# The Semantic Webscape: A View of the Semantic Web

Juhnyoung Lee
IBM T. J. Watson Research Center
Hawthorne, NY 10532
U.S.A.
jyl@us.ibm.com

Richard Goodwin
IBM T. J. Watson Research Center
Hawthorne, NY 10532
U.S.A.
rgoodwin@us.ibm.com

## ABSTRACT

It has been a few years since the semantic Web was initiated by W3C, but its status has not been quantitatively measured. It is crucial to understand the status at this early stage, for researchers, developers and administrators to gain insight into what will come in this field. The objective of our work is to quantitatively measure and present the status of the semantic Web. We conduct a longitudinal study on the semantic Web pages to track trends in the use of semantic markup languages. This paper presents early results of this study with two historical data sets from October 2003 and October 2004. Our results show that while it is very early stage of semantic Web adoption, its growth outpaces that of the entire Web for the period. Also, RDF (Resource Description Framework) has dominated among semantic markup languages, taking about 98% of all semantic pages on the Web. It has been used in a variety of metadata annotation applications. This study shows that the most popular application is RSS (RDF Site Summary) for syndicating news and blogs, which takes more than 60% of all semantic Web pages. It also shows that the use of OWL (Web Ontology Language) which was recommended by W3C in early 2004 has been increased 900% for the period.

## Categories and Subject Descriptors

H.1 [**Information Systems**]: Models and Principles; H.3.3 [**Information Systems**]: Information Search and Retrieval; H.3.1 [**Information Systems**]: Content Analysis and Indexing.

## General Terms

Measurement, Experimentation, Languages.

## Keywords

Semantic Web, Markup Languages, Ontology, RSS.

## 1. INTRODUCTION

It has been a few years since the Semantic Web was initiated by W3C [1]. It has been a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. For the past few years, we have seen significant progress in the current components of that framework, which are the RDF Core Model, the RDF Schema language and the Web Ontology language (OWL). (These languages all build on the foundation of URIs, XML, and XML namespaces.) We also have seen significant amount of research work going on for building tools such as programming interfaces, parsers, validators, editors and management systems. Furthermore, we have seen significant amount of interest from industry for the applications of the semantic Web technology in various areas including business information and process integration, life sciences, information search, and autonomic computing. The Gartner group recently reported that Semantic Web (with related technologies such as ontologies, metadata management, and taxonomies) is one of the top strategic technologies for 2005 [2].

The objective of this paper is to quantitatively measure and present the status of the Semantic Web. For this purpose, the questions we aim to answer include: Who is using the semantic markup languages? Which semantic markup languages are used, and how frequently? What applications of semantic Web are there? What subjects of ontology are described in the languages? What features of the languages are used, and how frequently? How the status is changing over time? We understand there are alternative ways to find answers to these questions. It is important to understand the status of the semantic Web at this early stage of the initiative, for researchers, developers, and administrators to gain insight into what will come in this field, and make an informed decision on where to go with their work. In our study, we attempt to find the answers by measuring the actual use of semantic Web languages on the Web. We directly collect data on actual semantic pages on the Web, instead of depending on an indirect survey. A detail description of our analysis method and a full study report can be found in [4].

## 2. HIGH-LEVEL OBSERVATIONS

We conducted a longitudinal study on the semantic pages on the Web to track trends in the use of semantic markup languages. This paper presents early results on our study with two historical data sets from October 2003 and October 2004. The input to this analysis is a set of links of all Web pages whose extension is .rdf, .daml, or .owl, indicating that the content were written in one of those semantic markup languages.

The first observation is that the number of Web pages written in semantic markup languages is very small. However, the number is growing rapidly overall, and significantly in some areas. As of October 2003, the number of semantic Web pages is 14,812 from some 7,000 servers. This number is out of over 5 billion links on some 30 million servers discovered by IBM WebFountain. The percentage of the semantic Web pages is less than 0.0003%. However, the growth of semantic Web pages outpaces that of the entire Web. As of October 2004, the number of semantic Web pages becomes 46,601, which is more than 300% growth. At the time, there are 7.5 billion links on some 77 million servers on the Web detected by IBM WebFountain. Figure 1 graphically shows the total number of semantic Web pages for the period.

*WWW 2005*, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

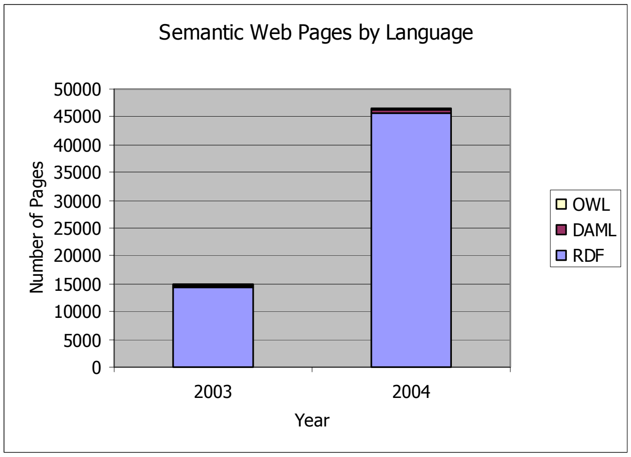

**Figure 1. Trend of semantic Web pages**

## 3. LANGUAGE ANALYSIS

Figure 1 also shows the classification of semantic Web pages by language. It is apparent that the great majority of semantic Web pages are written in RDF. As of October 2003, the number of semantic pages written in RDF is 14,240 out of the total 14,812. It is about 96%. As of October 2004, the number changes to 45,606 out of 46,601, which is almost 98%. The increase of RDF pages is about 220% for the period.

Compared to RDF, the numbers of semantic Web pages written in DAML and OWL are almost negligible. However, when closely examined, they show strong dynamics for the period, especially for OWL. As of October 2003, only 31 pages written in OWL were found in the entire Web. As of October 2004, the number became to 310, which is about 900% growth over a year period. DAML pages grew from 541 to 686 for the period, which is about 27% increase. Combined, semantic pages written in DAML and OWL increased about 74%. It is a significant number, although it is modest when compared to that of RDF.

## 4. APPLICATION ANALYSIS

The RDF specifications provide a lightweight ontology system to support the exchange of knowledge on the Web. RDF integrates a variety of applications from library catalogs and directories to syndication of news and content to personal collections of music, photos and events. This study discovered that a single RDF application which dominates among others is *RSS* (*Really Simple Syndication* or *RDF Site Summary* 1.0). It is a lightweight multipurpose extensible metadata description and syndication format proposed in August 2000 to the RDF Interest Group. RSS began catching on a couple of years ago, when Web logs or blogs, started using it to allow readers to know they had posted something new. Soon traditional publishers dove in. During the past year, The Wall Street Journal, National Public Radio, and Reuters Group among others have added RSS feeds [3]. RSS 1.0 uses RDF, but the current version RSS 2.0 is not based on RDF.

Another popular application of RDF discovered in this study is the Friend of a Friend (FOAF) project, which is about creating a Web of machine-readable homepages describing people, links between them and things they create and do. While applications of RDF are mostly metadata annotation of various resources, the counterparts of DAML and OWL are more semantically-rich ontologies, which are formal description of classes in a domain, their properties, and their relationships with other classes.

Figure 2 displays the RDF pages segmented by application. The RSS pages take more than 60% of the entire semantic pages in 2004. Its portion is actually decreased somewhat from 70% in 2003. However, it is still the dominating application. On the other hand, the number of pages involved in the FOAF projects grew more than 800% to 1,503 in 2004 from 161 in 2003. In 2004, FOAF takes about 3% of the entire semantic pages. The portion of other applications of RDF (e.g., library catalogs, directories, syndication of news and blogs, and personal collections of music, photos, and events) also grew more than 300% for the period.

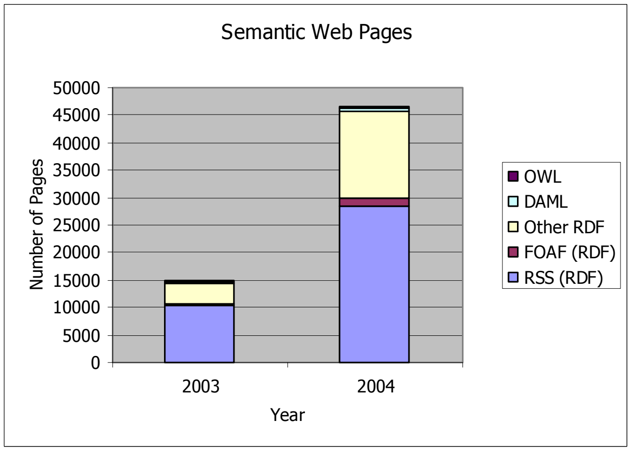

**Figure 2. Semantic Web pages by application**

## 5. CONCLUDING REMARKS

We measured and presented the status of the semantic Web. Our results show that it is very early stage of semantic Web adoption, but that there has been remarkable progress in the adoption over the last couple of years. This paper presents early results of our longitudinal study on semantic Web. The full study report with a detail description of the analysis method is available in [4].

## 6. ACKNOWLEDGMENTS

We thank David Gibson, Kevin McCurley, Andrew Tomkins, Runping Qi, Andrei Broder, Youngja Park, and Anca-Andreear Ivan for their generous technical and scientific support.

## 7. REFERENCES

[1] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific American, May 2001.

[2] The Gartner Group, "Top 10 Strategic Technologies for 2005," Gartner Symposium ITXPO, March 28 - April 1, 2004, San Diego Convention Center, San Diego, California.

[3] H. Green, "All the News You Choose – on One Page: RSS, which delivers customer-tailored bulletins to users, may shake up e-media" BusinessWeek, October 25, 2004.

[4] J. Lee and R. Goodwin, "The Semantic Webscape: a View of the Semantic Web," IBM Research Report, November 2004.