# Semantic Web Technologies: Overview

## History, Semantic Web Layers

# Rapid growth of information on the WWW – challenges and problems

- **Information overload**

In 1998, the size of the Web was estimated to exceed 300 million pages with a growth rate of about 20 million per month. The Google search index was measured around 500 million pages in 2000, 8 billion in 2004, and more than 27 billion in 2012.

While the early Web often suffered from a lack of high-quality relevant pages, the present Web now contains far too many relevant pages for any user to review. As an example, today (February 25, 2015) Google is returning about 43.8 million pages for the “World Wide Web” search phrase.

- **Poor retrieval and aggregation**

The explosion of Web documents and services would not be so critical if users could easily retrieve and combine the information needed.

Web documents are at best semi-structured in simple natural language text, therefore they are vulnerable to obstacles that prevent efficient content retrieval and aggregation.

An increasing problem is the number of languages used on the Web. Almost 65% of Web pages were in English in 1999. Data from Internet World Stats (www.internetworldstats.com) indicate a more balanced use of languages. The English using population at the end of 2009 constituted only 27.7% of total online users.

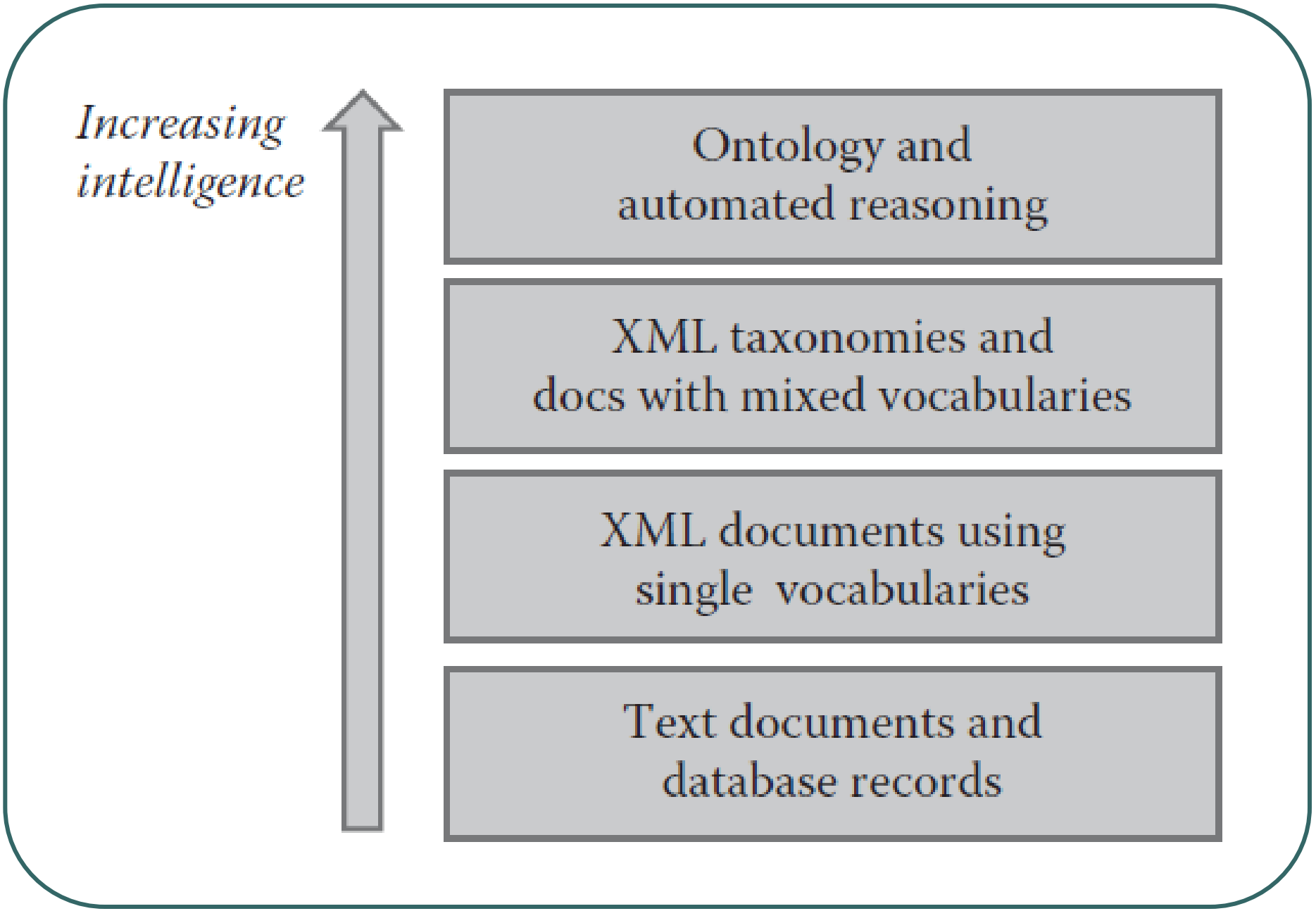
## History

The *Semantic Web* term was popularized by Tim Berners-Lee and later elaborated in 2001. The first part of his vision for the Semantic Web was to turn the Web into a truly collaborative medium – to help people share information and services and make it easier to aggregate data from different sources and different formats.

The second part of his vision was to create a Web that would be understandable and processable by machines. While humans can read and comprehend current Web pages, Berners-Lee envisioned new forms of Web pages that could be understood, combined, and analyzed by computers.

The ultimate goal was to enable humans and computers to cooperate in the same manner as humans do among each other.

Central to the Semantic Web vision is the shift from applications to data. The key to machine-processable data is to make the data smarter.


Increasing intelligence
Ontology and automated reasoning
XML taxonomies and docs with mixed vocabularies
XML documents using single vocabularies
Text documents and database records

- **Text and databases**

In this initial stage, most data is proprietary to an application. The application is responsible for interpreting the data and contains the intelligence of the system.

- **XML documents for single domains**

The second stage involves domain-specific XML schemas that achieve application independence within the domain. Data can flow between applications in a single domain but cannot be shared outside the system.

- **Taxonomies**

In this stage, data can be combined from different domains using hierarchical taxonomies of the relevant terminologies. Data is now smart enough to be easily discovered and combined with other data.

- **Ontologies and automated reasoning**

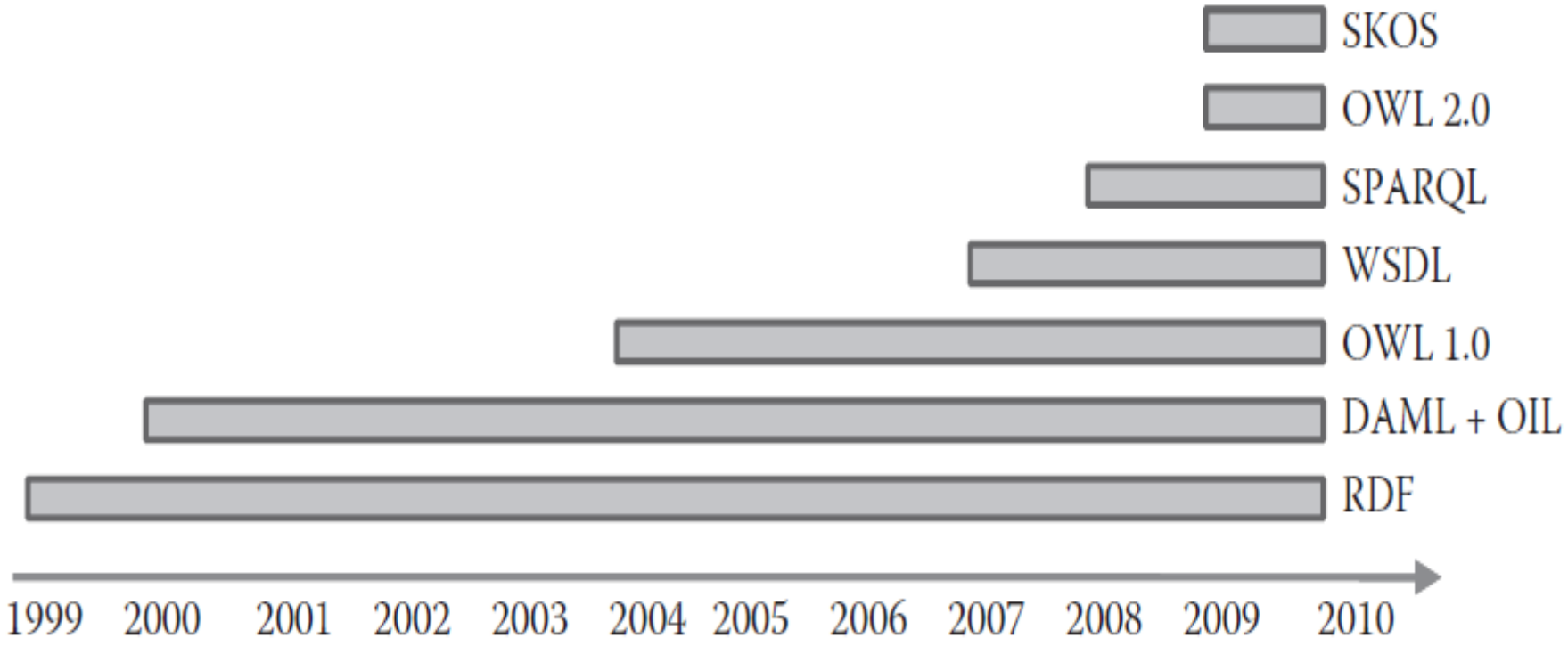
In the final stage, new data can be inferred from existing data and shared across applications with no human involvement or interpretation. Data is now smart enough to understand its definitions and relationships to other data.

In the Semantic Web these smart data are assumed to be application-independent, composable, classified, and comprise parts of a larger terminological structure.

Ontologies play a very important role in the Semantic Web community.

For the Semantic Web, ontologies enable us to define the terminology used to represent and share data within a domain. As long as the applications define their data with reference to the same ontology, they can interpret and reason others' data and collaborate without manually defining any mapping between the applications.

# Language Standards of Semantic Web



- The Web Service Definition Language (WSDL) is one of many languages for specifying Web services
- SPARQL is an RDF-based query language for accessing information in ontologies

- SKOS (Simple Knowledge Organization System) is a light weight data model for sharing and linking knowledge organization systems via the Web

# Semantic Web Layers

The assortment of tools, technologies, and specifications that lay the foundation for the Semantic Web can be broadly organized into four major layers: (1) data and metadata, (2) semantics, (3) enabling technology, and (4) environment.

| | | | |
|---|---|---|---|
| **Environment Layer** | • Security<br>• Privacy<br>• Trust | • Cryptography<br>• Integration<br>• Standardization | • Peer-to-Peer<br>• Semantic Grid<br>• Social Network |
| **Enabling Technology Layer** | • Agents<br>• Search<br>• Web Services | • Composition<br>• Visualization | • Personalization<br>• Repository Management<br>• Natural Language Processing |
| **Semantics Layer** | • Ontologies (OWL)<br>• Rules (RIF/RuleML/SWRL)<br>• Queries (SPARQL) | | • Logic (First Order, DL)<br>• Reasoning<br>• Trust |
| **Data and Metadata Layer** | • RDF and RDF Schema<br>• XML<br>• Unicode and URI | | |

- **Data and Metadata Layer**

The data and metadata layer is the lowest; it provides standard representations for data and facilitates the exchanges among various applications and systems.

The Unicode provides a standard representation for character sets in different languages used by different computers and the URI provides a standard way to uniformly identify resources such as Web pages and other forms of content.

The Unicode and URI together enable us to create content and make these resources available for others to find and use in a simple way. XML enables us to structure data using user-defined tags that have well defined meanings that are shared by applications. This helps improve data interoperability across systems.

Namespaces and schemas provide the mechanisms to express semantics in one location for access and utilization by many applications.

The next component in this layer is the Resource Description framework (RDF) that conceptually describes the information contained in a Web resource.

It can employ different formats for representing triplets (subjects, predicates, objects), can be used to model abstract concepts, and is effective for knowledge management.

RDF Schema is a language for declaring basic classes and types for describing the terms used. It supports reasoning to infer different types of resources.

- **Semantics Layer**

The semantics layer incorporates specifications, tools, and techniques that help add meaning or semantics to characterize the contents of resources. It facilitates the representation of Web content that enables applications to access information autonomously using common search terms.

The important ingredients of this layer are ontology language, rule language, query language, logic, reasoning mechanism, and trust.

OWL is the most popular ontology language used by applications for processing content from a resource without human intervention.

Thus, it facilitates machine interoperability by providing the necessary vocabulary along with formal semantics.

Rule languages help write inferencing rules in a standard way that can be used for reasoning in a particular domain.

A rule language provides kernel specification for rule structures that can be used for rule interchange and facilitates rule reuse and extension.

Among several standards, such as RIF (Rule Interchange Format), and Datalog RuleML, SWRL (Semantic Web Rule Language) is gaining popularity.

Querying Web content and automatically retrieving relevant segments from a resource by an application is the driving force behind the SPARQL Web query language. It provides both a protocol and a language for querying RDF graphs via pattern matching.

Logic and reasoning are also integral parts of the semantics layer. A reasoning system can use one or more ontologies and make new inferences based on the content of a particular resource. It also helps identify appropriate resources that meet a particular requirement.

Thus, the reasoning system enables applications to extract appropriate information from various resources.

Logic provides the theoretical underpinning required for reasoning and deduction.

First order logic, description logic, and others are commonly used to support reasoning.

- **Enabling Technology Layer**

This layer consists of a variety of technologies that can develop applications on the Semantic Web and accomplish different types of tasks or operationalize specific aspects of the Semantic Web.

For example, intelligent agents or multiagent systems can be used to access and process information automatically on the Semantic Web.

Some well-established technologies can be used synergistically to create valuable Semantic Web applications.

Some of the technologies relevant to this layer are agents, search, Web services, composition (information and service composition), visualization, personalization, repository management, and natural language processing.

- **Environment Layer**

The environment layer deals with the surroundings and the infrastructure in which the Semantic Web applications execute and meet the basic expectations of these applications in terms of data quality and information assurance.

It is also concerned with the operating environment and the degrees of interoperability of various domains. Some of the key aspects of this layer are security, privacy, trust, cryptography, application integration, standards, and environments such as peer-to-peer, semantic grid, and social networks.

# Future Research Directions

- **Ontology learning or creation**

These techniques allow semi-automatic or fully automatic creation of (parts of) ontologies from representative domain texts.

- **Performance**

With current technologies, retrieval, storing, and manipulation of ontologies are computationally too demanding for many large-scale applications. Initial work on more efficient methods for storing and reasoning over complex ontologies has started, but more progress is needed for semantic applications to scale up.

- **Ontology quality and selection**

As the number of available ontologies increases, their evaluation becomes more difficult. Existing approaches focus on the syntactic aspects of ontologies and do not take into account the semantic aspects and user contexts and familiarities.

While many research efforts address the issues of ontology searching and quality separately, none has considered ontology evaluation and selection together. We still have a great need for developing a semi-automatic framework for selecting the best ontology appropriate for a specific task within the Semantic Web.

- **Linked data**

In 2006, Tim Berners-Lee introduced the notion of linked data as a simplified approach to semantic applications. The approach is based on concepts and technologies for combining and integrating data using RDF triplets only.

Linked data allows more scalable applications to be built and has already been used in a number of small Web applications and enterprise data architectures. Whether the approach can handle functionally demanding tasks due to the limited expressiveness and lack of formality remains unclear.

- **Trust, Security, and Privacy**

Semantic Web applications assume and expect that the information content of resources is of high quality and can be trusted.

Similarly, security and privacy of sensitive information on the Semantic Web must be ensured.