Preface

There are clear indications that Semantic Web, if seen as a technology, has passed the early adoption phase of its technology adoption life cycle (Wikipedia Contributors, 2010). The adoption of Semantic Web is fuelled by convergence of a number of factors, including the following:

- accelerating growth of information and resources on the Web, and increasing heterogeneity (both in technological aspects such as representation and media, and in nontechnical aspects such as socio-cultural aspects)
- recognition on the part of not just the researchers but also practitioners and companies that syntactic and statistical solutions near the limit in effectiveness in dealing with scale and heterogeneity, and future gains will come from use of semantics
- good degree of consensus on and adoption of representation languages and core technologies for which W3C's Semantic Web initiative and its recommendations such as RDF, SPARQL, and OWL have played critical role
- availability of technologies, with plenty of open source tools and system exemplified by over 20 RDF stores and query systems, as well as broader ecosystem of available commercial service and product providers
- successful demonstration of its value proposition by a number of early adoption domains as demonstrated by deployed applications (Sheth & Stevens, 2007; Brammer & Terziyan, 2008; Cardoso et al., 2008) in several domains including healthcare (Sheth et al., 2006) and life sciences (Ruttenberg et al., 2009; Baker & Cheung, 2007), pharmaceuticals, financial services (Sheth, 2005), e-government and defense (Mentzas, 2007).

Early commercial use of Semantic Web approach was reported by Taalee (subsequently through acquisiton/merger Voquette, Semagix, Fortent) founded in 1999, the same year in which the term Semantic Web was coined by Tim Berners-Lee. A keynote given in 2000 gives clear examples of the semantic search and other applications that had paying customers (Sheth, 2000). This involves creation of ontologies or background knowledge in variety of domains, automatic semantic annotation of heterogeneous Web content, and applications including semantic search, semantic browsing, semantic personalization, semantic targeting/advertisement, and semantic analysis (Sheth et al., 2001). Those early efforts covered hundreds of websites and semantic processing at the rate of about million documents per hour per server, and was largely limited by the infrastructure available. A number of commercial products and services continued to increase that formed the basis of the innovation and early adoption parts of the technology life cycle.

Now let us see why we are in early majority phase of the lifecycle. A rapidly growing number of companies and organizations are offering products and servicesinvolving Semantic Web technologies or are using them for mission critical applications (Sheth & Stephens, 2007; [3] Herman, 2009). Example companies providing products and services (with example of one key Semantic Web application) include Adobe (internet and desktop application tools), Dow Jones (content delivery), General Electric (energy efficiency), Hakia (search), IBM (content analysis), Nokia (portal tools and services), OpenLink and Oracle (DBMS), WolframAlpha (search), and Reuters (semantic annotation service). A number of companies and organizations are using Semantic Web technologies for mission critical applications, including Office of Management and Budget, Pfizer, Eli Lily, Novartis, and Telefonica. Commercial interest in Semantic Web technology was most vividly demonstrated in the form of acquisitions of several startups and small companies by major Internet and technology companies, best exemplified by Microsoft's acquisition of Powerset (2008), Apple of Siri (April 2010), and Google of Metaweb (June 2010).

While use of Semantic Web technology on a full Web scale is yet to come, what we see is a concrete progress towards using and supporting semantic Web capabilities on the Web scale. The most concrete step taken by these Web scale systems, primarily search and other Web applications, is the creation and/or reuse of massive amounts of background knowledge, often involving a collection of domains, and each involving (a domain specific) sets of entities (also called objects, concepts, etc). All major search companies—Microsoft's Bing, Google, and Yahoo!—are known to be working towards this capability. Support of disambiguation is a litmus test of a semantic capability (as opposed to keyword/syntax centric approaches), which most of these systems are working hard to support. Equally important is adoption of RDFa and open sharing of metadata (such as Facebook'sOpenGraph).

Arguably, however, the most significant progress in Semantic Web has been that of Linked Data. *International Journal on Semantic Web & Information Systems* is proud to have had its first comprehensive special issue on the topic during 2010.

Let us now review the chapters in this book.

Interoperability is one the most challenging issues for cross-organizational Information Systems. Interoperability becomes very important and relevant for e-government Information Systems, which are capable to support cross-organizational communication in a cross-border setup. In "*Solving Semantic Interoperability Conflicts in Cross-Border E-Government Services*," Mocan, Loutas, Facca, Peristeras, Goudos, and Tarabanis propose seamless integration of Pan European e-services for citizens to resolve semantic interoperability, and it uses generic public service model of the Governance Enterprise Architecture and Web Service Modeling Ontology. The chapter discusses semantic interoperability conflicts at data-level and schema-level. Data mediation services and solutions are developed in EU funded SemanticGovproject to resolve semantic interoperability conflicts. The solution uses ontology mapping and involves creation of alignments among the domain ontologies at design time and their use at run-time.

Documents containing words not defined in the dictionary like WordNet and such undefined words are called "Unknown Word (UW)." Hwang and Kim in "*A New Similarity Measure for Automatic Construction of the Unknown Word Lexical Dictionary*" propose a new method to construct UW lexical dictionary through inputting various document collections scattered on the Web. To achieve true semantic information processing, the work searches for UWs and terms related to the UW. Bayesian probability is used to assign probabilistic weight and semantic weight based on WordNet is calculated to find the semantic relatedness between an UW and related term(s). The work uses newly designed word sense disambiguation (WSD) method to enable dictionary to have an accurate synset for related terms. Proposed WSD algorithm is designed to automatically construct an UW lexical dictionary with

an accuracy of 81% and it demonstrated efficient performance in comparison to SSI algorithm. Results show 15% improvement in performance in comparison to Dice Coefficient method.

Queries for any Web searching applications are likely to be ambiguous as words in queries usually carry several meanings. In "*Extracting Concepts' Relations and Users' Preferences for Personalizing Query Disambiguation*," Chen and Zhang present a cluster-based Word Sense Disambiguation (WSD) method to find out all appropriate interpretations for the query. Any ambiguous word is likely to have very close semantic relations; the work groups such similar senses together to explain the ambiguous word in one interpretation. In case of several contradictory interpretations for one ambiguous query, users' preferences retrieved from clickthrough data are obtained to determine suitable concepts or cluster of concepts. Experimental result shows better performance of the proposed method compare to case-based WSD and Adapt Lesk algorithms.

Web 2.0 platforms and systems are using RDF and RDFS as basic standards to store, query, update, and exchange the data. Reasoning of RDF data is a critical issue from performance and scalability point of view. There is an urgent need to improve reasoning algorithms to realize the capabilities of Semantic Web. Many researchers are aiming to improve the performance of reasoning algorithm while manipulating large scale RDF/OWL ontologies. SPARQL is used extensively to retrieve data from RDF stores. In "*The Berlin SPARQL Benchmark*," Bizer and Schultz propose a new benchmark to evaluate efficient performance of SPARQL features like OPTIONAL, ORDER BY, UNION, REGEX, and CONSTRUCT. The work compares the performance of three popular RDF stores to two SPARQL-to-SQL rewriters across architectures and uses e-commerce use case having 100M triple and a single client. The paper discusses design of the Berlin SPARQL Benchmark (BSBM) and compares performance of four popular RDF stores - Sesame, Virtuoso, Jena TDB, and Jena SDB with the performance of two SPARQL-to-SQL rewriters - D2R Server and Virtuoso RDF Views and performance of two RDBMS - MySQL and Virtuoso RDBMS. It employs benchmarking techniques such as executing query mixes, query parameterization, simulation of multiple clients, and system ramp-up. None of the benchmark results was found to be superior for a single client use case for all queries and dataset sizes and it justifies the need to improve the rewriting algorithms. Sophisticated optimization techniques should be developed to make SPARQL optimizers robust.

Hellmann, Lehmann, and Auer apply machine learning techniques to obtain complex class descriptions from objects in a very large knowledge base such as DBpedia, OpenCyc, GovTrack, et cetera. "*Learning of OWL Class Expressions on Very Large Knowledge Bases and its Applications*" aims to increase the scalability of OWL learning algorithms through intelligent pre-processing and develop, implement, and integrate a flexible method in the DL-Learner framework to extract relevant parts of very large knowledge bases for a given learning task.

Reasoning on Web based large scale RDF datasets is a highly challenging task. In "*Scalable Authoritative OWL Reasoning for the Web*" Hogan, Harth, and Polleres propose ter-Horst's pD fragment of OWL to compose a rule-based framework for application, which uses forward-chaining reasoning algorithm called Scalable Authoritative OWL Reasoner (SAOR). Forward-reasoning is used to avoid the runtime complexity of query-rewriting associated with backward-chaining approaches. The proposed system separates terminological data from assertional data, comprises of lightweight in-memory index, on-disk sorting and file-scans. It maintains a separate optimized T-box index to perform reasoning on OWL datasets. To keep the resulting knowledge-base manageable, SAOR algorithm considers only positive fragment of OWL reasoning, analyze the authority of sources to avoid hijacking of ontology

and uses pivot identifiers instead of full materialization of equality. Experiments are performed on a database collected from the Web with a billion statements.

To satisfy increase in the demand of services for smart phones/mobile devices, mobile and pervasive services should be capable of semantic reasoning. In "*Enabling Scalable Semantic Reasoning for Mobile Services*," Steller, Krishnaswamy, and Gaber propose an interesting strategy to optimize semantic reasoning for applications and services targeted for mobile devices. Proposed mTableaux algorithm optimizes description logic reasoning tasks so that large reasoning tasks can be scaled for small resource constrained mobile devices. The work presents comparative analysis of performance of proposed algorithm with semantic reasoners - Pellet, RacerPro and FaCT++ to demonstrate significant improvement in response time. Result accuracy is evaluated using recall and precision values.

Linked Data movement is a set of best practices to publish and connect structured data across the Web and can be considered as one of the pillars of Semantic Web. The number of linked data providers has increased significantly in last three years. In the chapter "*Linked Data: The Story So Far*," Bizer, Heath, and Berners-Lee publish linked data, and a review of applications based on linked data are described. Efforts related to linked data are classified into three categories: linked data browsers, linked data search engines, and domain specific linked data applications. SWSE and Falcons search engines are keyword based search engines, but compare to existing popular search engines, both exploit the underlying structure of the data, provide summary of the entity selected by the user, and additional structured data crawled from the Web and links to related entities. A number of services are being developed, offering domain-specific functionality by mashing up data from various linked data sources. Revyu, DBpedia Mobile, Talis Aspire, BBC Programmes and Music, DERI Pipes are few such domain specific linked data applications. To use the Web as a single global database, various research challenges: user interfaces and interaction paradigms, application architectures, schema mapping and data fusion, link maintenance, licensing, trust, quality, and relevance - are to be addressed.

In "*Community-Driven Consolidated Linked Data*," Shakya, Takeda, and Wuwongse propose an approach to enable people to share various data using easy-to-use social platform. The work has implemented social software, called StYLiD. It allows users with multiple perspectives to share various types of structured linked data and derive ontologies to provide online social platform to be used by ordinary people. Users have freedom to define their own concepts. StYLiD consolidates multiple schemas by mapping these schemas semi-automatically with the help of schema alignment techniques. Concepts are grouped semi-automatically based on proposed algorithm to calculate schema similarity. It generates informal ontologies to combine multiple perspectives and unify common elements. StYLiD is built upon Pligg - a Web 2.0 content management system and experiments are performed based on all user-defined schemas definitions or types, retrieved from Freebase.

"*Searching Linked Objects with Falcons: Approach, Implementation and Evaluation*" by Cheng and Qu presents a keyword-based search engine for linked objects called Falcon Object Search. For each object, it constructs comprehensive virtual document consisting of textual descriptions extracted from RDF description of an object. It builds inverted index based on terms in virtual documents. To execute keyword-based query, the system uses inverted index and compares the terms in the query with the virtual documents of objects to generate result set. The objects of result set are ranked by considering their relevance to the query and their popularity. For each resulting object, a query-relevant structured snippet is provided to show the associated literals and linked objects matched with the keyword query. The concept of PD-thread is used as the basic unit, a snippet. The method of ranking PD-threads into a snippet is devised. Type information of objects is expanded by executing class-inclusion reasoning over

descriptions of classes to implement class-based query refinement. The system recommends subclasses to allow navigation of class hierarchies to perform incremental result filtering.

In "*A URI is Worth a Thousand Tags: From Tagging to Linked Data with MOAT*," Passant, Laublet, Breslin, and Decker demonstrate how Web 2.0 content and linked data principles could be combined in order to solve issues of free-tagging systems, like ambiguity and heterogeneity of tags. It proposes MOAT ontology, based on quadripartite tagging model, in which each tag can be represented by a quadruple (<User>, <Resource>, <Tag>, <MeaningURI>). It helps to assign tags of choice to a resource while using the huge amount of authoritative URIs from the Web of data to narrow down the intended meaning.

In "*An Idea Ontology for Innovation Management*," Riedl, May, Finzen, Stathel, Kaufman, and Krcmar make an attempt to represent ideas using an ontology. It is difficult to obtain an accurate and formal definition of idea. The ontology is based on OWL, and it provides a common language to support interoperability between innovation tools to support full life cycle of an idea in an open innovation environment. This work defines its own definition of idea, and based on the detailed analysis of innovation management domain, ontology is designed. The ontology is aimed to capture the core concept of idea to support collaborative idea development, rating, discussing, tagging, and grouping of ideas in an open innovation environment.

In "*Inductive Classification of Semantically Annotated Resources through Reduced Coulomb Energy Networks*," Fanizzi, d'Amato, and Esposito propose an interesting method to induce classifiers from ontology to perform concept retrieval. Induced classifier can determine likelihood measure of the induced class-membership assertions to perform approximate query answering and ranking. The work proposes to use instance-based classifier to answer queries based on a non-parametric learning scheme; the Reduced Coulomb Energy (RCE) Network. The work extends classification algorithm using RCE networks based on entropic similarity measure for OWL. Experiments are performed to execute approximate query answering on a number of ontologies from public repositories. Results show induction classification to be competitive with reference to the deductive methods and are able to detect new knowledge assertions, which are not logically derivable.

In "*A Comparison of Corpus-Based and Structural Methods on Approximation of Semantic Relatedness in Ontologies*," Ruotsalo and Mäkelä compare the performance of corpus-based and structural approaches to determine semantic relatedness in light-weight ontologies. The work identifies the strength and weaknesses of the methods in various application scenarios. The experimental results show that neither corpus-based method nor structure-based measures is efficient and competitive. Latent Semantic Analysis (LSA) produces the best performance for the whole dataset. Structural measures produce better performance compare to LSA when cut-off values were applied. The performance of compared methods varies in case of different rank levels. LSA is found to be efficient in filtering out the non-relevant relations, and is able to find relations whereas structural measures fail. The work suggests using a combination of corpus-based methods and structural methods and identification of appropriate cut-off values based on the intended use case(s).

Amit Sheth
Wright State University, USA

REFERENCES

Baker, C. J. O., & Cheung, K.-H. (Eds.). (2007). *Semantic Web: Revolutionizing knowledge discovery in the life sciences*. Springer.

Brammer, M., & Terziyan, V. (Eds.). (2008). *Industrial applications of Semantic Web*. New York, NY: Springer-Verlag.

Cardoso, J., Hepp, M., & Miltiadis, D. (Eds.). (2008). *The Semantic Web, real-world applications from industry*. Springer.

Herman, I. (2009). What is being done today? Presentation given Deutsche Telekom, Darmstadt, Germany, December 14, 2009.

Mentzas, G. (2007). *Knowledge and semantic technologies for agile and adaptive e-government.* 7th Global Forum on Reinventing Government: Building Trust in Government, June 26-29, 2007.

Ruttenberg, A., Rees, J. A., Samwald, M., & Marshall, M. S. (2009). Life sciences on the Semantic Web: The Neurocommons and beyond. *Briefings in Bioinformatics*, *10*(2), 193–204. .doi:10.1093/bib/bbp004

Sheth, A. (2000). *Semantic Web & information brokering: Opportunities, commercialization and challenges*. Keynote talk at the International Workshop on Semantic Web: Models, Architecture and Management, Lisbon, Portugal, September 21, 2000.

Sheth, A. (2005). *Enterprise applications of Semantic Web: The sweet spot of risk and compliance*. IFIP International Conference on Industrial Applications of Semantic Web (IASW2005), Jyvaskyla, Finland, August 25–27, 2005.

Sheth, A., Avant, D., & Bertram, C. (2001). *System and method for creating a Semantic Web and its applications in browsing, searching, profiling, personalization and advertising.* (United States patent Number - 6311194), Taalee, Inc. Oct 30, 2001.

Sheth, A. P., Agrawal, S., Lathem, J., Oldham, N., & Wingate, Y. H.P., & Gallagher, K. (2006). *Active semantic electronic medical record*. 5th International Semantic Web Conference, Athens, GA, November 6–9, 2006.

Sheth, A. P., & Stephens, S. (2007). *Semantic Web: Technologies and applications for the real- world*. 16th World Wide Web Conference (WWW2007), Banff, Canada, May 8-12, 2007

Wikipedia contributors. (2010). *Technology adoption lifecycle*. Retrieved January 6, 2011, from http://en.wikipedia.org/w/index.php?title=Technology_adoption_lifecycle&oldid=386064217

ENDNOTES

1 Data from http://www.w3.org/2001/sw/sweo/public/UseCases/accessed in December 2010.
2 Data from http://www.w3.org/2005/04/swls/accessed in December 2010.
3 Data from http://esw.w3.org/CommercialProducts accessed in December 2010.