

Finding and Exploiting Value in Semantic Technologies on the Web

David W. Cearley, Whit Andrews, Nicholas Gall

This research note seeks to establish where Web-based semantic technologies will be useful in the enterprise. Enterprises will hear much hype about these technologies, and must evaluate their impact and define a strategic and pragmatic approach.

Key Finding

- During the next 10 years, Web-based technologies will improve the ability to embed semantic structures in documents, and create structured vocabularies and ontologies to define terms, concepts and relationships. This will offer extraordinary advances in the visibility and exploitation of information — especially in the ability of systems to interpret documents and infer meaning without human intervention. However, the grand vision of the Semantic Web will occur in multiple evolutionary steps, and small-scale initiatives are often the best starting points.

Recommendation

- Enterprises that selectively adopt individual Web-based semantic technologies with specific, pragmatic goals in mind will realize benefits long before those that attempt to solve comprehensive, or even multiproject, semantic problems. Maintaining focus on interoperability, instead of universality, will be critical. Semantic hypertext should be the focus for most Web projects, with extensive ontologies in Resource Description Framework (RDF) or Web Ontology Language (OWL) selectively used for more complex needs.

TABLE OF CONTENTS

Strategic Planning Assumption.....	3
Analysis	3
1.0 What You Need to Know	3
2.0 Semantic Standards	4
3.0 Taking Action	5
4.0 Appendix.....	6
4.1 The Semantic Web.....	6
4.2 Semantic Hypertext	7
4.3 Contrasting the Semantic Web and Semantic Hypertext	8
4.4 The Semantic Web Evolution	8
4.5 The Semantic Web Standards Evolution.....	9
4.6 Focusing Semantic Web Projects	10

LIST OF FIGURES

Figure 1. The Semantic Web.....	7
Figure 2. Contrasting the Semantic Web and Semantic Hypertext.....	8
Figure 3. The Semantic Web Evolution	9
Figure 4. Focusing Semantic Web Projects	11

STRATEGIC PLANNING ASSUMPTION

By 2012, 70% of public Web pages will have some level of semantic markup, but only 20% will use more extensive Semantic Web-based ontologies (0.6 probability).

ANALYSIS

1.0 What You Need to Know

Web-based semantic technologies, when properly focused on specific problems, can offer significant benefits, while a generalized and broad application will often have little impact. However, focused projects using both semantic hypertext and complete Semantic Web approaches must build toward a more comprehensive long-term semantic vision for the enterprise. Organizations should expect technology, vocabulary and ontology standards to evolve during the next five-plus years, and should focus effort on creating internal standards for metadata that can be leveraged in Web-based semantic technology efforts.

The Semantic Web is a grand vision for the future of the Web, as well as a collection of individual technologies to implement this vision. The Semantic Web was first described in detail by Tim Berners-Lee in 2000 (see Section 4.1), but its roots go back to the earliest days of the Web. Since its unveiling, the Semantic Web has been full of promise, but largely as an exercise for the academic and scientific communities. Standards have been slow to emerge, and demonstrations that clearly show its value are few. Meanwhile, another Web-based semantic technology has emerged, which we call "semantic hypertext" (see Section 4.2). Semantic hypertext in the form of microformats has gained significant momentum on the public Web, with bloggers and social-networking sites using it to add semantic tags to describe elements of their Web pages.

At first glance, the Semantic Web and semantic hypertext would appear to be at odds with each other (see Section 4.3). Microformats, the poster child for semantic hypertext, eschew use of RDF and focus only on semantic documents. In the classic definition, the Semantic Web is based on RDF for syntax and is focused on semantic data. Gartner believes this debate is ultimately counterproductive. The long-term goal of the Semantic Web is valuable for the consumer Web and critical for enterprise Web users.

However, the simple, focused, grassroots Web 2.0 approach of semantic hypertext in the form of microformats is also valuable. This generates immediate value and provides a simple first step to a Semantic Web. Meanwhile, RDF-based semantic hypertext models are emerging with an affinity to the classic Semantic Web described by Berners-Lee. In addition, technologies are emerging to convert microformats to RDF (see standards section below). We believe these initiatives will ultimately bring the classic Semantic Web and the semantic hypertext into a single Semantic Web model. Enterprises should view the "Semantic Web" as inclusive of semantic hypertext, and take pragmatic steps to implement the appropriate semantic technologies for either semantic documents or semantic data as needed.

The Semantic Web will not replace the current Web in a single step. Rather, we expect the Semantic Web to evolve slowly during the next 20 years as the Web evolves (see Section 4.4). Through 2012, we expect semantic hypertext approaches to be most broadly used on the Web itself, as Web site designers incrementally add simple semantic definitions to data on existing Web pages. More-extensive semantic Web ontologies will be created in specific domains where there is a need to define extensive formal vocabularies or complex data relationships, and to apply advanced techniques to infer relationships across datasets.

Industries where this is most likely to apply include life sciences, healthcare, library, defense, government, energy (the oil industry) and financial services. In many cases, these sophisticated semantic needs are being addressed today by a variety of proprietary tools that inherently have nothing to do with the Web. There are many established classification schemes that are not even in XML, much less RDF. However, we expect the Semantic Web standards (for example, RDF and OWL) will be embraced by these vendors and users over time as RDF and OWL capabilities mature. Use of a common model for describing semantic relationships (that is, RDF and OWL) will increase the ability to join ontologies that were created using different vendor systems. Indeed, enterprises would be well-served by insisting that any vendor providing sophisticated ontology-based systems provide at a minimum an RDF or OWL import/export capability.

By 2017, we expect the vision of the Semantic Web as defined by Berners-Lee to coalesce, as vocabularies and ontologies from various industries mature, and the majority of Web pages are decorated with some form of semantic hypertext. Maturing standards and use of bridge approaches such as Gleaning Resource Descriptions From Dialects of Languages (GRDDL) will allow semantic markup to be exposed as RDF data. In the longer term, the Semantic Web will likely require new methods such as the use of tuple spaces to deal with the need to integrate and reconcile changing RDF information across the Web in real time. By 2012, 80% of public Web sites will use some level of semantic hypertext to create Semantic Web documents (0.7 probability). By 2012, 15% of public Web sites will use more extensive Semantic Web-based ontologies to create semantic databases (0.6 probability).

2.0 Semantic Standards

Web-based semantic standards fall into two key categories: 1) standards for describing controlled vocabularies, taxonomies and ontologies; and 2) standards for mapping or attaching such semantic metadata to existing information resources.

The classic Semantic Web has focused on the first category of standards (see Section 4.5). RDF and OWL are used to create vocabularies, taxonomies and ontologies, which are stored in RDF repositories. To be useful, though, these ontologies must be accessible. SPARQL is emerging as a query language for RDF data, but is not yet a fully ratified standard. A variety of vendor tools have emerged during the past two years to create, access and process Semantic Web data (for example, RDF generators, ontology management, validators, stores or inference engines). However, these classic Semantic Web standards and related tools do not provide an easy means for Web site developers or publishers to tap into the existing universe of Web pages. The slow pace of standards and lack of focus on "legacy pages" have hampered Semantic Web adoption.

Created specifically to address the second category of standards, microformats provide a mechanism for embedding semantic metadata into existing HTML pages. "Folksonomy"-tagging services such as del.icio.us also address this category by providing a mechanism to map tags to resources, but they do not embed information into the original resource. Both approaches have proven popular, and microformats are arguably a de facto semantic hypertext standard.

RDF-based mechanisms (such as eRDF or RDFa) are emerging as potential standards for semantic hypertext with a greater affinity to the classic Semantic Web. However, embedded RDF (eRDF) is a W3C project largely driven by a single individual, and RDF attributes (RDFa) is an initiative led by a single vendor. Neither at this time is certified by a standards body, nor do they have enough momentum to be de facto standards.

GRDDL is another emerging standard (though not yet ratified) that addresses the second category. Using GRDDL provides a mechanism to convert information and metadata from microformat-tagged HTML documents into RDF, thus bringing semantic hypertext into the Semantic Web.

Enterprises will also see ontology standards (that is, the metadata describing specific terms, concepts and relationships) emerge from an industry perspective. For example, a number of ontologies are emerging from the academic, scientific and focused Web communities. Other efforts are emerging from standards groups working with key industries. Users should expect the evolution of standard cross-enterprise ontologies to be even slower than the evolution of more-simple semantic vocabularies. The exception will be in industries such as library science where the industry has historically created rich vocabularies and ontologies to organize information. Gartner will explore the evolution of these vocabularies and ontologies in future research.

3.0 Taking Action

In general, the notion is that any Web page, document or generated data record is more useful if machines can interpret it swiftly and crisply, if not with the same nuanced interpretation that a human might. Such benefits will drive enterprises and independent Web publishers to adopt Semantic Web technologies, even if they do so incrementally, as improving the likelihood that a document may be found increases the chance of deriving value from it. Reducing the length of time it takes to find a document makes its value more immediately realizable (because less time has been spent finding it). Allowing machine-driven resources to locate a document increases its value by making it available to a separate audience of applications and application users.

Web-based semantic technologies should be applied to specific targeted problems, and different technologies will be applied to different problems. Most of these benefits can be achieved using semantic hypertext (for example, microformats or eRDF) to tag specific documents. At a minimum, enterprises should be exploring the use of semantic hypertext to better define the semantics of page elements on both internal and externally focused Web pages. Adding this level of semantic information will improve the "mashability" of these sites.

The need to define more complex concepts and relationships, as well as to apply advanced techniques to infer relationships across documents and datasets, will lead to the use of more-sophisticated Semantic Web approaches. It is in the comparatively narrow context of enterprise and industry vocabularies, taxonomies, classification schemes and ontologies, as opposed to the World Wide Web at large, that we recommend enterprises pursue value from the full Semantic Web. Enterprises must also consider the maturity of existing metadata vocabularies. Where base metadata standards are mature, there is the potential to create more generalized ontologies that span multiple data sets, applications and even enterprises. Where these standards are immature, the scope of the project must be narrow, and the ontologies focused on a specific problem domain (see Section 4.6).

A pragmatic approach to adopting Web-based semantic concepts and technologies will enable enterprises to derive value today, while positioning themselves for the long-term emergence of the Semantic Web. Just as the broad Web will gradually improve as semantic technologies are applied to the pages — new and existing — within individual sites and families of sites, so will companies and government entities find benefits through discrete application of semantic technologies in specific projects that ultimately interconnect.

Enterprises that expect end users to manually create definitions and relationships and tag documents and data records will be disappointed. Too many perspectives and skill sets are necessary, and too much inconsistency will occur. User-created folksonomies are useful for personal categorization and social bookmarking, but they will not provide the degree of rigor needed for a business vocabulary or ontology that defines sophisticated concepts and relationships. An alternative is to present the user with a pre-defined taxonomy with which to tag documents. Folksonomies may be used as input toward creation and evolution of this taxonomy. However, any approach that requires significant user tagging is suspect, because users often fail to use the tags or use them improperly. A better approach is to use pre-defined vocabularies or

ontologies during the course of scanning a document to automatically generate semantic information about the document. This information can automatically be tagged to the document, presented to a human for review before tagging, or generated again when the need arises.

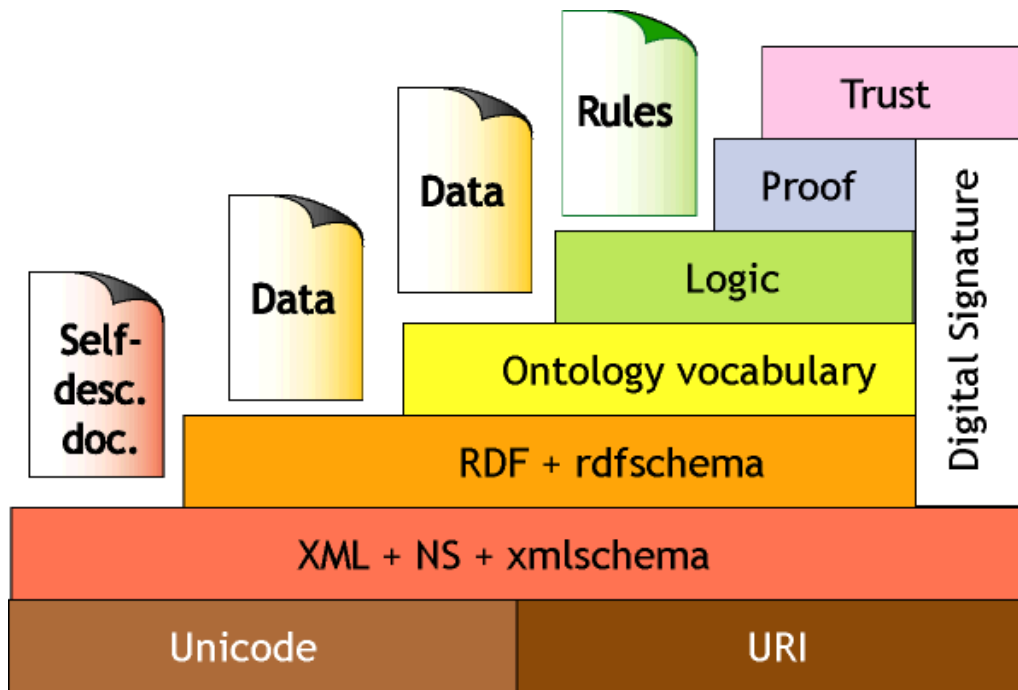
Technologies to accomplish semantic analysis do exist today, and more are emerging from established and new vendors. Acquiring software that serves the needs of classification, categorization, tagging, entity extraction, concept extraction, fact extraction, document abstraction or other capabilities directly linked to document definition and maintenance will significantly improve project success.

4.0 Appendix

4.1 The Semantic Web

The Web as it has evolved is geared primarily toward human consumption of information. Information on various Web sites can be understood by humans in the context of how the information is used by the site. Machines do not have the ability to intuit meaning like a human. In addition, ambiguity of meaning may still persist even in human-centric information access across multiple Web sites. To make meaning and context accessible to machines and remove ambiguity require a more explicit definition of metadata and semantic relationships. The Semantic Web vision attempts to address this issue by shifting away from the current model of a "Web of documents." In the Semantic Web, a "Web of data" describes all resources and provides a universal medium for exchange of information. Semantic standards and markup languages are used to describe meaning and establish relationships between data elements. The Semantic Web defines common formats for integration and combination of data drawn from diverse sources. It also defines a language for recording how the data relates to real-world objects. This allows a person or a machine to start off in one database, and then move through an unending set of databases, which are connected via semantic relationships (see Figure 1 from www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html).

Figure 1. The Semantic Web



Source: Tim Berners-Lee/W3C

4.2 Semantic Hypertext

Semantic hypertext approaches provide defined mechanisms for embedding semantics into HTML documents.

Microformats are the most widely used semantic hypertext model. As stated in the microformats wiki (see http://microformats.org/wiki/Main_Page), microformats are simple conventions for embedding semantic markup for a specific problem domain in human-readable XHTML or XML documents and RSS or Atom feeds. Structured blogging can best be seen as an implementation of microformats. It is implemented via a plug-in to enable publishing of structured XHTML with microformat-compatible page tags. The structured blogging plug-in also transports microcontent in XML feeds (RSS 1.0, RSS 2.0 or Atom) and provides automatic conversion into RDF and, in the future, Outline Processor Markup Language (OPML). In addition to being a container for RSS, OPML is a general-purpose outline structure for constructing hierarchies based on XML data. OPML 2.0 with namespaces enables any XML-formatted data to be incorporated into an OPML outline. OPML may be yet another mechanism for dealing with semantic hypertext.

RDFa is a mechanism for embedding RDF in XHTML. Created by the XHTML task force of the W3C Semantic Web best practices and deployment working group, in conjunction with the W3C HTML working group, it currently exists as a best-practice primer. A new working group was established in the fourth quarter of 2006 to address RDF embedding in more detail, but we do not expect formal W3C recommendations to appear before late 2007.

Embedded RDF is another approach to using RDF to annotate HTML documents. Although open to all, this model was created and promoted by Talis, a provider of products and services for

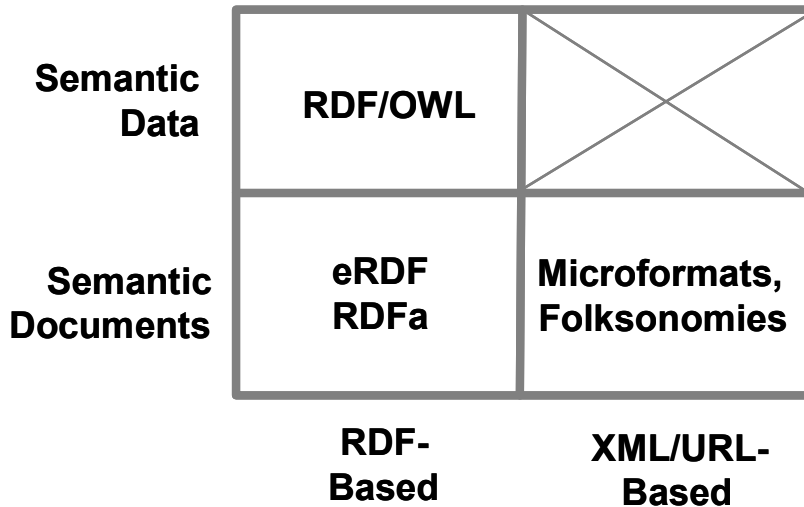
public and academic libraries in the U.K. and Ireland. Along with more formal RDF or OWL constructs, eRDF is actively being used by Talis in the development of its library systems.

4.3 Contrasting the Semantic Web and Semantic Hypertext

The Semantic Web envisions a shift from the current "Web of documents" to a future "Web of data," where information is richly described in data structures. It defines an array of technology standards (for example, XML, RDF or OWL) to enable this idealized future. It is focused on creating comprehensive vocabularies and ontologies to describe not only data elements and their relationships to one another, but also higher-level concepts that can only be inferred from these relationships. The Semantic Web deals with meaning across both unstructured and structured data, and emphasizes the creation of RDF and XML databases to store the semantic information. A primary goal of the Semantic Web is to make it easier for machines to understand and process information through the use of these semantic elements.

Semantic hypertext, on the other hand, starts with the current document-oriented Web model and uses one of the Web's basic technologies (HTML) to add simple semantic extensions. It is focused on creating simple vocabularies, using a Web 2.0 community model and adding simple elements to existing HTML pages to more fully describe the elements on these pages. Semantic markup is focused on Web pages, and the primary goal is to make it easier for Web designers to add slightly better descriptive information to these pages (see Figure 2).

Figure 2. Contrasting the Semantic Web and Semantic Hypertext

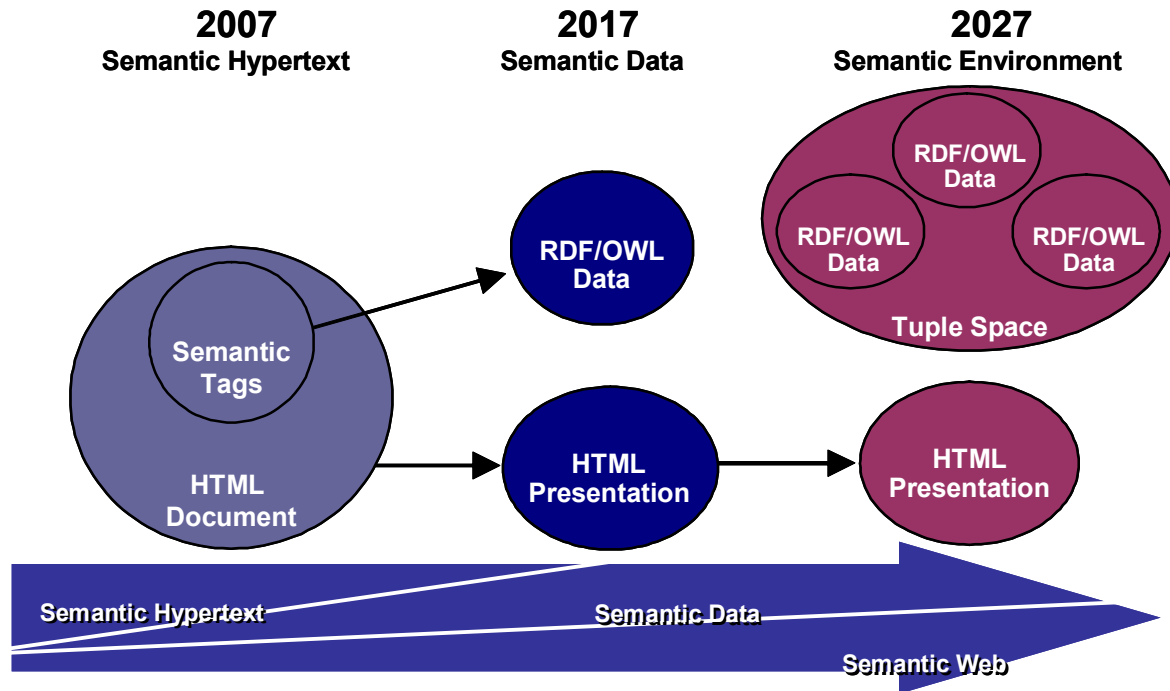


Source: Gartner (May 2007)

4.4 The Semantic Web Evolution

Figure 3 maps the evolution of the Semantic Web from 2007 to 2027.

Figure 3. The Semantic Web Evolution



Source: Gartner (May 2007)

4.5 The Semantic Web Standards Evolution

RDF was approved, in February 2004, as an XML-compliant syntax that uses an XML metamodel, grammar and namespace. RDF is a set of relationships with the schema following the triple pattern (subject, predicate and object), where the predicate defines the relationship between the subject and the object. Triples are used regardless of the underlying data structure (relational database, taxonomic, object and so on). RDF treats data and metadata equivalently, expressing both as triples. RDF and RDF schema (also called RDFS) are metadata (resource description) languages developed for content management. RDF is used for the description of the sheer instance data, whereas RDFS defines what kind of instance data can be defined.

The most important aspect of RDF is the triple pattern. While the XML is the primary syntactical model for RDF triples, it is not the only syntactical model. Notation3 (N3) is a compact and readable alternative syntax that is also more expressive than RDF or XML. N-triples is a line-oriented subset of N3, developed to express the result of test cases and be transmitted in Multipurpose Internet Messaging Extensions (MIME) format. Turtle and TriX are other alternative syntax options. RDF environments often understand multiple syntaxes, and increasingly, authoring tools hide the syntactical details altogether.

A query language and data access protocol for RDF, SPARQL works for any data source that can be mapped into RDF. The specification, managed by the W3C data access working group, has reached "candidate recommendation" status as of April 2006, and is targeted to become a full recommendation by the second quarter of 2007. For the past two years, lack of a standard query language for RDF has been a roadblock for RDF, and SPARQL will make RDF more attractive. The ability to move across and within various data sources is a key part of Web 2.0, and SPARQL has the potential to provide a common query mechanism for mashups as the amount of native RDF- or GRDDL-generated RDF expands. However, a hole remains in the Semantic Web

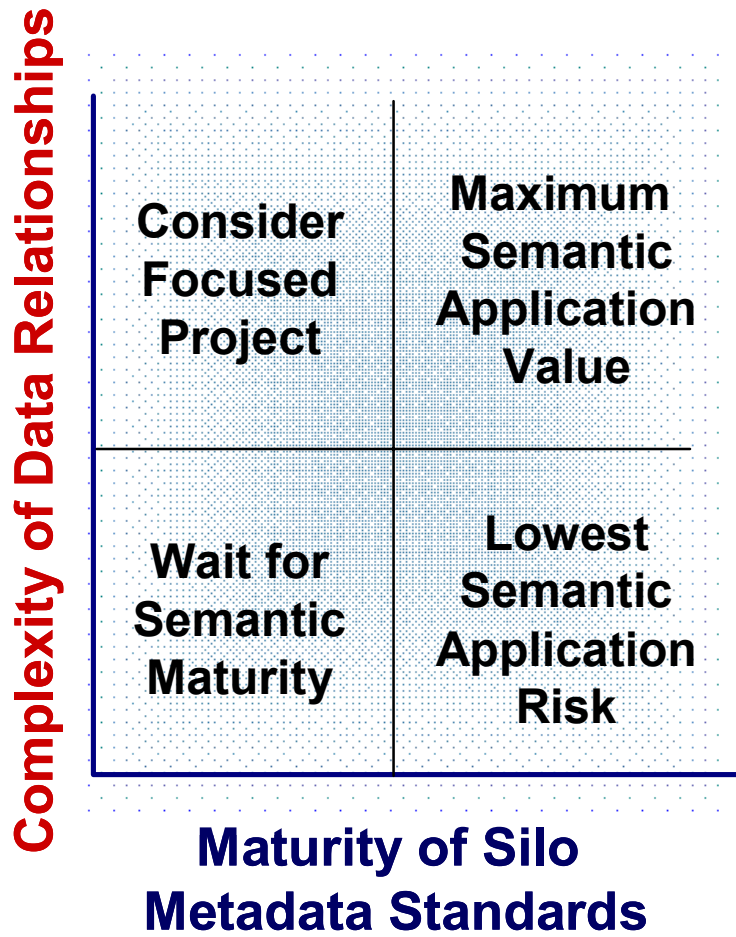
query model with regard to relational database management system (RDBMS) data. It is not feasible to convert all RDBMS data to RDF, given the size and complexity of some systems. "Bridges" are currently being defined that will provide a layer between RDF and the database, with RDBMS tables being mapped hastily to RDF graphs. In some cases, the mapping is generic, while in other cases, separate mapping files define the details.

4.6 Focusing Semantic Web Projects

Figure 4 plots the maturity of silo metadata standards versus the complexity of data relationships:

- In projects of high maturity and low complexity — Focus on aggressive use of semantic markup.
- In projects of high complexity and low maturity — Focus on the Semantic Web, but in a carefully bounded project.
- In projects of high complexity and high maturity — Consider more extensive and generalized Semantic Web infrastructure.
- In projects of low maturity and low complexity — Look for opportunistic use of semantic markup as needs arise.

Figure 4. Focusing Semantic Web Projects



Source: Gartner (May 2007)

REGIONAL HEADQUARTERS

Corporate Headquarters

56 Top Gallant Road
Stamford, CT 06902-7700
U.S.A.
+1 203 964 0096

European Headquarters

Tamesis
The Glanty
Egham
Surrey, TW20 9AW
UNITED KINGDOM
+44 1784 431611

Asia/Pacific Headquarters

Gartner Australasia Pty. Ltd.
Level 9, 141 Walker Street
North Sydney
New South Wales 2060
AUSTRALIA
+61 2 9459 4600

Japan Headquarters

Gartner Japan Ltd.
Aobadai Hills, 6F
7-7, Aobadai, 4-chome
Meguro-ku, Tokyo 153-0042
JAPAN
+81 3 3481 3670

Latin America Headquarters

Gartner do Brazil
Av. das Nações Unidas, 12551
9º andar—World Trade Center
04578-903—São Paulo SP
BRAZIL
+55 11 3443 1509