



An Introduction to the Semantic Web

Concepts, Platforms and Tools



Dennis Pierson
Senior Architect Financial Services

Contents

| | |
|--|-----------|
| Introduction | 5 |
| Purpose | 6 |
| Background | 7 |
| Foundational Definitions and Standards | 8 |
| Semantics | 8 |
| Metadata | 9 |
| Ontology | 9 |
| Ontology Models | 11 |
| Reasoning | 11 |
| The Semantic Web | 13 |
| RDF, the Resource Description Framework | 13 |
| RDFa | 14 |
| RDF Schema | 14 |
| SKOS | 14 |
| Linked Data | 14 |
| Turtle. | 14 |
| OWL, the Web Ontology Language | 14 |
| Quads, n-ary Relationships, and Reification | 15 |
| Query | 15 |
| SPARQL | 15 |
| Joseki. | 16 |
| Rules | 16 |
| RuleML | 16 |
| SWRL | 17 |
| RIF | 17 |
| SVBR | 17 |
| Open Data | 17 |
| OpenCyc | 17 |
| UMBEL | 17 |
| GoodRelations | 17 |
| YAGO | 18 |

| | |
|--|-----------|
| Open-Calais | 18 |
| The EDM Council | 18 |
| Fadyart. | 18 |
| Ontology + Linked Data = the Semantic Web | 20 |
| Platforms and Tools | 20 |
| Ontology Editors | 21 |
| SWOOP | 21 |
| Protégé 4 | 21 |
| NeOn toolkit | 21 |
| TopBraid. | 21 |
| Semantic Middleware | 21 |
| Jena | 21 |
| Sesame | 21 |
| Virtuoso | 22 |
| Storage | 24 |
| SDB | 24 |
| TDB | 24 |
| Oracle | 24 |
| OWLIM | 24 |
| Search Engines | 24 |
| Solr | 25 |
| SIREn | 25 |
| Nutch | 25 |
| Data Rationalization | 25 |
| Text Mining | 25 |
| Matching | 25 |
| S-Match | 25 |
| Mashups | 26 |
| Semantics in the Enterprise | 27 |
| In Conclusion | 29 |
| Links | 30 |



Introduction

Computers don't 'know' what data 'means'. They just follow orders. The step up to knowledge processing depends on a common framework for people and systems to agree on the meaning of the terms and properties that describe the entities and relationships that populate our business and personal domains.

A semantic layer where terms and relationships are formalized as metadata entities in semantic models provides a reference model for **agreement** so that people and automated processes can collect, join, and interpret data aligned to these models in the semantically same way. This concept is now enabling automated knowledge production, collection, and processing in business and across the internet.

The Semantic Web, linked data, ontologies, metadata models, taxonomies, vocabularies, semantic reasoners, dynamic and boundary-less integration, text mining, and smart search are some of the building blocks of a new generation of automated systems. They are showing up everywhere information is modeled: for data management and integration, service interfaces, web presentations, matching algorithms, business intelligence, research, litigation, publishing and advertising, social networks, news aggregation, and anywhere people want to know more about their businesses and their lives. In an accelerating and exponentially expanding universe of information, no competitive business can afford to ignore the advantages of knowledge-based, linked-data systems.

Purpose

This paper is intended for both business and technical people who would like an introduction to the universe of applied semantics, and for those who may have encountered some of these applications, but aren't familiar with the underlying technologies.

A semantic transition is upon us, and proceeding rapidly. The relatively conservative world of IT and data centers may not yet have taken notice. A close look, however, will reveal solutions to some of the most intransigent problems that plague the evolving enterprise: yesterday's solutions are today's headaches. Semantics technology inhabits the intersection of business and technical architectures, bridging the two, and offering an adaptable mechanism for establishing and maintaining alignment between them.

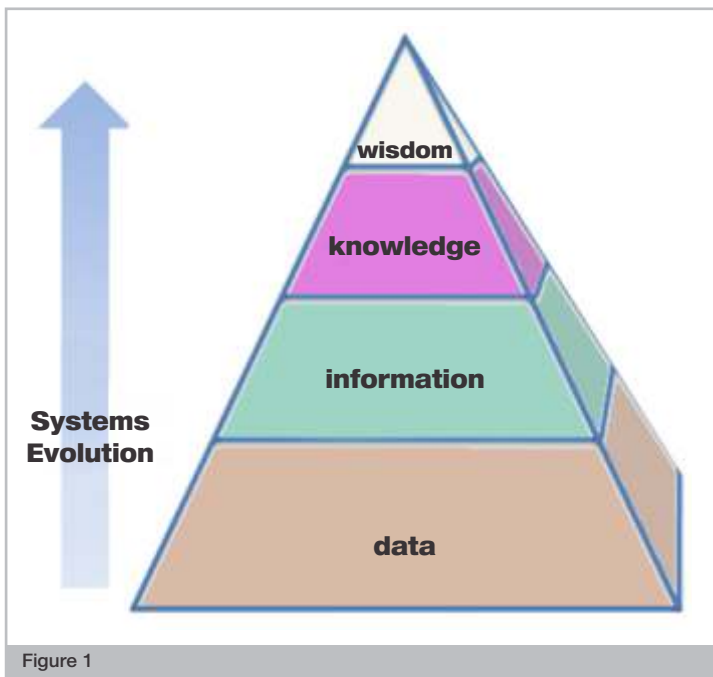


Figure 1

Semantic Web technologies emerged from the field of knowledge engineering. Before there was a world-wide-web, most knowledge research was conducted in universities, some very large companies, and specialty firms, such as defense intelligence. There has been a lot more activity in applied semantics since the introduction of linked-data standards. Commercial adoption of this technology is burgeoning, much of it based-on, contributing to, and integrated with open-source software. The sophistication and power of the tools that are available are impressive.

Our purpose here is to provide a snapshot survey and a general explanation of the concepts, methods, and standards underlying the current state of Semantic Web-based knowledge engineering tools, with a presentation of some of them and of the platforms and open resources that are available now for creating semantic applications. We also hope to introduce these concepts and technologies to a wider community of interest to engage a continuing discussion about their power and potential.

Subsequent papers will explore business applications.

Background

Information is processed data. We acquire, collect, organize, normalize, secure, optimize, and store data in various formats on a variety of storage media. But data is just the raw material of IT. We require application software to cleanse it, maintain its storage and access infrastructure, and to compose it into the next level in the knowledge pyramid: information. Information Processing has always been the purpose of our IT enterprise, even when we called it Data Processing.

Software presents our data to us in clear business contexts, organized and correlated as information in screens, reports, tables, graphs, and in ever more sophisticated human interaction technologies. Deriving knowledge from information requires an established base of knowledge as a frame of reference to correlate new information with what we already know, and a knowledge infrastructure that enables navigation and discovery along paths that are meaningful and well defined.

People digest information using human processing to supplement what they already know, adding to their personal and corporate knowledge bases to form decisions, policies, programs, products, and relationships that drive their businesses. Knowledge management is still primarily a human domain.

Much Knowledge is Hidden

There is a universe of knowledge latent in the many repositories and file systems scattered across our IT landscapes that is now accessible only in separate parts, and only through special programs, each designed for certain data types and formats. The same data (or is it the same?) may be presented through different systems in different contexts and formats. Sometimes they don't agree, and often we can only guess at the reasons. This knowledge - the semantics, of the storage, transformations, measures, computations, aggregations, and formats are all buried in storage schemas and application code - inaccessible outside of the applications' functional boundaries.

Relational databases are the standard architectural model for modern data storage technology. Their invention was a major breakthrough in storage and retrieval that formalized our understanding of optimal storage systems, allowing them to be decoupled into a generic layer independent of application models. Relational databases have thrived and endured because they are fast, secure, and they guarantee data integrity at a structural level. Much of this is due to the original relational logic concept that models business records as tables of rows of tightly-related data-element tuples referentially linked by keys allowing data management software to optimize access, concurrency, and security.

The advent of Object-Oriented programming into this relational universe caused some consternation while the industry tried to figure out how best to bridge the semantic gulf between sets of rows of tuples and coherent objects that spanned multiple tables and rows, but were not explicitly bounded in the relational model. Relational technology was flexible enough to accommodate this new paradigm, but even today after all the dust has settled, we tolerate solutions that intrude into our application software, forcing programmers to be aware of underlying storage models.

Object-oriented DBMS arose in response, but have not gained widespread use because they obscure the boundary between storage technology and application logic – much like stored procedures, and they have to rely on relational implementations to get the required scalability and service levels that we are used to getting from relational technology. Although the semantics of the OO model are more highly evolved and explicit, they are not exposed as models outside of their storage and application architectures. SOA takes a step in that direction.

Making the leap to knowledge-based processes requires a new layer of intelligence encoded as semantic business models that are independent of, span over, and are linked by metadata to applications and data storage models. This business-semantic layer acts as a stable, yet agile, system of reference that can link disparate data models, applications, repositories of documents, and streams of information such as emails, news feeds, twitter, and semantics-enabled internet search engines. Business intelligence is learning to integrate these dynamic and unstructured sources of information into a semantically rich environment capable of recognizing patterns of thought and behavior on these systems to process information in complex patterns that challenge human understanding far beyond current capabilities.¹

Events of the last few years in financial markets and litigation liability illustrate the urgency for engineering knowledge management into platforms that are affordable and accessible to well-informed, but not highly technical, business users who must stay abreast of changing conditions in ever more complex regulatory, legal, political, and business environments.

What are the technologies, algorithms, and methods that will enable us to use our complex processing platforms to bring us to a higher level of understanding of our businesses beyond what we are getting from the specialized programs that present us aggregated, transformed, formatted, and visualized data according to specifications that were developed in the context of specific business processes?

Are automated systems capable of achieving Knowledge? Can they acquire Information as Knowledge? Can they learn? Can they reason and draw conclusions? Can they plan? Apply heuristics? The short answer is yes with a lot of qualifications. Automated systems are deterministic – they do what we tell them to do. So how can they do that which we can't do ourselves, or that we can't predict? In simple terms, this is about managing complexity and scale. By applying semantic patterns and reasoning algorithms using a lot of computational horsepower, research at the cutting edge has produced machine understanding and learning that can compete with humans, as we've seen notably in IBM's Deep Blue chess playing and Watson natural-language processing systems.

Our systems will acquire wisdom as our models accumulate knowledge and our metamodels evolve.

Foundational Definitions and Standards

Semantics

Semantics is the study of meaning in language, a very broad and sometimes ambiguous domain, but one that has specific applications in computational knowledge. The notion of context is important: terms may have different, even opposite, meanings in different contexts, and so determining what is relevant to a term or in response to a query depends very much on the domain that is being modeled. Computability and decidability depend on the qualifications and restrictions applied to the terms and relationships in the models, and even though our domain focus will normally be restricted by business requirements, it is important to remember that the notion of context is critical to the semantics of a model at every level of granularity. Is Anne Hathaway related to Berkshire? Some automated trading programs seem to think so.²

A model for the collection, organization, and storage of information that can be processed semantically, as knowledge, requires characterization of the relationships among the elements of the model.

The relational model provides some inherent relationship modeling: the fact that data elements are co-located in a relation means that they are related, and that tables are referentially related demonstrates an integrity relationship. Parent-child relationships can be inferred in most cases, but the relational model does not make this explicit.

Semantic relationships must be modeled in application logic as interpretations of data. In a semantic model, these relationships can be arbitrarily characterized, and attributed logically as transitive, reflexive, same-as, setwise disjoint, closed, and so on, to facilitate reasoning across the entire knowledge base.

Metadata

Metadata is data about data. It's everywhere, but too often it's nowhere because we don't collect it, organize it, or use it effectively.

From the bottom up, we collect data as individual data elements from various streams, classify it, sort it, break it down, aggregate it, correlate it and store it. We use information about the data to accomplish these things. This information is metadata, implicit and explicit.

The table and column names in databases are data about data. The provenance of the data is metadata. Its intended use is metadata. Context is metadata. What is the purpose of the repository? What database stores the data? Which applications use it? What is its lifecycle? Dependencies? All of this is metadata. Most of it is implicit – in the storage structures, built into the processes that collect, transform and present it, and in the professional expertise of the users.

Metadata has metadata. Taxonomies are metadata structures. Ontologies are metadata structures enriched with relationships and classes of metadata.

To process data as knowledge, we have to make our metadata explicit: enrich it with relationships wherever these exist and are meaningful in the contexts that concern us. So we formalize metadata as classes, relationships, and attributes in our semantic models.

Explicit vs. Implicit Semantics

Our businesses are modeled conceptually in the implicit metadata of the data that we store as discrete elements in our databases, at minimal granularity - facts; we rely on application software to compose these elements into meaningful information for us.

A holistic, top-down analysis of the objects that compose our business models, down to that finest granularity of individual data elements, will reveal our semantic object-hierarchies. Often modeled explicitly in object-oriented programming models, these are typically instantiated by software, bottom-up, with data combined and composed into meaningful business-granularity objects.

Parent-child, class-hierarchy, semantic “is-a” relationships are built into semantic modeling, i.e. ontology, languages. A subclass entity is a member of a superclass, so it inherits the relationships and attributes of its ancestors. It is logically a member of the set of objects that have those attributes and relationships. Expressed as subsumption, class inheritance potentiates reasoning algorithms to draw inferences from data to discover new facts.

Business intelligence tools give us some ability to configure levels of compositional granularity so that we can visualize quantifiable dimensions of business-granularity elements. To automate processes that can derive knowledge beyond tables and graphs of numbers along any dimension that can be described semantically, that can infer facts, and that can span models and domains, requires models based on business-semantics class and relationship structures. This is what ontology languages and linked-data structures are designed to do.

Ontology

The semantic heart of the Semantic Web is a modeling concept called ontology. There are several perspectives on this concept, and different definitions from philosophy, computer science, information science, knowledge management, and library science. Here are a few that trace its usage from philosophy into the Semantic Web:

- The term ontology has long been used in the study of philosophy. From the Wikipedia: “Ontology is the philosophical study of the nature of being, existence, or reality as such, as well as the basic categories of being and their relations. Traditionally listed as a part of the major branch of philosophy known as metaphysics, ontology deals with questions concerning whether entities exist or can be said to exist, and how such entities can be grouped, related within a hierarchy, and subdivided according to similarities and differences.”³ This sounds a lot like very smart search, with some segmentation thrown in!
- In Information Science, the term is narrowed to “... an ontology is a formal representation of knowledge as a set of concepts within a domain, and the relationships between those concepts. It is used to reason about the entities within that domain, and may be used to describe the domain.”⁴
- From W3C's OWL page: “Ontologies are formalized vocabularies of terms, often covering a specific domain and shared by a community of users. They specify the definitions of terms by describing their relationships with other terms in the ontology.”⁵ W3C is the World Wide Web Consortium, a standards body.
- And a Semantic Web-specific concept from the TechWiki at openstructs.org, “Ontologies supply the structure for relating information to other information in the Semantic Web or the linked data realm. Because of this structural role, ontologies are pivotal to the coherence and interoperability of interconnected data.”⁶

In more detail - a superset of elements modeled in ontologies, also from the Wikipedia:⁷”

- **Individuals:** instances or objects (the basic or "ground level" objects)
- **Classes:** sets, collections, concepts, classes in programming, types of objects, or kinds of things.
- **Attributes:** aspects, properties, features, characteristics, or parameters that objects (and classes) can have
- **Relations:** ways in which classes and individuals can be related to one another
- **Function terms:** complex structures formed from certain relations that can be used in place of an individual term in a statement
- **Restrictions:** formally stated descriptions of what must be true in order for some assertion to be accepted as input
- **Rules:** statements in the form of an if-then (antecedent-consequent) sentence that describe the logical inferences that can be drawn from an assertion in a particular form
- **Axioms:** assertions (including rules) in a logical form that together comprise the overall theory that the ontology describes in its domain of application. This definition differs from that of "axioms" in generative grammar and formal logic. In those disciplines, axioms include only statements asserted as a priori knowledge. As used here, "axioms" also include the theory derived from axiomatic statements.
- **Events:** the changing of attributes or relations”

Ontology representations vary widely in their expressiveness and their complexity. These two qualities are directly proportional and closely bound.

Ontology Models

The primary structural components of a semantic model are entities and relations. We know about Entity-Relation modeling for relational databases. This is a reduced-expression subset of a general ontology. Semantic Web ontologies are richer, in that relations as well as entities are modeled semantically. They extend the familiar entity-relationship model to an entity-relationship-entity model. This allows the introduction of specific kinds of relationships between entities such as friend-of, or belongs-to.

In some modeling languages relationships are called properties, or roles: B is a friend-of A, A is the value of the property friend-of to B, or B plays the role friend-of to A. As in object-oriented languages, the entities are classes, and may be super-classes or sub-classes of other classes, inheriting attributes and properties (relationships), from their ancestry. Relationships also have class status, with inheritance. Ontology languages also allow for multiple-inheritance, that is, inheritance from more than one parent.

Ontology modeling provides generally for logical constraints on relationships such as cardinality, transitivity and reflexivity, and on sets of entity classes to determine closure in reasoning operations.

Entity-Relation-Entity tuples, called triples, are commonly represented in the W3C standard, the Resource Description Framework, or **RDF**, the lingua franca of the Semantic Web. The original intention was to provide a metadata annotation framework of content-based links embedded in web documents to facilitate content-based search, thus, the “Semantic Web”.

Reasoning

Reasoning was mentioned earlier in relation to ontology language expressivity and complexity - there is a tradeoff between the expressiveness of a representation language and the difficulty of reasoning over the representations built using that language.

Ontologies model facts and relationships in such a way that inferences can be drawn – facts that are not explicitly declared can be deduced with some certainty. The automation of reasoning with a theoretical-logical, and mathematical, foundation was pioneered in the development of logical and expert systems at the MIT AI Lab⁸. Ontology traversal along a class-structure path top-down, from general to specific, to collect instantiations of classes is known as backward-chaining. Bottom-up traversal is called forward-chaining. These are the core logical methods of automated reasoning that were the basis of expert systems.

From Wikipedia: “A semantic reasoner, reasoning engine, rules engine, or simply a reasoner, is a piece of software able to infer logical consequences from a set of asserted facts or axioms. ... The inference rules are commonly specified by means of an ontology language, and often a description language. Many reasoners use first-order predicate logic to perform reasoning; inference commonly proceeds by forward chaining and backward chaining.”⁹

As a very simple example of inference, consider a family-structure ontology. If Mark is parent of Amy and Amy and Bob are siblings, the reasoner can conclude that Mark is a parent of Bob. Reasoners also check ontologies for logical consistency - “women are smarter than men”, and “men are smarter than dogs” , and “dogs are smarter than people” may be asserted separately in different parts of an ontology.

Here’s a slightly richer and much tastier example from the Protégé tutorial available from the University of Manchester¹⁰:

This tutorial walks through the steps of creating an ontology of pizzas based on toppings. Pizza is a class related to the Topping class by the relation (property) HasTopping – this semantic triple, (subject,property,object), is (Pizza,HasTopping,Topping). Toppings are classes - for example, CheeseTopping has subclasses MozzarellaTopping and ParmesanTopping. After creating a Pizza subclass called NamedPizzas and several instances of that such as the AmericanHotPizza and the SohoPizza, that have (HasTopping) some (at least one) Topping that is a subtype of CheeseTopping:

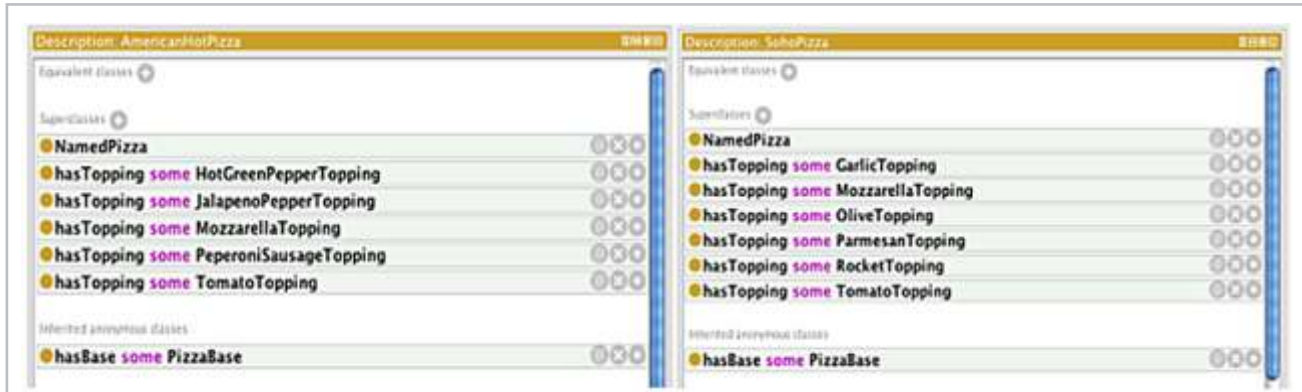


Figure 2

Then a subclass of Pizzas called Cheesy Pizza is defined as a restriction on all Pizzas that have cheese toppings. The reasoner will find (infer) all pizza classes that HasTopping CheeseTopping and infer that they are Cheesy Pizzas, and assign them to the Cheesy Pizza class as inferred subclasses:

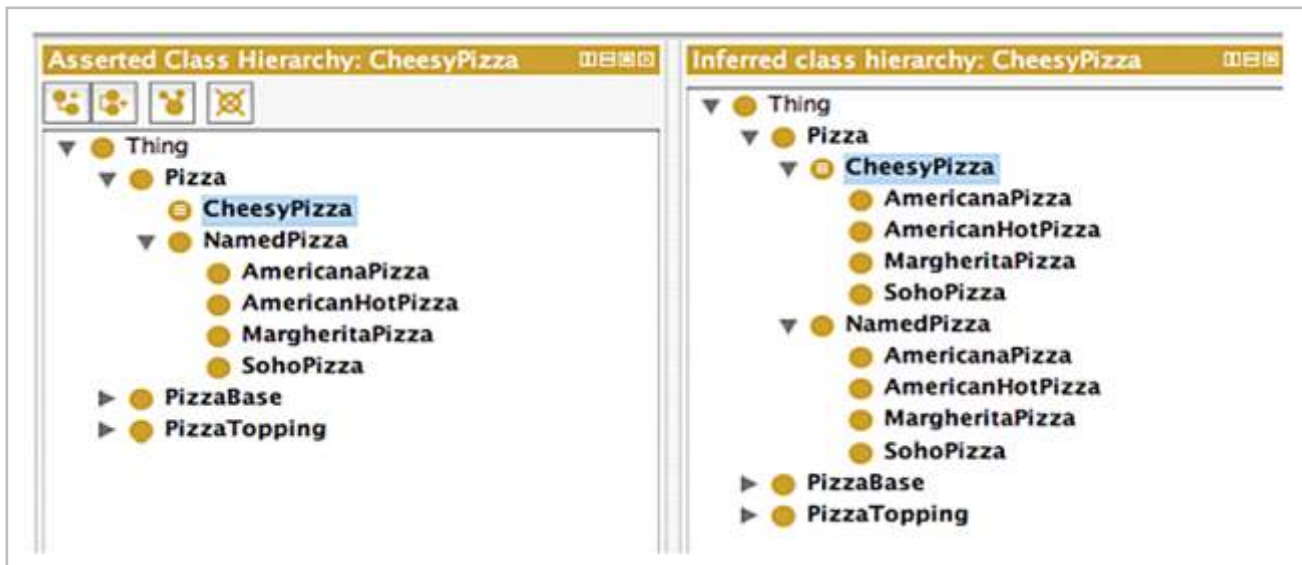


Figure 3

From A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools Edition 1.2¹¹

A practical working ontology will have hundreds, possibly thousands of classes, many attributes, and complex combinations of logical properties and constraints. A reasoner will ensure consistency and infer many relationships that are not explicitly declared. These can then be returned by searching.

The Semantic Web

Based on the standards and technologies described here, the Semantic Web is emerging as the standard infrastructure of information on the web, a skein of concepts embedded in web pages, databases, and documents across the world. Although a common term, it is without a formal definition.¹²

The Semantic Web is being created ad-hoc everywhere as content publishers tag their content with metadata as RDFa, which is linkable to any other data tagged using common URI vocabularies, in web pages, RSS feeds, and documents, social media, and yes, relational databases. This vast interlinked web of data is revolutionizing the planetary infosphere, enabling new applications that pull together and assemble information as it develops, and opening the gate for enterprises to publish in formats that can be automatically discovered and understood.¹³

As described at SemanticWeb.org: “The Semantic Web is the extension of the World Wide Web that enables people to share content beyond the boundaries of applications and websites. It has been described in rather different ways: as a utopic vision, as a web of data, or merely as a natural paradigm shift in our daily use of the Web. Most of all, the Semantic Web has inspired and engaged many people to create innovative semantic technologies and applications.”¹⁴

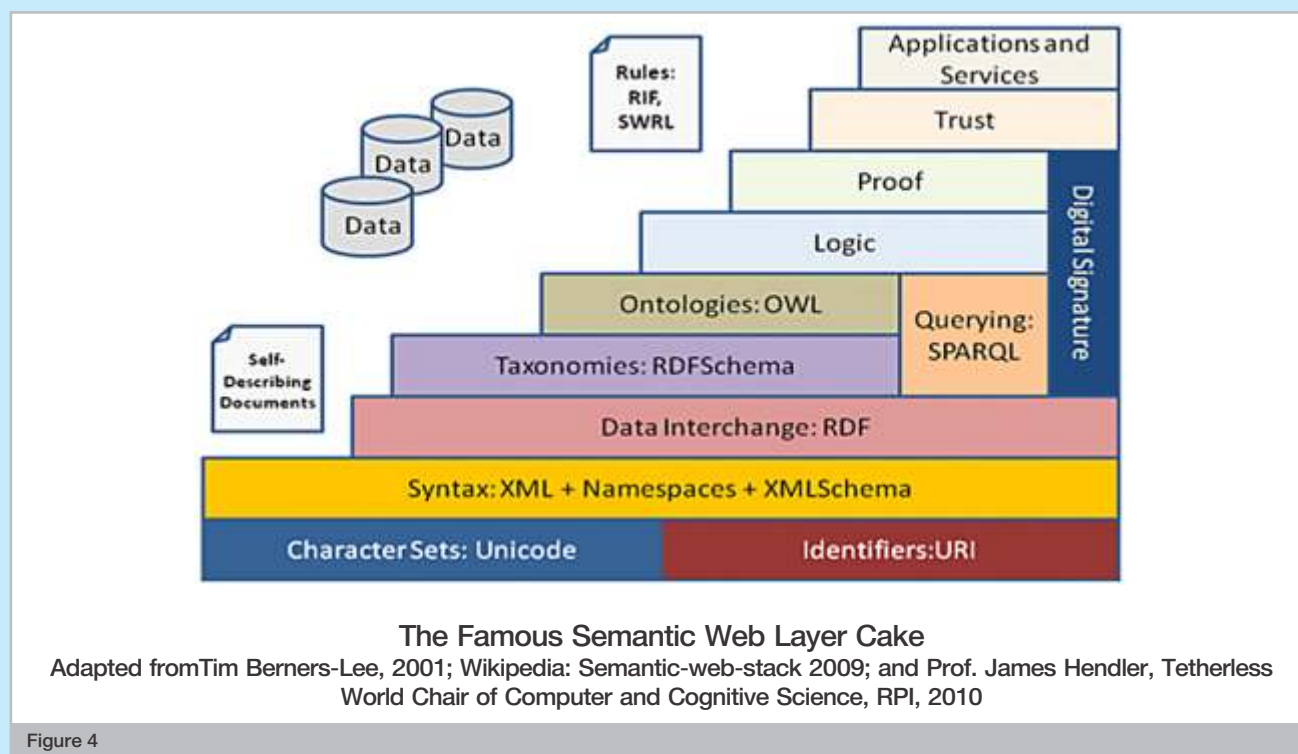


Figure 4

RDF, the Resource Description Framework

From the Wikipedia: “RDF is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modeling of information that is implemented in web resources, using a variety of syntax formats.”¹⁵

RDF is expressed in Subject, Predicate, Object. “The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the

object.”¹⁶ Resources - subject, object, and predicate are expressed as URIs; an object may also be a Unicode string literal. Subjects and objects may be empty. Predicates are also expressed as URIs.^{17,18}

This use of URIs is a foundational mechanism that asserts agreement on terms, a kind of semantic handshake that guarantees that if we use the same URI, we agree on the term of the reference. Agreement on meaning-in-context requires deeper semantic modeling. That’s as much as we can claim as a standard for the use of URIs – they are not required to be reference-able and when they are, there’s no standard yet for the meaning of the result; agreements on URI use are developing. The result could be an ontology reference.¹⁹

The W3C RDF model is an abstraction that is expressed in XML²⁰ modeled graphically as labeled, directed multi-graphs,²¹ and query-able using the W3C standard query language SPARQL. We’ll return to this a little later.

RDF is the foundational language of the Semantic Web, the framework for expressing RDFa, RDF Schema, SKOS, Linked Data, Turtle, and OWL, the Web Ontology Language, which is elaborated later in the ontology languages section.²²

- RDFa is an extension of RDF “that adds a set of attribute level extensions to XHTML for embedding rich metadata within Web documents.”²³ This provides explicit linking of machine-readable metadata to web pages that can be captured as RDF, and linked into semantic structures. RDFa is emerging as the fabric of the Semantic Web. “A little semantics and a lot of data” is putting cross-platform integration, across the web, into the hands of anyone who is willing to use it.²⁴
- RDF Schema is a modeling language extension of RDF that introduces class types, value constraints, and properties as predicate classes, with domain and range. It’s a core subset of OWL.²⁵
- SKOS is used primarily in the construction of vocabularies and thesauri, which are complementary to ontology models.²⁶
- Linked Data – From Tim-Berners-Lee, universally credited with inventing the World Wide Web: “The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.”²⁷
- Turtle - Terse RDF Triple Language is a textual syntax for RDF for composing graphs in a compact and natural text form with abbreviations for common usage patterns and datatypes. Turtle is popular among Semantic Web developers as a human-friendly alternative to RDF/XML.²⁸

OWL, the Web Ontology Language

Ontologies on the web, and as computational objects in general, have been the subject of a great deal of research. Several language models have been developed with varying degrees of expressive power. RDF and RDFS were mentioned earlier. OWL is generally recognized as the standard ontology language for the Semantic Web.

OWL is “a family of knowledge representation languages for authoring ontologies, characterized by formal semantics and RDF/XML-based serialization for the Semantic Web”²⁹.

OWL has been released in two major versions, OWL (2004) and OWL2 (2009), the latter superseding the former and thus the focus here. OWL2 is a refinement and extension of, and is completely backwards-compatible with, OWL, i.e. all valid OWL ontologies are valid OWL2 ontologies. OWL 2 is defined to use datatypes defined in the XML Schema Definition Language (XSD). RDF/XML must be supported by

OWL2 tools to provide a common underlying graph-structured representation for compatibility across tools, APIs and repositories.

The OWL family includes three sub-languages, in increasing expressivity: OWL Lite, OWL DL, and OWL Full. Greater expressivity requires more complex and difficult reasoning algorithms: the general graph search problem is NP-Complete³⁰. Expressivity and decidability must be balanced for applications of this technology to be practical. A summary of OWL sub-language properties and capabilities can be found at the species link on the Wikipedia OWL page³¹.

OWL 2 ontologies can be linked to, and will provide semantic depth to, information coded in RDF. OWL 2 ontologies themselves are usually exchanged as RDF documents³². The implications of this are very broad, and include the entire domain of published information in almost any format that is in any way semantically discoverable. Using graph manipulation algorithms, distributed information repositories tagged using RDF can be joined graphically and bound as instances to ontologies, enabling reasoning and semantic search across all of them.

Quads, n-ary Relationships, and Reification

As we stated above, RDF is typically, and by design, persisted as triples of (entity,relationship,entity). Sometimes, and increasingly, we need to associate information such as source, trust, or date with an RDF triple - capturing and attaching provenance to data is becoming very important. This kind of annotation, also called reification, is modeled as a quad (entity,relationship,entity,provenance). The fourth member could be anything, however the semantic intent is an annotation of the triple. Reification in OWL entails another order of complexity involving the creation of a new class to capture the complexity of n-ary (greater than three) relationships that are not amenable to reasoning algorithms.³³

Query

Ontologies are normally persisted as RDF triples or quads. These are typically modeled in a directed, labeled graph data format. Searches over semantic repositories, ontologies and the instance data they describe are implemented as graph searches.

SPARQL, the SPARQL Protocol and RDF Query Language ('sparkle'), is a W3C semantic query language specification. It uses a SQL-like syntax, but allows the use of variables as objects of the search.³⁴

For example:

```
SELECT ?title
WHERE
{
  <http://example.org/book/book1> <http://purl.org/dc/elements/1.1/title> ?title
}
```

SPARQL was designed to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. It is capable of querying graph patterns along with their conjunctions and disjunctions. SPARQL also supports extensible value testing and constraining queries by source RDF graphs. The results of SPARQL queries can be results sets or RDF graphs.

From an interview with Jeen Broekstra, one of the lead designers of the Sesame semantic framework: "The importance of SPARQL in general cannot be overestimated ...: SPARQL defines a standard way in which to communicate with RDF-based services on the Web. ... Having SPARQL can make sure that a Sesame-based application can freely communicate with, say, a Jena-based application, over the Web ..."³⁵ Sesame and Jena are presented below in the tools section.

Most forms of SPARQL query contain a set of triple patterns called a basic graph pattern. Triple patterns are like RDF triples except that each of the subject, predicate and object may be a variable. A basic graph pattern matches a subgraph of the RDF data when RDF terms³⁶ from that subgraph may be substituted for the variables and the result is an RDF graph equivalent to the subgraph.³⁷

SPARQL query processors are implemented in semantic middleware tools, such as Jena and Sesame, that manage RDF graph structures. Oracle 11g offers a semantic triple-store repository with a SPARQL-like query language³⁸. Oracle also offers Sesame and Jena plugins. There is more discussion on these topics in the Tools section.

SPARQL 1.1 is a W3C working draft of SPARQL that adds subqueries and updates and includes federation extensions. “SPARQL can be used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware”³⁹

There are some good SPARQL examples in the W3C SPARQL 1.1 working draft.⁴⁰

Joseki⁴¹, a Jena subproject, is a RESTful RDF server that sits over an RDF repository or semantic middleware. It processes SPARQL queries delivered using the SPARQL protocol⁴² returning the results as RDF graphs. SPARQL 1.1 update via a REST POST operation adds a whole named graph to the repository; via a REST UPDATE, update adds a subgraph to a named graph.⁴³

Rules

Rules are expressed in different ways at different levels of the semantic stack: query and search, ontology-embedded pattern-based rules, inference, and policy. Rules are used elsewhere to drive workflow, interoperability and integration, determine authorization and entitlement, service policies, evaluate risk and compliance, and to make decisions involving complex conditions among thousands of variables. All of these interact somewhere in the universe of IT. Approached holistically they can be modeled semantically, managed centrally, and harnessed to govern processes at enterprise level.

There are emerging standards for rules on the Semantic Web. Rules-ML has been around a while. SWRL has evolved from RuleML. Commercial rules engines are taking advantage of this. Semantic Web applications are looking at rules that govern web services such as policies, access, trust management, privacy, pre and post condition evaluations and much more. Contracts are, in-effect, rulesets. RIF, the Rules Interchange Format has recently been released by W3C. It formalizes the patterns and formats required to communicate rules among different systems.

Ontologies are rulesets. They make statements of facts: they assert relationships as true with certain constraints. Patterns of behavior can be asserted as sets of facts and modeled as ontology. A very interesting example of this can be found in the paper “Ontological Constructs to Create Money Laundering Schemes” authored by Murad Mehmet and Duminda Wijesekera⁴⁴.

There are a number of ways that rules implicit within an ontology and rules expressed externally can work together. As functionality spreads to intranets, internets, and clouds, the reuse and interoperability of rules become critical; these are primary goals of Semantic Web rules initiatives.

Some tools:

- RuleML, the Rulemarkup language, is a family of XML-serialized rules languages including derivation rules, transformation rules, and reaction rules. It focuses on rules interoperation, and can be used for “queries and inferences in Web ontologies, mappings between Web ontologies, and dynamic Web behaviors of workflows, services, and agents”.⁴⁵

- SWRL (**Semantic Web Rule Language**) is a proposal for a Semantic Web rules-language. It combines the expressiveness of the Horn sublanguages of the OWL Web Ontology Language (OWL DL and Lite) with expressions from RuleML.⁴⁶
- RIF the Rule Interchange Format is a standard for exchanging rules among rule systems, in particular among Web rule engines. RIF focuses on exchange rather than trying to develop a single one-fits-all rule language because single language would not satisfy the needs of many popular paradigms for using rules in knowledge representation and business modeling.⁴⁷
- SVBR- Semantics of Business Vocabulary and Business Rules. In this specification, the OMG – the Object Management group, a standards consortium, defines the vocabulary and rules for documenting the semantics of business vocabularies, business facts, and business rules; as well as an XMI schema for the interchange of business vocabularies and business rules among organizations and between software tools. This specification is interpretable in predicate logic with a small extension in modal logic.⁴⁸ SVBR is actually a well-featured semantic modeling language.

Open Data

Companies, open-data organizations, consortiums, ad-hoc projects, and individuals are collecting and publishing tagged data on the internet. Though there is no regulating authority to standardize or control this, people are using it to collect and integrate information and to build knowledge applications.

The Semantic Web is complemented by several ambitious projects that are building general purpose ontologies, known as upper ontologies. These describe “very general concepts that are the same across all knowledge domains. The most important function of an upper ontology is to support very broad semantic interoperability between a large number of ontologies accessible under this upper ontology.”⁴⁹

Public Ontologies

Some notable ones:

- **OpenCyc** is the open-source version of the Cyc ontology and reasoning engine. “... the world's largest and most complete general knowledge base and commonsense reasoning engine”.⁵⁰ OpenCyc contains the full set of (non-proprietary) Cyc terms as well as millions of assertions about them.⁵¹

Cycorp also makes a commercial version of Cyc, including many more assertions and additional Natural Language capabilities, available at no cost for research purposes. Check out ResearchCyc for licensing and downloading information.⁵²

- The **UMBEL** Vocabulary and Reference Concept Ontology is an “Upper Mapping and Binding Exchange Layer”, designed to help content interoperate on the Web. It provides a “vocabulary for the construction of concept-based domain ontologies, designed to act as references for the linking and mapping of external content, and its own broad, general reference structure of 21,000 concepts, which provides a scaffolding to orient other datasets and domain vocabularies.” Umbel 1.0 was released on Feb 14, 2011.⁵³
- **GoodRelations** is “a language (data dictionary, others prefer schema or ontology) that can be used to describe very precisely what your business is offering. ... you can use GoodRelations (on your web page) to create a small data package that describes your products and their features and prices, your stores and opening hours, payment options and the like.”⁵⁴ Google recommends GoodRelations for their rich-snippets program. Rich-snippets is page metadata that tells the Google search engine how to format and display search result.

- **YAGO** is a huge semantic knowledge base, a free and downloadable ontology from the YAGO-NAGA project at the Max Planck Institute for Informatics in Saarbrücken/Germany. Derived from Wikipedia, WordNet⁵⁵ and GeoNames⁵⁶ it contains more than 10 million entities such as persons, organizations, and cities, and more than 80 million facts about them.⁵⁷
- **Open-Calais** from Thompson-Reuters is a large fact repository and RDF tagging service as a web service that can be integrated into semantic platforms. It offers up to 50,000 free transactions per day. Paid services offer more volume with SLA guarantees. Your metadata, not your content, becomes part of their knowledge base.⁵⁸ The open-source CMS and web platform project Drupal has integrated the Open-Calais service into its platform.⁵⁹
- **DBPedia** “currently describes more than 3.5 million things, out of which 1.67 million are classified in a consistent Ontology, including 364,000 persons, 462,000 places, 99,000 music albums, 54,000 films, 17,000 video games, 148,000 organisations, 169,000 species and 5,200 diseases. The DBpedia data set features labels and abstracts for these 3.5 million things in up to 97 different languages; 1,850,000 links to images and 5,900,000 links to external web pages; 6,500,000 external links into other RDF datasets, 633,000 Wikipedia categories, and 2,900,000 YAGO categories. The DBpedia knowledge base altogether consists of over 672 million pieces of information (RDF triples) out of which 286 million were extracted from the English edition of Wikipedia and 386 million were extracted from other language editions.”⁶⁰
- **Freebase** is an open, Creative Commons licensed repository of structured data of almost 22 million entities. It is a very dynamic open and crowd-sourced knowledge base with a growing community of contributors and a healthy collection of applications. It was developed by Metaweb, a stealth company founded by Danny Hillis of Thinking Machines legend, and acquired by Google in 2010.⁶¹

The Financial Services sector has faced enormous challenges tracking the variety and complexity of investment strategies and instruments, in addition to fast-changing markets and new regulatory environments. Their IT environment is dynamic – much of the competitive pressure in the industry is IT-based. Information models play an important role in their business. Expect to see widespread adoption of semantic technology throughout the industry.

Publicly available financial domain ontologies:

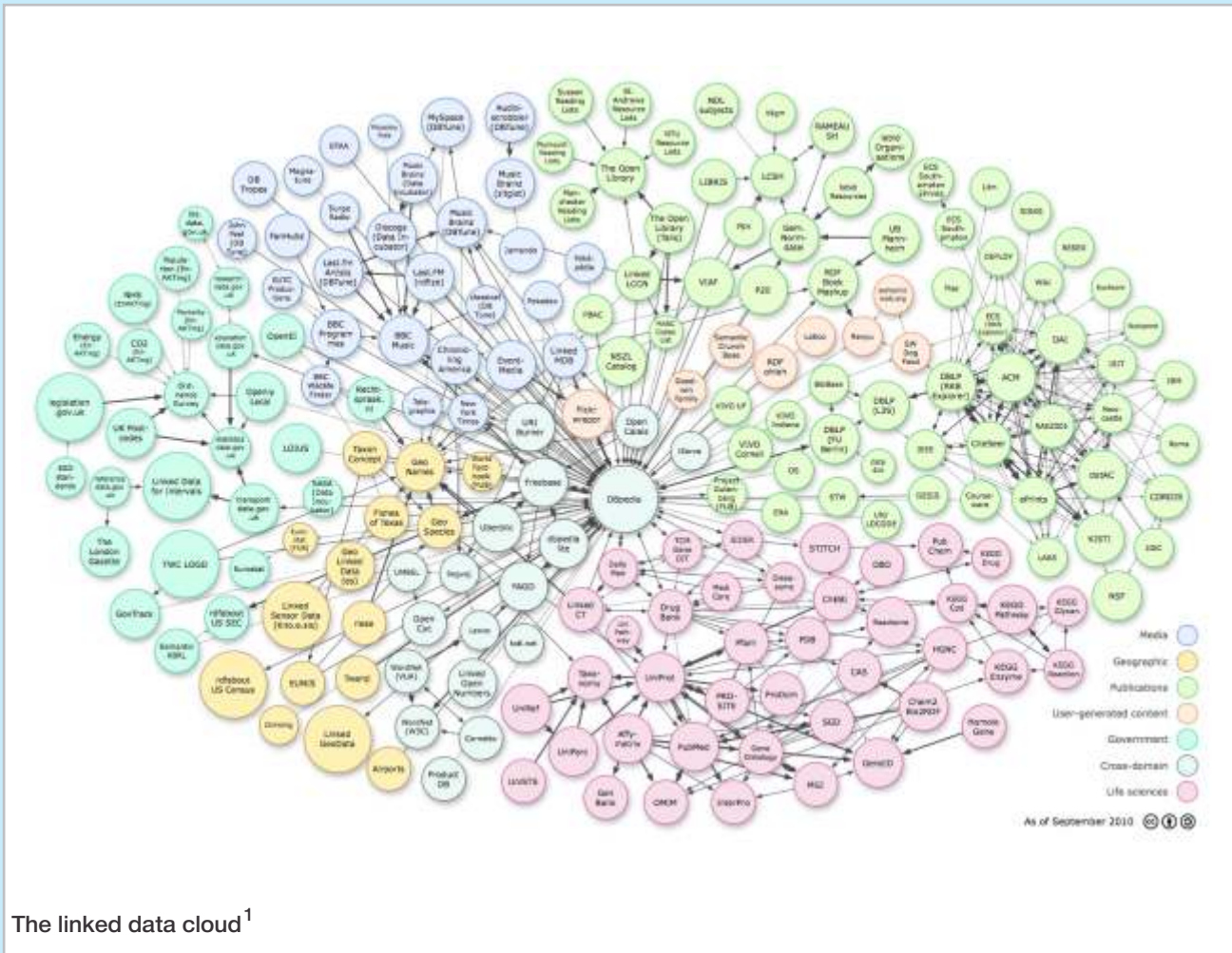
- **The EDM Council**, an industry consortium is developing comprehensive semantic reference models for financial services in a collaborative multi-year effort among industry domain experts and semantic modelers. These models are publicly accessible at their website.⁶² Aligned with financial data reference models and industry vocabularies, they will be a powerful weapon in our semantic arsenal.
- **Fadyart**⁶³ was developed using the TopBraid Composer⁶⁴, and so has some issues loading into Protégé, but is a relatively simple example of what’s required to construct a domain ontology. Their hyperbolic tree visualization is fun⁶⁵.

Some of these ontologies will become standards, and their usefulness will continue to grow into rich resources for the application of natural-language-based reasoning.

Open Data

RDF-tagged and microformatted data is available on the internet from hundreds of sources, public and private. Governments are actively publishing records in these formats. www.data.gov, the US government open data offering, alone includes

- 379,931 raw⁶⁶ and geospatial⁶⁷ datasets
- 934 government apps⁶⁸
- 236 citizen-developed apps⁶⁹
- 172 agencies and subagencies⁷⁰



The linked data cloud¹

Figure 5

These datasets can be pulled, queried, linked into graphs, sifted, sliced, and mapped into mashups and visualizations with little or no manual handling. They can be joined with domain ontologies describing bodies of knowledge and patterns of behavior for reports, research, or any other personal or business purpose.

Regulations governing corporate reporting and behavior could be accessed dynamically this way, joined to a semantic model of company data, and monitored without modification even when regulations change. Vocabularies and Ontologies would have to be maintained as they evolve, but such maintenance would be less frequent, much simpler, and could be centralized at a high level. Standardized industry ontologies such as EDMC's, aligned with regulatory bodies might persist with a minimum of maintenance.

There are too many open data sources to name them here. One primary point of reference is the linked open **data (LOD)** cloud, as depicted above, and documented on the W3C Linking Open Data community task force page⁷¹.

Ontology + Linked Data = the Semantic Web

Web 2.0 brought us interactive web applications, search engines, and social networking with crowd-sourcing. Mixing linked data into this brew gives us the ability to gather related information easily from around the world, join it into structured information maps, extract detailed points-of-view into interactive graphical presentations, and to publish these back to the world, all using open-source tools and free web channels.

But, how do we “know” the data we pull together is relevant? Anyone can tag data. Spammers will play this furiously. How do we sift through the cloud to get just what we want; how do we know that the vocabularies we see out there are in synch with the vocabularies we use? Trusted sources will help, but we still need agreement on terms and rules; we want to ascertain the provenance of our far-flung data.

We need semantic structures to identify relevance and reliability for us, to parse out just the information we want from the vast sea of the internet or the not-quite-so-vast but still quite large stream of data that comes through our firewalls. Yes, we need domain-specific semantic models – ontologies – that give us a framework of contexts and definitions for agreement on terms.⁷² “... ontologies are a technology to make a minimal commitment to agreement while being as clear as possible. ... clarity comes from careful specification, with at least some of the specification document couched in a formal language that forces one to be explicit about assumptions and the meanings of terms... it is a pragmatic and not a theoretical choice to specify a common conceptualization.” – Tom Gruber on the TagCommons.org homepage.⁷³ The page is a little dated, but Tom lives where linked-data meets ontology.

Platforms and Tools

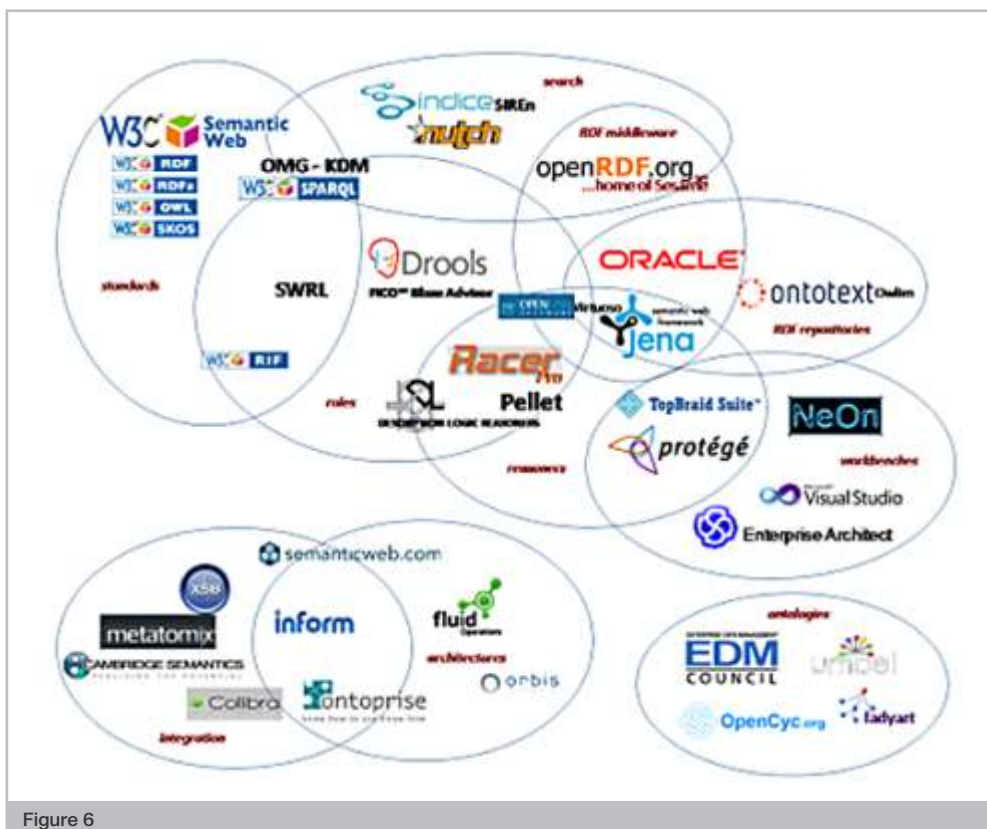


Figure 6

Ontology Editors

Three open-source ontology IDEs are available:

- SWOOP⁷⁴, a lightweight, open-source browser-like tool for creating, editing, and debugging OWL ontologies, originally produced by the MIND lab at University of Maryland, College Park, now hosted by Google.
- Protégé 4⁷⁵, a solid tool from Stanford University, has a broad-based and active user community, although maintenance seems a little spotty at times - it's non-commercial open-source. Protégé is mature, easy to use, and the Manchester Tutorial⁷⁶ is fun and easy to follow. We took an example from a Protégé-OWL tutorial in the Reasoning section. Download Protégé 4 for a good introduction to Protégé, OWL, and reasoners. The Pellet reasoner from Clark-Parsia⁷⁷ is available as an OWL plug-in.
- The NeOn toolkit⁷⁸ is an ontology engineering environment based on Eclipse with an extensive set of plugins. NeOn provides comprehensive support for the ontology engineering life-cycle. And, not open-source, but notable with a free version:

Notable commercial ontology modeling IDEs:

- Knoodl, from Revelytix is a wiki-based collaborative modeling environment "... for creating, managing, and analyzing RDF/OWL descriptions. Its many features support collaboration in all stages of these activities... Knoodl is hosted in the Amazon EC2 cloud and can be used for free. It may also be licensed for private use as MyKnoodl."⁷⁹
- TopBraid Composer from TopQuadrant is "... an enterprise-class modeling environment for developing Semantic Web ontologies and building semantic applications. Fully compliant with W3C standards, Composer offers comprehensive support for developing, managing and testing configurations of knowledge models and their instance knowledge bases. TopBraid Composer is the leading industrial-strength RDF editor and OWL ontology editor, as well as the best SPARQL tool on the market." Their Free Edition is downloadable with wiki support.⁸⁰

Semantic Middleware and Integration Tools

Jena -developed in the HP Labs Semantic Web Research⁸¹ program, Jena is an open-source Java framework for building Semantic Web applications. It provides a programmatic environment for RDF, RDFS, OWL, and SPARQL and includes a rule-based inference engine.⁸²

The Jena Framework includes:

- An RDF API
- Reading and writing RDF in RDF/XML, N3 and N-Triples
- An OWL API
- In-memory and persistent storage
- Scalable native RDF databases
- A SPARQL query engine

The Jena API is built exclusively to handle RDF triples. Implementers are required to construct the graph models and traverse class and property subsumption hierarchies.

Sesame is an open-source framework for storage, inferencing and querying of RDF data. Sesame was designed to be flexible by using a layered API approach for adapting RDF graph processing to different storage models such as relational databases, in-memory, filesystems, and keyword indexes. Sesame provides tools to developers of RDF and RDF Schema applications, including an API supporting both local and remote access, and several query languages.⁸³

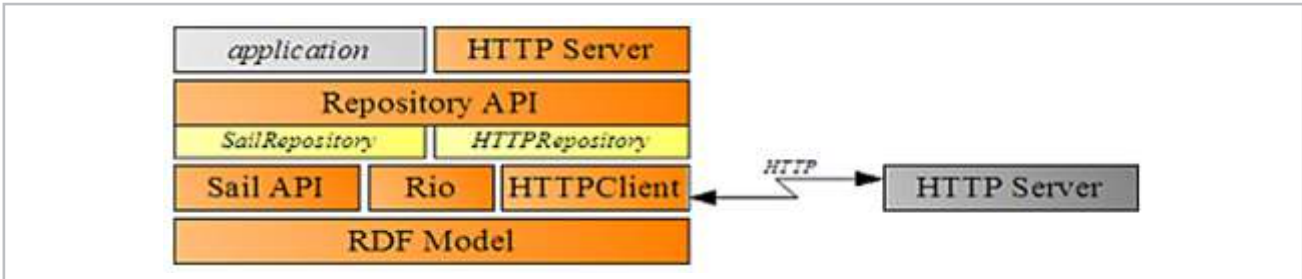


Figure 7

Sesame Architecture⁸⁴

Sesame implements subsumption over property and class hierarchies in the SAIL – Sesame access and inferencing layer API, with domain and range restrictions. SAIL offers efficient inferencing capability because implementations “understand” their storage models, the schema of a relational model, for example, so that semantic querying can be tailored in a SAIL layer to minimize the computational effort.⁸⁵

Virtuoso is a leading-edge commercial semantic platform that integrates backend semantic data stores, SPARQL, internet links, streaming data, free text, and web services via APIs and its Universal Server. This product gives us a clear picture of a successful platform architecture.⁸⁶

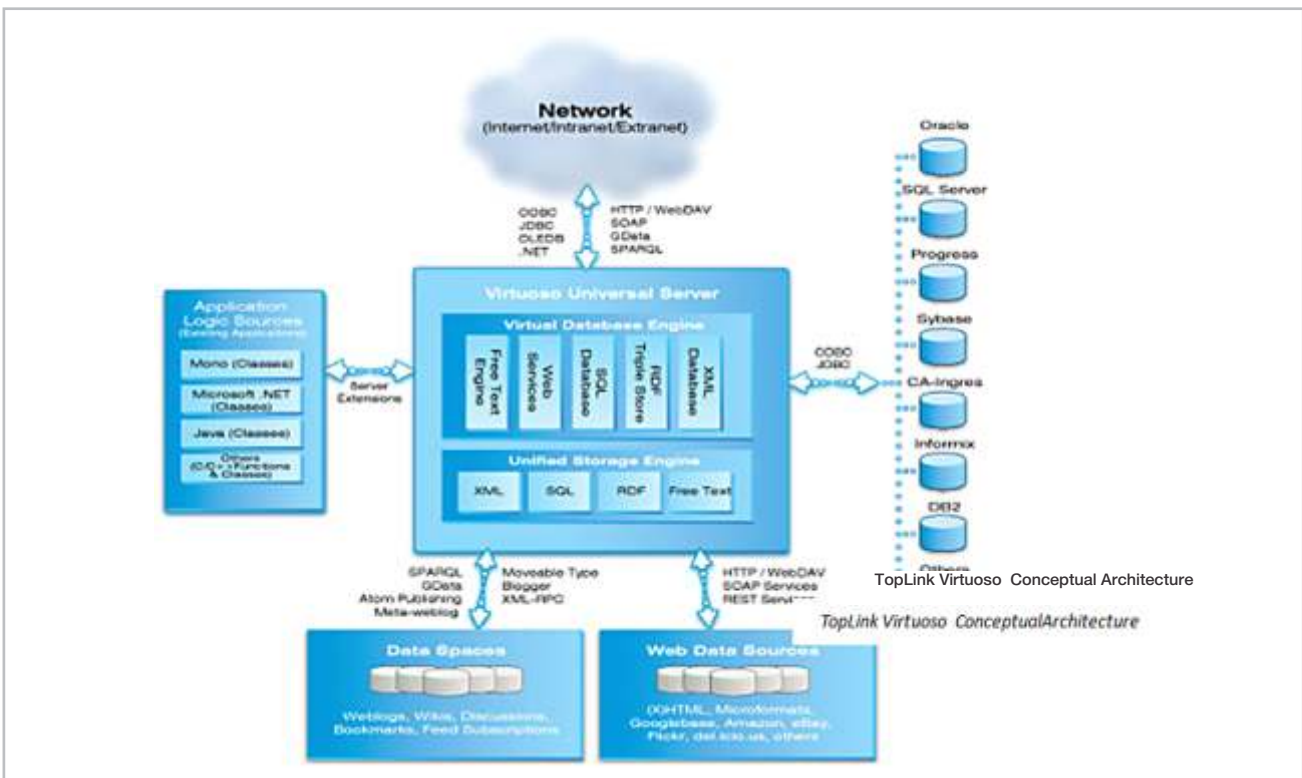


Figure 8

Revelytix has a suite of products for federated semantic information integration across the enterprise, including

Knoodl is a web based ontology modeling and collaboration platform using OntVis, a graphical ontologies visualization tool. OntVis allows business people to see the complexity of ontologies graphically. This kind of feature is essential to demystify the inherently complex nature of ontologies as they grow.

SPYDER is a mapping tool that reverse-engineers database schemas and metadata descriptions to RDF using the new R2RML W3C specification. This facilitates linking of relational metadata to domain ontologies and provides SPARQL querying of relational databases, essential for semantic integration.

SPINNER is a SPARQL query federation and optimizer that makes any number of SPARQL endpoints appear as a single graph.

REX: This is a rules execution engine for RIF, the Rules Interchange Format, that runs rules against ontologies at query-time, enabling run time inferencing and the application of ad hoc rules to models and data.⁸⁷

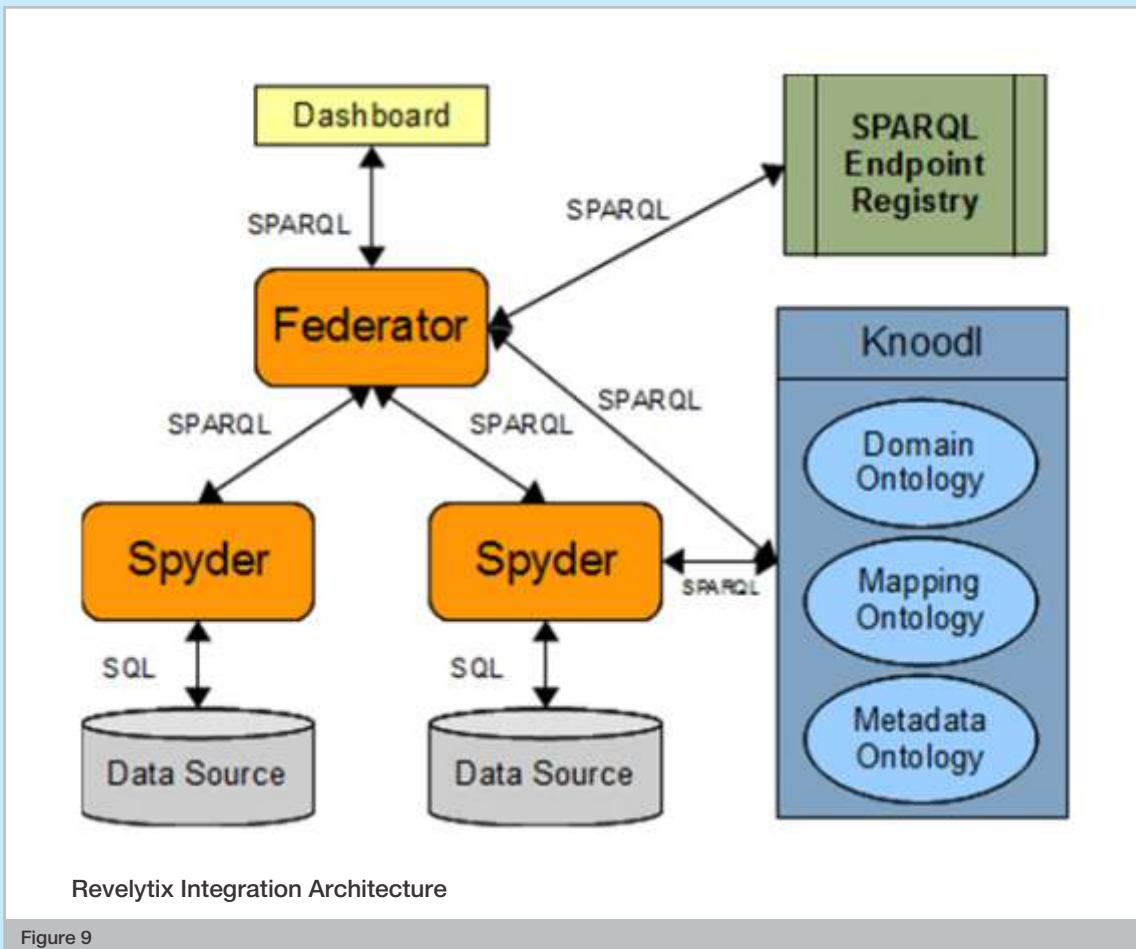


Figure 9

Storage

Jena includes a couple of triple-stores:

- **TDB** provides for large scale storage and query of RDF datasets. TDB supports SPARQL.⁸⁸
- **SDB**, designed specifically to support SPARQL using SQL databases to take advantage of their scalability, load balancing, security, clustering, backup and administration features.⁸⁹

Oracle Semantic Technologies is an open, standards-based, scalable, secure, reliable, and high performing RDF management platform. Based on a graph data model, RDF triples are persisted, indexed, and queried like other object-relational data types. Oracle products, of course, are not open-source and are not free for commercial use. This product does offer plug-ins for both Jena and Sesame.⁹⁰

The **OWLIM** Semantic Repository – from OntoText, is a family of commercially available semantic repositories or RDF database management systems with:

- native RDF engines, implemented in Java and compliant with Sesame
- robust support for the semantics of RDFS, OWL Horst and OWL 2 RL
- scalable loading and query evaluation performance
- SwiftOWLIM is free for any use. They claim it is the “fastest semantic repository in the World: it supports non-trivial inference with tens of millions of statements on contemporary desktop hardware.”⁹¹

Ontotext claims that their commercial product, BigOwlLim is the most scalable semantic repository in the world: “... it can load tens of billions of RDF statements, using non-trivial inference and [it] delivers outstanding multi-user query performance. BigOWLIM is a robust engine packed with advanced features that bring unmatched efficiency to a huge variety of application scenarios:

- cluster configuration that supports load-balancing and automatic fail-over to provide even greater query performance and resilience
- optimized OWL:sameAs handling that delivers dramatic improvements in performance and usability when huge volumes of data from multiple sources are integrated
- hybrid querying capabilities that combine SPARQL with efficient full-text search and ranking of query results”⁹²

Big Data is here - keep an eye on Hadoop⁹³, a file-system-based fully distributed DBMS that uses parallel processing for scalability far beyond the reach of relational DBMS. The Apache incubator project **Heart**⁹⁴ (Highly Extensible & Accumulative RDF Table) will develop a “planet-scale RDF data store and a distributed processing engine based on Hadoop & Hbase⁹⁵”.

Performance and scalability claims are, of course, subject to normal marketplace skepticism and the passing of time.

Search Engines

Lucene search technology has been extended in open-source projects Solr, SIREn, and Nutch, adding capabilities useful for semantic applications.

Solr is An open-source Apache search engine project built on top of Lucene. It is mentioned here as reference for Nutch and SIREn, following.

Solr features “full-text search, hit highlighting, faceted search, dynamic clustering, database integration, and rich document (e.g., Word, PDF) handling. Solr is highly scalable, providing distributed search and index replication.”⁹⁶

SIREn is a commercial Apache-license open-source semantic search engine that extends the open-source search products Lucene and Solr with semi-structured search features, including RDF search, and is much more scalable than conventional triplestore search tools. SIREn uses the Lucene and Solr libraries, so that their rich feature set is available under the covers.

Nutch is a Solr Web Crawler with an ontology (OWL) plugin and RDF parser that parses and indexes HTML, Plain Text , XML, JavaScript for extracting links, OpenOffice ODF, Microsoft Power Point, Microsoft Word,Adobe PDF, RSS, RTF, MP3 song information like title, artist, album, comments, etc., and ZIP files.

Data Rationalization

Data rationalization is a semantic approach to aligning the various layers of data modeling, Conceptual, Logical, and Physical, to a business-centric semantic model – a Managed Metadata Environment. As we touched on earlier, metadata is implicit everywhere in enterprise data and applications.

This is an approach to formalizing business metadata in ontologies to align architectural models semantically with important implications for Master-Data Management.⁹⁷

Text Mining

Text mining uses natural language techniques to classify documents with semantically-defined vocabularies. Primary drivers for the technologies used for text mining are eDiscovery in scientific research – specifically genetics and pharmaceuticals, litigation, and of course, government intelligence where huge volumes of text have to be sifted for criminal and national security reasons. One shocking result of text mining tools can be seen in a recent article in the New York Times: “Armies of Expensive Lawyers, Replaced by Cheaper Software”.⁹⁸

These techniques are for classifying and aligning unstructured information semantically with structured information in specific domains as part of an enterprise information strategy to bring the entire corpus of enterprise information into scope as discoverable knowledge. GATE⁹⁹ is one prominent open-source project dedicated to “text engineering”.

Matching

Matching algorithms with various levels of sophistication have been around for a long time. This is a process that shows up in such disparate business as dating services, clinical trials, and financial trading. A semantic approach using the techniques and tools we describe here can deliver a level of sophistication, over many disparate data sources, which previously had to be encoded in specialized applications using proprietary data assembled in traditional ways.

Semantic matching is still a nascent sector of the semantic web, but it will find a place. Applied to applications like market segmentation and customer targeting, semantic matching could produce

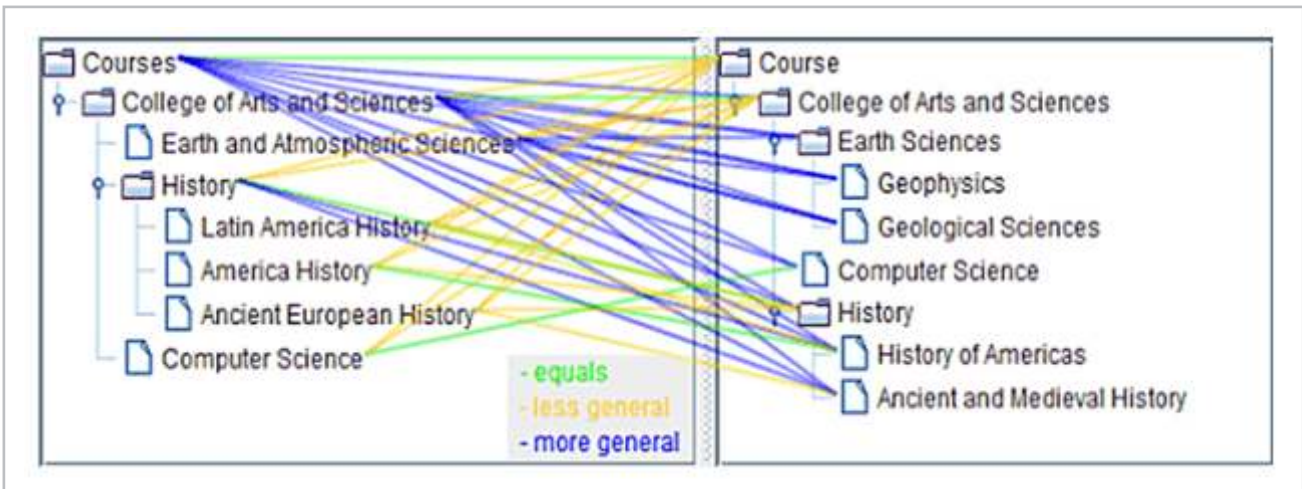


Figure 10

Semantic matching result when comparing two example course catalogs. © 2010 Knowdive @ DISI

results of refinement that surpass all but the best current custom-coded rules-based algorithms, especially, in consumer markets, with information pulled from social-network data. The open-source S-Match project on SourceForge¹⁰⁰ is an intriguing example. It is also hosted at S-Match.org¹⁰¹.

Mashups

Mashups are old news - RSS protocols and formats enable subscriber-clients to pull content from the internet using RSS readers, now built into email clients and browsers. Because it is delivered as XML, content from various sources can be spliced together into novel presentations with some ease.

RDF publishing brings a new level of knowledge utility into this realm. Content exposed as RDF over the internet provides the capability to search across the internet, integrate the results in RDF graphs, align to a domain ontology, query the joined graphs with SPARQL, convert the results into other formats (e.g. JSON, CSV), and then mash them up using open-source tools and free web services such as Yahoo Pipes¹⁰², IBM's many eyes¹⁰³, Microsoft's Web n-gram Service¹⁰⁴, Google Visualization¹⁰⁵, and MIT's exhibit¹⁰⁶ to create interactive graphical visualizations.

The following picture is from a mashup presentation about the Linking Open Government Data project from Rensselaer Polytechnic Institute:¹⁰⁷

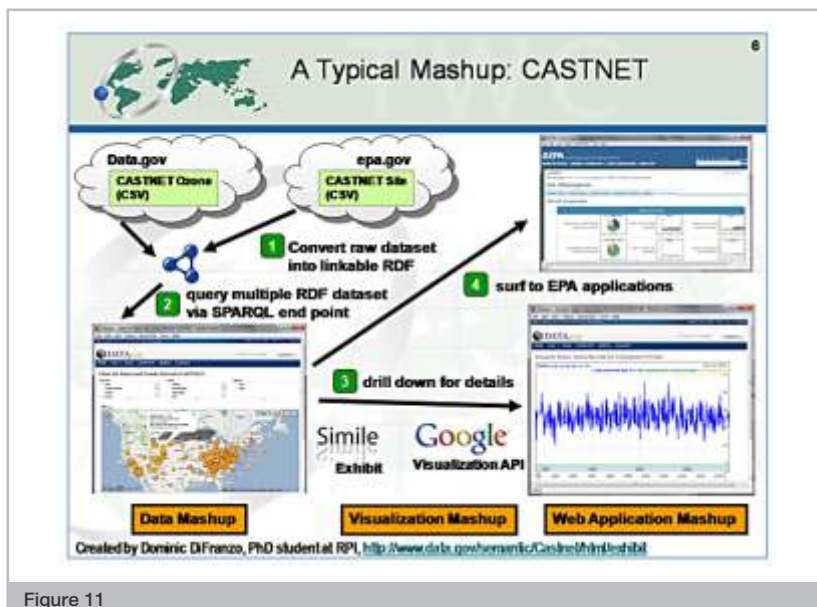


Figure 11

Semantics in the Enterprise

The complexity of everything seems to grow faster than we can comprehend, with real-world consequences that are sometimes catastrophic. We all were witness to, and affected to some degree by the recent, nearly global implosion of the global banking sector. This will continue to be felt for years, if not generations to come. While there were some clear-headed individuals who tried to warn our government and corporate leaders that disaster was imminent, the most common public reaction we heard was “who could have predicted?”, or “nobody saw this coming.” Actually, some smart well-positioned short-sellers made fortunes, and some of the biggest players were hedging both sides. But that's not this topic. The question we're asking here is, “how could we have predicted?”

“Assessment and analysis of systemic risk not only requires Legal Entity and Financial Instrument Identifiers but also a deeper understanding of their complex hierarchies and traceability across front-middle-back office systems. Such capability can be better achieved by modeling the business events and concepts, their contextual nuances and inter-relationships supported by standard interchange formats across heterogeneous IT systems” said Eric Chacon, Global Head of Data Standards at Citi Chief Data Office and Dr. Harsh W. Sharma of Citi Chief Data Office, OMG Finance Task Force Co-Chair¹⁰⁸

From another point of view, where more research is likely being applied, is “how can we identify opportunity amidst the natural chaos of the markets while minimizing our risk?” This is, of course, a primary business objective of industry, and a lot of resources are applied to that end.

How can we use knowledge tools in combination with business intelligence and event-driven business processes to make our business smarter?

“Semantic precision is now recognized as the prerequisite for automating business processes and essential to ensure confidence in analytical objectives of both financial institutions and regulators”, said Michael Atkin, Managing Director of the EDM Council.¹⁰⁹

Integration: Heterogeneous data models and storage technologies abound. IT landscapes are littered with disparate and incompatible repositories integrated using several generations of tools – ftp, ETL, hubs, buses, and services. Mapping and maintaining all of this plumbing is daunting. Semantic technologies offer a path to lightweight, agile integration within the enterprise and across the web, and they offer a knowledge-based model for creating views across document repositories, emails, news and opinion feeds, social media, and the web. The open enterprise is emerging, and new companies are leveraging semantics into a new generation of business models.

Compliance: Rules and the nuances of their application shift and evolve continually. Much of the damage wrought in the latest financial cataclysm was due to exotic derivatives without any regulatory oversight. Is it possible that some common-sense ruleset could have been formulated, adapted dynamically, and applied to filter and correlate required reporting and publically available information streams? Shouldn't the industry have noticed that there was some fishy business transpiring, such as the fact that there was some 30 trillion dollars worth of mortgage-based derivatives in circulation?

As a colleague asked me, “how do you recognize a Ponzi scheme?” It would seem simple, right? One especially notorious fraudster collected almost 20B USD over twenty-plus years in an obviously too-good-to-be-true investment operation. Recognition of fraudulent schemes distributed among cooperating parties, each of whom appears to be acting lawfully, can only be recognized by identifying and tracing the relationships among them, and the pattern of their interactions over time and across multiple jurisdictions.

Risk: This is a problem similar to compliance in financial markets, but with different rules. What is risk? How do we identify cumulative or combinatorial scenarios within and across markets and financial instruments that expose us to risk? Can we detect early warning signals? Can we project trends? Can we recognize the potential for counterparty contagion to infect us? As web-service interfaces are exposed across the enterprise and out to the internet, how do we best fortify our access points using automated decisioning?

Market Research: Across markets and instruments, where are the opportunity differentials? What are the risks? We may ask “how can we identify opportunity amidst the natural chaos of the markets while avoiding risk?” This is, of course, the primary business model of business, but in our expanding information universe, this entails sifting through masses of data, sometimes not obviously related, stored in heterogeneous systems across institutional boundaries in a variety of media to identify patterns according to complex and dynamic rulesets. How do we map market potential along dimensions that aren’t captured in hypercubes?

Forensics: Litigation liability can be very costly, even enterprise-threatening. Similar strategies of rules-driven semantic pattern recognition are being applied to this area as well. Smart enterprise text-mining and enterprise search tools based on knowledge representation are active in this arena.

Search, risk, compliance, research, and forensics applications of this technology are all just different faces on the same problem set. A core semantics platform would be essentially content-neutral. Customization for a specific domain will require acquiring, creating, and adapting ontologies, i.e. semantic data models, with the attributes and relationships that bring each domain into operational focus.

Actionable Enterprise Architecture: As we have seen, relational database and software-design UML models are relatively simple ontologies with some rules built in. SQL is a rules language. Rich semantic modeling using ontologies and reasoners can provide us with a master semantic framework from which these other models and specifications can be derived. Semantic modeling can also form a core methodological framework for software design and systems evolution.

Business ontologies must be maintained at a high level, somewhere in the intersection of IT and business. Design specifications will naturally align to business models if these are projections of semantic models. Semantic technologies will eventually permeate enterprise IT and form a knowledge layer over systems architectures that will bridge business and technical models, affording the possibility of automating architectural alignment.

Semantic models will increasingly form the essential modeling infrastructure for Enterprise Architecture, providing agile alignment between business models and the various granularities of architectural models. This technology is still in an early-adopter stage, but its inevitability is rapidly gaining recognition. Semantic capabilities will embed the power of Knowledge throughout Enterprise Information Systems.

Semantic Web technologies are being applied now in such areas as knowledge-based search, rules consolidation and reuse, master data, metadata and content management, publishing, eCommerce, internet advertising, service arbitration, text mining and eDiscovery, social networking, clinical trials, segmentation and targeting, and real-time semantic cross-domain RDBMS integration for analytics and BI.

Semantic technologies are a conceptual leap from the current state of IT, and most organizations have not positioned themselves to apply the resources required to take this technology to fruition measured in bottom-line business value.

In Conclusion

Bringing this to a close isn't easy. A lot has happened while it was under construction. New applications keep popping up. More for another day.

Let's just end with this:

Businesses today compete in a globally networked world of exponentially increasing complexity and scale. Service integration is supplanting traditional monolithic application models, and many companies are eager to push their IT into the cloud, where service-and-event-based rules-driven process automation will dominate. The necessity of applying ever smarter process controls in this environment is apparent, but will propel us into the next generation.

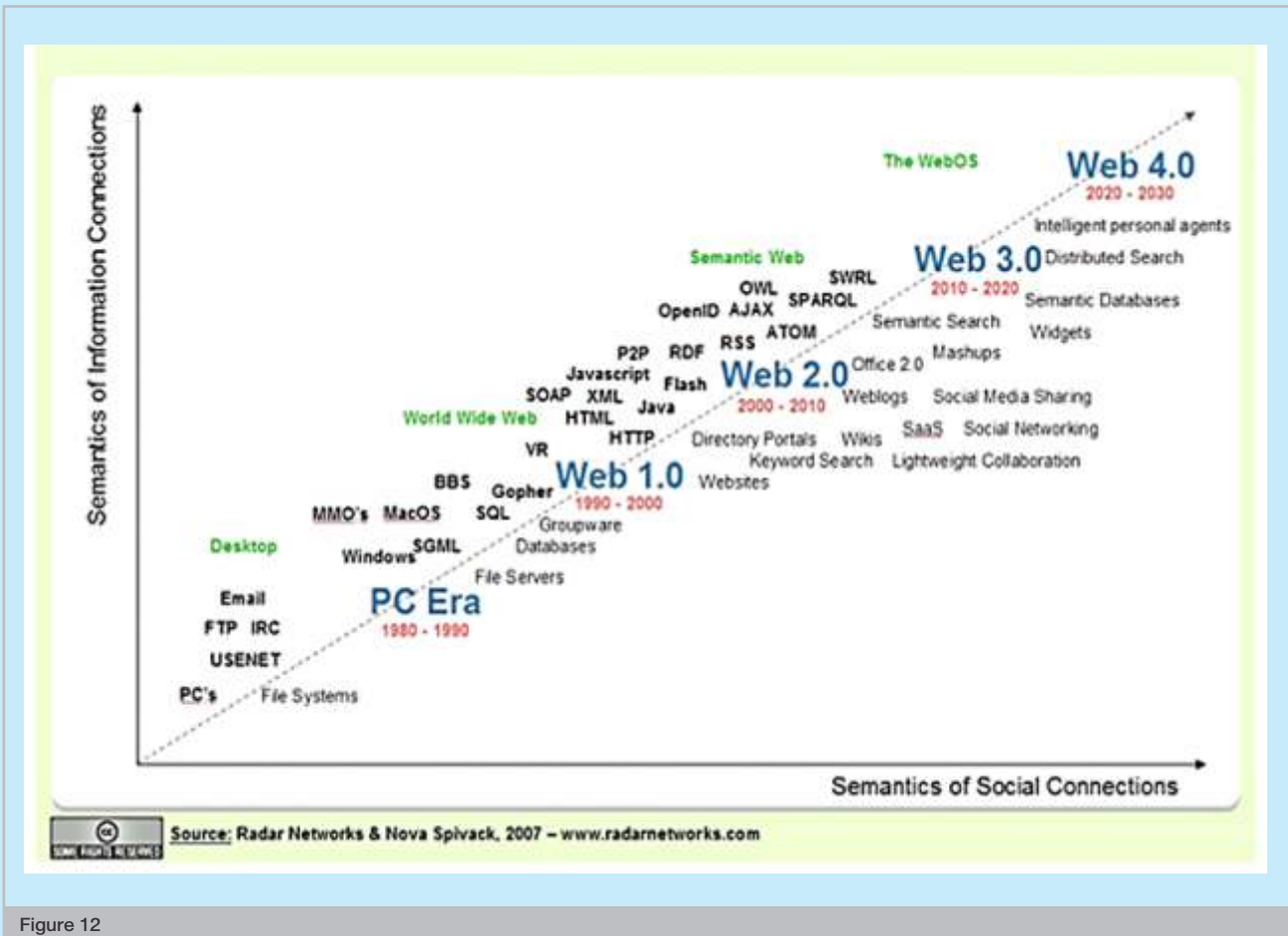


Figure 12

Web 3.0 – Nova Spivak –Minding the Planet, NovaSpivack.com¹¹⁰

This world-wide-web of information is self-integrating bottom-up. Linked data and semantic models are the information infrastructure and the knowledge fabric of the emerging planetary knowledge-sphere. Web 3.0 is upon us.

Today there are methods, standards, tools, and platforms rich enough and mature enough to rationalize vast, complex, and dynamic systems of heterogeneous platforms, and to lift us into the world of possibilities of the Semantic Web.

Links

1. For an example see Ontological Constructs to Create Money Laundering Schemes Murad Mehmet and Duminda Wijesekera, George Mason University Fairfax, undated, retrieved March 15, 2011
2. <http://www.ft.com/cms/s/2/5b68adac-5721-11e0-9035-00144feab49a.html#axzz1HjIH822Y> - retriever March 29, 2011
3. [http://en.wikipedia.org/wiki/Ontology_\(philosophy\)](http://en.wikipedia.org/wiki/Ontology_(philosophy)) retrieved March 17, 2011
4. [http://en.wikipedia.org/wiki/Ontology_\(information_science\)](http://en.wikipedia.org/wiki/Ontology_(information_science)) retrieved March 17, 2011
5. <http://www.w3.org/TR/owl2-overview/> Recommendation 10/27/2009 - May Be Superseded - retrieved March 15, 2011
6. http://techwiki.openstructs.org/index.php/Intro_to_Ontologies - retrieved March 15, 2011
7. [http://en.wikipedia.org/wiki/Ontology_\(information_science\)](http://en.wikipedia.org/wiki/Ontology_(information_science)) - retrieved March 21, 2011
8. <http://mitpress.mit.edu/e-books/Hal/chap2/two1.html> - retrieved March 21, 2011
9. http://en.wikipedia.org/wiki/Semantic_reasoner - retrieved March 15, 2011
10. Ibid University of Manchester Pizza Tutorial - retrieved March 15, 2011
11. Ibid
12. <http://www.w3.org/2001/sw/Copyright> © 1994-2011 W3C © (MIT, ERCIM, Keio), All Rights Reserved. - retrieved March 15, 2011
13. Three examples among many: <http://mashable.com/>, <http://www.opencalais.com/applications/wirecatch-news-aggregator>, and <http://realtravel.com/> - all retrieved March 30
14. http://semanticweb.org/wiki/Main_Page retrieved March 17, 2011
15. http://en.wikipedia.org/wiki/Resource_Description_Framework retrieved March 17, 2011
16. <http://www.w3.org/RDF/> retrieved March 17, 2011
17. ibid
18. http://en.wikipedia.org/wiki/Uniform_Resource_Identifier
19. <http://dfdf.inesc-id.pt/tr/web-arch> - Xiaoshu Wang November 12, 2007, retrieved March 17, 2011
20. <http://www.w3.org/TR/rdf-syntax-grammar/> W3C Recommendation 10 February 2004, Copyright © 2004 W3C® (MIT, ERCIM, Keio), All Rights Reserved- retrieved March 17, 2011
21. http://en.wikipedia.org/wiki/Glossary_of_graph_theory - retrieved March 17, 2011
22. See <http://www.w3.org/TR/> for all of the W3C standards. retrieved March 17, 2011
23. <http://en.wikipedia.org/wiki/RDFa> retrieved March 17, 2011
24. <http://www.w3.org/TR/2008/WD-xhtml-rdfa-primer-20080620/#id84395> W3C Working Draft 20 June 2008, Copyright © 2008 W3C® (MIT, ERCIM, Keio), All Rights Reserved.- retrieved March 15, 2011

25. <http://www.w3.org/TR/2004/REC-owl-features-20040210/#s3.1> W3C Recommendation 10 February 2004, Copyright © 2004 W3C® (MIT, ERCIM, Keio), All Rights Reserved. - retrieved March 17, 2011
26. http://en.wikipedia.org/wiki/Simple_Knowledge_Organization_System retrieved March 17, 2011
27. <http://www.w3.org/DesignIssues/LinkedData.html> Tim Berners-Lee - last updated June 18, 2009, retrieved March 17, 2011
28. <http://www.w3.org/TeamSubmission/turtle/#sec-intro> W3C Team Submission 14 January 2008, David Beckett and Tim Berners-Lee W3C Copyright © 2008 W3C® (MIT, ERCIM, Keio), All Rights Reserved, retrieved March 17, 2011
29. http://en.wikipedia.org/wiki/Web_Ontology_Language - retrieved March 15, 2011
30. <http://theory.stanford.edu/~megiddo/pdf/searchgraph.pdf> - retrieved March 15, 2011
31. http://en.wikipedia.org/wiki/Web_Ontology_Language#Species retrieved March 16, 2011
32. <http://www.w3.org/TR/2009/REC-owl2-xml-serialization-20091027/> W3C Recommendation 27 October 2009, Copyright © 2009 W3C® (MIT, ERCIM, Keio), All Rights Reserved - retrieved March 16, 2011
33. [http://en.wikipedia.org/wiki/Reification_\(computer_science\)#Reification_on_Semantic_Web](http://en.wikipedia.org/wiki/Reification_(computer_science)#Reification_on_Semantic_Web) - retrieved March 16, 2011
34. <http://www.w3.org/TR/2010/WD-sparql11-query-20100126/#WritingSimpleQueries> a Working Draft. Copyright © 2010 W3C® (MIT, ERCIM, Keio), All Rights Reserved. retrieved March 16, 2011
35. <http://www.semantic-web.at/1.36.resource.90.jeen-broekstra-x22-the-importance-of-sparql-cannot-be-overestimated-x22.htm> - retrieved March 18, 2011
36. http://www.w3.org/TR/rdf-sparql-query/#defn_RDFTerm - retrieved March 31, 2011
37. Ibid W3C sparql11-query
38. <http://www.oracle.com/technetwork/database/options/semantic-tech/semtech11gr2-featover-131765.pdf> Oracle Database Semantic Technologies - Oracle Feature Overview - retrieved March 18, 2011
39. <http://www.w3.org/TR/2010/WD-sparql11-federated-query-20100601/> retrieved March 23, 2011
40. <http://www.w3.org/TR/2010/WD-sparql11-query-20100126/#WritingSimpleQueries>- W3C Working Draft 26 January 2010 Copyright © 2010 W3C® (MIT, ERCIM, Keio), All Rights Reserved. - retrieved March 23, 2011
41. <http://www.joseki.org/> - retrieved March 18, 2011
42. <http://www.w3.org/TR/rdf-sparql-protocol/> - W3C Recommendation 15 January 2008, Copyright © 2006-2007 W3C® (MIT, ERCIM, Keio), All Rights Reserved. retrieved March 21, 2011
43. <http://www.vimeo.com/14569996> includes a complete update on the Jena project. - retrieved March 21, 2011
44. Ibid Mehmet and Wijesekera
45. <http://ruleml.org/> - retrieved March 25, 2011

46. <http://www.w3.org/Submission/SWRL/> - W3C Member Submission 21 May 2004, Ian Horrocks, Peter F. Patel-Schneider, , BBN Copyright © 2004 National Research Council of Canada, Network Inference, and Stanford University. retrieved March 25, 2011
47. <http://www.w3.org/TR/rif-overview/> W3C Working Group Note 22 June 2010, Copyright © 2010 W3C® (MIT, ERCIM, Keio) - retrieved March 25, 2011
48. <http://www.omg.org/spec/SBVR/> - retrieved March 25, 2011
49. [http://en.wikipedia.org/wiki/Upper_ontology_\(information_science\)](http://en.wikipedia.org/wiki/Upper_ontology_(information_science)) - retrieved March 16, 2011
50. <http://researchcyc.cyc.com/> - retrieved March 16, 2011
51. <http://opencyc.org/> - retrieved March 16, 2011
52. Ibid ResearchCyc
53. <http://www.umbel.org/> - retrieved March 16, 2011
54. <http://www.heppnetz.de/projects/goodrelations/> - retrieved March 25, 2011
55. <http://wordnet.princeton.edu/> - retrieved March 30, 2011
56. <http://www.geonames.org/> - retrieved March 30, 2011
57. <http://www.mpi-inf.mpg.de/yago-naga/yago/> - retrieved March 30, 2011
58. <http://www.opencalais.com/> - retrieved March 22, 2011
59. <http://drupal.org/project/opencalais> – retrieved March 22, 2011
60. <http://www.edmcouncil.org/default.aspx> - retrieved March 28, 2011
61. <http://dbpedia.org/About>– retrieved August 8, 2011
62. http://wiki.freebase.com/wiki/What_is_Freebase%3F – retrieved August 8, 2011
63. <http://fadyart.com/Treebolic/Tfi.html> - retrieved March 17, 2011
64. <http://www.data.gov/catalog/raw> – retrieved March 22, 2011
65. <http://www.data.gov/catalog/geodata> – retrieved March 22, 2011
66. <http://www.data.gov/catalog/tools> – retrieved March 22, 2011
67. <http://www.data.gov/developers/showcase> – retrieved March 22, 2011
68. <http://www.data.gov/> – retrieved March 22, 2011
69. <http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets> - retrieved March 17, 2011
70. http://videlectures.net/iswc06_gruber_wswms/ Tom Gruber presentation “Where the Social Web Meets the Semantic Web” at The 5th International Semantic Web Conference November 2006 – retrieved March 30, 2011
71. <http://tagcommons.org/> – retrieved March 31, 2011
72. <http://code.google.com/p/swoop/> retrieved March 22, 2011

73. <http://protege.stanford.edu/> - Copyright © 2011 Stanford Center for Biomedical Informatics Research retrieved March 17, 2011
74. A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools Edition 1.2 Matthew Horridge, et al. March 13, 2009, ©The University Of Manchester - retrieved March 17, 2011
75. <http://clarkparsia.com/pellet/> - retrieved March 18, 2011
76. http://neon-toolkit.org/wiki/Main_Page retrieved March 17, 2011
77. <http://www.eclipse.org/> retrieved March 17, 2011
78. A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools Edition 1.2 Matthew Horridge, et al. March 13, 2009, ©The University Of Manchester - retrieved March 17, 2011
79. Collaboration among semantic modelers and subject matter experts in a key aspect of enterprise semantic implementations and model maintenance This tool is designed for that.
80. <http://jena.sourceforge.net/> - retrieved March 18, 2011
81. <http://www.openrdf.org/> - retrieved March 18, 2011
82. <http://www.openrdf.org/doc/sesame2/users/ch03.html#figure-sesame-components> - retrieved March 18, 2011
83. <http://www.openrdf.org/doc/sesame/users/ch04.html#d0e659> - retrieved March 18, 2011
84. <http://virtuoso.openlinksw.com/vdb-conceptual-architecture/>
85. <http://openjena.org/TDB/> retrieved March 12, 2011
86. <http://www.openjena.org/SDB/> retrieved March 12, 2011
87. <http://revelytix.com/category/tags/products> - retrieved August 8, 2011
88. <http://www.ontotext.com/owlim/> – retrieved March 31, 2011
89. <http://www.ontotext.com/owlim/> – retrieved March 31, 2011
90. <http://hadoop.apache.org/> – retrieved March 25, 2011
91. <http://wiki.apache.org/incubator/HeartProposal> – retrieved March 25, 2011
92. <http://hadoop.apache.org/hbase> – retrieved March 25, 2011
93. <http://lucene.apache.org/solr/> - retrieved March 18, 2011
94. <http://dl.orchestranetworks.com/sita/MAG/Download/passport/MAGPassportMDM.pdf> - Pierre Bonnet, Creative Commons by MDM Alliance Group – retrieved March 30, 2011
95. <http://www.nytimes.com/2011/03/05/science/05legal.html?hpw>
96. <http://gate.ac.uk/> – retrieved March 30, 2011
97. <http://sourceforge.net/projects/s-match/> – accessed March 25, 2011
98. <http://s-match.org/> – retrieved March 24, 2011, also see S-Match: an open source framework for matching lightweight ontologies, Fausto Giunchiglia, Aliaksandr Autayeu and Juan Pane June 30, 2010

99. <http://pipes.yahoo.com/pipes/> – retrieved March 24, 2011
100. <http://www-958.ibm.com/software/data/cognos/manyeyes/> – retrieved March 24, 2011
101. <http://research.microsoft.com/en-us/collaboration/focus/cs/web-ngram.aspx> – retrieved March 30, 2011
102. <http://code.google.com/apis/visualization/documentation/gallery.html> – retrieved March 30, 2011
103. <http://www.simile-widgets.org/exhibit/> – retrieved March 24, 2011
104. <http://data-gov.tw.rpi.edu/2010/data-gov-talk-2010-suny-ctg.ppt> by Li Ding, and Professors James Hendler and Deborah McGuinness from the Tetherless World Constellation project at RPI. – retrieved March 30, 2011
105. <http://www.wallstreetandtech.com/data-management/229301007> – retrieved March 18, 2011
106. http://news.yahoo.com/s/usnw/20110315/pl_usnw/DC65807 – retrieved March 21, 2011
107. Nova Spivak - Minding the Planet – retrieved March 30, 2011

About the Author



Dennis Pierson

Senior Information Systems Architect and head of our semantics practice

With a career spanning 30 years in the software industry, Dennis has worked as a consultant and in product development in Financial Services, Pharmaceuticals, Insurance, Defense, and Manufacturing. He heads Mphasis' Semantics Lab, using Domain Modeling as a core architecture practice, and bringing Semantic Web knowledge-based solutions to enterprise IT.



ABOUT MPHASIS.

Mphasis is a \$1 billion global service provider, delivering technology based solutions to clients across the world. With currently over 41,000 people, Mphasis services clients in Banking and Capital Markets, Insurance, Manufacturing, Communications, Media & Entertainment, Healthcare & Life Sciences, Transportation & Logistics, Retail & Consumer Packaged Goods, Energy & Utilities, and Governments around the world. Our competency lies in our ability to offer integrated service offerings in Applications, Infrastructure Services, and Business Process Outsourcing capabilities. To know more about Mphasis, log on to www.mphasis.com

For more information, contact: sales@mphasis.com.

USA: 460 Park Avenue South, Suite #1101, New York, NY 10016, USA
Tel.: +1 212 686 6655, Fax: +1 212 686 2422

UK: 88 Wood Street, London EC2V 7RS, UK
Tel.: +44 20 85281000, Fax: +44 20 85281001

AUSTRALIA: 9 Norberry Terrace, 177-199 Pacific Hwy, North Sydney, 2060, Australia
Tel.: +61 2 99542222, Fax: +61 2 99558112

INDIA: Bagmane Technology Park, Byrasandra Village, C.V. Raman Nagar, Bangalore 560 093, India
Tel.: +91 80 4004 0404, Fax: +91 80 4004 9999



0811