Characterizing the Semantic Web on the Web*

Li Ding[1] and Tim Finin[2]

[1] Knowledge Systems Laboratory
Stanford University, Stanford CA 94305
ding@ksl.stanford.edu

[2] Computer Science and Electrical Engineering
University of Maryland, Baltimore County, Baltimore MD 21250
finin@umbc.edu

Abstract. Semantic Web languages are being used to represent, encode and exchange *semantic* data in many contexts beyond the Web – in databases, multiagent systems, mobile computing, and ad hoc networking environments. The core paradigm, however, remains what we call the *Web aspect* of the Semantic Web – its use by independent and distributed agents who publish and consume data on the World Wide Web. To better understand this central use case, we have harvested and analyzed a collection of Semantic Web documents from an estimated ten million available on the Web. Using a corpus of more than 1.7 million documents comprising over 300 million RDF triples, we describe a number of global metrics, properties and usage patterns. Most of the metrics, such as the size of Semantic Web documents and the use frequency of Semantic Web terms, were found to follow a power law distribution.

1 Introduction

Unpacking the phrase *Semantic Web* immediately produces its two constituent concepts: it is (i) a semantic framework to represent the meaning of data that is (ii) designed for use on the Web. Most current research, both basic and applied, has focused on the first of these and largely ignored the second. An obvious lesson from the last ten years of Web-based developments is we must not underestimate the impact of the (still emerging) Web on technology and society.

Reviewing recent papers in journals and conferences one finds many on all aspects of RDF and OWL as knowledge representation languages – complexity, scalability, completeness, efficient reasoning algorithms, integration with databases, rule extensions, expressing uncertainty, human friendly encodings, etc. Developing systems and tools that use these languages for ontology engineering, visualization, manual markup, etc. is also a popular topic. Finally, application papers typically center on using RDF based representations to express the knowledge and data needed for particular problem domains, such as workflow models, action descriptions, healthcare records, policy enforcement, or user preferences. For the most part, this work touches little on issues that stem for the (initial) intended use of Semantic Web languages for publishing and using ontologies and data on the World Wide Web.

* Partial support was provided by NSF awards ITR-IIS-0326460 and ITR-IDM-0219649.

A great deal of practical work has been done, of course, on developing Web appropriate standards for the Semantic Web and harmonizing them with existing Web standards and practices. Many applications and testbeds have also focused on core Web paradigms, such as semantically enhanced Web services and policy-driven negotiation for Web resource access. Our claim is that we need more research on modeling and understanding how Semantic Web concepts and technology is and can be used on the Web. In this respect, we stand on the shoulders of those who call for *"Creating a Science of the Web"* [1].

There are also many useful and important applications of Semantic Web languages and systems that do not involve the Web. RDF and OWL are used in agent communication languages [2], instant messaging [3], and in GIS systems [4], to name just a few. We believe that the *Web aspect* of the Semantic Web remains as the common, unifying vision, one in which millions of people, agents and applications publish and consume knowledge and data using the evolving Web standards and protocols [5].

In this paper, we focus on characterizing the Semantic Web on the Web, i.e., as a collection of loosely federated knowledge bases that are semantically encoded in Semantic Web languages but are physically published and consumed on the Web by independent agents. Our work consists of three parts:

- **designing a conceptual model.** Instead of using the current model of the Semantic Web, i.e., one universal RDF graph, our new model covers both structure (RDF graphs) and provenance (Web documents and associated agents).
- **creating a global catalog.** A global catalog of online Semantic Web data has long been desired but missing; therefore, we have developed effective harvesting methods and have accumulated a significant dataset.
- **measuring data.** Using our conceptual model, we measure the collected dataset to derive interesting global statistics and implications.

Related work. While some research has tried to characterize the reach and patterns of use of the Semantic Web on the Web, they have not attempted to be systematic and have used limited datasets.

Harvesting and simple summary. A number of simple systems have been designed to find and collect RDF documents on the web, including Eberhart's RDF crawler [6], OntoKhoj [7], the DAML Crawler [8]. Several repositories for Semantic Web documents have been created and maintained using a combination of manual and automatic techniques. These include the DAML Ontology Library [9] which collected a modest number of Semantic Web documents (at most 22,000) with a limited summary of document properties such as parse error types, document size, documents per website, and namespace building usage. Additional relevant work can be found in Web characterization literature [10, 11] which studies global distributions of document properties such as the average size of web documents.

Characterizing the universal RDF graph. Gil et al. [12] analyzed the structure of a RDF graph that results from merging nearly 200 documents from the DAML Ontology Library. The dataset is too limited to be a representative of the entire Semantic Web and even the subset of ontologies on the Semantic Web.

Rating Semantic Web ontologies. Several studies have tried to measure the quality of Semantic Web ontologies, i.e. Semantic Web documents that define or contribute

to the definition of classes and properties. Most [13, 14] employ content analysis on ontologies with various foci, such as building a comprehensive evaluation framework [15], qualifying concept consistency [16, 17], quantifying the graph structure of class and property taxonomy hierarchy [18–21], and measuring the structure and the instance space of a given ontology [22]. These studies have been limited in two ways. First, they have only analyzed ontologies, which we estimate account for only about 1% of the RDF documents on the Web. Second, the empirical evaluations are based on very small datasets, typically of fewer than 30 documents.

Characterizing social networks in FOAF. One of the most successful application of RDF is the use of the FOAF ontology to encode social networks. Several studies [23–26] have analyzed large amounts of FOAF data, typically by collecting FOAF documents via specialized crawlers and then making statistical measurements on vocabulary usage and network structure. Although the evaluation datasets are large, their sources and vocabularies are limited. Most FOAF documents are obtained from a few portal websites such the *www.livejournal.com* blogging system.

Contributions. Our work is a systematic study of the semantic aspect and the web aspect of the Semantic Web. It is highlighted by contributing a new conceptual model of the Semantic Web on the Web, harvesting a significant dataset that is much larger and more diverse than other existing work, and inheriting and introducing wide spectrum of measurements for global properties on both semantic structure and knowledge provenance of the Semantic Web.

In section two of this paper we explain our conceptualization of the Semantic Web on the Web. Section three briefly illustrates our harvesting methods and evaluates the significance of harvest result. Sections four and five elaborate our metrics and findings about the global properties of the Semantic Web and section six offers some concluding remarks. In this paper, we assume, for simplicity's sake, that the following namespaces are defined: *rdf* for RDF, *rdfs* for RDF schema, *owl* for OWL, *foaf* for FOAF, *dc* for Dublin Core Element and *wn* for WordNet.

2 The Conceptual Model of the Semantic Web on the Web

The foundation for our Semantic Web characterization is the Web Of Belief Ontology which captures not only the semantic structure of RDF graph but also its provenance in terms of the Web and the agent world. This paper only covers the essential notions from the model and readers are invited to see [27] for details.

A **Semantic Web document** (SWD) is an atomic Semantic Web "data transfer packet" on the Web. It is both a Web page addressable by a URL and an RDF graph containing Semantic Web data. It can be a static or dynamic web page, for example one generated by a database query. In particular, SWDs can be divided into *pure SWDs* (PSWDs), which are completely written in Semantic Web languages, and *embedded SWDs* (ESWDs), which embed RDF graphs in their text content, e.g., HTML documents containing Creative Commons license metadata.

The *URI reference* (URIref) of an *rdfs:Resource* conveys dual semantics: (i) a unique identifier for the resource, and (ii) the Web address of the SWD defining the resource. URIrefs are widely used to merge RDF graphs distributed on the Semantic Web. A

resource's semantics depends on its usage in an RDF graph. In particular, we are interested in **Semantic Web terms** (SWTs), i.e., named resources that have **meta-usages** (being used as classes or properties) in SWDs. Six types of use are defined below and illustrated in Figure 1. For a given RDF graph, a resource X is:

- **defined as a class (DEF-C)** if there exists a triple of the form *(X, rdf:type, C)* where C is *rdfs:subClassOf rdfs:Class*. For example, *foaf:Person* is defined as a class in triple *t3*.
- **defined as a property (DEF-P)** if there exists a triple *(X, rdf:type, P)* where P is *rdfs:subClassOf rdf:Property*. For example, *foaf:mbox* is defined as a property in triple *t1*.
- **populated (or instantiated) as a class (POP-C)** if there exists a triple *(_a , rdf:type , X)* where $\_a$ can be any resource. For example, *rdfs:Class* has been populated as a class in triple *t3*.
- **populated (or instantiated) as a property (POP-P)** if there exists a triple *(_a , X , _b)* where $\_a$ and $\_b$ can be any resource (or literal). For example, *rdf:type* has been populated as a property in triple *t3*.
- **referenced as a class (REF-C)** if X is of type *rdfs:Class* according to the ontology constructs from Semantic Web languages except *rdf:type*. For example, *foaf:Person* is referenced as a class in triple *t2*.
- **referenced as a property (REF-P)** if X is of type *rdf:Property* according to ontology constructs from Semantic Web languages except *rdf:type*. For example, *foaf:mbox* is referenced as a property in triple *t2*.

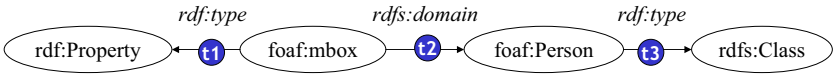

Fig. 1. This RDF graph adapted from the FOAF ontology illustrates some of the relations defined in the *Web of Belief* ontology.

Note that we may find multiple types of meta-usage of a URI in different SWDs, including some rare and undesired cases: the SWT *rdfs:subClassOf* is defined as a property by the RDFS ontology and also as a class by another SWD[3].

Two additional concepts are used studying ontologies. **Semantic Web Ontology** (SWO) is a sub-class of Semantic Web document and physically groups definitions of SWTs. An SWO is identified by containing (i) DEF-C, DEF-P, RDF-C, REF-P meta-usages or (ii) instances of *owl:Ontology*[4]. **Semantic Web Namespace** (SWN) is a sub-class of *rdfs:Resource* and logically groups SWTs and enables distributed definition (i.e., users can define the SWTs using the same SWN in different SWOs). An SWN is identified as the namespace part of an SWT.

[3] http://ilrt.org/discovery/2001/09/rdf-schema-tests/rdf-schema.rdfs

[4] The Swoogle system has experimented with different heuristics for identifying a SWD as an SWO and is currently using this very liberal one.

3 Creating a Global Catalog

In order to build a global catalog of the Semantic Web on the Web, we need to harvest publicly accessible SWDs. There are two primary difficulties: (i) SWDs are sparsely distributed on the Web and found on sites in varying density, e.g. *www.cnn.com* hosts no SWDs but *www.liverjournal.com* has millions; and (ii) Confirming that a document contains RDF content requires RDF parsing which entails high cost when done for millions of documents.

3.1 Estimating the number of online SWDs

The scale and complexity of harvesting task is dominated by the number of online SWDs, which we have estimated using the Google search engine [5] Since Google does not index all SWDs and its estimated total result is coarse, we use it to derived an *order of magnitude* estimate of the total number of online SWDs.

In theory, the search query *"rdf"* would suffice because the RDF namespace is declared by virtually all SWDs. In pracice, however, this simple Google query has two problems. First, it does not cover all indexed SWDs. For example, many RSS 1.0 files, which are RDF documents, are not matched by it. Second, it matches many documents that are not SWDs. For example the query "rdf filetype:html" identifies more than 38 million HTML documents. Based on queries run on 12 May 2006, we estimate that there are between 10^7 and 10^9 Semantic Web documents online.

- For a conservative estimate we emphasize precision and use a query where most results will be SWDs. The query "*rdf filetype:rdf*" produced 4.91M estimated matches. The constraint "filetype:rdf" was chosen because it is the most common file extension used among SWDs, and more than 75% web documents using it are SWDs [6]. This yields a conservative estimate of 10^7 SWDs.
- For an optimistic estimate we emphasize recall using a query whose results will include most online SWDs. The query "*rdf OR inurl:rss OR inurl:foaf -filetype:html*" produces about 205M results. This derives an optimistic estimate of 10^9 SWDs.

3.2 A Hybrid Semantic Web Harvesting Framework

Most existing harvesting methods are limited in significance or diversity. Conventional Web crawling approaches [6, 7] are inefficient because most hyperlinks in Web documents (including SWDs) point to conventional Web documents. Similarly, brute-force sampling, i.e., testing port 80 of reachable IP addresses [11], introduces prohibitive cost in validating millions of web documents. Meta-search based approaches [28] are limited by the inability to filter out conventional web documents from search engine results and the fact that some search engines intentionally ignore SWDs. Manual submission based approaches, such as that used for the DAML ontology library [9] and SchemaWeb [29]

[5] We have found the Google and Yahoo search engines to have the most RDF documents indexed, with Google having more than twice as many as Yahoo.

[6] Other constraints usually returns fewer results, e.g. "owl filetype:owl" returns 55K results.

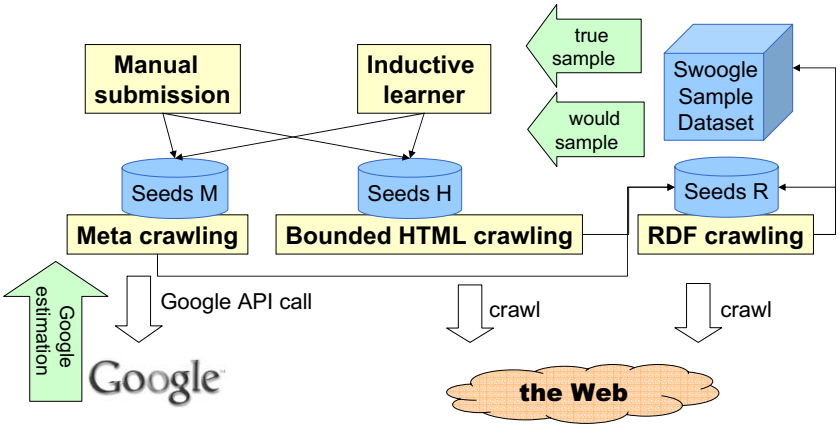

Fig. 2. The Swoogle system uses an adaptive Semantic Web harvesting framework with three different kinds of crawlers.

scale poorly and are difficult to maintain. RDF crawlers (also known as scutters[7] or Semantic Web crawlers) [30, 31] are limited because the seeding URLs (i.e., the starting points of crawling) are hard to obtain and surfing heuristics (i.e., patterns for selecting hyperlinks to SWDs) are often biased.

In order to effectively harvest as many as possible SWDs on the Web with minimum cost, we developed a automatic, hybrid Semantic Web harvesting framework [27] that integrates several harvesting methods. Figure 2 illustrates its work-flow, which has the following major steps.

1. **Bootstrapping**. Manual submissions are used to bootstrap the harvesting, providing seeds for Google-based meta-crawling and bounded HTML crawling.
2. **Google-based Meta-crawling.** *Meta crawling* [32] involves directly harvesting URLs from search engines without crawling the Web. Google is chosen because it indexes the largest number of Web documents and offers richer query constraints than others. We collect seeds from manual bootstrapping input and the *inductive learner* that selects "good" seeds from the harvested *Swoogle sample dataset*. A "good" seed is a Google query whose results contain high percentage of SWDs, e.g., most URLs returned by the query *rdf filetype:rdf* are indeed SWDs.
3. **Bounded HTML crawling.** *HTML crawling* (i.e., conventional Web crawling) harvests web documents by extracting and following hyperlinks, and is useful in harvesting clusters of SWDs on the Web. Our *bounded HTML crawling* imposes some thresholds (e.g., search depth, maximum number of URLs, and minimum percentage of SWD) to limit search space and ensure efficiency. For example, we have harvested many PML documents[8] by a bounded HTML crawl starting at http://iw.standford.edu/proofs. Again, manual submission and automated inductive learner are involved in collecting seeding URLs.

[7] See the Scutter specification at http://rdfweb.org/topic/ScutterSpec.
[8] SWDs that populate instances of the Proof Markup Language(PML) ontology (http://inferenceweb.stanford.edu/2004/07/iw.owl).

4. **RDF crawling.** The *RDF crawler* enhances conventional HTML crawling by adding RDF validation and hyperlink extraction components. It visits newly discovered URLs and periodically revisits pages to keep metadata current. For each URL, it tries to parse an RDF graph from the document using RDF parsers (e.g. Jena). If successful, it generates document level metadata and also enqueues the new discovered URLS that may link to SWDs.
5. **Inductive learner and Swoogle Sample dataset.** The sample dataset covers the metadata of the SWDs confirmed by RDF crawling. Based on the features (e.g. URL, term frequency, the source website) of harvested documents and their labels (e.g. whether they are SWD, embedded SWD or non-SWD), an automated inductive learner is used to generate new seeds for Google-based Meta-crawling and Bounded HTML crawling.

The crawler schedules its methods using the following harvesting strategies: (i) SWO harvesting has the highest priority since they are critical for users to encode and understand Semantic Web data; (ii) PSWDs are harvested with higher priorities than ESWDs because the former usually contain more Semantic Web data than the latter; and (iii) we delay harvesting URLs from websites where more than 10,000 SWDs have already been found (e.g., liveJournal) to avoid having the catalog dominated by SWDs from a few websites.

3.3 Harvesting Result and Performance

The dataset **SW06MAY** resulted from harvesting data between January 2005 and May 2006. It has 3,675,153 URLs, including 1,448,504 (40%) confirmed as SWDs, 13% confirmed as non-SWDs, 9% unreachable URLs, and 38% unpinged (not yet visited) URLs. The confirmed SWDs are from 162,245 websites[9] and contribute 279,461,895 triples. Although *SW06MAY* is much smaller than the Web with its 11.5 billion documents [33], it is much larger than any existing datasets, including:

- (2002) Eberhard [6] reported 1,479 valid SWDs out of nearly 3,000,000 URLs.
- (2003) OntoKhoj [7] reported 418 ontologies out of 2,018,412 URLs after 48-hour crawling.
- (2004) DAML Crawler reported 21,021 DAML files out of 743,017 URLs.

Significance of ontology discovery. SW06MAY contributes 83,007 SWOs including many unintended ones, such as (i) instance data with unnecessary class or property definitions or references, e.g., 55,565 (66.9%) *PML documents* from *onto.stanford.edu*, and 882 (1.1%) *semantic blog documents* from *lojjic.net*, and (ii) instance data that has unnecessary instances of *owl:Ontology*, e.g., 4,437 (5.3%) *publication metadata pages* from *www.aifb.uni-karlsruhe.de* and more *web portal metadata pages* from *ontoware.org*. Therefore, the "true" number of SWOs in SW06MAY is just 22,123 (26.7%) SWOs after removing the "unintended" ones. Moreover, this number can further reduced to 13,012 (15.7%) since there are many duplications [10].

[9] A website is uniquely identified by its domain name (host name part of a URL) but not it's IP address. Virtual hosting can result in one IP address hosting many web domains.

[10] We are currently detecting duplicate SWDs by simply comparing the md5sum of two target documents. While crude, the method is efficient and useful. For example, we have found

Significance of dataset growth. The significance of *SW06MAY* can be verified by its fast growth trend. Figure 3a shows the numbers of total URLs (*url), pinged URLs (*ping), confirmed SWDs (*swd) and confirmed pure SWDs (*pswd) discovered before the date on x-axis, and it exhibits a steady growing trend. The "ping" curve touches the "url" curve because our harvesting strategy delays harvesting URLs from websites hosting more than $10,000$ URLs until all other URLs have been visited. The increasing gap between "ping" curve and "swd" curve indicates that harvesting recall increases at the expense of the decrease of precision.

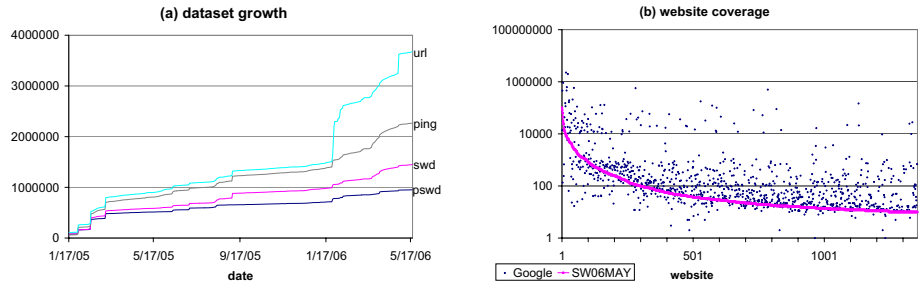

Fig. 3. The SW06MAY dataset has nearly 4M URLs collected from more than 160K sites. An analysis of the dataset demonstrates the growth in Semantic Web documents (left) and also provides evidence that our hybrid harvesting framework is sound (right).

Significance analysis on website coverage. We further evaluate the significance of *SW06MAY* by comparing its *website coverage* (i.e., the number of pure SWDs per website) with Google's estimation. In Figure 3b, each dot on the curve denotes the website coverage of one website that hosts at least ten pure SWDs. For each of the 1,355 websites in the graph, we use "Google" dots to show the optimistic Google estimation of website coverage with an additional "site" constraint, e.g., "(rdf OR inurl:foaf OR inurl:rss) -filetype:html site:www.cs.umbc.edu". The figure shows that Google's estimate, even with high variance, exhibits a trend similar to *SW06MAY*'s estimate. We conclude that the *SW06MAY* provides evidence in the basic soundness of our harvesting approach. Moreover, we suggest three causes of the variance: (i) Google's estimation may be too high since it is optimistic; (ii) The Google query site constraint searches all sub-domains of the site (e.g., site:w3.org also returns results from www4.w3.org), but *SW06MAY*'s results only return results from the specified site; and (iii) our harvesting framework may index fewer SWDs (see Google dots above the curve) because it uses far less harvesting seeds than Google and keeps a long "unpinged" list, or index more SWDs (see Google dots below the curve) because it complements Google's crawling limitation.

166 different SWDs having the same md5sum as the SWO *http://purl.org/dc/terms*. Trying to proving semantic equivalence is in general, not an option.

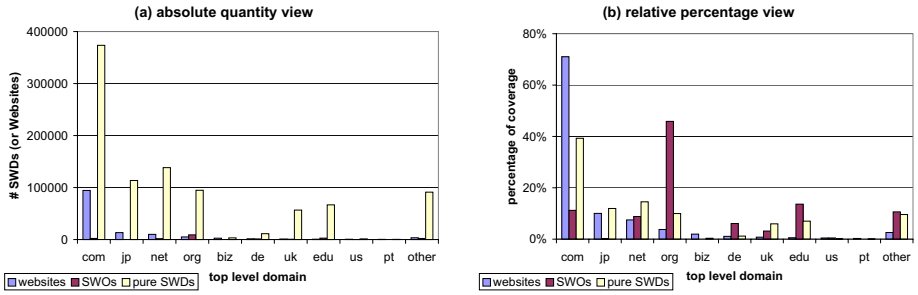

Fig. 4. An analysis of the SW06MAY dataset shows the distribution of SWDs and SWOs (after removing unintended ones) over selected top-level domains. Codes used are jp:Japan, de:Germany, uk:United Kingdom, us:United States, pt:Portugal, and other:remaining TLDs.

4 Measuring Semantic Web documents

SWD Top-level Domains. Analyzing the top-level domains (TLDs) of SWDs suggests the degree to which Semantic Web data is published by region and type of organization. Using SW06MAY we calculated the number of websites, SWDs and pure SWDs for the top ten TLDs as shown in Figure 4. The TLDs are ordered by the number of websites. Figure 4a shows that pure SWDs dominate the Semantic Web while SWOs are few in number. Figure 4b reveals several points. First, the ".com" domains have contributed the largest portion of hosts (71%) and pure SWDs (39%). Examining the data indicated two reasons: ".com" sites make heavier use of virtual hosting technology and publish many RSS and FOAF documents. Second, most SWOs are from ".org" domains (46%) and "edu" (14%). This is likely due of the deep interests in developing ontologies from academic and non-profit organizations.

SWD Source Websites. Figure 5 depicts the cumulative distribution of the number of PSWDs per website. The curves do contain skewed parts: (i) the sharp drop at the tail of curve (near 100,000 on x-axis) is caused by our harvesting strategy that delays harvesting websites after finding more than 10K SWDs; and (ii) the drop at the head of curve is due to virtual hosting technology[11]. Interestingly, *livejournal.com* is involved in both. Both curves in Figure 5 show power law distribution and the similar parameters of the two regressed equations support the conclusion that the distribution is invariant.

Table 1 lists the ten domains hosting the largest number of pure SWDs. The "content" column shows the topic of website, and the "unpinged" column indicates that we intentionally delay crawling some giant websites. SWDs from these websites are automatically generated and well inter-linked. The 6th and 9th websites are recently promoted to this list.

SWD Age. We measure an SWD's age by its last-modified time extracted from the HTTP response header. Figure 6a shows cumulative distribution of last-modified time, i.e., the number of PSWDs and SWOs with a last-modified before the date on X-axis. SWD's with no reported last-modified time are excluded. Note that the "pswd" curve exhibits an exponential distribution, indicating that many new PSWDs have been added

[11] Many social networking sites offer each user a unique virtual host name.

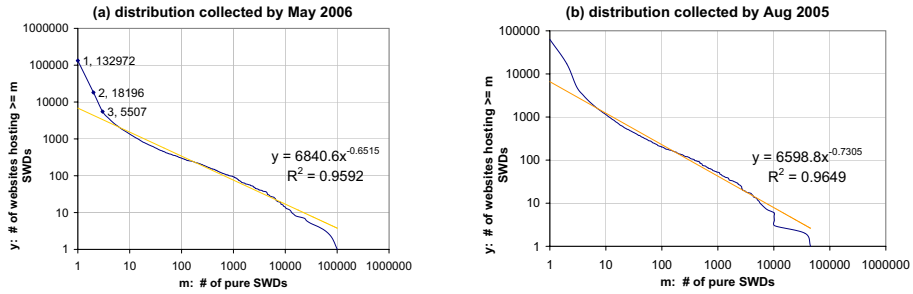

Fig. 5. Data from SW06MAY shows that the distribution of the number of websites hosting more than m pure SWDs follows a power law. The straight lines correspond to regression function with the given equations. The R^2 values close to one indicate good regressions.

rank	website	# PSWDs	# unpinged	content
1	www.livejournal.com	100,518	88,962	foaf, personal profile
2	www.tribe.net	80,402	25,234	foaf
3	www.greatestjournal.com	62,453	849	foaf
4	onto.stanford.edu	45,278	403	pml, portal proof
5	blog.livedoor.jp	31,741	12,776	foaf
6	r622-1.mpiwg-berlin.mpg.de	25,733	136	vml annotation
7	www.ecademy.com	23,242	3,308	foaf
8	www.hackcraft.net	16,238	0	dc, book annotation
9	open.bbc.co.uk	14,544	350,473	dc, BBC program annotation
10	www.uklug.co.uk	13,263	2	rss

Table 1. This table lists the ten largest source websites of pure Semantic Web documents (PSWDs) from May 2006. The *unpinged* column gives the number of URLs discovered on the site that are suspected of also being Semantic Web documents but have not yet been processed.

to the Semantic Web or that many old ones are being actively modified. The “swo” curve additionally excludes PML documents and exhibits exponential distribution with a flat tail, which we interpret as indicating a more active ontology development earlier in the time period transitioning to more reuse later.

Figure 6b shows two distributions of last-modified time collected in Aug 2005 and May 2006 respectively. The difference before August 2005 represents a loss of 155,709 PSWDs and is due to documents going offline (25%) and being updated (75%). The difference after that is caused by updated documents and newly discovered PSWDs. The non-trivial at which PSWDs go offline significantly affects the growth of Semantic Web data.

SWD Size. We measure an SWD’s size as the number of triples in the SWD’s RDF graph. Figure 7a shows the distribution of SWD’s size, i.e., the number of SWDs having exactly m triples, and Figure 7b the corresponding cumulative distribution. Figure 7c depicts the distribution of ESWD’s size. Most ESWDs are very small with 62% having exactly three triples and 97% having ten or fewer triples. These contribute significantly to the big peak in Figure 7a. Figure 7d shows the distribution of the size of PSWDs, with most (60%) having five to 1000 triples. The peaks in the curve are

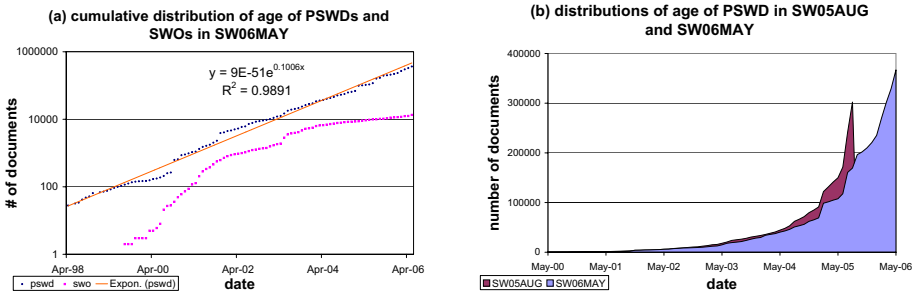

Fig. 6. Distributions of the last-modified time of PSWDs and SWOs.

caused by automatically generated SWDs which publish Semantic Web data in fixed patterns. For example, many PML documents have exactly 28 or 36 triples, and many RSS documents have exactly 130 triples[12]. The large number of SWOs with fewer than four triples are mainly RDF and OWL test documents. SW06MAY's largest SWO[13] has 1,013,493 triples and defines 337,831 classes and properties.

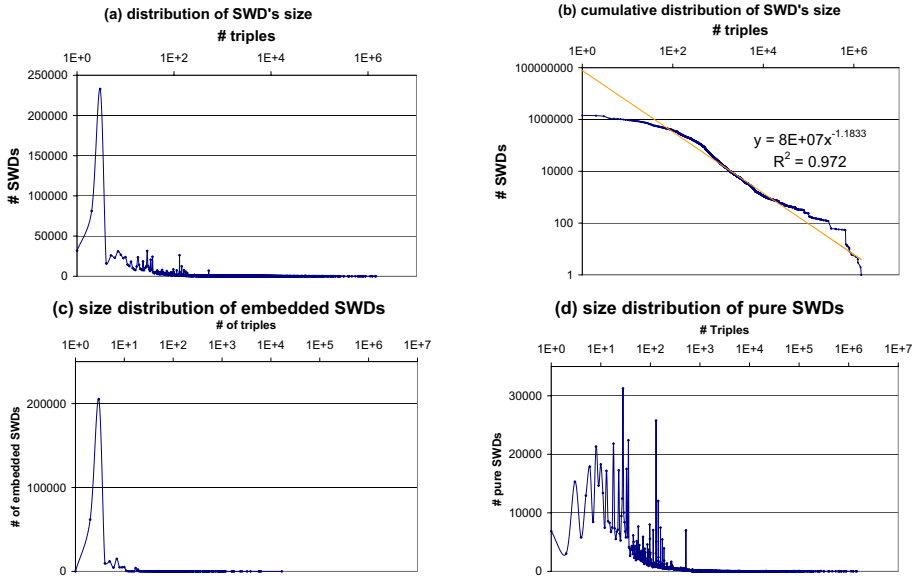

Fig. 7. The distributions of the number of triples per SWD

SWD Size Change. Updating a SWD usually result in in a change in its size. We have investigated this by tracking the size changes for different versions of an SWD. The *SW06MAY* dataset has 183,464 PSWDs that are alive (sill online) and for which we

[12] A typical RSS file has one *rss:channel* with eight triples, fifteen *rss:item* instances each with seven triples, and one *rdf:Seq* with seventeen triples connecting the *rss:channel* to the item instances.

[13] http://www.fruitfly.org/~cjm/obo-download/obo-all/ncbi_taxonomy/ncbi_taxonomy.owl

have at least three versions. For these, 37,012 (20%) lost a total of 1,774,161 triples; 73,964 (40%) gained a combination of 6,064,218 triples, and the rest 72,488 (40%) maintained their original size[14]. The statistics also show that the total number of triples keeps increasing; therefore, we hypothesize the volume of Semantic Web data is increasing.

5 Measuring Semantic Web Terms

Semantic Web Terms (SWTs) are classes and properties that are named by non-anonymous URIrefs. The *SW06MAR* dataset has 1,576,927 distinct Semantic Web terms defined with respect to 14,488 Semantic Web namespaces. We derive four SWT-usage patterns by analyzing the combination of six basic types of meta-usages.

- Only a few classes (1.5%) and properties (1.0%) have both explicit definitions and instances.
- Most SWTs (95.1%) have no instances, and some SWTs (2.2%) have no definitions.
- Some SWTs (0.08%) mistakenly have both class and property meta-usage.
- Some SWTs (0.08%) only have REF-C or REF-P meta-usages. While some are *XMLSchma* terms and not RDF, others appear to be due to errors or misuse.

SWT Definition Complexity. A simple way to measure the complexity of a SWT is to count the number of triples used to define it. Figure 8a shows the cumulative distribution of the size of SWT definitions in the curve labeled "all". This follows a power law distribution with the deviations at the head and tail reflecting a preference for defining SWTs using a manageable number of triples, two to ten triples in most cases. Terms that can be defined in just a few triples are not very useful, and the definitional size of complex terms can be reduced by defining and using auxiliary definitions. One observed definition has nearly 1000 triples[15]. We've divided definitional triples into two classes: annotation and relation triples, whose *rdf:object*s are *rdfs:Literal*s and *rdf:object*s, respectively. Note that relation triples are more common. We also noticed that 104,152 SWTs have been defined in more than one SWOs.

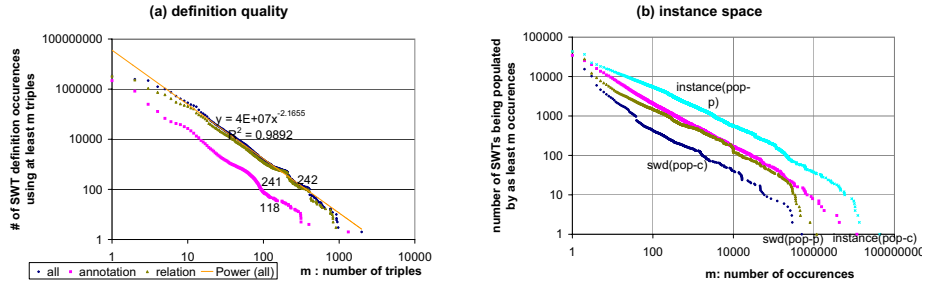

Fig. 8. The cumulative distribution of meta-usages of SWT

[14] Most of the PSWDs maintaining their size are RSS documents.

[15] The SWD http://elikonas.ced.tuc.gr/ontologies/DomainOntologies/middle_ontology defines the *MOSemanticRelationType* class using 973 triples.

SWT Instance Space. Since Semantic Web data include both definitions and instance data, we measure the instance space of the Semantic Web by counting POP-C and POP-P meta-usages of SWTs[16]. Figure 8b shows the cumulative distribution of the number of SWTs populated as a class (or property) by at least m instances (or SWDs). All four curves follow a power law distribution. For both classes and properties, most are defined but never directly used. Only 423 classes have been instantiated by more than 100 SWDs and just 2,115 have more than 100 instances. The number of properties used is somewhat higher, with 1,489 SWTs used to define data in more than 100 SWDs and 5,404 properties used in more than 100 assertions.

Table 2 lists popular classes and properties. The number of an SWT's class-instances is usually proportional to the number of SWDs populating the SWT; however, exceptions exist. For example, while the *wn:Noun* class has significant number of instances, they are mostly in a few huge SWDs. In general, the Semantic Web's instance space is dominated by three categories: (i) instances of meta-ontologies such as OWL, (ii) instances of a small number of very popular ontologies such as DC, FOAF, and RSS; and (iii) instances from giant data files, such as WordNet and National Library of Medicine's Medical Subject Headings (MeSH) ontology.

resource URI	**#swd**	**#instance**
Most instantiated classes ordered by #swd		
`http://xmlns.com/foaf/0.1/Person`	499,671	11,686,519
`http://www.w3.org/1999/02/22-rdf-syntax-ns#Seq`	290,321	308,907
`http://purl.org/rss/1.0/channel`	282,677	289,160
`http://purl.org/rss/1.0/item`	259,220	4,277,868
`http://xmlns.com/foaf/0.1/Document`	223,510	247,311
Most instantiated classes ordered by #instance		
`http://xmlns.com/foaf/0.1/Person`	499,671	11,686,519
`http://purl.org/rss/1.0/item`	259,220	4,277,868
`http://www.cogsci.princeton.edu/~wn/schema/Noun`	56	3,697,400
`http://www.w3.org/2002/07/owl#Class`	68,053	1,795,941
`http://www.nlm.nih.gov/mesh/2004#Concept`	38	1,551,046
Most instantiated properties ordered by #swd		
`http://www.w3.org/1999/02/22-rdf-syntax-ns#type`	1,170,975	43,291,848
`http://purl.org/dc/elements/1.1/title`	801,254	13,448,548
`http://xmlns.com/foaf/0.1/mbox_sha1sum`	462,198	2,633,739
`http://purl.org/dc/elements/1.1/description`	453,826	2,874,327
`http://www.w3.org/2000/01/rdf-schema#seeAlso`	432,288	12,330,223

Table 2. This table shows the most popular Semantic Web classes and properties based on the number of Semantic Web documents (SWDs) that use them and, for classes, also on the number of immediate instances.

RDFS and OWL usage. To what degree does the current Semantic Web make use of RDFS and OWL? One simple way of addressing this question is to examine the

[16] Since no RDFS or OWL inferencing is done, the statistics reflect immediate class instances.

number of SWDs that use the RDFS and OWL namespaces. The OWL namespace has been declared by 112,870 SWDs (8%) and actually used by 108,059 (7%). The RDFS namespace enjoys more use, being declared by 677,049 (47%) and used by 537,614 (37%) SWDs.

What about their terms? Not surprisingly, *owl:Class* is the most used term from the OWL namespace with 1,795,941 instantiations in 68,053 SWDs. Contrasting this with *rdfs:Class*, which has 327,485 instantiations by 8,572 SWDs, seems to suggest that OWL is being more heavily used than RDFS. However, the relationship is not so simple. When examining properties, rdf:Property has 529,052 immediate instantiations from 58,598 SWDs, considerably more than the OWL property terms owl:ObjectProperty (169,885 assertions in 8,041 SWDs) and owl:DatatypeProperty (48,386 assertions in 4,557 SWDs).

For RDFS and OWL properties, the most used properties is *rdf:type*, followed by some annotation properties such as *rdfs:seeAlso* and *rdfs:label*. Among those properties that are used as ontology constructs, *owl:sameAs* and *rdfs:subClassOf* are the most used. We also noticed significant use of two OWL equality assertions: *owl:sameAs* (279,648 assertions in 17,425 SWDs) and *owl:equivalentClass* (69,681 assertions in 4,341 SWDs). Their common use may be an indication of increased ontology alignment. We have found limited use of properties that require OWL DL or OWL FULL reasoning support. The most common one in our dataset was *owl:unionOf* which is used in only 2,527 SWDs.

Instantiation of *rdfs:domain*. Semantic Web data is published asynchronously by autonomous and distributed agents which may use, and misuse, a variety of ontologies. Given enough data, we can attempt to reverse-engineer the definitions of classes and terms introduced by ontologies. Consider instances of the *rdfs:domain* relation which associates a class with properties that describe its instances. We have observed 111,071 unique instantiations of *rdfs:domain*, and the number of instantiations that have been observed in at least m instances (or SWDs), again, follows a power law distribution.

The highly instantiated *rdfs:domain* relations are mainly from popular instance space such as FOAF and RSS documents. An interesting observation is that *rdfs:seeAlso* property has been frequently used as *instance property* of *foaf:Person*. This corresponding definition cannot be found in the RDFS or FOAF ontologies although it has been informally mentioned in FOAF specification. The popularity of instantiation is usually determined by the number of SWDs that has the instantiation; moreover, we also noticed a popular instantiation – the domain of *wn:wordFrom* is *wn:Noun* which has over 6.5 million occurrences in only 56 SWDs.

We can use data on the instantiations of *rdfs:domain* relation to derive the most used properties of a given class. For example, for immediate *foaf:Person* instances, the most common properties used are *foaf:mbox_sha1sum* (461,922 SWDs), *rdfs:seeAlso* (385,516), and *foaf:nick* (361,901). We can also find strong co-occurrence association among properties of a class. The properties *geo:lat* (85,742) and *geo:long* (85,741) are virtually always used together in modifying a class *geo:Point*. This kind of information can be used to help publishers choose a good set of properties, which may be from different ontologies, for a given class. Moreover, we can use such information in on-

tology revision, e.g., adding the missing *rdfs:domain* definition or revise incompatible definition.

6 Conclusions

The Semantic Web is not just one universal RDF graph but a federated collection documents distributed on and accessed via the World Wide Web. It must be studied from both the *Web perspective* and the *semantic perspective*. In order to characterize the Semantic Web on the Web and guide Web-scale data access, we estimated the size of the Semantic Web using Google, implemented a hybrid framework for harvesting Semantic Web data, and measured the results to answer questions on the Semantic Web's current deployment status.

The statistics where characterized by power law distributions and "complex system" behavior in many cases and, in general, support several conclusions about the emerging Semantic Web. (i) Semantic Web data is growing steadily on the Web even when many documents are only online for a short-while. (ii) The space of instances is sparsely populated since most classes ($>$97%) have no instances and the majority of properties ($>$70%) have never been used to assert data. (iii) Ontologies can be induced or amended by *reverse engineering* the instantiations of ontological definition in instance space [27].

Our work raises question about the current paradigm for ontologies and URIrefs. Is the concept of an "ontology" as a collection or container for Semantic Web terms needed or even useful? An ontology object encourages self consistency but introduces some limitations as well. Recent work on ontology partitions argues against large, monolithic ontologies in favor of having many interconnected components. We might even eliminate namespaces as boundaries. For example, the Dublin Core Element ontology has been widely used together with terms from many other semantic web ontologies. Another debatable item is the URIref. We use triples to annotate an URIref that is an identifier of a resource. Multiple RDF graphs from different documents describing the same URIref can introduce inconsistency. Integrating these definitions may encounter several questions: (i) are URIrefs good enough for grouping the triples describing it; (ii) can we ensure that all of the graphs are accessible to consumers; and (ii) should all be used or should some be rejected as untrustworthy.

References

1. Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., Weitzner, D.J.: Creating a science of the web. Science **313** (2006) 769–771
2. Zou, Y., Finin, T., Ding, L., Chen, H., Pan, R.: Using Semantic web technology in Multi-Agent systems: a case study in the TAGA Trading agent environment. In: Proceeding of the 5th International Conference on Electronic Commerce. (2003)
3. Franz, T., Staab, S.: Sam: Semantics aware instant messaging for the networked semantic desktop. In: Proceedings of the ISWC 2005 Workshop on The Semantic Desktop - Next Generation Information Management and Collaboration Infrastructure. (2005)
4. Visser, U., Stuckenschmidt, H., Schuster, G., Vogele, T.: Ontologies for geographic information processing. Computers and Geoscience **28** (2002) 103–117

5. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American **284** (2001) 35–43
6. Eberhart, A.: Survey of rdf data on the web. Technical report, International University in Germany (2002)
7. Patel, C., Supekar, K., Lee, Y., Park, E.K.: OntoKhoj: a semantic web portal for ontology searching, ranking and classification. In: WIDM'03. (2003)
8. Dean, M., Barber, K.: Daml crawler. http://www.daml.org/crawler/ (August 2006) (2002)
9. DAML: The DAML ontology library. http://www.daml.org/ontologies/ (August 2006) (2004)
10. Pitkow, J.E.: Summary of www characterizations. Computer Networks **30** (1998)
11. Lawrence, S., Giles, C.L.: Accessibility of information on the web. Nature **400** (1999)
12. Gil, R., Garca, R., Delgado, J.: Measuring the semantic web. SIGSEMIS Bulletin **1** (2004)
13. Hartmann, J., Sure, Y., Giboin, A., Maynard, D., del Carmen Surez-Figueroa, M., Cuel, R.: Methods for ontology evaluation. Technical report, University of Karlsruhe (2004)
14. Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: A theoretical framework for ontology evaluation and validation. In: Proc. of the 2nd Italian Semantic Web Workshop. (2005)
15. Lozano-Tello, A., Gomez-Perez, A.: ONTOMETRIC:a method to choose the appropriate ontology. Journal of Database Management **15** (2003)
16. Welty, C.A., Guarino, N.: Supporting ontological analysis of taxonomic relationships. Data Knowledge Engineering **39** (2001)
17. Parsia, B., Sirin, E., Kalyanpur, A.: Debugging owl ontologies. In: WWW'05. (2005)
18. Magkanaraki, A., Alexaki, S., Christophides, V., Plexousakis, D.: Benchmarking RDF schemas for the semantic web. In: ISWC'02. (2002)
19. Supekar, K., Patel, C., Lee, Y.: Characterizing quality of knowledge on semantic web. In: FLAIRS'02. (2002)
20. Alani, H., Brewster, C.: Ontology ranking based on the analysis of concept structures. In: K-CAP'05. (2005)
21. Yao, H., Orme, A.M., Etzkorn, L.: Cohesion metrics for ontology design and application. Journal of Computer Science **1** (2005)
22. Tartir, S., Arpinar, I.B., Moore, M., Sheth, A.P., Aleman-Meza, B.: Ontoqa: Metric-based ontology quality analysis. In: Proc. of Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources. (2006)
23. John C. Paolillo and Elijah Wright: The Challenges of FOAF Characterization. In: Proc. of the 1st Workshop on Friend of a Friend, Social Networking and the (Semantic) Web. (2004)
24. Grimnes, G.A., Edwards, P., Preece, A.: Learning meta-descriptions of the foaf network. In: ISWC'04. (2004)
25. Mika, P.: Social Networks and the Semantic Web: An Experiment in Online Social Network Analysis. In: Proc. of International Conference on Web Intelligence. (2004)
26. Ding, L., Zhou, L., Finin, T., Joshi, A.: How the semantic web is being used:an analysis of foaf. In: Proceedings of the 38th International Conference on System Sciences. (2005)
27. Ding, L.: Enhancing Semantic Web Data Access. PhD thesis, UMBC (2006)
28. Zhang, Y., Vasconcelos, W., Sleeman, D.: Ontosearch: An ontology search engine. In: Proc. of 24th Conf. on Innovative Techniques and Applications of Artificial Intelligence. (2004)
29. Lindesay, V.: The schemaweb repository. http://www.schemaweb.info/ (August 2006) (2005)
30. Biddulph, M.: Crawling the semantic web. In: XML Europe. (2004)
31. Apsitis, K., Staab, S., Handschuh, S., Oppermann, H.: Specification of an RDF Crawler. http://ontobroker.semanticweb.org/rdfcrawl/help/specification.html (March 2006) (2005)
32. Sherman, C.: Metacrawlers and metasearch engines. http://searchenginewatch.com/links/-article.php/2156241 (March 2006) (2004)
33. Gulli, A., Signorini, A.: The indexable web is more than 11.5 billion pages. In: WWW'05 (poster). (2005)