fully harness the power of information semantics."[2]

These new capabilities in information technology will not come without significant work and investment by early pioneers. Semantic computing is like moving from hierarchical databases to relational databases or moving from procedural programming techniques to object-oriented approaches. It will take a bit of time for people to understand the nuances and architectures of semantics-based approaches. But as people grasp the full power of these new technologies and approaches, a first generation of innovations will produce impressive results for a number of existing IT problem areas. Successive innovations will ultimately lead to dramatic new capabilities that fundamentally change the way we share and exchange information across users, systems, and

networks (Fromm and Pollock, 2004). When taken within a multi-year view, these innovations hold as much promise to define a new wave in computing much the same as did the mainframe, the personal computer, Ethernet, and the first version of the World Wide Web.

2.0 Introduction to Semantic Computing

> People are starting to realize that their information outlives their software.
>
> *Tim Berners-Lee*

> Information meaning is too tightly coupled to its initial use or application. Thus it is very difficult for either (a) machines to reuse information or (b) for people to query on concepts (instead of just on terms).
>
> *Jeffrey T. Pollock*

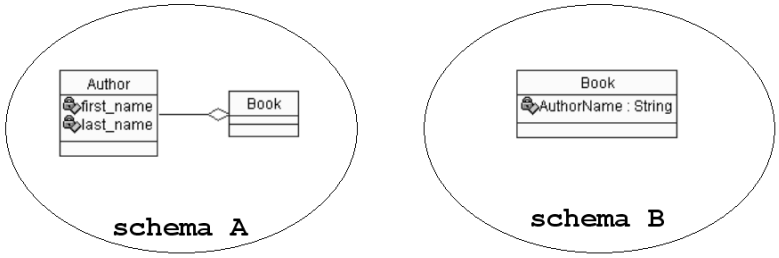
Illustrating the need for better information technology solutions to data management challenges faced by the government is not difficult. Information sharing is just one example. The challenge in sharing and making sense of information contained within federal, state, and local agencies – whether it is in the context of law enforcement, marine transportation, environmental protection, child support, public health, or homeland security, to name just a few – is a daunting one. Agencies can expend a large amount of time and money creating common vocabulary standards and then systems integrators can laboriously work to get each data-store owner to adopt and adhere to these standards. Unfortunately, this approach (if it even reaches the point of creating a standard vocabulary) quickly devolves into problems and delays in implementation. The real challenge in sharing information among disparate sources is not in creating a common language but in addressing the organizational and cultural differences that all too often prevent adherence or adaptation to a particular vocabulary standard (Fromm and Pollock, 2004).

2.1 Semantic Conflicts within the Enterprise

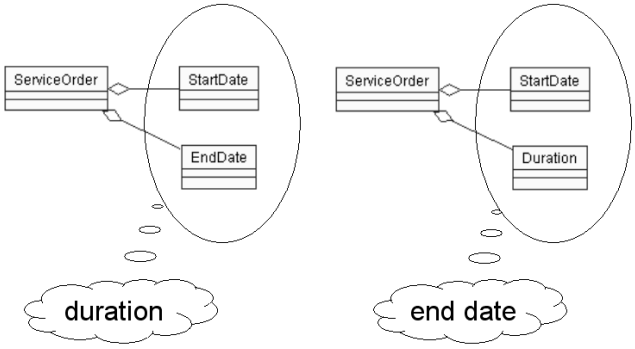
Structural and cultural differences embedded within organizational IT systems reflect their unique missions, hierarchies, vocabularies, work flow, and work patterns. “Price” may appear in one system; “cost” in another. A “Captain” in the Army is equivalent to a “Lieutenant” in the Navy; a “Captain” in the Navy is a “Colonel” in the Army. (These differences extend beyond the armed forces. Many state police organizations use ranks modeled after the marines; many public health organization use ranks modeled after the navy; many police and investigative bodies have their own unique command structures.) Similarly, an “informant” in a law enforcement organization might be termed an “information source” in an intelligence organization (the latter of which might include sources other than just people.) These are relatively simple differences in naming. The more complex and abstract a concept, the more differences there are in syntax, structure, and most importantly, meaning. One challenge for the system developer and/or information modeler is to determine whether differences in naming reflect a deeper underlying difference in concepts and meaning. Differences in naming can be handled relatively simply using readily available tools such as look-up tables or thesauri. Differences in concepts and definitions, however, require a much deeper alignment of meaning.

For instance, different systems may use of the same term for different concepts or stages within a value chain. The term “cost” in many systems is a reference to the price for which a consumer purchases an item, and yet “cost” might simultaneously be used in other systems as a reference to the price at which a supplier might sell an item to a distributor. Meanings also change contextually over time. Personnel changes, organizational history, organizational politics/culture, and corporate-driven mandates are just several of the forces that could alter meanings over time. (It goes without saying that terminologies also frequently change for much the same reasons.) These types of complex conflicts typically require more extensive semantics-based solutions of one sort or another.

Figure 1 shows two examples of semantic conflicts that can found across various data sets. (Appendix C contains a more expansive table of semantic conflicts.) These types of conflicts are common across most data sets, occurring almost as a natural byproduct of data modeling – whether due to isolated development, changing needs, organizational or structural differences, or any number of other reasons.



Aggregation conflict – difference in structure



Value representation conflict

Figure 1: Types of Semantic Conflicts

(Adapted from Pollock and Hodgson, 2004)

Syntactical, structural, and semantic conflict issues are becoming increasingly apparent within both corporate enterprises and government agencies. With messaging and transport solutions becoming increasingly commonplace and commoditized and with XML becoming a basic building block for exchanging data, it is readily apparent to most that these steps only partially complete the picture.

Additional technologies are needed in order to effectively rationalize the processes and information sets between and among organizations – without requiring point-to-point data and terminology mappings, processes that are both time- and personnel-intensive. Part of the promise of semantic technologies is the ability to employ logical languages that expose the structures and meanings of data more explicitly, thereby allowing software to reconcile whether terms and definitions are equivalent, different, or even contradictory.

2.2 Semantic Issues within the World Wide Web

The problem is not only within and between organizations and their business or operating partners; it also exists in various forms on the World Wide Web. Information on the Web is becoming increasingly fragmented and varied in terms of appropriateness, timeliness, and trustworthiness. Search engines are wonderful tools but, increasingly, fault lines are appearing. These fault lines manifest themselves in doubts about completeness of search; the growing use of script-like search commands such as "filetype"; or the rise in search engines focusing on specific types of data or media such as RSS feeds, images, or music. Federal and state governments have expended enormous resources in making information available to the public online, and yet the current state of the World Wide Web has placed limiting factors on the accessibility and applicability of this information.

2.3 Key Capabilities of Semantic Technologies

Fortunately, just as Internet and World Wide Web protocols helped connect vast amounts of information for human consumption, new approaches are emerging that help connect equal or greater amounts of information for machine manipulation and processing. These innovations will simplify information interoperability and provide better information relevance and confidence within the enterprise and on the World Wide Web. Over time, they will pave the way for new intelligent brokering and knowledge reasoning capabilities across the field of collected information. Figure 2 contains a table of the key capabilities of semantic computing and the resulting impact for stakeholders.

Capability	Purpose	Stakeholders	Impact	Take-away
Near-term				
Semantic Web Services	Provides flexible look-up and discovery and schema transformation	System Developers and System Integrators	Reduced friction in web services adoption and deployment	More automated and flexible data connections
Information Integration and/or Interoperability	Reduces integration complexity from n^2 to n	Data and Metadata Architects	Reduced cost to integrate heterogeneous data sources	Increased interoperability at improved speed and reduced cost
Intelligent Search	Provides context sensitive search, queries on concepts, and personalized filtering	Business and Technology Managers, Analysts, and Individuals	Reduced human filtering of search results, more relevant searches	Increased search accuracy translates into greater productivity

Capability	Purpose	Stakeholders	Impact	Take-away
Longer-term				
Model-Driven Applications	Enables software applications to process domain logic from actionable models	Software Developers	Less coding required, faster changes to domain logic	Less code maintenance and faster change responsiveness
Adaptive and Autonomic Computing	Provides the ability for applications to diagnose and forecast system administration	System Administrators	Increased reliability and reduced cost through self diagnostics and planning of complex systems	Reduced cost to maintain systems and lessened human intervention
Intelligent Reasoning	Supports machine inference based on rich data and evolvable schemas	Applications and Cognitive Agents	Reduced requirements for embedding logic and constraints apart from domain models	Reduced application development cost

Figure 2: Computing Capabilities Assessment

(Adapted by Richard Murphy)

Two new data and logic structures recently approved by the World Wide Web consortium (W3C) are making it possible to make information richer and more autonomous and, ultimately, far more accessible and adaptive. These new constructs – Resource Description Framework (RDF) and Web Ontology Language (OWL) – make extensive use of knowledge representation principles to add additional functionality and compatibility to existing W3C markup languages. RDF provides a framework for establishing relationships between data, whereas OWL enhances RDF with the ability to specify constraints on different data elements and their relationships to one another. These standards – in conjunction with new tools and infrastructure components built to support them – are driving the development of adaptive computing within the enterprise as well as the growth of the next generation of the web, called the Semantic Web.

The vision of the Semantic Web is to extend the current web by enriching the information transmitted and accessed over the Internet with well-defined meaning, thus enabling computers to do more of the work in assembling and processing data in order to turn it into highly relevant information and knowledge. In other words, the initiatives underlying the Semantic Web establish a set of protocols and technologies that promise to improve the categorization and association of data thereby enhancing the ability to create relationships and to generate inferences among diverse systems and data.

For example, asking a librarian for a map of Gettysburg at the time of the Civil War will typically lead to books containing maps from that era. A search in a search engine, however, will include many results with text concerning maps of Gettysburg, but these may or may not contain actual maps. Additionally, citations may be missed that did not match the exact form of date specified in the search string. Likewise, a search for networking security events in the Washington, DC, area might miss an anti-spam talk in McLean, VA, because the relationship between networking security and anti-spam

and the concept of McLean, VA, being in the Washington, DC, area are not yet fundamental associations within the realm of the World Wide Web. The steps taken by the W3C are targeted toward filling this gap in data association and collective understanding.

Building on the foundation provided by XML and related data-serialization efforts, RDF and OWL are beginning to be woven into the fabric of web-based tools and the World Wide Web. Figure 3 illustrates how the W3C's standards address the syntax, structure, and semantics of data and information and how they fit within the spectrum of semantic computing.

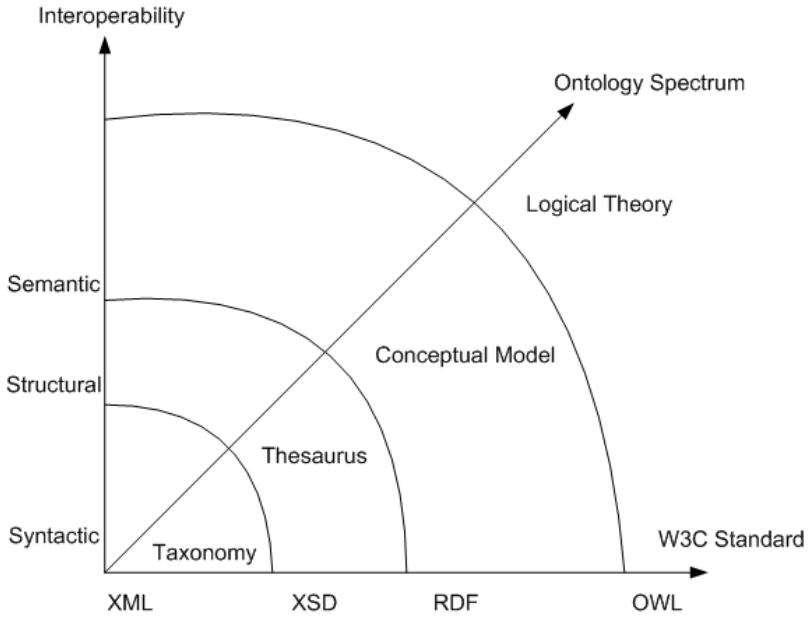


Figure 3: Three Dimensions of Semantic Computing

(From Daconta, Obrst, Smith 2003; Adapted by Richard Murphy)

In addition to defining these data and logic structures, the W3C has also defined initial architectures and logic required to implement semantic solutions alongside existing applications and data sets. Many companies have adopted semantic approaches and the vision of the Semantic Web and are actively pursuing technology strategies that further advance the field (Pollock and Hodgson, 2004). These technologies and approaches are being used today by a growing number of early adopters. Initial applications clearly demonstrate that they can be implemented incrementally and can deliver ROI-supported value. Several government agencies are planning or beginning pilot programs that use these newly approved standards to address complex challenges within narrowly defined problem spaces. Other modules in this set of white papers will provide specific information on these agencies and programs.

2.4 Semantic Technologies vs. Semantic Web Technologies

Many of the semantic technologies mentioned in this document predate the Semantic Web. (These technologies, however, may not have been termed “semantic” at the time.) Some have roots in artificial intelligence; others are extensions of markup language efforts; while others are logical outgrowths of enterprise application integration. The W3C’s effort to formalize a collection of data and logic languages has been an important catalyst in bringing many of the fields and technologies on common ground. Not all semantic technologies, however, make use of W3C-approved languages and frameworks, and so this paper makes a distinction between “semantic technologies” – technologies that make use of semantic concepts per se– and “Semantic Web technologies” – technologies that are fully compliant with W3C Recommendations. The former term is used predominately throughout this paper not only to provide a wider range of discussion of this emerging discipline but also to better differentiate the technologies in place now from a vision that might be several years down the road.

3.0 The Vision of the Semantic Web

> "The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation."
> *Tim Berners-Lee, James Hendler, Ora Lassila (2001)*

According to the World Wide Web Consortium (W3C), the Web can reach its full potential only if it becomes a place where data can be shared, processed, and understood by automated tools as well as by people. For the Web to scale, tomorrow's programs must be able to share, process, and understand data even when these programs have been designed independently from one another.

Still in its definition stage, the term Semantic Web is perhaps new to many people, even to those within IT circles. But the problems it aims to address are ones we have been struggling to solve for decades – issues such as information overload, stovepipe systems, and poor content aggregation (Daconta, Orbst, and Smith, 2003). The fundamental roots of these problems are the lack of semantic definitions in individual systems, the lack of semantic integration among data sets, and the lack of semantic interoperability across disparate systems. The Semantic Web extends beyond the capabilities of the current Web and existing information technologies, enabling more effective collaborations and smarter decision-making. It is an aggregation of intelligent websites and data stores accessible by an array of semantic technologies, conceptual frameworks, and well-understood contracts of interaction to allow machines to do more of the work to respond to service requests – whether that be taking on rote search processes, providing better information relevance and confidence, or performing intelligent reasoning or brokering.

Figure 4 shows a conceptual stack for the Semantic Web, illustrating how semantic technologies can be added to extend the capabilities of the current web.

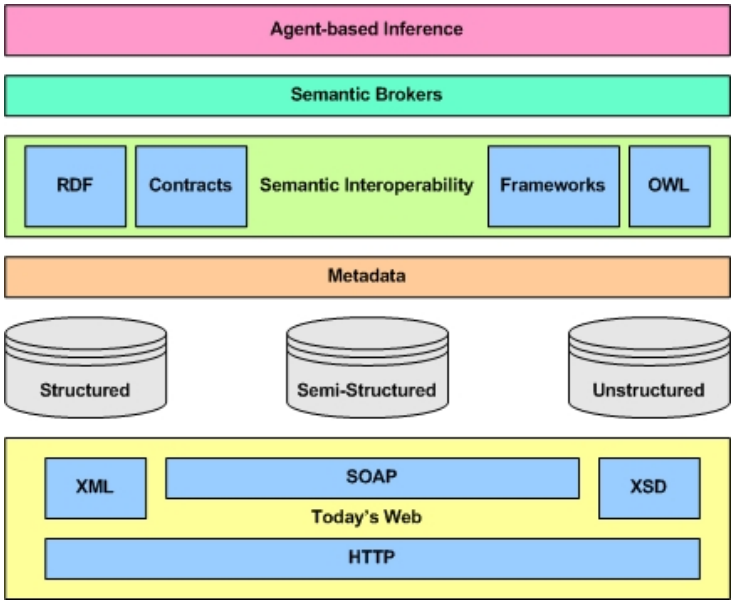


Figure 4: Semantic Web Conceptual Stack

The steps to reach this state, however, are not likely to be accomplished in a few short years. Certainly, rapid progress will be made on some ends, just as numerous websites appeared soon after the introduction of low-cost/no-cost web servers and free graphical browsers. But the progression in the development of websites moved relatively chaotically over the course of a half-dozen years – starting from an ad hoc set of scripting languages, low-end tools, and custom-built server components, and steadily progressing to a relatively unified set of core languages, application servers, content management systems, e-commerce engines, web services, and other enterprise-worthy components and offerings. The growth of the Semantic Web is likely to go through a similar progression in market dynamics. Although the business models of a connected world are better understood and the level of awareness of emerging web technologies more greatly heightened, there will nevertheless be a significant time lag until many of the pieces of the vision are assembled.

3.1 What the Semantic Web Is and Is Not

1. The Semantic Web is not a new and distinct set of websites.

The Semantic Web is an extension of the current World Wide Web, not a separate set of new and distinct websites. It builds on the current World Wide Web constructs and topology, but adds further capabilities by defining machine-processable data and relationship standards along with richer semantic associations. Existing sites may use these constructs to describe information within web pages in ways more readily accessible by outside processes such as search engines, spider searching technology, and parsing scripts. Additionally, new data stores, including many databases, can be exposed and made available to machine processing that can do the heavy lifting to federate queries and consolidate results across multiple forms of syntax, structure, and semantics. The

protocols underlying the Semantic Web are meant to be transparent to existing technologies that support the current World Wide Web.

2. The Semantic Web is not being constructed with just human accessibility in mind.

The current Web relies mainly on text markup and data link protocols for structuring and interconnecting information at a very coarse level. The protocols are used primarily to describe and link documents in the forms presentable for human consumption (but that have useful hooks for first-order machine searching and aggregation). Semantic Web protocols define and connect information at a much more refined level. Meanings are expressed in formats understood and processed more easily by machines in ways that can bridge structural and semantic differences within data stores. This abstraction and increased accessibility means that current web capabilities can be augmented and extended – and new, powerful ones introduced.

3. The Semantic Web is not built upon radical untested information theories.

The emergence of the Semantic Web is a natural progression in accredited information theories, borrowing concepts from the knowledge representation and knowledge management worlds as well as from revised thinking within the World Wide Web community. The newly approved protocols have lineages that go back many years and embody the ideas of a great number of skilled practitioners in computer languages, information theory, database management, model-based design approaches, and logics. These concepts have been proven within a number of real-world situations although the unifying set of standards from the W3C promises to accelerate and broaden adoption within the enterprise and on the Web.

With respect to issues about knowledge representation and its yet-to-be-fulfilled promise, a look at history shows numerous examples of a unifying standard providing critical momentum for acceptance of a concept. HTML was derived from SGML, an only mildly popular text markup language, and yet HTML went on to cause a sea change in the use of information technology. Many in the field point to the long acceptance timeframes for both object-oriented programming and conceptual-to-physical programming models. According to Ralph Hodgson, “knowledge representation is a fundamental discipline that now has an infrastructure and a set of supporting standards to move it out of the labs and into real-world use.”[3]

4. The Semantic Web is not a drastic departure from current data modeling concepts.

According to Tim Berners-Lee, the Semantic Web data model is analogous to the relational database model. “A relational database consists of tables, which consist of rows, or records. Each record consists of a set of fields. The record is nothing but the content of its fields, just as an RDF node is nothing but the connections: the property values. The mapping is very direct – a record is an RDF node; the field (column) name is RDF propertyType; and the record field (table cell) is a value. Indeed, one of the main driving forces for the Semantic Web has always been the expression, on the Web, of the vast amount of relational database information in a way that can be processed by machines." (Berners-Lees, 1998) That said, the Semantic Web is a much more expressive, comprehensive, and powerful form of data modeling. It builds on traditional data modeling techniques – be they entity-relation modeling or another form – and transforms them into much more powerful ways for expressing rich relationships in a more thoroughly understandable manner.

5. The Semantic Web is not some magical piece of artificial intelligence

The concept of machine-understandable documents does not imply some form of magical artificial intelligence that allows machines to comprehend human mumblings. It only indicates a machine's ability to solve a well-defined problem by performing well-defined operations on existing well-defined data (Berners-Lee, Handler, and Lassila, 2001). Current search engines perform capabilities that would have been magical 20 years ago, but that we recognize now as being the result of IP protocols, HTML, the concept of websites, web pages, links, graphical browsers, innovative search and ranking algorithms, and, of course, a large number of incredibly fast servers and equally large and fast disk storage arrays. Semantic Web capabilities will likewise be the result of a logical series of interconnected progressions in information technology and knowledge representation formed around a common base of standards and approaches.

6. The Semantic Web is not an existing entity, ready for users to make use of it.

The Semantic Web currently exists as a vision, albeit a promising and captivating one. Similar to the current Web, the Semantic Web will be formed through a combination of open standard and proprietary protocols, frameworks, technologies, and services. The W3C-approved standards – XML, RDF, and OWL – form the base protocols. New data schemas and contract mechanisms, built using these new protocols, will arise around communities of interest, industry, and intent; some will be designed carefully by experienced data architects and formally recognized by established standards bodies; others will appear from out of nowhere and gain widespread acceptance overnight. A host of new technologies and services will appear such as semantically-aware content publishing tools; context modeling tools; mediation, inference, and reputing engines; data-cleansing and thesaurus services; and new authentication and verification components. Although various elements of the vision already exist, rollout of these technologies, coordination amidst competitive forces, and fulfillment of the vision will take many years.

3.2 Near-term Benefits

While the full vision of the Semantic Web may be a bit distant, there are, on the near horizon, capabilities that many think will make enterprise software more connectable, interoperable, and adaptable as well as significantly cheaper to maintain. The use of semantic approaches in combination with the existing and emerging semantics-based schemas and tools can bring immediate and/or near-term benefits to many corporate enterprise and government agency IT initiatives.

Semantic interoperability, for example, represents a more limited or constrained subset of the vision of the Semantic Web. Significant returns, however, can still be gained by using semantic-based tools to arbitrate and mediate the structures, meanings, and contexts within relatively confined and well-understood domains for specific goals related to information sharing and information interoperability. In other words, semantic interoperability addresses a more discrete problem set with more clearly defined endpoints (Pollock and Hodgson, 2004).

Semantic technologies can also provide a loosely connected overlay on top of existing Web service and XML frameworks, which in turn can offer greater adaptive capabilities than those currently available. They can also make immediate inroads in helping with service discovery and reconciliation, as well as negotiation of requests and responses across different vocabularies. Considering the depth

and difficulty of issues the federal, state, and local agencies have in these regards, semantic technologies may provide the first flexible, open, and comprehensive means to solve them.

4.0 Key Concepts

A little semantics goes a long way.
James Hendler

Semantic computing is an emerging discipline being formed and shaped as this is written. As such, there are many definitions and interpretations, and even a few low-intensity philosophical wars being waged among thought-leaders and practitioners. That said, the release of RDF and OWL as W3C Recommendations earlier in the year has created a greater commonality in expression.

Because semantic computing makes use of various forms of abstraction and logical expression, it can be difficult to see how the languages provide many of the powerful capabilities expressed in earlier sections. But just as the Internet and World Wide Web are built upon layers of protocols and technologies, so too is the Semantic Web. Understanding several key concepts and becoming familiar with the core building blocks of the Semantic Web will form a basis for visualizing how higher order tools, components, and technologies can deliver on the promise of richer and more flexible machine-processable data. Understanding some of the foundational concepts will also allow readers to better understand the state of the technologies and the areas that still need to be refined in order to reach the full vision of the Semantic Web.

4.1 Richer Data,[4] More Flexible Associations, and Evolvable Schemas

Semantic technologies differ from database schemas, data dictionaries, and controlled vocabularies in an important way: They have been designed with the connectivity in mind allowing different conceptual domains to work together as a network. The “subway map” shown in Figure 5 is a canonical Semantic Web diagram illustrating how concepts can be connected or associated with related and/or non-related concepts.

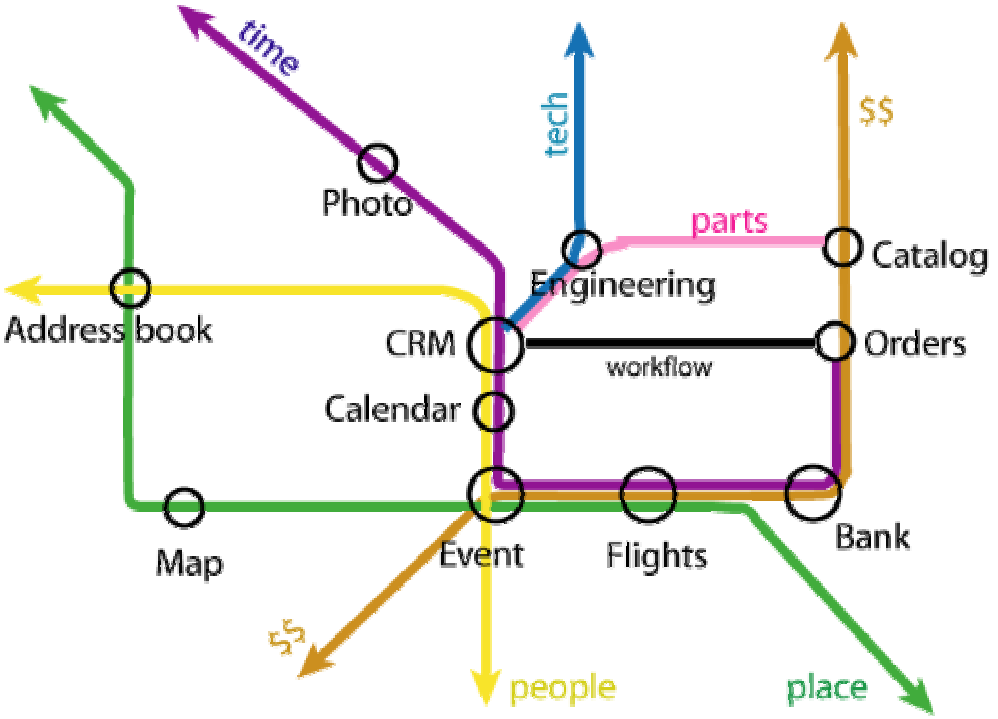


Figure 5: Semantic Web Subway Map[5]

(From Berners-Lee, ISWC, 2003)

Fundamental concepts can be seen as lines in the diagram, each identifying a particular atomic form of data such as person, price, time, or place. The intersection of one or more of these fundamental concepts forms an entity with some higher level of associated meaning. For example, the concept of an Address Book is a combination of people and addresses (and other contact information); the concept of a Catalog is a collection of parts and prices.

Although simple in nature, the diagram shows that when data is more richly described – a date or a location, for example – it can then be related in ways that are greater than the specific form or representation of data. In other words, using a search for a map of Gettysburg in 1863 as an example, if the data associated with the map is "richer" (meaning that a date is tagged or identified as a date), then intelligent searches can be made using flexible representations of date that would include a variety of date representations (July 1863; 1863; or even 1860s, for example) as well as associations with concepts (such as "Battle of Gettysburg" or "Civil War," which in and of themselves have a date ranges associated with them – July 1-3, 1963 and April 1861 to April 1865, respectively).

Likewise, in the example of the anti-spam seminar in McLean, VA, the concept of place carries with it associations that can put cities (McLean, VA) within larger more flexible boundary areas (Washington, DC, area), or connect cities, zip codes, parks, monuments, and other location-based information in a transversable conceptual domain. Intelligent searches using semantic approaches use the

combination of knowing the data type (such as a date or a location) along with flexible models of associations (many of which are still in progress) that can bridge between syntaxes, structural representations, or contexts. This idea of "decentralized, but connectable" (and some would add "evolvable") is fundamental to the vision for the Semantic Web.

4.2 Forms of Data

The structure of the data has direct bearing, at least at this point in the evolution of the technology, on what approaches are used to provide data with greater ability to describe itself to non-native processes. Enterprise data sets have many formats and structures. Not only is enterprise data different in internal binary formats, such as the difference between a text file and an object, but the information is also organized within a particular structure and representation. The continuum of data structure formats range from highly unstructured to highly structured.

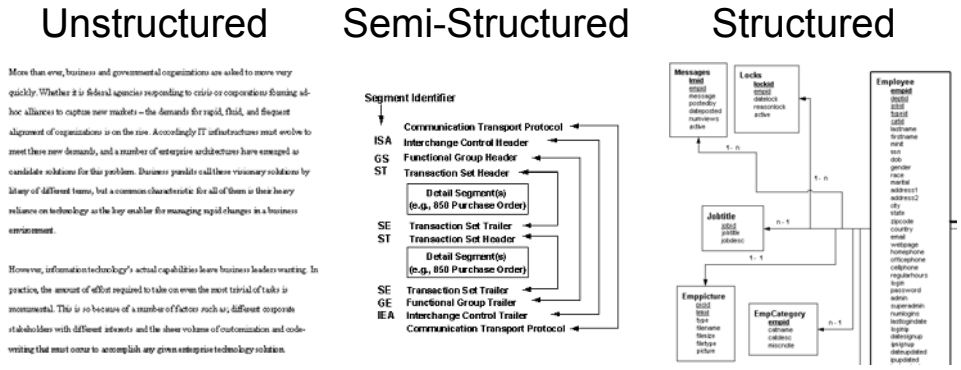


Figure 6: Data Structure Continuum

(From Pollock and Hodgson, 2004)

Structured data is the most organized of this continuum. Typically it will have definitions of metadata such as type, length, table, and constraints. Examples of structured data include relational database models, object models, and XML documents. Structured data is typically created for machine processing and consumption.

Unstructured data is data that possesses no inherent structure or relationships (aside from certain layout conventions) that can communicate its meaning aside from linear progression or other general form of organization. The canonical example is a document containing free-form text that is arbitrary in presentation and lacking in any meta-data or structure that can be useful in describing its relevance to other documents aside from title and author. Unstructured data is most commonly created for human interpretation although machines (especially in the age of the Internet) are able to do some powerful things with unstructured data.

Semi-structured data is the area in between the two but its boundaries are a little fuzzy in that there are no specific delineations as to where to draw them. For the most part, it is acceptable to think of semi-structured data as data that is organized but not explicitly defined in a highly associative way.

Traditional examples of semi-structured data include positional text messages, such as EDI, comma separated value (CSV) proprietary data files, PostScript files, or HTML files. Semi-structured data tends to be transitional data (data in transit from one system to another) or data created for a specific processing purpose and not intended as a data store in its own right (Pollock and Hodgson, 2004).

The form of the data has direct implications on what approaches are best used to describe a particular data item or body of data, to expose information about the structure and meaning of the data, and to make associations with other data elements or bodies. The body of semantic technologies, though, can work across the various forms. Within a homeland security information-sharing environment, for example, conflicting terms, definitions, and/or data representations within e-mails, RSS feeds, memos, or reports could be reconciled with data stored or obtained from structured databases and other highly structured forms.

4.3 Metadata

One of the earliest forms of supplemental data description is metadata. Metadata is quite literally "data about data." In its simplest form, metadata can be the label of a data field. Other items of metadata can include the data type or data length. XML makes use of the concept of metadata by establishing a protocol for using descriptive terms for data fields and facilitating the creation of logical schemas and namespaces around associated data elements.

The existence of a certain amount of metadata is almost a given within the concept of highly structured data sets. Less structured data sets, however, have less inherent metadata, and so a growing practice is to provide metadata by tagging data with information about itself, such tags commonly being expressed in XML. Tagging a photo as a "photo" or a map as a "map" adds tremendous value when searching a set of image files. Going further, a photo can be annotated with information concerning the subject of the photo and the date and location it was taken. A map can be categorized as a type of map such as a street map, topographical map, or battle command map, and can include a date or location associated with it.

One form of metadata, called meta tags, was included as part of the specification of well-formed Web pages and intended to provide better information about Web page content. Meta tags have fallen out of favor because search engines stopped using the tags due to issues about tricking search engines and concerns about trustworthiness. But the use of metadata in other forms is making a comeback as a fundamental data association approach within the enterprise and on the Web. New approaches for assessing the reputation or trustworthiness of a data source are also being developed, which will help increase relevance and improve confidence.

A prime example of the growing popularity of metadata is the dramatic increase in RSS feeds in 2003 and 2004. RSS stands for Real Simple Syndication and is a format for syndicating news and news-like content. Simply put, RSS is a metadata standard (expressed in XML) that is used to describe news headlines and item information (such as author and creation timestamp) within news distribution channels. RSS is a relatively lightweight metadata description form but one that is both multipurpose and extensible. The standard has been in existence for several years, but only since 2003 has it found widespread use, especially within the blogging community. Over 900,000 RSS channels exist

within the Web (as of September 2004), with thousands being added every day. Some users of this very popular standard include Reuters, W3C News, Slashdot, XML News, and others. Increased media coverage and emerging RSS development strategies within technology circles validate the viability of this technology.

The use of metadata within the enterprise has also grown steadily throughout the last several years, one impetus being the emergence of XML and common metadata schemas. When the form and meaning of metadata are agreed upon, XML is a simple yet powerful tool for making information independent of the system and application it was originally created in or resides in. Problems arise, however, when organizations or individuals implement metadata in a proprietary manner that goes undocumented and/or insufficiently described for others to understand. This proprietary approach often results in a situation in which basic information becomes unknown and largely inaccessible by anyone other than the data-store owners.

Handling inconsistencies and reconciling disparities in terminology, structure, and semantics within metadata, however, is one of the early applications for semantic technologies. For instance, when a large federal agency tried to determine the best manner through which it would be possible to share health and pollutant information brought together into a single Web portal from a variety of sources, it clearly understood the importance of metadata but faced challenges in bringing various data forms together. The overriding challenge of this project was the consolidation of disparate information in terms of both format and source (including sources not within the agency's control or circle of influence). After analyzing the problem, system designers concluded that the need to reconcile diverging terminological inconsistencies and discrepancies in meaning could best be accomplished by leveraging a metadata management tool equipped for handling such scenarios.

This tool contains, at its core, capabilities to reconcile semantic conflicts and provide normalized and consistent queries and views of the various data sources. Semantic technologies accelerate the use of metadata within the enterprise for a variety of reasons. These technologies make metadata (a) useful, (b) easily manageable, and (c) reusable. Metadata that can be reused by developers, accessed more than once by users, and guaranteed to be accurate by analysts, is metadata that improves performance and productivity. Additionally, metadata that is deemed relevant for specific needs – as well as something which can contribute to the organization as a whole – is metadata that will be invested in by employees and others.

4.3.1 Standards

Metadata standards include the Dublin Core Metadata Initiative (DCMI). DCMI is "dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems."[6] ISO 16642[7] specifies a guiding framework on the basic principles for representing data recorded in terminological data collections. This framework includes a meta-model and methods for describing specific terminological markup languages in XML. ISO/IEC 11179[8] is a 6-part standard on standardization of data elements. It specifies rules and guidelines for constructing definitions for data elements. PRISM is a publishing industry initiative developing a standard metadata vocabulary.[9] The Object Management Group offers a number of modeling and metadata specifications for use within

application and systems development. One noteworthy standard is the Meta-Object Facility (MOF). MOF is an extensible model-driven integration framework for defining, manipulating, and integrating metadata and data in a platform independent manner.[10] MOF-based standards are in use for integrating tools, applications, and data. Another highly relevant example is the OMG's Common Warehouse Metamodel (CWM) for metadata interchange.[11] (OMG and W3C are currently exploring ways of working more closely together.)

Metadata standards and/or standardization efforts also exist for a number of industries ranging from the geospatial information and healthcare to general consumer markets. Notable efforts within the geospatial realm include ISO 19115[12] and the Federal Geographic Data Committee's work with Digital Geospatial Metadata.[13] A standard gaining popularity in commercial use is called XMP (Extensible Metadata Platform) and was developed by Adobe Systems. XMP facilitates embedding metadata in files using a subset of RDF. Most notably, it supports PDF and many image formats although it is designed to support nearly any file type. Many Adobe applications can write XMP schemas plus Adobe offers an extensive XMP software development kit. Creative Commons, a nonprofit organization that facilitates digital rights management of web content, uses XMP – as well as several other metadata formats including native RDF, SMIL, and several audio formats – to embed digital rights management information in machine-processable formats.[14] This capability helps automate the management and negotiation of digital rights. By way of illustration, a query could be performed on a set of image files looking for not only specific subject matter but also for images that could be used free of charge for non-commercial use, for example.

4.4 Semantic Models (Taxonomies and Ontologies)

> It's possible to use the term 'ontology' these days and have people know what you mean.
>
> *Michael Daconta*

The pursuit of data models that can adequately and accurately describe the vast array of relationships within an organization, body of information, or other knowledge domain space is an ongoing one. The challenge is heightened when trying to arrive at approaches that are machine computational, meaning that the models can be used by computers in a deterministic and largely autonomous way. Numerous knowledge representation technologies have been devised, some successfully and some not. As a result of these efforts, computer scientists have made significant progress toward finding out the most appropriate manner in which to express highly descriptive relationships and logical concepts existing within business environments, organizational interactions, and, to a larger extent, everyday society.

Overcoming the communication gaps resulting from reliance on numerous vocabularies remains a challenge. Technical challenges have until recently had to do with overlapping and redundant terminological inconsistencies. Without knowing it, business units, individuals, and others have expended scarce resources referring to identical elements using different terminologies and different relationship models, causing confusion and limiting communication possibilities. Identifying and reconciling these semantic distinctions is a fundamental reason for using semantic models. Figure 7 displays a spectrum of commonly used semantic models.

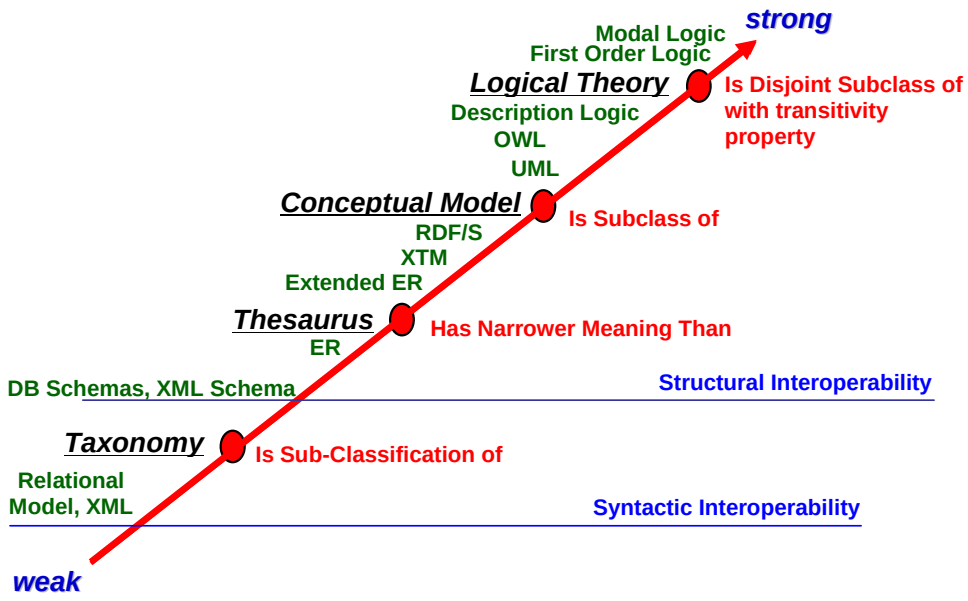


Figure 7: The Ontology Spectrum

(From Daconta, Orbst, and Smith, 2003)

Comment: This figure is unclear. The text (in red) immediately to the right of the red progression arrow is incomplete. For example: Adjacent to "Taxonomy" is the statement, "Is Sub-Classification of." Sub-Classification of WHAT?

This diagram shows a range of models, from models on the lower left with less expressive or "weak" semantics to models on the upper right with increasingly more expressive or "strong" semantics. In general, the progression from the lower left to the upper right also indicates an increase in the amount of structure that a model exhibits, with the most expressive semantic models having the most structure. Included in the diagram are models and languages that may be familiar to the reader such as the relational database model and XML on the lower left. These models are followed by XML Schema, Entity-Relation models, XTM (the XML Topic Map standard), RDF/S (Resource Description Framework/Schema), UML (Unified Modeling Language), OWL (Web Ontology Language), and up to First Order Logic (the Predicate Calculus), and higher. In truth, the spectrum extends beyond modal logic but any such discussion is still largely theoretical as well as outside the scope of this document.

One of the simplest forms of semantic model is a taxonomy. A taxonomy might be thought of as a way of categorizing or classifying information within a reasonably well-defined associative structure. The form of association between two items is inherent in the structure and in the connections between items. A taxonomy captures the fact that connections between terms exist but does not define their nature. All the relationships become hierarchical "parent-child" links. Sometimes this hierarchical structure is called a "tree," with the root at the top and branching downward. In hierarchies, there is an ordered connection between each item and the item or items below it. A common example of a taxonomy is the hierarchical structure used to describe fauna and flora within the biological sciences.

Figure 8 shows a portion of the taxonomy describing government concepts that are part of Federal Enterprise Architecture (FEA). Because of the hierarchical nature of a taxonomy, some concepts

have to be grouped under more then one category. For example, “Programs” is shown twice: once under “Agencies” and again under “Partnerships.” Taxonomies are useful for classifying things. They are not, however, useful for modeling the meanings of things.

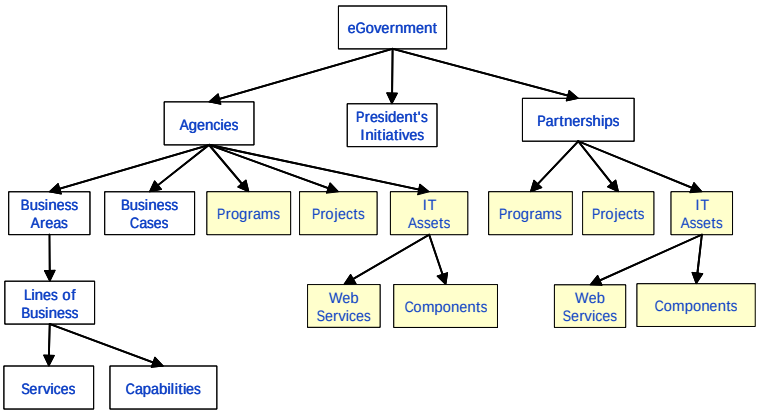


Figure 8: Example of a Taxonomy for e-Government

A thesaurus is a higher order form of semantic model than a taxonomy because its associations contain additional inherent meaning. In other words, a thesaurus is a taxonomy with some additional semantic relations in the form of a controlled vocabulary. The nodes in a thesaurus are “terms,” meaning they are words or phrases. These terms have “narrower than” or “broader than” relationships to each other. A thesaurus also includes other semantic relationships between terms, such as synonyms.

Taxonomies and thesauri are limited in their semantic expressiveness because they offer only a one-dimensional axis on which to define relationships. As such, they are typically used to create a classification system, but they fall flat when trying to represent multidimensional and/or varied conceptual domains. Concepts are the bearers of meaning as opposed to the agents of meaning. They are largely abstract and therefore more complex to model. Concepts and their relationships to other concepts, their properties, attributes, and the rules among them cannot be modeled using a taxonomy. Other more sophisticated forms of models, however, can represent these elements.

A semantic model in which relationships (associations between items) are explicitly named and differentiated is called an ontology. (In Figure 7, both conceptual models and logical theories can be considered ontologies, the former a weaker ontology and the latter a stronger ontology). Because the relationships are specified, there is no longer a need for a strict structure that encompasses or defines the relationships. The model essentially becomes a network of connections with each connection having an association independent of any other connection. Unlike a taxonomy, which is commonly shown as a “tree,” an ontology typically takes the form of a “graph,” i.e., a network with branches across nodes (representing other relationships) and with some child nodes having links from multiple parents. This connective variability provides tremendous flexibility in dealing with concepts, because many conceptual domains cannot be expressed adequately with either a taxonomy or a thesaurus. Too many anomalies and contradictions occur, thereby forcing

unsustainable compromises. Moreover, moving between unlike concepts often requires brittle connective mechanisms that are difficult to maintain or expand.

Using the map of Gettysburg as an example, the idea of using a concept such as “Battle of Gettysburg” or “Civil War” to infer a date range is difficult if not impossible using taxonomies. Having associations whereby the associations can be defined independent of an ordered relationship structure makes it possible to include a “date” or “date range” association between “Battle of Gettysburg” and “July 1-3, 1863.” As a result, an inference can be made within a search engine about a date range if it has the ability to “walk” any associations within an ontology of a concept having to do with dates. As noted previously, none of this implies some magical artificial intelligence that allows machines to comprehend human mumblings. It only indicates a machine's ability to solve a well-defined problem by performing well-defined operations on existing well-defined data.

Figure 9 shows an ontology for part of an FEA Capabilities Manager produced by TopQuadrant. The model below could be used to infer that a specific IT component has been developed in support of a given President’s initiative. The model also identifies agencies that partnered in developing a specific component.

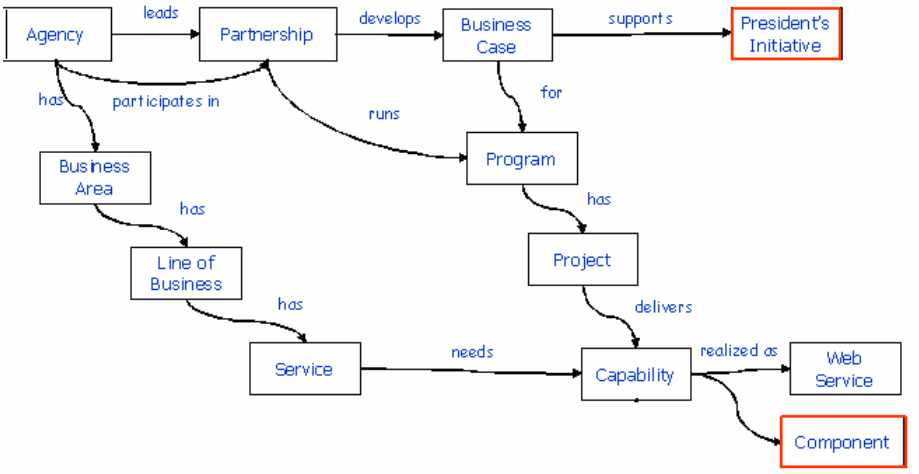


Figure 9: Part of the FEA Capabilities Manager Ontology Model

(Adapted by TopQuadrant)

Comment: This figure is blurry and difficult to read.

Simple ontologies are mere networks of connections; richer ontologies can include, for example, rules and constraints governing these connections. Just as improvements in languages and approaches to model-based programming increased the ability to move from conceptual models to programmatic models without the need for human coding steps, similar advancements have taken place within ontological development. Whereas once ontologies were created primarily for human consumption, the development of robust protocols for expressing ontologies along with a growing infrastructure that support such models, provides increased capabilities for models to deduce the underlying context and draw logical conclusions based on these associations and rules.

4.4.1 Standards

The current state of the art on representing and using ontologies has grown out of several efforts that started in the 1980s. Early semantic systems initially suffered from a lack of standards for knowledge representation along with the absence of ubiquitous network infrastructures. With the advent of the World Wide Web and the acceptance of XML as a de facto standard for exchange of information on the Web, ontology efforts have started to converge and solidify. RDF, OWL, and Topic Maps (an ISO standard for representing networks of concepts to be superimposed on content resources) all use XML for serialization. This results in strongly typed representations (with public properties and fields contained in a serial format), making it easy to store and transport these models over the Web as well as integrate them with other web standards such as Web services.

A cautionary note expressed by some in the knowledge management community is that there may be a proliferation of competing ontologies, which may in turn mean continued friction in achieving seamless sharing of structure and meaning across systems. Whereas different ontologies can be aligned for automated transformation from one model to another, it typically requires a good deal of human modeling to get to that point. (Aligning ontologies of any significant size can be similar to aligning large databases, a task that often requires significant planning and effort.) These knowledge management professionals stress that significant benefits can result from using a widely shared foundational ontology, a subject that will be addressed in a later section.

5.0 Core Building Blocks

5.1 Semantic Web Wedding Cake

Tim Berners-Lee published a description of the Semantic Web Wedding Cake (or “layer cake”) in a conference talk he presented at the XML 2000 conference. The description has garnered widespread interest within the Semantic Web community and has been cited by numerous other writers and analysts.

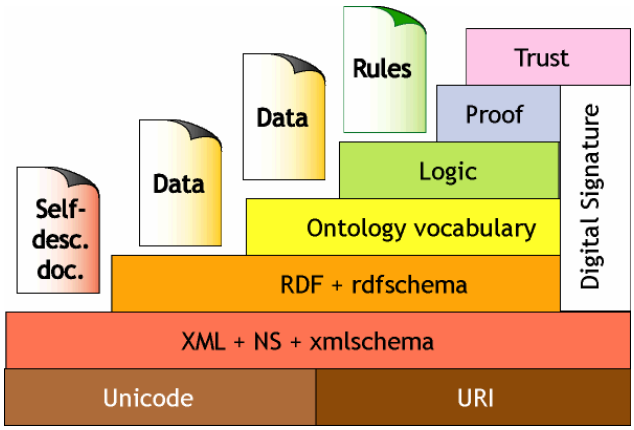


Figure 10: Semantic Web Wedding Cake

(From Berners-Lee, XML 2000 Conference)

Figure 10 serves as a corollary to the Semantic Web Conceptual Stack shown previously in Figure 4. Here, the emphasis is on the protocols and languages that will be used as foundations to technical components. The bottom of the Wedding Cake shows standards that are well defined and widely accepted. Unicode is the 16-bit character set representation is the almost universally adopted successor to ASCII. URI stands for Universal Resource Identifier and is the W3C's codification for describing the name and location of current and future objects within the Internet. It is an expansion on the concept of Universal Resource Locator or URL, which is the commonly known identifier for websites and webpages.

It should be noted that Figure 10 is a relatively informal and illustrative melding of several distinct issues. For example, ontologies are terms and definitions stated in a particular language. Logic refers to making logical inferences across a set of associated data items. Proof comes about if one keeps track of the steps in logical inference, whereas Trust refers to the origin of data, that is, whether the origin data and/or the methods used to manipulate it are trustworthy.[15] All these items are relatively distinct concepts and do not necessarily have to build on one another or even appear in the order illustrated. One can have a logic statement without an ontology. Likewise, one can have trust without logic. That said, this diagram provides a blueprint for a set of protocols and languages that will provide information technology professionals with expansive capabilities for bringing about truly adaptive computing.

5.2 Languages

5.2.1 XML (eXtensible Markup Language)

XML stands for eXtensible Markup Language and is a standard way of describing, transporting, and exchanging data that was pioneered by the W3C in the late 1990s. XML serves as a mechanism for marking up data through the use of customized "tags" in such a way as to describe the data. XML is not necessarily related to HTML and, in fact, the two were designed for entirely different purposes. Despite this fact, the two can complement each other in various ways, depending on a user's needs.

The tags are typically the labels for the data such as "FirstName" or "StreetAddress." When trying to use XML to define a standard interchange format, it is important to have agreement on the tags. For example, two book suppliers might wish to formalize a partnership involving data exchange. Specifying at the outset that Supplier A's definition of "Author" is identical to Supplier B's definition of "Writer" and codifying that in the XML structure would be an essential part of formulating proper data agreement. Additional terms that overlap and have the same meaning would also need to be formally identified, usually in something called a DTD or XML Schema. (XML Schema is a mechanism for defining XML documents in a formal way, thereby ensuring the accurate exchange of information.)

Examples of XML Schemas in working use can be found in many government and industry registries.[16] According to the U.S. CIO Council XML Working Group, "The full benefits of XML will be achieved only if organizations use the same data element definitions and those definitions are available for partners to discover and retrieve. A registry/repository is a means to discover and retrieve documents, templates, and software (i.e., objects and resources) over the Internet. The registry is used to discover the object. It provides information about the object, including its location. A repository is where the object resides for retrieval by users."[17]

In the context of semantics and the Semantic Web effort, XML is a set of syntax rules for creating semantically rich markup languages in particular domains. XML allows users to add arbitrary structure to their documents but says nothing about what the structures mean (Berners-Lee, Hendler, and Lassila, 2001). In other words, whereas IT systems, databases, and content management systems have become good at describing things, they have not done so well at describing associations. More concrete and faithful descriptions are needed that provide better senses of words, terms, and domains.

5.2.2 RDF (Resource Description Framework)

RDF stands for Resource Description Framework and has been specifically designed to provide this associative information. RDF offers ways to make data richer and more flexible, and therefore able to exist in environments outside those explicitly defined by system programmers and data modelers. RDF encodes information in sets of triples, each triple being rather like the subject, verb, and object of an elementary sentence. (This same model can also represent resource, property, and value constructs.) RDF provides an infrastructure for linking distributed metadata and also serves in conjunction with OWL as a core language for describing and representing ontologies.

One of the primary benefits of using RDF to describe data associations is the scalability and flexibility it provides. Explicit database tables can be created that do much the same thing but the unique nature of RDF provides a flexible mechanism that allows far greater associative capabilities, thereby increasing the ability to query and make inferences on topic matters not explicitly hard-wired into tables. The benefits only increase when trying to integrate new data sources, especially when they have different structures or semantics or, more importantly, when they cross conceptual domains (as in the case of environmental and public health data or, alternatively, law enforcement and intelligence data).

RDF triples are serialized in XML, providing a way to describe relationships between data elements using XML tags or other syntax in a format that can be easily processed by machines. In an effort to support a loosely coupled and/or virtual architecture, a Universal Resource Identifier (URI) is used to identify each of the triple elements. The purpose of a URI is to uniquely identify a concept in the form of subject, verb, or object by linking to the origin where the concept is defined.

RDF Schema (sometimes written as RDFS or RDF-S) offers a way of semantically describing and extending RDF. It provides mechanisms for describing groups of related resources and the relationships among these resources. RDF Schema does the same thing for RDF that DTD and XML Schema do for XML.

A number of query languages for RDF have been developed within academic and industry circles. In October 2004, the W3C RDF Data Access Working Group released a draft specification for SPARQL (pronounced "sparkle"), a query language for RDF that seeks to unify the way developers and end users write, and to consume RDF search results across a wide range of information.[18]

5.2.3 OWL (Web Ontology Language)

OWL stands for Web Ontology Language. (The acronym is purposely transposed from the actual name – OWL instead of WOL – as a conscious link to the name of the owl in the book Winnie the Pooh.) Whereas RDF's primary value can be seen in enabling association and integration of distributed data, OWL's main value is in enabling reasoning over distributed data.

OWL is a highly expressive modeling language that is compatible with existing data stores and modeling constructs including XML, Rational, and object-oriented approaches. OWL also provides loosely-coupled “views” of data which makes federated knowledge bases easy to build and evolve. Most importantly, OWL has machine-actionable semantics. Run-time and design-time software tools can do “things” with models, data, metadata, rules, and logic without human assistance or highly specific application code. (Pollock, 2004)

OWL is derived from a number of efforts to develop a set of flexible and computational logic constructs, many of which go back many years. It is the next generation of the ontology language called DAML+OIL, which in turn integrated two efforts, DAML, the DARPA Markup Language, an effort that was based in the United States, and OIL, the Ontology Inference Layer (or Language), an effort that was based in Europe. It also has roots in SHOE (Simple HTML Ontology Extensions), an effort led by James Hendler at the University of Maryland, created specifically for incorporating machine-readable knowledge into web documents thereby facilitating intelligent agent capabilities. There are three levels of OWL defined (OWL Lite, OWL DL, and OWL Full) each having progressively more expressiveness and inferencing power. These levels were created to make it easier for tool vendors to support a specified level of OWL.

RDF and OWL can operate together or separately. In some cases, supporting the distributed nature of data may be the primary objective, in which case only RDF may be used. In other cases, distribution plus reasoning capabilities may be desired, and so both RDF and OWL may be featured. In other instances, just reasoning capabilities are desired, and so OWL may suffice.

5.2.4 Other Language Development Efforts

Other languages are currently being developed to address additional layers within the Semantic Web vision. For example, a Rule language will provide the capability to express certain logical relationships in a form suitable for machine processing. This language will allow the expression of business rules and will provide greater reasoning and inference capabilities. RuleML was initially proffered as a rule language, although efforts to formalize the Semantic Web Rules Language (SWRL) are currently underway at W3C. A Logic language will conceivably provide a universal mechanism for expressing monotonic logic and validating proofs. A long-term hope is to eventually make use of assertions from around the Web to derive new knowledge. (The problem here is that deduction systems are not terribly interoperable. Rather than designing a single overarching reasoning system, current activities are focused on specifying a universal language for representing proofs. Systems can then digitally sign and export these proofs for other systems to use and incorporate.)

Likewise, constructs, schemas, and architectures for inferring reputation and trust are also being developed, both within the W3C and by the larger web community. These approaches are being looked at not just to infer reputation and trust by and among individuals, but also of groups of people (such as companies, media sources, non-government organizations, and political movements), inanimate objects (such as books, movies, music, academic papers, and consumer products), and even ideas (such as belief systems, political ideas, and policy proposals). (Masum and Zhang, 2004)

One challenge faced by practitioners in the field is to create frameworks and languages with sufficient expressiveness to capture the knowledge that can be described in ambiguous human languages. At issue is how to create languages, tools, and systems that will support the easy expression of simple things, while making it possible to express complex things. Another challenge is to maintain compatibility with existing syntax standards such as HTML, XML, and RDF while dealing with issues pertaining to the readability and human accessibility of the syntax. Ultimately, better tools will be developed that will minimize these issues but in the meantime complexities within some of the higher order languages may make it more difficult to develop fully compliant implementations using current editing and modeling tools.

6.0 Semantic Tools and Components

"Semantic Web tools are getting better every day. New companies are starting to form. Big companies are starting to move."
James Hendler[19]

Several models exist that describe the lifecycle or stages of maturity that technologies go through. Typically these have four stages: entry (or definition), growth (or validation), maturity (or refinement), and decline (or consolidation). By most measures, the Semantic Web, as experienced in a publicly available format, is still in the entry/definition phase. Many of the semantic technologies, however, are well into the growth/validation phase. (The shift into maturity is often elusive; the tipping point being visible only after the fact, and at times passing through a period of hype and unmet expectations.)

Leaders in technology applications across government and private industry have been forging new paths and obtaining successful results from their semantic implementation projects. There are semantic research projects in a number of federal agencies. Semantic products are available from large and established companies such as Adobe, Hewlett Packard, and IBM, as well as from many small pioneering companies such as Unicorn, Network Inference, and Semagix. In addition, there are a number of open source and publicly available tools created by public and private research institutions and organizations.

What follows is a brief survey of commercial and open source tools that can be used to create applications powered by semantic technologies. One way to understand how these tools work together is to view them as either design-time tools or run-time tools. Design-time tools are used by document authors, system designers, and others as part of the creation, design, or authoring process. Examples include tools to create metadata or to create or populate ontologies. Other software components are used as run-time components to process queries, transform data, or otherwise produce operational results. Examples include mediation servers and inference engines.

Many of the tools are used as a set within an implementation process – for example, modeling and mapping tools during design-time in partnership with query facilities and mediation servers at run-time.

6.1 Metadata Publishing and Management Tools

The process for creating metadata about a document or data item can occur when that item is created or authored, when it is imported into a content management system or a website, or when it is viewed or read by users. It can also be added by some other explicit or implicit action at any point over the course of the existence of that data item. In other words, metadata creation is not just a one-time occurrence. Metadata can continue to accummulate and can be modified at any time by conceivably any number of people.

At content creation, authors typically connect information such as the subject, creator, location, language, and copyright status with a particular document. This information makes the document much more searchable. RSS consists essentially of this type of information, providing newsreading applications with significantly expanded capabilities for searching and filtering information. Moveable Type from a company called SixApart is one of the more popular tools in the blogging community for creating RSS-compliant documents. The increasing popularity and simplicity of RSS is causing its use to extend outside of the blogging community into the general media and even into the enterprise. Other vendors of desktop and web-authoring tools are also moving quickly to provide RSS publishing capabilities.

The creation of metadata is only one step in the process. Metadata management tools are needed in order to maintain metadata vocabularies, perform metadata-driven queries, and provide visualization tools for monitoring changes in areas of interest. An example of a website that uses metadata as a key aspect of creating a collaborative and shared system of data is Flickr, a site for people to easily upload and share digital photos. What sets it apart from other digital photos services is that it provides photo-tagging capabilities as well as an innovative interface for viewing the categories of photos. (The tags are contained in a map and vary in size depending on the frequency of the tag within the data store.) What distinguishes it from earlier metadata implementations is that the feedback loop is extremely tight, meaning that the assignment of tags is bound closely to their use. As soon as photos and sets of photos are tagged, users see clusters of items carrying the same tag. Users can easily change tags to refine the clusters.[20]

In terms of tools for querying metadata, the components are not much different than current search engines, although the inclusion of metadata makes for richer data and therefore more precise and relevant searches. Query scripts and languages will likely adapt to allow users more precision although the balance between simplicity and features is constantly in flux, especially in more publicly available search engines. As with the Flickr example above, however, new visualization tools are being developed to help users navigate through complex fields of related data.

6.2 Modeling Tools (Ontology creation and modification)

Modeling tools are used to create and modify ontologies. Knowledge modelers used them to create and edit class structures and model domains.[21] The tools often have an interface that is similar to a

file system directory structure or bookmark folder interface. They also tend to offer the ability to import, transform, and re-purpose, in whole or in part, existing ontological structures that are often in the form of database schemas, product catalogues, and yellow pages listings. Other prominent feature includes advanced mechanisms for organizing, matching, and associating similar terms and concepts.

Also, because it is a common practice for modelers to create smaller interconnected ontologies instead of a single large monolithic model – primarily for better reusability and ease of use – support for splitting, merging, and connecting models can be an important capability in the ontology editor. Some editors even support collaborative work methods and rich visualization and graphical interaction modes.

Protégé-2000 is a free ontology editor from Stanford University with a large and active user community. It features an open architecture that allows independent developers to write plug-ins that can significantly extend Protégé capabilities. Commercial modeling tools are available from a number of vendors including Network Inference, Language and Computing, and Intelligent Views.

IBM's Ontology Management System (also known as SNOBASE, for Semantic Network Ontology Base) is a framework for loading ontologies from files and via the Internet and for locally creating, modifying, querying, and storing ontologies. Internally, SNOBASE uses an inference engine, an ontology persistent store, an ontology directory, and ontology source connectors. Applications can query against the created ontology models and the inference engine deduces the answers and returns results sets similar to JDBC (Java Data Base Connectivity) result sets. As of the time of publication of this paper, SNOBASE is not, however, compatible with OWL. The Sigma ontology development and reasoning system is also a fully formed design and run-time ontology management system. It can be freely licensed although it, like SNOBASE, is not compliant with OWL.

6.3 Ontologies

Arriving at the right ontology is often a critical element of successful implementation of semantics-based projects. Even more so than database design, ontology creation is a highly specialized field. Not only are there not as yet a sizeable number of skilled practitioners, it can take considerable time to arrive at an ontology that successfully captures a conceptual domain. As a result, it is important to look at existing bodies of work that can be used (and reused) in lieu of having to create something from scratch. Likely sources of existing ontologies can typically be located in close association with ontology modeling tools, several of which are named above. Use of proprietary ontologies may be contingent upon licensing of the modeling tools, a practice which is not unreasonable considering the efforts expended to develop the ontologies. Other ontologies, however, may be open and free for use for commercial and non-commercial purposes, much in the vein of Linux, JBoss, Wikipedia, Musicbrainz, and other open source software and data repositories.

Current ontology development efforts vary in scope and size. Some ontologies have been developed specifically in answer to localized implementations such as reconciling charts of accounts or health care records, areas where the emphasis is primarily on information interoperability – arbitrating between syntaxes, structures, and semantics – and less on logic programming. Other ontology

development efforts take a more top-down approach under the assumption that a shared view of a wide knowledge domain is critical to widespread proliferation of adaptive computing and intelligent reasoning capabilities. There is significant advocacy in these latter circles on the establishment of an enterprise-wide common upper ontology under the belief that it will provide the foundation for any number of domain ontologies. New domain ontologies could be extensions of, and fully compliant with, this upper ontology. Existing ontologies and legacy data models could be mapped to this upper ontology, which theoretically would constitute a significant number of the steps toward achieving greater semantic interoperability across domains. (It should be noted, however, that additional development and engineering is still needed to demonstrate the feasibility and scalability of this approach.)

Several candidate upper ontologies now exist, including DOLCE (Gangemi, et al., 2002), Upper Cyc (Lenat, 1995), and SUMO (Niles and Pease, 2001), but none of these as yet has gained significant market adoption. Proponents of this upper ontology approach believe that were the U.S. Department of Defense and/or the Federal Government to adopt one of these candidates, there is a good chance industry would follow, after which the US could then propose it as a standard to the International Standards Organization.

Even where domain-specific ontologies do not exist, it is possible to jumpstart development by making use of existing taxonomies, XML standards, or other lower order data models. At the federal level, the Knowledge Management working group (http://km.gov) has made significant progress in sharing information about taxonomy projects across agencies. XML.Gov (http://xml.gov) has a mission to facilitate efficient and effective use of XML across agencies in order that seamless sharing of documents and data can be achieved. Many government agencies have existing taxonomies, or have begun to develop taxonomies for their information domains. JusticeXML, for example, is an impressive body of work that could be extended and enhanced by RDF and OWL to provide a more flexible data model, an effort that could pave the way for far easier access to federal, state, and local law enforcement information by other agencies.

6.4 Mapping Tools (Ontology population)

Once an ontology model is created, it needs to be populated with data (referred to as class instances in “ontology speak”). This process is usually accomplished by linking various data sources to the concepts in an ontology using a mapping tool. Once “maps” have been created, a query in one data source could be transformed by its map to the ontology and then from the ontology to the other data sources using their maps. The corresponding data could then be returned in the same manner without any of the data stores knowing or caring about the others. In other words, each data source may have a unique “map” to an overarching ontology that acts as a pivot table among the various sources and targets. Providing this abstraction layer requires some effort on the part of creating the ontology and then creating the data maps, but once this has been done each data source can interoperate with other data sources strictly within run-time processes. Bringing new data sources onboard will, in most cases, have little or no effect on existing data sources.

This process drastically reduces the amount of data value mapping and semantic conflict resolution that typically takes place using current enterprise application approaches – approaches that up to

now typically require n-squared mappings (mapping from and to each and every data source) or alternatively, exporting to hard-coded, inflexible, and explicit standards. The modeling and mapping process makes the process far less political and far more flexible and adaptable. Anomalies specific to a single data source, for example, can be handled almost transparently, whereas addressing such anomalies within the typical standards process would entail expending significant time and energy. Most of the tools used to handle structured data forms have features that automate the process of mapping database fields to ontologies. Network Inference and Unicorn are two vendors with tools of this type. Tools that aggregate, normalize, and map unstructured data forms to ontologies typically work with a variety of unstructured data forms including Word, RTF, text files, and HTML. Semagix is a leading vendor for unstructured data.

6.5 Data Stores

Ontologies and other RDF data models can be stored in native RDF data stores or in relational databases that have been customized to support associative data techniques. Native RDF-data stores are inherently designed to support the concept of triples and can offer an efficient out-of-the-box approach to storing ontologies. RDF native databases are available from companies such as Tucana Technologies and Intellidimensions. Several high-quality open source RDF data stores also exist, including Kowari, Redland, Sesame, and 3Store. To use a relational database, the database must be designed in a somewhat non-traditional way. Instead of having a table that describes each major concept, the database design typically mimics the concept of triples by using a single table containing four columns. Three of the columns store the triple while the forth column is used to store its identification tag. (A report entitled “Mapping Semantic Web Data with RDBMSes” is an excellent resource for finding out more about implementing triple stores in relational databases.) [22]

Issues related to representing, storing, and querying using triples (i.e., RDF) versus traditional relational approaches, as well as the use and/or co-existence of the two types of data stores within implementations, are still working themselves out within industry and the marketplace. Each store-and-query facility provides unique capabilities that the other, at present, does not. RDF is great for situations when it is difficult to anticipate the types of queries that will be performed in the future. It is also terrific for handling metadata and for making queries that require inferences across imprecise or disparate data. For example, a query along the lines of, “How many energy producers qualify for ‘green’ status this year?” is much easier to perform using an RDF query language than in SQL (once the models have been created to tie together various data stores). At the same time, queries that are trivial in SQL, such as, “Which energy producers reduced their CO_2 output the most this year?” can be quite complicated using an RDF query language.

It is important to note that RDF query languages are still evolving, which may to some extent explain this limitation. Other limitations of RDF relate to performance issues. Because queries can be broadened, for example, to include concepts instead of just terms, the search space can be dramatically increased. Because RDF data stores are relatively new and the number of implementations relatively small, system developers need to iterate over their designs, paying particular attention to queries and functions that could have negative effects on performance. In terms of industry growth, it is difficult to predict how RDF will affect the database industry. RDF data stores

may remain a distinct data storage category in their own right or their capabilities may be subsumed into relational databases in a manner similar to what happened with object-oriented databases.

6.6 Mediation Engines

Mediation engines are automated tools that can dynamically transform data among different syntaxes, structures, and semantics using models instead of hard-wired transformation code. They are critical components of any interoperability architecture. Using data maps, ontologies, and other forms of conceptual models, mediation engines are run-time processes that provide an abstraction layer between heterogeneous data sets, allowing organizations to essentially agree to disagree about how data and information should be represented. Mediation engines typically work with highly structured data. Unstructured and semi-structured data must first be bound to a schema prior to creating the mediation maps (Pollock, 2004).

6.7 Inference Engines

Inference engines (sometimes referred to as reasoners) are software tools that derive new facts or associations from existing information. It is often said that an inference engine emulates the human capability to arrive at a conclusion by reasoning. In reality, inferencing is not some mythical artificial intelligence capability but, rather, a quite common approach in data processing. One can think of a complex data mining exercise as a form of inferencing. By creating a model of the information and relationships, we enable reasoners to draw logical conclusions based on the model. A common example of an inference is to use models of people and their connections to other people to gain new knowledge. Exploration of these network graphs can enable inferences about relationships that may not have been explicitly defined.

Note that with just RDF and OWL, inferences are limited to the associations represented in the models, which primarily means inferring transitive relationships. With the addition of rule and logic languages, however, greater leaps in conceptual understandings, learning, and adaptation can take place, although implementations with these types of capabilities are, as yet, few and far between.

Both free and commercial versions of inference engines are available. For example, Jena, an open source Java framework for writing Semantic Web applications developed by HP Labs, has a reasoner subsystem. Jena reasoner includes a generic rule based inference engine together with configured rule sets for RDFS and for the OWL-Lite subset of OWL Full. JESS is a popular OWL inference engine from Carnegie Melon University. Network Inference offers a commercial reasoner based on description logic (OWL-DL).

6.8 Other Components

Ordinary web pages are a good source of instance information; many tools for populating ontologies are based on annotation of web pages. W3C Annotea project offers free annotation tools. Commercial vendors include Ontoprise and Lockheed-Martin. Several software vendors, including Semagix, Siderian Software and Entopia, offer products that use ontologies to categorize information and to provide improved search and navigation.

7.0 Applications of Semantic Technologies

> Semantic technologies can solve problems that, using current technologies, are unsolvable at any price.
> *Don Hall*[23]

There are a wide variety of applications where semantic technologies can provide key benefits. At their core, semantic approaches are an infrastructure capability that, when combined with other key technologies, represent the next wave of computing. When taken with a multi-year view, there is great promise that these technologies will help the IT industry reach the ever-elusive goal of truly adaptive computing. In some respects, though, the future is already happening. Commercial enterprises and government agencies are implementing production-level programs using existing semantic data stores, ontologies, toolsets, and applications. A few of these near-term project areas include Semantic Web services, semantic interoperability, and intelligent search.

7.1 Semantic Web Services

A Web service is a software system designed to support interoperable machine-to-machine interaction over a network. A Web service has an interface described in a machine-processable format using Web Services Description Language (WSDL).[24] The combination of WSDL, UDDI,[25] and SOAP[26] form a triad of technologies that will shift the entire market toward service-oriented architectures (SOA). Together, these technologies provide directory, component lookup, and exchange protocol services on top of an HTTP or SMTP network protocol.

Microsoft, IBM, and most other large software vendors have embraced the concepts and languages that underlie the Web services model, and an increasing number of books and industry articles point to the benefits of adopting a service oriented architecture. Web services, however, are not without shortcomings. Security issues have long been a concern but several solutions that address these issues have been introduced over the last several years. Perhaps the most significant improvement opportunities for Web services that remain are in the areas of (a) flexible look-up and discovery and (b) information management and schema transformation. Fundamentally, Web service technologies handle messages in a loosely coupled manner but they do not currently bridge differences in description terminologies nor do they inherently enable the recipient to understand a message that has been sent.[27] With Web services, these parts of the exchange rely on custom-coded solutions and/or widespread community agreement upon some kind of document exchange standard (the latter is rarely achieved).

This difficulty in ensuring flexible discovery and service initiation, as well as seamless operational use of information exchanged with Web services, has led to W3C's efforts to incorporate semantic technologies as part of its Semantic Web Services initiative. Semantic Web Services are a Web Service implementation that leverages the Web Ontology Language Service specification (OWL-S) to provide a flexible framework for describing and initiating web services. OWL-S supplies Web service providers with a core set of markup language constructs for describing the properties and capabilities of their Web Services in unambiguous, computer-interpretable form. OWL-S markup of Web services will facilitate the automation of Web service tasks, including automated Web service discovery,

execution, interoperation, composition, and execution monitoring. Following the layered approach to markup language development, the current version of OWL-S builds upon W3C's standard OWL.

7.2 Semantic Interoperability

Office of Management and Budget (OMB) directives and guidance call for the unification and simplification of business processes and information technology across the federal government. In order to achieve this goal, each agency must ensure that its information can be readily shared across the federal government. In an environment where agencies must collaborate but have diverse terminology and definitions, information sharing requires interpreting the meaning of data in different contexts. Semantic technologies support this requirement by offering a framework for connecting distributed data and describing it in a context-sensitive way.

Formally put, the use of semantic technologies makes it possible to describe the logical nature and context of the information being exchanged, while allowing for maximum independence among communicating parties. The results are greater transparency and more dynamic communication among information domains irrespective of business logic, processes, and workflows (Pollock and Hodgson, 2004).

The technical vision is one where flexible information models, not inflexible programs or code, are used to drive dynamic, self-healing, and emergent infrastructures for the sharing of mission critical data in massively scaleable environments. Recent advances in taxonomy and thesaurus technology, context modeling approaches, inferencing technology, and ontology-driven interoperability can be applied in a cohesive framework that dramatically changes the way information is managed in disperse, decentralized communities of knowledge (Pollock and Hodgson, 2004).

NASA views semantic interoperability as an extremely promising way to make information available to all stakeholders without having to standardize on a particular format or vocabulary or re-key databases to conform to a uniform model. One example where NASA is using these concepts is to address serious and ongoing maintenance problems related to the aging wiring systems within the Space Shuttle fleet. The existing set of wiring system databases contains information about part specifications, bills of materials, drawings, change orders, standard practices, test procedures, test reports, inspection reports, failure tracking and reporting information, work orders, and repair disposition documentation. Tens of diverse databases and systems – each supporting different but related aspects of engineering and design work – are in use within NASA with related data dispersed among several contracting companies that support the Space Shuttle program. Troubleshooting wiring problems requires timely access to many cross-organizational systems, databases, and knowledge repositories, the breadth of which is enormous. The situation becomes especially critical for diagnosing and troubleshooting in-flight anomalies whereby a timely resolution is mission-critical as well as life-critical.[28] The work to make these sets of data richer and more accessible to the numerous parties who need access to them is still in its early stages, but as highlighted in the quote at the beginning of this chapter, semantic technologies represent one of the more promising ways to address what is largely an unsolvable problem using current data integration approaches.

A highly distilled version of how such a project works is as follows. Design-time tools are used to develop RDF and OWL models that encompass a particular domain. These models could be based on existing XML standards or defined via other means. Other design-time tools can be used to flexibly map specific data representations to these models, thus eliminating the need to explicitly convert applications to adopt a certain data standard. Run-time processes can then use these models and maps as pivot tables to transform data from source to target or to perform federated queries from a single query statement. Semantic interoperability frameworks of this type can provide a solid basis for better resolving differences in syntax, structure, and semantics – ushering in a future where organizations can agree to disagree, yet still share data and interoperate without having to change their current methods of operation.

One of the key advantages of using semantic interoperability approaches is that they do not necessarily require the replacement of existing integration technologies, databases, or software applications. A semantic framework made up of various semantics-based components and application program interfaces (APIs) can be deployed with web services or traditional middleware APIs to leverage existing infrastructure investments, and yet still provide massive benefits by virtually centralizing the query, transformation, and business rules metadata that flows through the network infrastructure's pipes. As such, the software will fit into the customer's existing IT ecosystem with low overhead for installation, minimal coding, and maximum reusability (Pollock and Hodgson, 2004).

7.3 Intelligent Search

Related in some regards to semantic interoperability is the area of intelligent search. As mentioned above, semantic interoperability techniques can allow queries native to one system to be federated to other non-native systems. This eliminates the need to convert systems to a universal query language and enables systems to continue maintaining the information they have in their current formats. By overlaying a virtual layer on top of the data sources, queries can be defined in a universal manner, thereby enabling access to all mapped assets. Federated searching can also be made smarter by making searches more semantically precise. In other words, searches can be broadened to include concepts, or narrowed to include only specific key words. The depth – or granularity – of such searches enables the specification of the search that the individual desires.

Another aspect of intelligent search is the ability to make searches more relevant to the person searching by making use of identity and relationship information. Relationships among people and information about them can be key links to greater relevance and confidence. Despite investments in knowledge management systems, many people still rely on their personal network of friends, neighbors, co-workers, and others to locate experts or find trusted information. Personal relationships are also useful in sales situations and in many organizational interactions. Social networking schemas and software are making broad use of this.

An example of how this information can be used on a larger scale is the case of a telephone company exploring technologies for providing more intelligent phone number look-up. Instead of providing a generic list of matched names, the telecommunications company is looking at combining information about the person searching and the list of possible names, in order to provide a more intelligent match. For example, inferring relationships between social networks could provide information on

whether a person is known, or could be expected to be known, to the other person (by employing friend-of-a-friend forms of calculations). Other information such as locations or schools attended or past or current jobs could be used to infer matches. To be sure, there are significant privacy issues involved; many believe, however, that techniques such as hashing and encryption[29] of personally identifiable information and progressive disclosure will likely resolve many privacy concerns.

Semantic approaches for enabling intelligent search are beginning to find their ways into knowledge management systems. Whereas current knowledge management systems tend to exist within their own silos and have difficultly crossing organization boundaries, intelligent search techniques can be added as overlays to existing information infrastructures, thereby bridging physical data formats, knowledge domains, and organizational structures.

8.0 Additional Topics

Additional modules are in development that will explore additional topics about semantic technologies and the Semantic Web. The second module in this white paper series will examine the business case for the semantic interoperability in the federal government. It will use business scenarios and storyboarding approaches to describe why and how semantic interoperability can deliver ROI-supported value. These business scenarios will contain detailed descriptions of the business problem, expressed both in business and in architectural terms.

A third module will provide a roadmap for agencies on how they can take advantage of semantic technologies and begin to develop Semantic Web implementations. New technologies, applications, and services are being developed to take advantage of these new advances. This module will provide steps and implementation recommendations, whereby agencies can map their progress and schedule future projects in ways that optimize adoption time and minimize development friction.

9.0 References

Bedford, Denise. “Charter Statement of Taxonomy and Semantics Special Interest Group.” 2004. <http://www.km.gov>

Berners-Lee, Tim, James Hendler, and Ora Lassila. “The Semantic Web.” ScientificAmerican.Com, May 17, 2001.
< http://www.scientificamerican.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21 >

Berners-Lee, Tim. “Semantic Web: Where to Direct Our Energy?” Invited Talk at the International Semantic Web Conference (ISWC), October 2003.
<http://iswc2003.semanticweb.org/>

Berners-Lee, Tim. “Semantic Web Road Map.” W3C. September 1998.
<http://www.w3.org/2003/Talks/0521-www-keynote-tbl/ >

Berners-Lee, Tim. “Web Services – Semantic Web.” W3C. 2003.
<http://www.w3.org/2003/Talks/0521-www-keynote-tbl/ >

Berners-Lee, Tim. "What the Semantic Web Can Represent." W3C. September 1998.
<http://www.w3.org/DesignIssues/RDFnot.html>

Daconta, Michael C., Leo J. Obrst, and Kevin T. Smith. *The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management.* Indianapolis: Wiley, 2003.

Daconta, Michael C. "10 Practical Reasons Why You Need an Ontology.”
<http://www.daconta.net/briefs/why-ontology-general.pdf>

Daconta, Michael C. “The Semantic Web Foundations of Net-Centric Warfare.”
<http://www.daconta.net/Articles/Net-centricWarfareWhitepa.html>

Gangemi, Aldo, Nicola Guarino Claudio Masolo, Alessandro Oltramari, and Luc Schneider. “Sweetening Ontologies with DOLCE.” EKAW 2002, Siguenza, Spain, October 1-4, 2002, Proceedings. Lecture Notes in Computer Science 2473. Springer, 2002.
<http://www.loa-cnr.it/Papers/DOLCE-EKAW.pdf>

Fromm, Kenneth. “The Semantic Web in the Enterprise - EAS Speakers Shed Some Light on SemanticTechnologies and Their Roles in Web Service and XML Frameworks.” Enterprise Architect, Summer 2004.
<http://www.ftponline.com/ea/default.aspx>

Fromm, Kenneth, and Jeffrey T. Pollock. “Semantic Computing’s Building Blocks.” Enterprise Architect, April 15, 2004 (Vol.2, No.8).
<http://www.ftponline.com/ea/default.aspx>

Lenat, Douglas B. “CYC: A Large-Scale Investment in Knowledge Infrastructure.” Communications of the ACM, November 1995 (Vol.38, No.11).

Masum, Hassan, and Yi-Cheng Zhang. “Manifesto for the Reputation Society.” First Monday, July 2004 (Vol.9, No.7).
<http://www.firstmonday.org/issues/issue9_7/masum/index.html>

Niles, I., and A. Pease. “Towards A Standard Upper Ontology.” Proceedings of Formal Ontology in Information Systems (FOIS 2001), October 17-19, Ogunquit, Maine, USA, pp 2-9. See also <http://www.ontologyportal.org>.

Obrst, Leo. “Ontologies for Semantically Interoperable Systems.” MITRE, Center for Innovative Computing & Informatics. Presentation to the KM.Gov Semantics Interoperability Community of Practice. April 14, 2004.

Obrst, Leo. “Ontologies and the Semantic Web: An Overview.” MITRE, Center for Innovative Computing & Informatics. July 13, 2004.

Obrst, Leo J., Howard Liu, and Robert Wray. “Ontologies for Corporate Web Applications.” *AI Magazine* (Fall 2003): pp. 49-62.
<http://www.findarticles.com/p/articles/mi_m2483/is_3_24/ai_110575583/>

Pease, A. “The Sigma Ontology Development Environment.” In working notes of the IJCAI-2003 Workshop on Ontologies and Distributed System, Heiner Stuckenschmidt, ed, Acapulco, Mexico, August 2003.

Pollock, Jeffrey T, and Ralph Hodgson. *Adaptive Information: Improving Business Through Semantic Interoperability, Grid Computing, and Enterprise Integration.* Wiley-Interscience, September 2004.

Pollock, Jeffrey T. “The Evolution of Integration to EAI, EII, and Onward to Semantics.” ManTech EII Workshop, July 2004
<www.meiim.com/files/Jeffrey_Pollock.ppt>

Sonntag, William. Submission to “Problem Statements for Semantic Technology Panels – Interactive Discussion Session" in the One-Day Conference on "Semantic Technologies for eGov" White House Conference Center, Monday, September 8th, 2003. United States Environmental Protection Agency, Office of Environmental Information.
<http://www.topquadrant.com/documents/Sept%208th.%20-%20Collected%20Problem%20Owner%20Statements-less%20DIA,%20DCMA.pdf>

TopQuadrant. “Harnessing the Value of Semantic Integration For Your Business.” TopQuadrant Whitepaper. June 15, 2004.

TopQuadrant. “Semantic Technology, Version 1.2.” TopQuadrant Technology Briefing. March 2004.
<http://www.topquadrant.com/documents/TQ04_Semantic_Technology_Briefing.PDF>

Udell, Jon. “Collaborative Knowledge Gardening.” Infoworld.com, August 20, 2004.
<http://www.infoworld.com/article/04/08/20/34OPstrategic_1.html>

W3C. “Resource Description Framework.” World Wide Web Consortium. August 2004.
<http://www.w3.org/RDF/>

W3C. “Web Ontology Language (OWL).” World Wide Web Consortium. August 2004.
<http://www.w3.org/2004/OWL/>

10.0 Endnotes

[1] The World Wide Web Consortium's Semantic Web Activity Statement: <http://www.w3.org/2001/sw/Activity - intro>

[2] The World Wide Web Consortium's Semantic Web Activity page: <http://www.w3.org/2001/sw>/

[3] Ralph Hodgson, Semantic Web in the Enterprise Panel, Enterprise Architect Summit, June 2004.

[4] The term "rich data" is used to described data that has greater fidelity and independence from the system or application in which it resides. The term "smart data" is a term popularized by Michael Daconta and others. While "smart data" is sometimes used analogously to "rich data," its definition has another condition and that is that it contains logical constructs that enable inference and higher order processing. Rather than noting the subtleties between the definitions, however, readers are better served by focusing on their commonality, which is that the more autonomous and self-describing that data is, the greater its use outside of its native system or application.

[5] Tim Berners-Lee, Semantic Web Status and Direction ISWC2003 Keynote, October 2003. <http://www.w3.org/2003/Talks/1023-iswc-tbl/slide10-0.html>

[6] Information about Dublin Core: <http://www.dublincore.org/about/>

[7] ISO 16642:2003: Computer Applications in Terminology – Terminological Markup Framework <http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=32347&ICS1=1&ICS2=20&ICS3=>

[8] ISO/IEC 11179: Information Technology – Metadata Registries <http://metadata-standards.org/11179/>

[9] Information about PRISM: <http://www.prismstandard.org/>

[10] Information on OMG's Meta-Object Facility (MOF): <http://www.omg.org/technology/documents/formal/mof.htm>

[11] Information on OMG's Common Warehouse Metamodel (CWM): <http://www.omg.org/cwm/>

[12] ISO 19115:2003: Geographic Information – Metadata <http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020&ICS1=35&ICS2=240&ICS3=70>

[13] Information on the Federal Geographic Data Committee (FGDC) and the Content Standard for Digital Geospatial Metadata (CSDGM): <http://geology.usgs.gov/tools/metadata/tools/doc/faq.html> <http://www.fgdc.gov/metadata/contstan.html>

[14] Information about Creative Commons and Creative Commons metadata-supported formats: <http://www.creativecommons.org/> and <http://creativecommons.org/technology/usingmarkup>

[15] Trust and trustworthiness in the context of data and users within the World Wide Web and the Semantic Web are rapidly evolving concepts. Traditional IT definitions often allude to performance characteristics of data such as integrity (completeness), timeliness (currency, up to date), reliability, or accuracy. Real world implementations are increasingly taking into account the reputation of the data provider via methods such as the accumulation of karma points (Slashdot), ratings by other users (eBay), or other implicit or explicit actions of data providers or data consumers. Social and organizational mechanisms are evolving to help streamline massively distributed collaborative developments all the while ensuring high quality output. Open source

development, in the case of Linux, and open content creation, in the case of Wikipedia, are two primary areas where these mechanisms can be seen in action.

[16] An XML registry for the exchange of environmental data can be found at:
<http://oaspub.epa.gov/emg/xmlsearch$.startup>

[17] Information about XML.gov XML Registry/Repository efforts:
<http://xml.gov/registries.asp>

[18] SPARQL Query Language for RDF:
<http://www.w3.org/TR/2004/WD-rdf-sparql-query-20041012/>

[19] James Hendler, Closing Keynote, Semantic Technologies for E-Government, September 2003.

[20] "Flickr, as I would explain it to my friends and family, is a way to easily upload and share digital photos. And del.icio.us does the same thing, only for Web bookmarks. To CTOs, though, I'd say that both are collaborative systems for building a shared database of items, developing a metadata vocabulary about the items, performing metadata-driven queries, and monitoring change in areas of interest. In the case of Flickr, an item is a photo; in the case of del.icio.us, it's a URL. But the same methods could apply to any of the shared digital artifacts that we create, find, and use in the course of our daily work." (Udell, 2004)
Information on Flickr:
<http://www.flickr.com/>

[21] The paper addresses tools for ontology development but not the skill sets or processes of those working with ontologies and other logic models. This may give a false impression that the current state of modeling and mapping tools largely automate the creation and use of ontologies with little thought process required by humans. This belief is incorrect. As mentioned earlier in the paper, using semantic models is like moving from flat-file databases to relational databases or like moving from procedural programming techniques to object-oriented approaches. It will take a bit of time for people to understand the nuances and architectures of semantics-based approaches as well as time for the tools to mature, specifically as they relate to modeling and mapping of ontologies and data structures.

[22] Mapping Semantic Web Data with RDBMSes
<http://www.w3.org/2001/sw/Europe/reports/scalable_rdbms_mapping_report/>

[23] Don Hall, Program Director, Logistics Enterprise Support Program, in support of the Assistant Deputy Under Secretary of Defense for Logistics Systems Management, in numerous conversations and presenations.

[24] From W3C Working Note on Web Services Architecture:
<http://www.w3.org/TR/ws-arch/#whatis>

[25] Universal Description, Discovery, and Integration (UDDI):
<http://www.uddi.org/>

[26] Simple Object Access Protocol (SOAP):
<http://www.w3.org/TR/soap/>

[27] For additional background, see "Semantic Discovery for Web Services" in Web Services Journal, April 2003.
<http://www.sys-con.com/webservices/article.cfm?id=507>

[28] Extracted from NASA presentations and reports.

[29] Hashing means using an algorithm to convert a string (a user's name and other signature information, for example) into a mathematical summary. That summary is then encrypted using various public key encryption systems.

Appendix A: Organizational Charters

The Semantic Interoperability Community of Practice (SICoP) is established by a group of individuals for the purpose of achieving semantic interoperability and semantic data integration in the government sector. SICoP seeks to enable Semantic Interoperability, specifically the "operationalizing" of these technologies and approaches, through online conversation, meetings, tutorials, conferences, pilot projects, and other activities aimed at developing and disseminating best practices. The individuals making up this community of practice represent a broad range of government organizations and the industry and academic partners that support them. SICoP, however, claims neither formal nor implied endorsements by the organizations represented.

SICoP is a Special Interest Group within the Knowledge Management Working Group (KMWG) sponsored by the Best Practices Committee of the Chief Information Officers Council, (CIOC) in partnership with the XML Working Group, among others. Both the SICoP and its parent KMWG serve as interagency bodies to bring the benefits of the government's intellectual assets to all Federal organizations, customers, and partners. SICoP will communicate its actions and findings through the KM Working Group to the Best Practices Committee, the CIO Council, and its member agencies, although its main purpose to support SICoP members in their efforts to introduce semantic technologies and evolve the Semantic Web capabilities within their agencies.

Appendix B: Glossary

Term	Definition/Description	Source
Application Program Interface (API)	An application programming interface (API) is a set of definitions of the ways in which one piece of computer software communicates with another. It is a method of achieving abstraction, usually (but not necessarily) between lower-level and higher-level software.	Wikipedia, a free-content encyclopedia <http://www.wikipedia.com/>
Blog	A weblog, or simply a blog, is a web application that contains periodic, reverse chronologically ordered posts on a common webpage. Such a Web site would typically be accessible to any Internet user. The term "blog" came into common use as a way of avoiding confusion with the term server log. Blogs run from individual diaries to arms of political campaigns, media programs and corporations, and from one occasional author to having large communities of writers. The totality of weblogs or blog-related webs is usually called the blogosphere. The format of weblogs varies, from simple bullet lists of hyperlinks, to article summaries with user-provided comments and ratings. Individual weblog entries are almost always date and time-stamped, with the newest post at the top of the page. Because links are so important to weblogs, most blogs have a way of archiving older entries and generating a static address for individual entries; this static link is referred to as a permalink. The latest headlines, with hyperlinks and summaries, are offered in weblogs in the RSS XML-format, to be read with an RSS feedreader. A weblog is often run through a content management system or CMS.	Wikipedia <http://www.wikipedia.com/>
Controlled Vocabulary	A finite set of standard terms for use in taxonomy categories. Within an organization, there can be multiple controlled vocabularies, e.g. a core vocabulary for the entire organization and a sub-controlled vocabulary specific to each business unit within the organization. Controlled vocabulary can be used in notation for taxonomy categories, information cataloging, and tagging, as well as for labels for a Web site navigation interface.	“Taxonomy Analytical Brief”, Department of State, IRM Business Center, May 27, 2003
Data	A collection of raw facts, instructions, or statements in isolation.	“Taxonomy Analytical Brief”, Department of State, IRM Business Center, May 27, 2003
Information	A set of related facts, instructions, or statements	“Taxonomy Analytical Brief”,

	about something in a given context (i.e., a specific place and time) of which you are uncertain of its truth or value.	Department of State, IRM Business Center, May 27, 2003
Information Architecture	Information architecture, in the broadest sense, is simply a set of aids that match information needs with information resources. A well implemented architectural design structures information in an organization through specific formats, categories, and relationships. It needs to consider business context, content (information) and users.	"Taxonomy Analytical Brief", Department of State, IRM Business Center, May 27, 2003
Knowledge	A set of related facts, instructions, or statements about something in a given context of which you are certain of its truth and value.	"Taxonomy Analytical Brief", Department of State, IRM Business Center, May 27, 2003
Metadata	The simplest definition of metadata is " structured data about data." Metadata is descriptive information about an object or resource whether it be physical or electronic. While metadata itself is relatively new, the underlying concepts behind metadata have been in use for as long as collections of information have been organized. Library card catalogs represent a well-established type of metadata that has served as collection management and resource discovery tools for decades. Metadata can be generated either "by hand" or derived automatically using software.	Dublin Core Metadata Initiative, Frequently Asked Questions <http://dublincore.org/resources/faq/ - whatismetadata>
Namespace	In many programming languages, a namespace is a context for identifiers. In general, a namespace is an abstract zone that is or could be populated by names, or technical terms, or words. A namespace uniquely identifies a set of names so that there is no ambiguity when objects having different origins but the same names are mixed together. In a namespace, each name must be unique. The namespace is the context, and in the namespace each word can uniquely represent (map to) a real-world concept. Each language is a namespace, whether it is a natural (ethnic) language, a constructed language, the technical terminology of a profession, a dialect, a sociolect, or an artificial language (e.g. a programming language).	Wikipedia <http://www.wikipedia.com/>
Ontology	"An ontology is a specification of a conceptualization. In the context of knowledge sharing, I use the term ontology to mean a specification of a conceptualization. That is, an ontology is a description (like a formal specification of a	Tom Gruber, Stanford University <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>

	program) of the concepts and relationships that can exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set-of-concept-definitions, but more general. And it is certainly a different sense of the word than its use in philosophy. What is important is what an ontology is for. My colleagues and I have been designing ontologies for the purpose of enabling knowledge sharing and reuse. In that context, an ontology is a specification used for making ontological commitments. The formal definition of ontological commitment is given below. For pragmetic reasons, we choose to write an ontology as a set of definitions of formal vocabulary. Although this isn't the only way to specify a conceptualization, it has some nice properties for knowledge sharing among AI software (e.g., semantics independent of reader and context). Practically, an ontological commitment is an agreement to use a vocabulary (i.e., ask queries and make assertions) in a way that is consistent (but not complete) with respect to the theory specified by an ontology. We build agents that commit to ontologies. We design ontologies so we can share knowledge with and among these agents."	
Schema	The word schema comes from the Greek word "σχήμα" (schema) that means "shape" or, more generally, "plan." The word schema can represent any of several different things: 1. In computer science, a schema is a model. 2. In formal logic, a rule (usually recursively definable) describing a set (usually infinite) of statements. For example, the axiom schema of replacement is a schema of axioms in axiomatic set theory. 3. A description of the structure of a database; or: a defined part of a database. See software architecture, conceptual schema, Sowa's conceptual graph, semantic network, Berners-Lee's semantic web. 4. An XML schema provides a means for defining the structure, content and to some extent, the semantics of XML documents. 5. Part of a formal specification written in the Z formal specification language. 6. A minimal and specialized ontology, i.e., a list of questions, answers to which describe what exists in the world. This includes only what is required for some narrow range of actions; e.g., a library card catalogue schema asks librarians only to provide enough information about the book to help library users decide if	Wikipedia <http://www.wikipedia.com/>

Comment: This definition needs to be replaced by one that actually defines the term.

	they want to browse through it, and if so, how to find it. By contrast, an ontology enables a much broader range of actions, e.g., all of those normally associated with a working trade or profession.	
Semantics	Semantics are at base the processes that use or create values for taxonomies. Without semantics, taxonomies are simple or elaborate but empty structures. Officially, semantics is a branch of linguistics that deals with the study of meaning, changes in meaning, and the principles that govern the relationship between sentences or words and their meanings. Semantics involved in creating meaning for simple taxonomies are different from those that are used to create meaning for network or faceted taxonomies. Semantics involves the study of the relationships between signs and symbols. From an information perspective, semantics also involves effective information communication within and across languages, information surrogation, information organization, and discovery.	Extracted from the Mission Statement of the Taxonomies and Semantics Special Interest Group <http://km.gov/>
Semantic Integration	Semantic integration is often used as a synonym for semantic interoperability, although some vendors use it to refer to a less comprehensive solution that builds on existing XML integration efforts.	Editor
Semantic Interoperability	Semantic interoperability is an enterprise capability derived from the application of special technologies that infer, relate, interpret, and classify the implicit meanings of digital content, which in turn drive business process, enterprise knowledge, business rules and software application interoperability.	"Adaptive Information: Improving Business Through Semantic Interoperability, Grid Computing, and Enterprise Integration" by Jeff Pollock and Ralph Hodgson, Wiley Publishing 2004
Semantic Web	Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.	"The Semantic Web", By Tim Berners-Lee, James Hendler and Ora Lassila, Scientific American, May 2001
Semantic Web Services	Semantic Web Services are a Web Service implementation that leverages the Web Ontology Language Service specification (OWL-S). OWL-S supplies Web service providers with a core set of markup language constructs for describing the properties and capabilities of their Web Services in unambiguous, computer-intepretable form. OWL-S markup of Web Services will facilitate the automation of Web service tasks including automated Web service discovery, execution, interoperation, composition and execution monitoring. Following the layered approach to markup language development, the current version of OWL-S builds on top of W3C's standard OWL.	"Adaptive Information: Improving Business Through Semantic Interoperability, Grid Computing, and Enterprise Integration" by Jeff Pollock and Ralph Hodgson, Wiley Publishing 2004
Taxonomy	Taxonomies are defined simply as the structures used to organize information. When people think of	Extracted from the Mission Statement of the Taxonomies and

	taxonomies they typically understand hierarchical structures like those created in the biological sciences. From an information science perspective, though, taxonomies may take on one or a combination of several types of structures – they may be simple flat structures, hierarchies, network/plex structures or faceted taxonomies. Each of these kinds of structures serves a different kind of information management and access purpose. All are critical for supporting today's complex information solutions and are integral components of today's complex information systems.	Semantics Special Interest Group <http://km.gov/>
Taxonomy Structure	Taxonomy structure represents the underlying hierarchical structure of the concepts within a defined scope and context, similar to the Library of Congress classification system. It is used by content managers to categorize information within the content management workflow process.	"Taxonomy Analytical Brief", Department of State, IRM Business Center, May 27, 2003
Taxonomy View	Taxonomy view is the visual view of taxonomy structure presented to the end users. It could be the same as taxonomy structure or it could be completely different. Taxonomy view organizes Web content into logical groupings, similar to Yahoo's hierarchical directory listing. Sample deliverables of taxonomy view include conceptual navigation model, information access points, broad information categories and associated standards and guidelines.	"Taxonomy Analytical Brief", Department of State, IRM Business Center, May 27, 2003
Thesaurus	A set of related terms describing a set of documents. This is not hierarchical: it describes the standard terms for concepts in a *controlled vocabulary*. Thesauri include synonyms and more complex relationships, such as broader or narrower terms, related terms and other forms of words.	"Taxonomy Analytical Brief", Department of State, IRM Business Center, May 27, 2003
Topic Maps	This is an ISO standard for the representation and interchange of knowledge, with an emphasis on the findability of information. The standard is formally known as ISO/IEC 13250:2003. A topic map can represent information using topics (representing any concept, from people, countries, and organizations to software modules, individual files, and events), associations (which represent the relationships between them), and occurrences (which represent relationships between topics and information resources relevant to them). Topics, associations, and occurrences can be typed, but the types must be defined by the creator of the topic maps, and is known as the ontology of the topic map. There are also additional features, such as merging and scope. The concept of merging and identity allows automated integration	Wikipedia <http://www.wikipedia.com/>

	of topic maps from diverse sources into a coherent new topic map. Topic maps have a standard XML-based interchange syntax, as well as a de facto standard API, and query and schema languages are being developed within ISO.	
Weblog	See *Blog*.	
Web Service	A web service is a collection of protocols and standards used for exchanging data between applications. Software applications written in various programming languages and running on various platforms can use web services to exchange data over computer networks like the Internet. This interoperability is due to the use of open standards. OASIS and the W3C are the steering committees responsible for the architecture and standardization of web services. To improve interoperability between web service implementations, the WS-I organization has been developing a series of profiles to further define the standards involved.	Wikipedia <http://www.wikipedia.com/>

Appendix C: Types of Semantic Conflicts

CONFLICT	DESCRIPTION	FOR EXAMPLE
Data Type	Different primitive or abstract types for same information	SSN as a VARCHAR vs. a NUM
Labeling	Synonyms/antonyms have different text labels	When ORG_NAME and COMPNY tables have data that mean the same thing
Aggregation	Different conceptions about the relationships among concepts in similar data sets. Collections or constraints have been modeled differently for same information	Does a "motorcycle" have 1, 2, 3, 4 or more wheels, how are the constraints modeled in your schema?
Generalization	Different abstractions are used to model same domain	Are "cars" and "trucks" kinds of "vehicles" or are they top-level classes themselves?
Value Representation	Different choices are made about what concepts are made explict	"StartTime" plus "Duration" equals "endTime" …or does "endTime" minus "startTime" equal "duration?" – how is it modeled?
Impedance Mismatch	Fundamentally different data representations are used	Relational to Object mappings (key migrations, multiplicity, etc)
Naming	Synonyms/antonyms exist in same/similar concept instance values	"Company" table has many entries: "DaimlerBenz," "Mercedes," "Chrysler," etc. but they are refer to the same thing
Scaling and Unit	Different units of measures with incompatible scales	4 point grade scale vs. a 5 point grade scale
Confounding	Similar concepts with different definitions	"EarningsPerShare" object for a NASD application vs. a NYSE system
Domain	Fundamental incompatibilities in underlying domains	"MainAssembly" object in a Ford product system vs. a brake supplier system
Integrity	Disparity among the integrity constraints	Does an airline ticket have a primary key that uniquely IDs a passenger? (most don't)

Provided by Jeff Pollock, December 2004.
<http://jtpollock.us/semanticconflicts/chart_semantic_conflicts.pdf>