Preprint for Tim Finin and Li Ding, Search Engines for Semantic Web Knowledge, Proceedings of XTech 2006: Building Web 2.0, Amsterdam, 16-19 May 2006.

Search Engines for Semantic Web Knowledge

Tim Finin and Li Ding

University of Maryland, Baltimore County
Baltimore MD 21250 USA

finin@umbc.edu, dingli1@cs.umbc.edu

Introduction

Web search engines like Google have made people "smarter" by providing ready access to the world's knowledge whenever they need to look up a fact, learn about a topic or evaluate opinions. The W3C's Semantic Web effort aims to make such information more accessible to computer programs by encoding it on the Web in machine understandable form. The Semantic Web languages RDF [BEC04] and OWL [DEA04] are being used to encode knowledge and to build a new layer of services, tools and applications supporting "semantic interoperability" in distributed systems. As the volume of RDF encoded knowledge on the Web grows, software agents will need their own search engines to help them find the relevant and trustworthy knowledge required to carry out their tasks. This paper discusses the general issues underlying the indexing and retrieval of RDF based information and describes Swoogle, a crawler based search engine whose index currently contains information on over a million RDF documents. Swoogle also serves human knowledge engineers by helping them to find Semantic Web ontologies, terms and instance data and to understand how and by whom they are being used. We will present some statistics derived from Swoogle's databases that characterize the current state of the Semantic Web.

Search on the Semantic Web differs from conventional Web search for several reasons. First, Semantic Web content is intended to be published by machines for machines—tools, Web services, software agents, information systems, and so forth. Although Semantic Web annotations and markup can help users find human-readable documents, there will likely be an "agent layer" between human users and Semantic Web search engines.

Second, knowledge encoded in Semantic Web languages such as RDF differs from both the largely unstructured free text found on most Web pages and the highly structured information found in databases. Such semi-structured information requires using a combination of techniques for effective indexing and retrieval. RDF and the Web Ontology Language (OWL) introduce aspects beyond those used in ordinary XML, allowing users to define terms (for example, classes and properties), express relationships among them, and assert constraints and axioms that hold for well-formed data.

Third, a single Semantic Web document can be a mixture of concrete facts, class and property definitions, logical constraints, and metadata. Fully understanding the document can require substantial reasoning, so developers must face the design issue of

how much reasoning search engines can do and when they should do it. This reasoning produces additional facts, constraints, and metadata that may also need to be indexed, potentially along with the supporting justifications or provenance. Conventional search engines do not try to understand document content because the task is just too difficult and requires more research on text understanding.

Finally, the graph structure formed by a collection of Semantic Web documents differs significantly from the structure that emerges from a collection of HTML documents. This difference influences both the development of effective strategies for automatically discovering Semantic Web documents and the establishment of appropriate metrics for ranking their importance.

The Semantic Web Model

The Semantic Web is a framework that allows publishing, sharing, and reusing data and knowledge on the Web and across applications, enterprises, and community boundaries. The W3C's approach is based on the layered set of standards shown in Figure 1. The bottom two layers provide a foundation, using XML for syntax and uniform resource identifiers (URIs) for naming. The middle three layers provide a representation for concepts, properties, and individuals based on the RDF, RDF Schema (RDFS) [BRI04] and OWL. The top-most layers, still under development, extend the semantics to represent inference rules, a logic framework, proofs, and trust.

These languages and concepts can be used in contexts other than as Web documents, including storing information in databases, exchanging data in messages and even describing the contents of networking packets. For our purposes, however, we are interested in the use of RDF to encode information on Web pages, what we will call the *Semantic Web on the Web*. In this view, the Semantic Web consists of Semantic Web documents (SWDs) typically published as Web pages encoded in XML or one of several other encoding languages. Figure 2 shows a simple SWD encoded using the RDF/XML syntax and figure 3 depicts its representation as a graph.

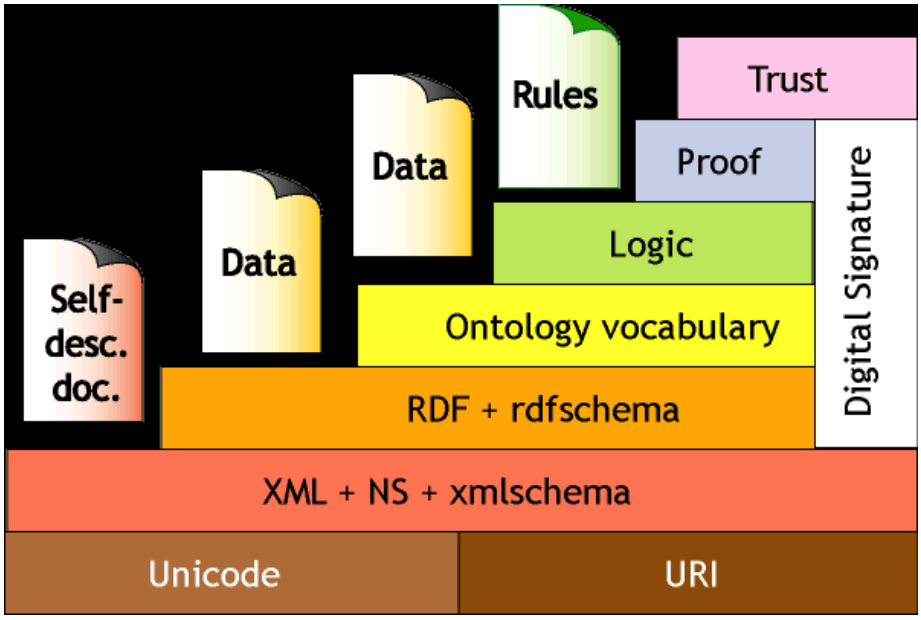

Figure 1: Tim Berners-Lee's layer cake of enabling Semantic Web standards and technologies.

```
1:<?xml version="1.0" encoding="utf-8"?>
2:<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3:          xmlns:owl="http://www.w3.org/2002/07/owl#"
4:          xmlns:foaf="http://xmlns.com/foaf/0.1/" >
5: <foaf:Person>
6:    <foaf:name>Li Ding</foaf:name>
7:  <foaf:mbox rdf:resource="mailto:dingli1@umbc.edu"/>
8:    <owl:sameAs rdf:resource="http://www.csee.umbc.edu/~dingli1/foaf.rdf#dingli"/>
9: </foaf:Person>
10:</rdf:RDF>
```

Figure 2. *An example Semantic Web document written in RDF/XML. The SWD is available at http://ebiquity.umbc.edu/get/a/resource/134.rdf.*

Line 1 declares that this is an XML document. Lines 2-4 further define the content to be an RDF document and provide abbreviations for three common "namespaces" for RDF, OWL, and Friend of a Friend (FOAF), which defines classes and properties for describing people, their common attributes, and relations among them. The SWD's vocabulary consists of literals ('Li Ding' in line 6), URI-based resources (mailto:dingli1@umbc.edu in line 7), and anonymous resources (lines 5-9). Users assert statements using RDF triples such as the one starting at line 5, which has an anonymous resource as the subject, rdf:type as the predicate, and foaf:Person as the object. A higher level of granularity is class-instance, which RDFS's object-oriented ontology constructs offer. Lines 5-9 assert that "there is an instance of a foaf:Person having foaf:name Li Ding, foaf:mbox mailto:dingli1@umbc.edu, and this instance is owl:sameAs, identified by http://www.csee.umbc.edu/~dingli1/foaf.

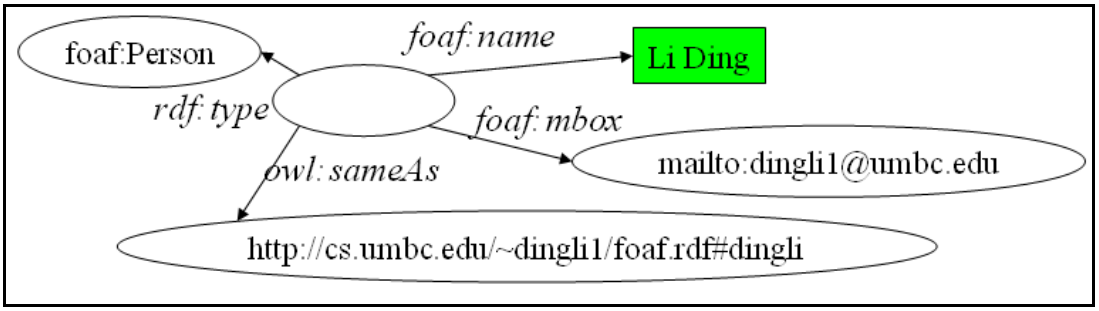

Figure 3: The RDF document in figure 2 has a simple representation as a graph.

A variation on this model is starting to become more popular – embedding Semantic Web content in HTML or XHTML pages. The microformats [KHA06] idea is being used by many as a way to include "semantic" information in HTML links. A more principled approach is being taken by a W3C working group which has developed RDF/A as a set of attributes used to embed RDF in XHTML. Currently this approach is still largely experimental. Microformats are supported by a somewhat informal standardization process and RDF/A has not yet reached the status of a recommended standard.

The Semantic Web on the Web can also be thought of as a collection of loosely federated databases that separates physical Web storage (enforced by online SWDs) from the logical representation (enforced by the RDF graph model). In this view, the Semantic Web represents a large, universal RDF graph whose parts are physically serialized by SWDs distributed across the Web. However, the formal semantics associated with Semantic Web languages support generating new facts from existing ones, while conventional databases only enumerate all facts.

Searching the Semantic Web

Search engines for both the conventional Web and the Semantic Web involve the same set of high-level tasks: discovering and harvesting documents, processing search queries from users and agents, ranking search results, caching and archiving documents, and providing human interfaces and software APIs. Figure 4 shows the high-level architecture of Swoogle. We'll discuss how we've approached each of these in turn for the Swoogle Semantic Web search engine.

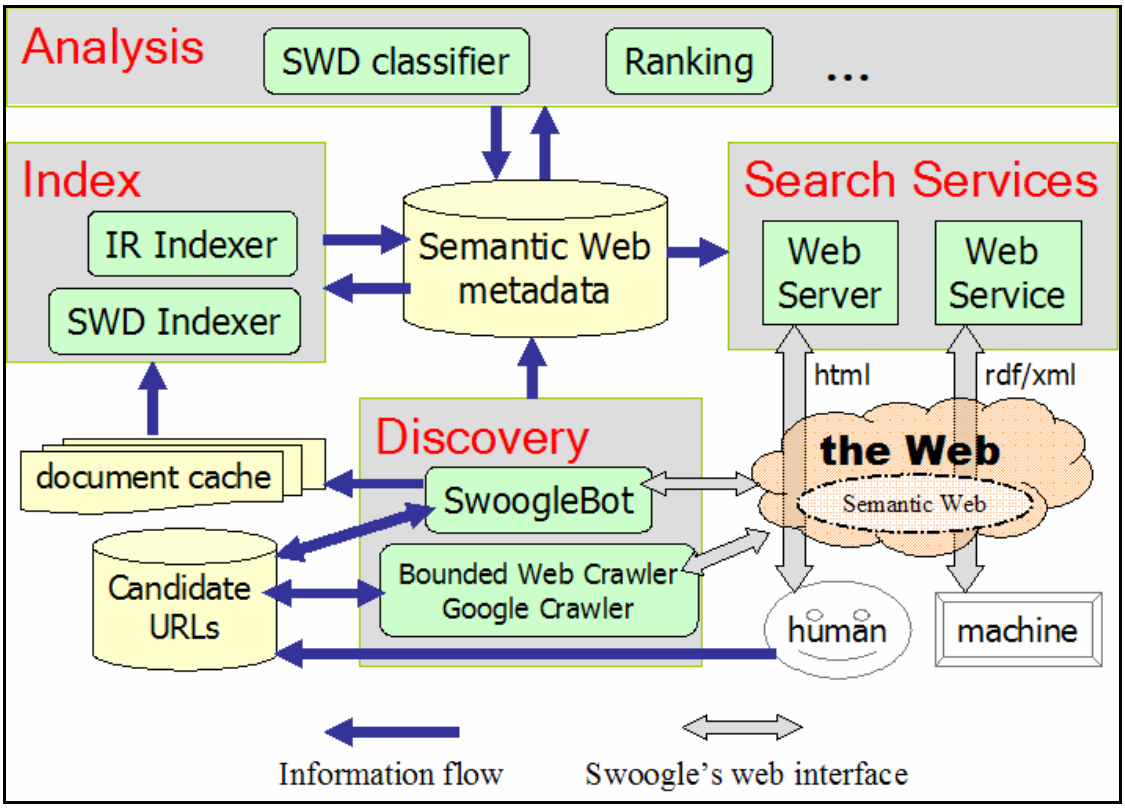

Figure 4: Swoogle's high-level architecture reveals its major components for harvesting, search, archiving and interfacing.

Harvest

Conventional search engines employ crawlers to harvest new Web documents. A typical crawler starts from a set of seed URLs, visits Web documents, and traverses the Web by following the hyperlinks found in the visited documents. The fact that the Web forms a well-connected graph and that people can manually submit new URLs make this an effective process. A Semantic Web crawler must deal with several problems. SWDs are still needles in the Web's haystack, so an exhaustive crawl of the Web is not an efficient approach. Moreover, the global SWD graph is not yet as dense and well-connected as that formed by conventional Web pages. Finally, many of the URLs found in SWDs reference conventional Web resources. Following these links can be computationally expensive, so heuristics to limit and prune candidate links are beneficial.

The URLs of SWDs be collected by manual submission or meta-search on conventional Web search engines. However, these sources usually have partial view of the Semantic Web. Conventional HTML crawling usually generates huge overhead, but it is useful in harvesting SWDs linked by certain hubs. RDF crawlers (also known as scutters) can extract links from the parsed RDF graph, but the link indicators should not be limited to rdfs:seeAlso [BID04]. Finally, using conventional Web search engines like Google to find documents with filetypes suggesting Semantic Web content has problems with both

precision and recall.

Swoogle implements a hybrid approach with several components, including a Google meta-search engine, an RDF crawler and a focused HTML crawler. Manual submission of URLs is used to bootstrap the seeds for Google and bounded HTML crawling. The two crawlers are used to automatically collect the seeding URLs of RDF crawling. The RDF crawler visits and revisits URLs to maintain an up-to-date picture of the Semantic Web, and selectively harvests new seeding URLs for itself using syntactic and semantic parsing results. The harvested SWDs are then used as training data to inductively generate new seeds for Google and HTML crawling.

Search

A search engine's core task is processing queries against the data it has indexed. While queries to Web search engines return documents, the results of a Semantic Web search query can be more or less than a document. As Figure 5 shows, a Semantic Web can aggregate data at several levels of granularity, ranging from the universal graph of all RDF data on the Web to a single RDF triple or even the constituent terms such as a URI. These levels of granularity results in the following frequently encountered search targets:

- *URIs having class/property usage by metadata.* For example, "Find classes which are immediate subclasses of foaf:Person". For Semantic Web content, these terms are analogous to words in natural language. This search helps users to generate Semantic Web data and queries.
- *URLs of SWDs by RDF graph.* For example, find documents that have a foaf:Person instance with a foaf:mbox equal to "mailto:dingli1@umbc.edu" and foaf:name equal to "li ding". This search helps users find documents on the Semantic Web that provide (partial) evidences for a given RDF graph.
- *Search for URLs of SWDs by metadata*. For example, find documents that use the OWL namespace and define properties with local names including 'before' or 'after'. This search shows users' interest in the physical storage of Semantic Web data since an SWD is the basic data transfer packet on the Web and its URL made the data addressable. This level of granularity helps improve efficiency in filtering out huge amounts of irrelevant knowledge. Some documents, such as those representing consensus ontologies, are intended for sharing and reuse. Discovering and using them is essential to achieving the goal of *semantic interoperability*.

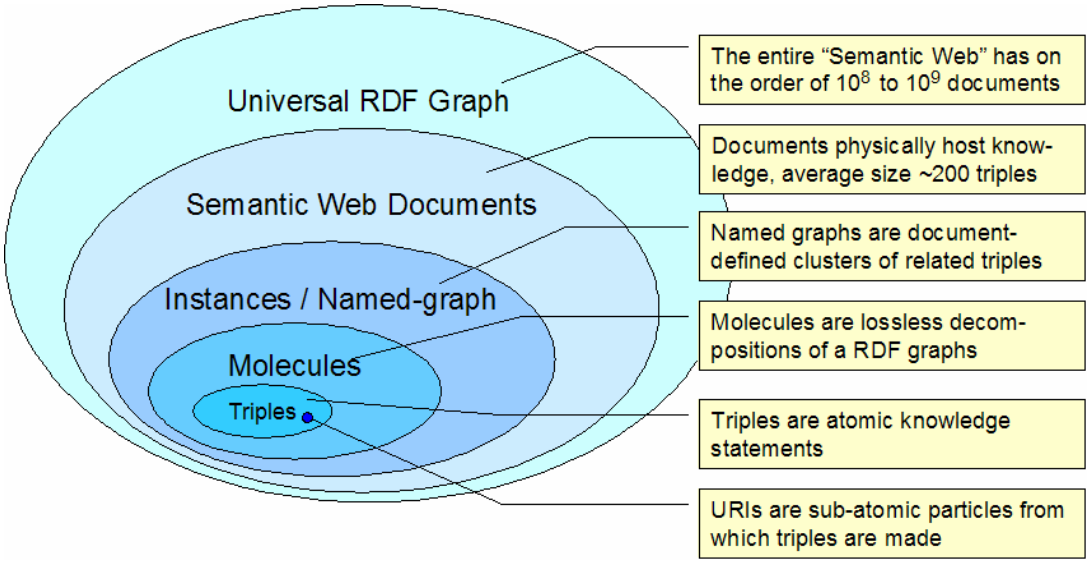

Figure 5: The granularity levels range from the universal graph comprising all RDF data on the Web to individual triples and their constituent resources and literals.

The search targets are essentially references, but the search constraints vary. The second one uses RDF graph as search constraint, and requires an RDF database storing all triples and their provenance. The first and third scenarios are similar to dictionary lookup and Web search, respectively. Using a compact metadata model can avoid the prohibitive space cost for storing all triples.

The annotation metadata of URI includes the namespace and local-name extracted from the term's URI; the literal description of the term from different SWDs. The annotation metadata of SWDs includes metadata about itself (such as document URL and last-modified time) and its content (such as terms being defined or populated and ontology documents being imported). Moreover, Swoogle maintains relational metadata that let users to combine keyword search and surfing to locate search targets.

Rank

Google was the first search engine to order its search results based in part on a Web page's "popularity" as computed from the Web's graph structure. This idea has turned out to be enormously useful in practice and is equally applicable to Semantic Web search engines. However, Google's PageRank [PAG98] algorithm, which is based on the "random surfer model", cannot be directly used in the Semantic Web for several reasons. URIs in a document are not merely hyperlinks but semantic symbols referencing classes, Semantic Web instances, ontology documents, normal Web resources, etc. Semantic Web surfing is not merely random hyperlink-based surfing but *rational surfing* that requires understanding the semantic content of documents.

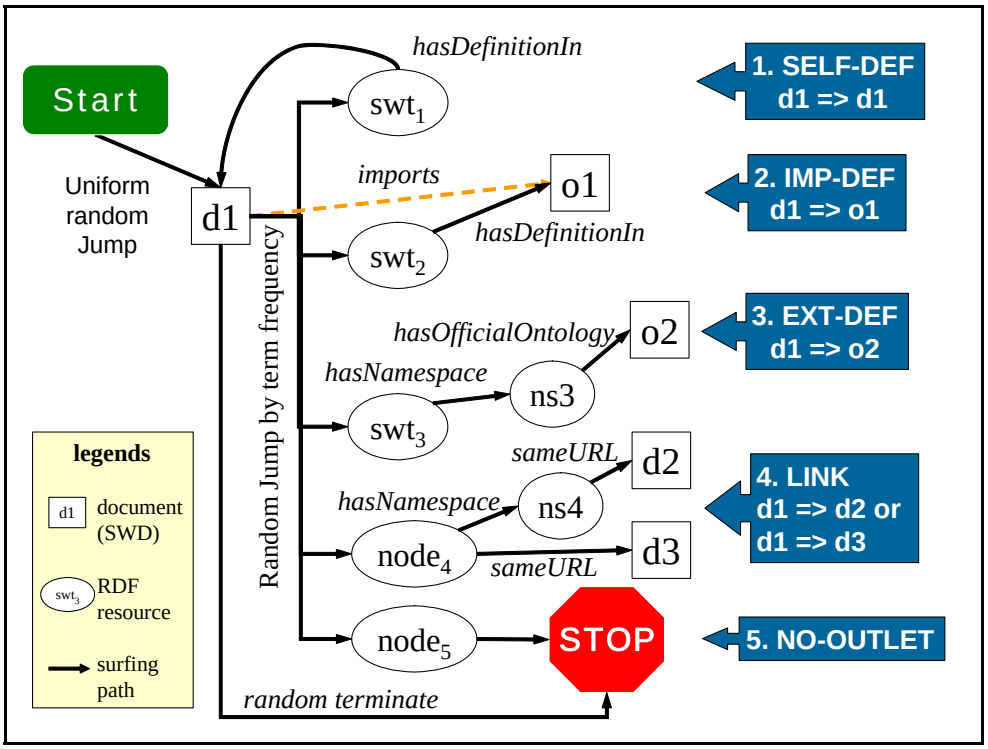

Figure 6: Swoogle's ranking algorithm for is based on a "rational surfer model" that captures how a program might access links in processing Semantic Web documents.

In order to rank the popularity of Semantic Web documents, we adopt the surfing model in which a rational surfer always recursively pursues the definition of classes and properties for complete understanding of a given RDF graph. Figure 6 illustrates the rational surfing behavior of a software agent, which unfolds as follows. The agent jumps randomly to one of the accessible SWDs with uniform probability. It either terminates surfing with constant probability or chooses one RDF node in the RDF graph of the document, and the node is chosen based on its term frequency in the N-Triples version of the document. The agent either surfs to another document or terminates surfing based on the semantics of the chosen node. Paths 1, 2 and 3 represent the agent pursuing a definition. If the node is not anonymous and is used as a class or property usage in the present document, the agent pursues its definition from the present document, the imported ontologies, or the ontology addressed by the namespace parted of the node's URI. Path 4 shows the hyper-link based surfing behavior: if the node is not anonymous and is not used as a class or property, the surfer follows the URL obtained from its URI or namespace to another Semantic Web document. Path 5 includes all cases when no further surfing path starts from the present node, e.g., the present node is literal or anonymous, or the present node's URI links to a normal Web document.

Archive

Like most search engines, Swoogle keeps a cache of the Semantic Web documents it indexes. Swoogle goes beyond this, however, in two ways. First, it also maintains a copy of each documents representation as a set of triples, a more useful form for

programs and agents. Second, and more significantly, Swoogle maintains an archive of all of the current and old versions of each Semantic Web document in its index. The resulting Semantic Web Archive can be used by researchers to study how ontologies evolve, to track the growth of documents containing RDF data or to investigate the natural life cycle of Semantic Web documents.

Applications

To explore what services a Semantic Web search engine can provide and evaluate how well Swoogle provides them, we have used Swoogle to support several applications and use cases. These projects include helping researchers find ontologies and data, semantic search over documents representing proofs, and finding and evaluating semantic associations in large graph databases.

In the NSF-supported SPIRE project [FIN04][PAR06] , a group of biologists and ecologists is exploring how to use the Semantic Web to publish, discover, and reuse models, data, and services. Researchers need to find appropriate ontologies and terms for annotating their data, and they also need resources for discovering data and services others have published.

With Swoogle's *ontology search interface*, users can search for existing ontology documents that define terms in which user-supplied keywords are the substring of their local-name. For example, to find an ontology to use for describing temporal relations, the search might use the keywords before, after and interval. Swoogle's *ontology dictionary* provides definitions of properties or classes for a given set of keywords. It can assemble and merge definitions from multiple sources, list terms sharing the same namespace or the same local-name, and list associations between classes and properties. Those associations can either be "ontological" (for example, the foaf:knows property is defined as existing between instances of foaf:person), or "empirical" (for example, applying the dc:creator property to an instance of foaf:Person). Judging the rank or popularity of terms and ontologies is also relevant. Community consensus models as reflected in ontologies tend to be ranked highly, thus searches use them more often.

Researchers are using Swoogle in conjunction with the Inference Web (IW) [PIN03] which explicitly represents proofs using the PML ontology [PIN04]. One IW component, IWSearch (http://iw4.stanford.edu/iwsearch/IWSearch/), uses Swoogle to discover newly published or updated PML documents on the Web and itself is powered by a specialized instance of Swoogle to index and search instances found in a corpus of more than 50,000 PML documents. Indexing the conclusion part of a proof NodeSet instance can lead to the discovery of additional NodeSets sharing the same conclusion as the one from the given justification tree, thus expanding the justification tree with additional proofs.

SEMDIS, an NSF project jointly conducted with researchers at the University of Georgia is also using Swoogle. This project is automating the discovery, merging, and evaluation of semantic associations in data drawn from a variety of information sources. SEMDIS augments information collected from the Semantic Web with additional data

extracted from text documents and databases [ALE06]. The result, encoded as a large RDF graph along with provenance assertions and trust information, is processed to discover and evaluate "interesting" semantic associations. SEMDIS conducts two kinds of Semantic Web searches: searching for a semantic association (i.e., a connected subgraph) in the large-scale RDF graph, and searching for additional SWDs that (partially) support a given semantic association. The first kind of search finds paths between two nodes in a graph, a common issue in RDF databases. The second is a provenance search to find a set of SWDs that (partially) imply a hypothesized semantic association. Researchers have prototyped this type of search as a RDF molecule-based approach [DIN05b].

State of the Semantic Web

How big is the Semantic Web? Is it widespread or being used by a small number of academic sites? How fast is its use growing? How many ontologies have been published and which are the most popular ones? A Semantic Web search engine like Swoogle can help answer such questions through studies on its collection of documents.

A single, metric such as the number of public RDF documents on the Web is an overly simple measure by which to chart the adoption and evolution of the Semantic Web vision. Nonetheless, it is worth computing, at least as an initial measure. For various reasons, we have prevented Swoogle from trying to find and index every published RDF document. We have, however, developed a methodology to estimate upper and lower bounds for the number of accessible Semantic Web documents using Google queries.

We estimate the lower bound using Google's filetype query feature. Since most Web documents having special filetype extensions such as "rdf" and "owl" are mainly SWDs and the keyword "rdf" is present in almost all SWDs, the Google query

> rdf filetype:rdf OR filetype:owl OR filetype:rss OR filetype:xml OR filetype:n3 OR filetype:nt

returns results that are mostly SWDs. This query, which currently returns 5.9 million results, indicates that at least this many SWDs are available on the Web and known to Google.

The upper bound is hard to estimate for several reasons. First, Google does not index all SWDs. For example, Google has indexed several hundred SWDs from the LiveJournal blogging community while LiveJournal publishes a FOAF document for each of its 10 million users. Second, our simple Google query misses some SWDs indexed by Google. Searching for "*inurl:rss –rdf -filetype:html"* finds many files that use the RSS RDF standard [RSS01]. For a rough upper bound, we use the Google query

> rdf OR inurl:foaf OR inurl:rss -filetype:html

which currently returns about 240 millions results and suggests that there are on the

order of 10^8 to 10^9 SWDs available on the Web.

Swoogle has harvested three millions candidate URLs and confirmed that 1.3 million of these are SWDs as of March 2006. Swoogle considers a document to be a SWD if it can be successfully parsed by Jena [MCB02] and produces triples. This number is less than the lower bound because Swoogle has limited access to Google's index and we have intentionally limited the number of documents collected from LiveJournal and several other sites with a large number of SWDs. We consider Swoogle's current collection, which is continually growing, to be the largest and least biased collection of Semantic Web Documents available. The following statistics are based on our analysis on these 1.3 million SWDs.

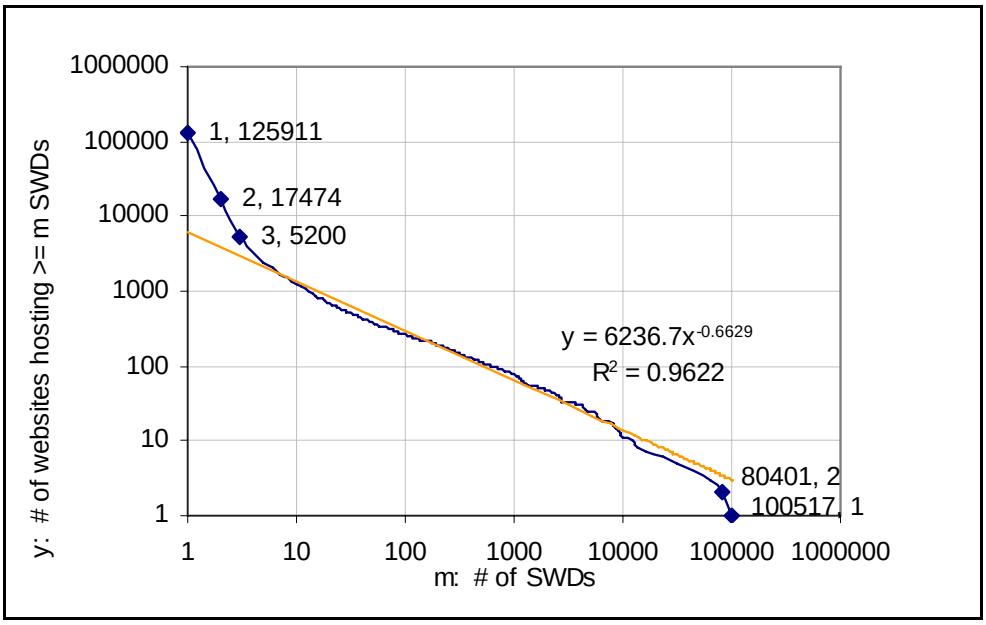

Figure 7: The cumulative distribution of the number of Semantic Web documents across Web sites follows the power law.

The 1.3 million SWDs are distributed among 125,911 websites, where a website is identified by a triple of (protocol, domain name, port number). Figure 7 shows the cumulative distribution of the number of SWDs per Web site, which generally follows power law distribution. The tail of the curve has a sharp drop when approaching 100,000 in x-axis because our crawling strategy prefers harvesting the first 10,000 SWDs from each website. The head of the curve also has a sharp drop due to virtual hosts: some content publishing websites automatically offer users a unique virtual host name under its domain.

Figure 8 shows the ten Web sites hosting the largest number of pure SWDs, i.e., RDF documents as opposed to those with embedded RDF content. The *unpinged urls* have not yet been accessed and categorized. Each of these websites is specialized in publishing one type of SWDs, such as personal profiles (e.g., FOAF documents), personal blog RSS feed documents, portable proofs (e.g., PML documents) and publication information.

Website	***# pure SWDs***	***# unpinged URLs***	***Content type***
www.livejournal.com	100,518	79,331	foaf
www.tribe.net	80,402	25,151	foaf
www.greatestjournal.com	62,453	835	foaf
onto.stanford.edu	45,278	206	pml
blog.livedoor.jp	31,741	6,733	foaf
www.ecademy.com	23,242	3,281	foaf
www.hackcraft.net	16,238	0	dc, book
www.uklug.co.uk	13,263	2	rss
users.livejournal.com	12,783	40,211	foaf
ch.kitaguni.tv	11,931	3,010	rss

Figure 8: The ten internet domains with the largest number of Semantic Web Documents.

We further count domain names, semantic web ontologies and pure semantic web documents for each top-level domain. Figure 9 shows that the .com domain has the largest contribution to Semantic Web data and Semantic Web websites while the .edu domain has the largest contribution to Semantic Web ontologies. The number of SWDs we consider to be ontologies (SWOs) is 28,564 when we filter out PML documents which contains ontological definition but are not intended to be ontologies. Swoogle considers an SWD to be an ontology if the number of its triples contributing to term definitions exceeds a threshold value.

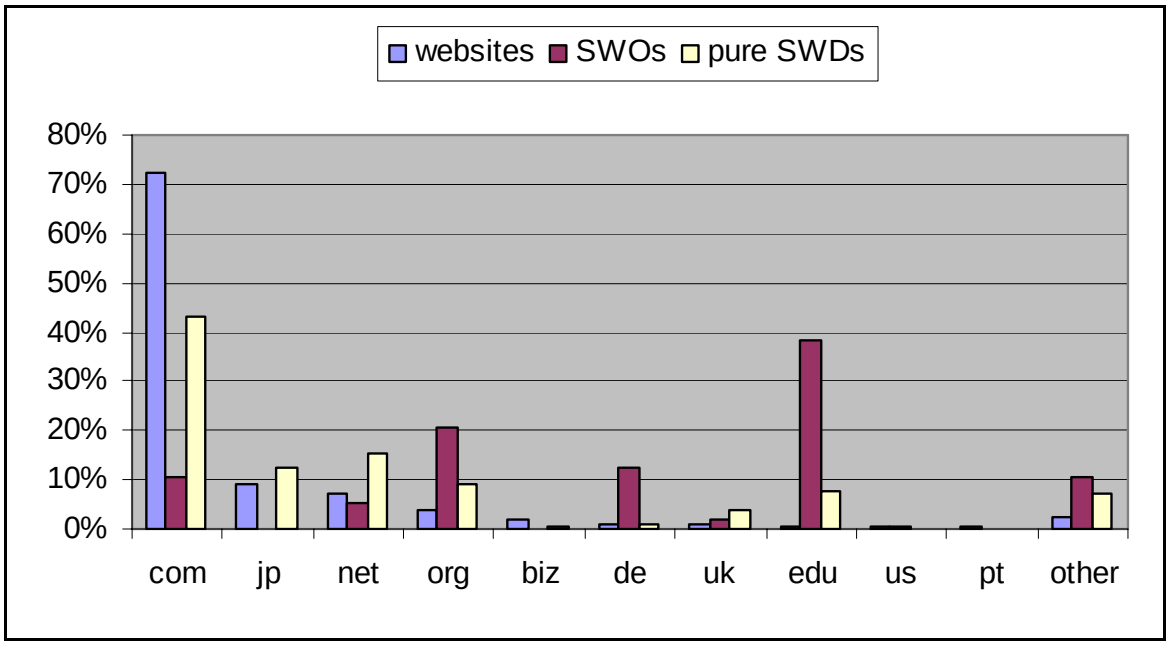

Figure 9: The distribution of Semantic Web ontologies (SWOs) and documents (SWDs) differs across the various Internet top level domains.

The size of an SWD is usually computed using the number of triples in the SWD's RDF graph. The mean size of an SWD in Swoogle's collection is 181 triples. Since many SWDs are generated by software under certain structure, some sizes are frequently observed among SWDs, for example, many PML documents have exactly 28 or 36 triples and many RSS documents have exactly 130 triples. Further investigation shows that the size of embedded SWDs are usually quite small -- 69% have exactly three triples and 96% have no more than ten triples. The size of pure SWDs varies considerably, with 60% having between five and 1000 triples.

The age of an SWD is measured by the difference between the current time and the last-modified time of the SWD. Figure 10 plots the cumulative distribution of the number of pure SWDs and SWO having been last modified before the date in X-axis. The plot excludes SWDs which does not have last-modified time specified in HTTP header and the 100K SWDs that have gone offline before March 2006.

The 'pswd' curve exhibits exponential growth; intuitively, the growth implies that either the many new PSWDs have been added to the Semantic Web and/or many old PSWDs have been updated recently. This statistics supports a promising hypothesis that the semantic web is growing rapidly. The 'swo' curve is plotted after filtered PML documents. Swoogle's data shows that the growth in the number of ontologies continues but appears to be slowing. This can be explained by an increase in the reuse of existing ontologies as the Semantic Web matures – a good sign.

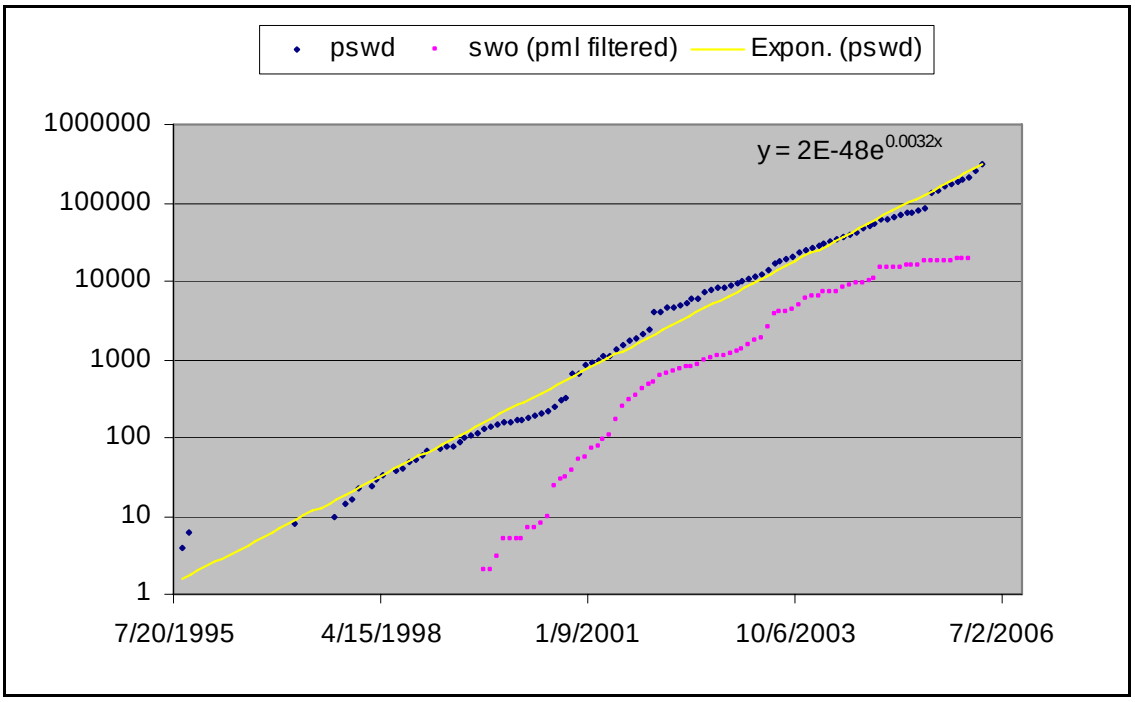

Figure 10: The distribution of the number of pure SWDs and SWOs having been last modified before the given date.

The 1.3 million SWDs contribution 237,645,189 triples (we simply sum up triples from each SWDs without merging equivalent ones), and 1,415,054 distinct Semantic Web terms (which has class/property usage in SWDs) using 11,648 namespaces.

Figure 11 shows the usage pattern of Semantic Web terms (SWTs) and introduces some interesting observations. Most SWTs (95%) are defined or referenced without being actually populated, while some SWTs (1.8%) are populated without being defined. Some SWTs (0.1%) are mistakenly used as both a class and as a property. A significant number (5.3%) are used even though they have not been defined. While some of these are undefined because they are from XMLSchma, most are due to typographic errors, misspellings, inaccessible defining documents, or other problems.
A common question posed by Semantic Web knowledge consumers is what kind of knowledge is available. One way to answer this question is by analyzing the instances found on the Semantic Web, i.e., how SWTs, classes and properties, are used to create instances and make assertions about them. We have examined the cumulative distribution of the number of SWTs associated with at least *m* instances/SWD. And found that the graphs for both classes and properties follow a power law distribution. Swoogle's collection also shows that relatively few SWTs have been used to define large amounts of data. For example, 370 classes have been used to create instances in more than 100 SWDs and 1700 classes have more than 100 instances. The same is true for properties with about 1200 properties used to assert values for instances in than 100 SWDs and about 4600 properties used in more than 100 assertions.

SWT usage pattern	***# swd***	***comments***
defined as class but not populated	1,001,571	ontology
defined as property but not populated	91,238	ontology
referenced as class only	59,289	mistakes
defined and populated as class	19,000	good practice
populated property without definition	14,266	bad practice
class defined w/o description or instances	12,929	bad practice
defined and populated as property	12,326	good practice
property defined w/o description or instances	11,291	bad practice
populated class without definition	7,761	bad practice
referenced as property only	5,672	mistakes
property defined & populated w/o description	1,940	ok practice
class defined & populated w/o description	711	ok practice
property usage w/o explicit definition	67	bad practice
class usage w/o explicit definition	449	bad practice
used/defined/referenced as class & property	1159	mistakes

Figure 11: Swoogle's collection reveals some interesting observations about how Semantic Web Terms (SWTs) are used and abused.

The ten terms most often associated with instances are shown in Figures 12 and 13, ordered by the number of documents and instances, respectively. The first number of SWDs indicates the class's popularity and the number of instance indicates the richness of instance space. Although the number of an SWT's class instances is often proportional to the number of SWDs populating the SWT, exceptions exist, for example, the WordNet Noun class has over 2 million instances but these come from just 36 large documents. Similar observations can be made about the use of properties. Swoogle's collection demonstrates that a small number of ontologies (e.g., FOAF, DC, RSS) and schema files (e.g., RDF, RDFS, OWL) dominate current Semantic Web vocabulary. Some database dumps from several authorities, such as WorldNet, NIH, CYC, IEEE, also contribute significant amount of Semantic Web instance data using giant instance document. For example, tag:govshare.info,2005:rdf/vote/option contributed 1,965,182 property instances in 14 SWDs.

Our initial studies of Swoogle's collection leads us to believe that the Semantic Web has already reached a significant size, as measured by the total number of documents and the number of sites over which they are distributed.

resource URI	*SWDs*	*instances*
http://xmlns.com/foaf/0.1/Person	467,806	11,040,981
http://www.w3.org/1999/02/22-rdf-syntax-ns#Seq	267,603	277,608
http://purl.org/rss/1.0/channel	259,417	265,700
http://purl.org/rss/1.0/item	241,984	3,971,918
http://xmlns.com/foaf/0.1/Document	220,064	242,994
http://xmlns.com/foaf/0.1/PersonalProfileDocument	178,946	178,975
http://www.w3.org/2003/01/geo/wgs84pos#Point	85,695	107,859
http://www.w3.org/2002/07/owl#Class	62,867	1,075,220
http://www.w3.org/1999/02/22-rdf-syntax-ns#Property	57,561	503,829
http://www.w3.org/1999/02/22-rdf-syntax-ns#List	53,726	54,491

Figure 12: The top ten SWTs ranked by the number of documents containing instances.

resource URI	**SWDs**	**instances**
http://xmlns.com/foaf/0.1/Person	467,806	11,040,981
http://purl.org/rss/1.0/item	241,984	3,971,918
http://www.cogsci.princeton.edu/~wn/schema/Noun	36	2,376,900
http://xmlns.com/wordnet/1.6/Person	2,823	1,138,374
http://xmlns.com/foaf/0.1/chatEvent	2,693	1,138,182
http://www.w3.org/2002/07/owl#Class	62,867	1,075,220
http://www.nlm.nih.gov/mesh/2004#Concept	18	734,706
http://www.daml.org/2002/02/telephone/1/areacodes-ont#Exchange	768	614,400
http://www.w3.org/1999/02/22-rdf-syntax-ns#Property	57,561	503,829
http://www.cogsci.princeton.edu/~wn/schema/Verb	36	436,572

Figure 13: The top ten SWTs ranked by the number of instances.

Conclusions

As the Web has grown in size, search engines have become a critical component of its infrastructure, and there is an increasing need for search engines that can efficiently handle Semantic Web content. While we can't be sure what form this content will take in the future, the current standard is based on Semantic Web documents. We are continuing to use Swoogle to study the growth and characteristics of the Semantic Web and the use of RDF and OWL. We are also developing new features and capabilities and exploring how it can be used in novel applications. Many open issues remain.

One set of open problems involves scale. Techniques that work today with 5×10^6 documents may fail when the Semantic Web has 5×10^8 documents. Extending Swoogle to index and effectively query large amounts of instance data remains a challenge. Some of these problems could potentially be solved by moving away from the open source software we are using and creating custom-designed index stores and distributed systems—analogous to what Google has done for conventional Web searches. It remains to be seen, however, if that alone would suffice.

We also need to experiment with how much and where a Semantic Web search engine should reason over the contents of documents and queries. In previous work [FIN05] we experimented with expanding documents using reasoning prior to indexing. A complementary approach is to expand queries containing RDF terms [VOO94]. This is related in part to the problem of scale—the larger the collection becomes, the less efficient it is to reason over it.

Other issues involve trust and the use of local knowledge that is not part of the Semantic Web. Information encoded in RDF is now being embedded in other documents, such as PDF and XHTML documents, JPG images, and Excel spreadsheets. When techniques for such embedding become standard, we expect the growth of Semantic Web content on the Web to accelerate dramatically. This will add a new requirement for hybrid information retrieval systems that can index documents based on words as well as RDF content.

Our experience with Swoogle has given us a chance to see the growth of the Semantic Web on the Web over the past two years. The number of RDF documents has grown steadily while the number of underlying ontologies has grown more slowly, as might be expected. While the numbers are still much less than the number of conventional pages, the growth we observe makes us optimistic that the Semantic Web has a strong future.

Bibliography

[ADI06] B. Adida and M. Birbeck (eds.), RDF/A Primer 1.0 -- Embedding RDF in XHTML, W3C Working Draft, 10 March 2006. http://www.w3.org/TR/xhtml-rdfa-primer/

[ALE03] B. Aleman-Meza, C. Halaschek, I. Arpinar and A. Sheth, “Context-Aware Semantic Association Ranking,” Proceedings of the First International Workshop on Semantic Web and Databases, September 2003.

[ALE06] B. Aleman-Meza, M. Nagarajan, C. Ramakrishnan, A. Sheth, B. Arpinar, L. Ding, P. Kolari, A. Joshi, and T. Finin, “Semantic Analytics on Social Networks: Experiences in Addressing the Problem of Conflict of Interest Detection", WWW 2006, Edinburgh, Scotland, May 2006.

[BEC04] D. Beckett, RDF/XML Syntax Specification (Revised), Feb. 2004; http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/.

[BER01] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” Scientific American, May 2001, pp. 35-43.

[BID04] M. Biddulph, Crawling the Semantic Web, Proceedings of XML Europe, April 2004.

[BRI04] D. Brickley and R.V. Guha, RDF Vocabulary Description Language 1.0: RDF Schema, Feb. 2004; http://www.w3.org/TR/rdf-schema/.

[CAI04] M. Cai and M. Frank, “RDFPeers: A Scalable Distributed RDF Repository Based on a Structured Peer-to-Peer Network,” Proceedings of the 13th International Conference on the World Wide Web, pp. 650-657, 2004.

[CAR04] J.J. Carroll, C. Bizer, P. Hayes and P. Stickler, Named Graphs, Provenance, and Trust, Proceedings of the 14th international conference on World Wide Web, pp 613-622, May 2005.

[DEA04] M. Dean and G. Schreiber, Web Ontology Language Reference, Feb. 2004; http://www.w3.org/TR/2004/REC-owl-ref-20040210/.

[DIN04] L. Ding et al., “Swoogle: A Search and Metadata Engine for the Semantic Web,” Proc. 13th ACM Conference on Information and Knowledge Management, ACM Press, 2004.

[DIN05a] L. Ding, T. Finin, A. Joshi, Y. Peng, R. Pan and P. Reddivari, Search on the Semantic Web, IEEE Computer, volume 10, number 38, pp 62-69, October 2005.

[DIN05b] L. Ding, T. Finin, A. Joshi, Y. Peng, P. da Silva, and D. McGuinness, Tracking RDF Graph Provenance using RDF Molecules, (poster paper), Proceedings of the 4th International Semantic Web Conference, November 2005.

[DIN05c] L. Ding, R. Pan, T. Finin, A. Joshi, Y. Peng and P. Kolari, Finding and Ranking Knowledge on the Semantic Web, Proceedings of the 4th International Semantic Web Conference, November 2005.

[DIN06] L. Ding, Enhancing Semantic Web Data Access, Ph.D. Dissertation, University of Maryland, Baltimore County, May 2006.

[EBE02] A. Eberhart, Survey of RDF Data on the Web, Proceedings of the 6th World Multiconference on Systemics, Cybernetics and Informatics, July 2002.

[FIN04] T. Finin and J. Sachs, “Will the Semantic Web Change Science?” Science Next Wave, Sept. 2004; http://nextwave.sciencemag.org.

[FIN05] T. Finin, J. Mayfield, A. Joshi, R. Cost and C. Fink, "Information Retrieval and the Semantic Web," Proceedings of the 38th International Conference on System Sciences, January 2005.

[GUO04] Y. Guo, Z. Pan, and J. Heflin, "An Evaluation of Knowledge Base Systems for Large OWL Datasets, Proceedings of the Third International Semantic Web Conference, pp. 274-288, 2004.

[KHA06] R. Kahare, Microformats: The Next (Small) Thing on the Semantic Web? IEEE Internet Computing, volume 10 issue 1, pp 68-75, January 2006.

[KLY04] G. Klyne and J.J. Carroll, Resource Description Framework (RDF): Concepts and Abstract Syntax, Feb. 2004; http://www.w3.org/TR/rdf-concepts/.

[MCB02] Brian McBride, Jena: A Semantic Web Toolkit, IEEE Internet Computing, pp. 55-59, volume 6, issue 6, November 2002

[PAG98] L. Page, S. Brin, R. Motwani and T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, technical report, Stanford Digital Library Technologies Project, Stanford University, 1998.

[PAR06] C. Parr, A. Parafiynyk, J. Sachs, L. Ding, S. Dornbush, T. Finin, T. Wang and A. Hollender, Integrating Ecoinformatics Resources on the Semantic Web, poster paper, Proceedings of the 15th International World Wide Web Conference, May 2006.

[PIN03] P. Pinheiro da Silva, D. McGuinness and R. McCool, "Knowledge Provenance Infrastructure," Data Eng. Bulletin, vol. 26, no. 4, 2003, pp. 26-32.

[PIN04] P. Pinheiro da Silva, D. McGuinness and R. Fikes, A Proof Markup Language for Semantic Web Services, technical report KSL04-01, Knowledge Systems Laboratory, Stanford University, 2004.

[RSS01] RDF Site Summary (RSS) 1.0, http://web.resource.org/rss/1.0/spec

[SHE05] A. Sheth et al., "Semantic Association Identification and Knowledge Discovery for National Security Applications," Journal of Database Management on Database Technology, volume 16, number 1, 2005.

[VOO94] E. Voorhees, "Query Expansion Using Lexical-Semantic Relations," Proceedings of the 17th International Conference Research and Development in Information Retrieval, 1994.