Chapter 2

The Semantic Web

This chapter embeds the present thesis in its broader context of related research. It introduces the Semantic Web, which is an extension of the current World Wide Web, and is intended as global-scale collection of machine-readable statements. The chapter introduces the original visions regarding the Semantic Web and complements the visions with some considerations about its current status. It also discusses the characteristics and features of Semantic Web applications, and lists some commercial and academic projects.

2.1 Overview

The Semantic Web promises to solve some of the problems that exist regarding the current Web. Section 2.1.1 gives an overview of the basic ideas, design goals and the current status of the endeavor. Section 2.1.2 introduces features that distinguish the Semantic Web from the current Web as well as from traditional knowledge-based systems, for example how intelligent behavior emerges on the Semantic Web. Popular misconceptions about the Semantic Web presented in Section 2.1.3 help to clarify the concepts involved.

2.1.1 Background and Vision

While being the result of an unprecedented success story, the current Web is often inconsistent, disconnected, and out of sync. It feels like it is “a mile wide, but only an inch deep” [8, p 10]. An update of a bit of information in one place leaves the other places untouched, causing inconsistency. That is one of the reasons why many modern websites rely on relational database systems to generate website content on the fly. Database normalization tech-
Techniques provide consistency locally, but databases usually do not integrate with third-party websites. Content on Web pages is made for human consumption, it is stored as HTML, or in application dependent formats such as files created with office programs. Distributed systems can hardly process and integrate content with other systems automatically, which leads to disconnectedness, and sometimes to frustrating effects for the user. Information, for example address data, has no explicit representation that can be processed by a machine, and therefore the consumer has to manually transform and transfer the content to use it in another service.

The Semantic Web in contrast is designed to provide a layer that makes smarter applications perform to their potential. Data in the Semantic Web is intended to be modeled and described in a way that makes it possible to integrate it on a global level – a “Web of data” [85] or a “Web of actionable information” [168, p 96]. As an example, if two different companies producing computer parts and exposing their product data on the web, a third company or service should be able to automatically understand and use that data [58]. “The main idea of the Semantic Web is to support a distributed Web at the level of data rather than at the level of presentation. Instead of having one webpage point to another, data items point to one another using global references called Uniform Resource Identifiers (URI).” [8] The applications or underlying database systems no longer hold the coherent data model used by the applications themselves, but is part of the Web infrastructure. The data items on the Semantic Web are described in a machine-readable, distributable way upon a single and distributed data model – making the Web less dumb.

Van Harmelen [185], referring to Marshall and Shipman in [115], presents a more diversified view on the Semantic Web – he distinguishes two types of goals: (i) In the first interpretation the Semantic Web aims towards the integration of structured and semi-structured data sources over the Web in order to federate and re-use those data sets. (ii) The second interpretation focuses on the enhancement of the current Web content with additional semantic metadata – where techniques such as concept extraction, named-entity recognition, automatic classification extract the metadata automatically. These conflicting assumptions also lead to some of the fallacies and criticisms about the Semantic Web presented below in the paragraph Misconceptions and Criticism. But a central aspect, which both interpretations agree on, is that the Semantic Web is a global-scale collection of formal, ontology-based and machine-readable statements about Web resources and other entities.
2.1. OVERVIEW

The World Wide Web Consortium's (W3C) Semantic Web working groups are the major force supplying the Semantic Web's vision as well as its design principles, formal specifications and enabling technologies. Those specifications include work on RDF, RDFS, OWL, SPARQL and others - these technologies will be covered in Sections 3.2 to 3.4.

The Semantic Web is also known under the names Deep Web, Smart Web or sometimes as Web 3.0, although Tim Berners Lee described it as a part of Web 3.0: "I think maybe when you've got an overlay of scalable vector graphics - everything rippling and folding and looking misty - on Web 2.0 and access to a Semantic Web integrated across a huge space of data, you'll have access to an unbelievable data resource." [169]

Furthermore, the Semantic Web is a way to tackle a traditional problem in Artificial Intelligence (AI) research: the so-called knowledge acquisition bottleneck [55]. The knowledge acquisition bottleneck is concerned with the difficulty of the acquisition, representation and maintenance of an intelligent system's knowledge base [47]. Some people see it from a rather epistemological view, as the difficulties in formalizing knowledge to make it processable for machines, but pragmatically it is more an economic problem: the cost of acquiring and maintaining a knowledge base must be less than the economic benefits derived from the system. The knowledge acquisition bottleneck is a crucial problem in AI research, because after a phase focusing on general methods for problem solving and efficient theorem providing in the mid 1970s [47], many in the community realized that the fundamental problem of understanding intelligence is how to represent large amounts of knowledge in a way that permits their effective use [70]. Over the last 20 years more researchers developed robust and cost-effective knowledge-engineering processes, including technologies for specifying reusable model components (ontologies) and reasoning components - which have a strong influence on current Semantic Web components. The Semantic Web has a strong connection to AI research, but its key advocates argue that it Web is not AI. AI is concerned with engineering intelligent machines, while the Semantic Web is a technological infrastructure to enable large scale data interoperability [47]. Compared to classical knowledge-based systems with their closed domains it could open a way to new intelligent applications exploiting the large-scale and distributed knowledge supplied by the infrastructure.

Although much effort has been invested in tools and technologies, especially the formalisms, standards and languages provided by the W3C, with the effect that those technologies are quite mature nowadays, d’Aquin et al. [47] state that the Semantic Web, from an applicational point of view,
is still in an "embryonic state". The reason is that most of the existing applications only consume their own data, rather than the Semantic Web as a large scale information source. Motta and Sabou [123] present a number of criteria, which applications ought to satisfy to move away from this "First Generation" of Semantic Web applications to a new generation, most of which will be discussed in the next paragraph. According to Lee and Goodwin [101], the Semantic Web is mirroring the growth of the Web in the early nineties – and indicator that a large-scale adoption will become reality sooner or later. The Semantic Web search engine Swoogle\(^2\), for example, found 705,406,123 triples of semantic data as of April 2010. Similar numbers from other Semantic Web search engines also indicate that there already exists a useful knowledge source for intelligent applications. Several projects work on bringing more data online in order to increase its usefulness and applicability, most prominently the global Linking Open Data community project.\(^3\) Collectively, their data set consist of more than 13.1 billion RDF triples (April 2010).

### 2.1.2 Features

Viewing the Semantic Web as a very large Knowledge-Based System (KBS), d'Aquin et al. [47] present several key differences between classical KBS and the Semantic Web in the areas of (i) heterogeneity, (ii) quality, (iii) scale and (iv) reasoning. (i) KBS are built around small sets of carefully designed and integrated ontologies, whereas the Semantic Web makes the non-trivial effort of integrating very heterogeneous ontologies necessary, heterogeneous in terms of ontology encoding, quality, complexity, modeling and views. (ii) In KBS a small team of knowledge engineers builds ontologies in a centralized fashion. On the Semantic Web information stems from different sources and strongly varies in quality – so trust is a key issue. (iii) The Semantic Web with its millions of documents and billions of triples calls for a totally new way to locate and process data. (iv) Instead of sophisticated reasoning mechanisms used on generic tasks as applied by traditional KBS, Semantic Web applications rather draw their intelligence from scale, i.e. the sheer amount of data available.

The *Network effects* create a virtuous cycle[8] of content creation: The more people participate and put information or data online, the more attractive it is for new people to join. Metcalfè's law [170] describes this observation more formally. Another feature is the so-called *Open World Assumption*,

\(^2\)[http://swoogle.umbc.edu](http://swoogle.umbc.edu), Statistics from 2009-06-27

\(^3\)[http://linkeddata.org](http://linkeddata.org)
2.1. OVERVIEW

which implies that at any point new information can come to light, and that no conclusion may be drawn relying on the fact that the information available at a point is all information existing. The Non-unique Naming, the final feature, describes the fact that some Web resources may be referred to using different names by different people – so distinct URIs need not refer to distinct resources.

2.1.3 Misconceptions and Criticism

Van Harmelen [185] lists four popular fallacies or misconceptions about the Semantic Web. The first one is that the Semantic Web, or its standards, enforce meaning from the top onto users with formalisms such as OWL. Van Harmelen counters that those standards are there for users to express their own meaning freely, and that they can assign their meaning to terms in vocabularies. Fallacy number two refers to the popular opinion that the Semantic Web requires everybody to conform to a single predefined meaning of terms – but in fact the motto is rather “let a thousand ontologies bloom”. This is also a reason why much research effort is invested in the area of ontology mapping (see Section 3.1.3). The third fallacy is that the Semantic Web requires users to understand the details of formalized knowledge representation. Although the details of ontology languages are complicated matters, not every user need to known them, as a user doesn’t have to know HTML or CSS to navigate the current Web. The last of the misconceptions is that the Semantic Web people will demand the manual markup of all existing Web pages. The Semantic Web relies on automation of large-scale markup extraction from current Web representations, mostly with lightweight semantics. Many modern Web applications address this issue by creating annotations in machine-readable formats upon the publishing of data, for example as microformats.4

Alani et al. [5] present a few misconceptions about the Semantic Web from the viewpoint of adoption and application of its technologies in organizations, some of which overlap with the fallacies described in the paragraph above. The misconception that ontologies are typically large and complex, and that they are expensive to design, build and maintain is countered with the argument that applications don’t always require heavyweight and complex ontologies of domain knowledge, but that lightweight ontologies often suffice. Lightweight ontologies (see Section 3.1) can have a wide applicability, and they are cost effective to build in terms of overall utility to the community. Some decision makers worry that existing data has to be expen-

sively converted to Semantic Web formats, and current technologies replaced. However, simple scripts or conversion languages can often automatically accomplish the conversion – data is kept in the current format and exported when needed. Many organizations suspect that providing public access to their data only benefits the public; but as the current document Web has shown, there are economic gains for the owners of information, too. The last cause for worries covered by Alani et al. [5] is the fear that the promiscuous release of data and information will be a privacy nightmare. In fact there are standards being developed for access control, and in the meantime, as with conventional database and Web technologies, organizations can choose which data they share publicly.

Peter Gärdenfors [65] criticizes the Semantic Web effort on a different level. Relating to arguments by Shirky [171] he states that the Semantic Web with its “neat ontologies and syllogistic logic” is not that effective in the real world where a shared world view is hard to create. Reducing semantic content to first order logic or set theory, he doubts that Web Ontology Language (OWL) can express important notions like similarity in a natural way. He also refers to the symbol grounding problem [76], which is about the concern how a symbolic expression can obtain any meaning that goes beyond the formal language itself – and be grounded in the external world in terms of meaning. Gärdenfors argues that John Locke already brought up this problem in the year 1690 in his Essay Concerning Human Understanding [106] where he described the difficulty of agreeing on the precise number of simple ideas belonging to any sort of thing, or its qualities.

2.2 Applications

The later parts of this section discuss some applications that make extensive use of Semantic Web technologies. If applications also integrate the massive amounts of Semantic Web data and documents that are available on the Internet, then d’Aquin et al. [47] call them “next generation Semantic Web applications”. Section 2.1.2 describes the set of features that distinguish next generation applications from classical knowledge-based systems. Because the Semantic Web combines heterogeneous sources, variable data quality and global-scale distributed data, those applications will derive their intelligent behavior rather from the capability to exploit large amounts of data than from complex inferencing – intelligence comes as a side effect of scale. Other types of reasoning, partly on non-semantic data, become crucial: reasoning based on machine learning and on linguistic and statistical techniques. In contrast to next generation Semantic Web applications the first generation
2.2. APPLICATIONS

typically uses just a single ontology that supports the integration of a set of

data sources fixed at design time.

D'Aquin et al. [47] present the features of next generation applications:

(i) The application needs to be able to find relevant information on the Web

for the task at hand dynamically. (ii) The application has to select appro-

priate information (in terms of quality, etc.) from the documents found in

(i). (iii) As the application must be able to exploit heterogeneous knowledge

sources, it cannot make assumptions about the ontological nature of target

information. (iv). Ontologies and resources must be combined – as it cannot

be expected that one single source provides all necessary information. To be

able to leverage the power of online semantics, it is crucial to have a single

access point to the data. This access point collects, analyzes, and indexes Se-

mantic Web data and provides it to the applications. As current access points

such as Swoogle\textsuperscript{5} and Sindice\textsuperscript{6} have limitations, d’Aquin et al. developed

Watson [48] as a new Semantic Web gateway to provide mechanisms for ex-

tracting semantic documents with keyword search, retrieving their metadata,

and querying the content (e.g. with SPARQL). “Watson offers applications

all the necessary elements to select and exploit online semantic resources”.

Among the applications that build on the Watson gateway are PowerMagpie,

PowerAqua and Scarlet [48]. PowerMagpie helps users to interpret arbitrary

Web content by extracting and summarizing important conceptual entities

relevant to a page, it highlights those entities and puts them in context with

dynamically retrieved ontologies. PowerAqua is a question-answering sys-
tem based on an unlimited number of ontologies, which is able to combine

various ontologies at runtime. Scarlet explores ontologies to automatically

retrieve relations between two input concepts – Scarlet will be discussed in

more detail in Section 4.3, as it is integrated into the system developed for

the present thesis.

Corporations still use Semantic Web applications quite rarely, Alani et

al. [5] state that “it’s probably fair to say that many organizations still

view the Semantic Web with some scepticism. In part, they may suspect

that they’re expected to pioneer an approach in which quick wins are few”.

Furthermore, they worry about cost and privacy issues when linking ever-

increasing amounts of data to the Web. Some of the misconceptions have

already been addressed in Section 2.1.3, Alani et al.\textsuperscript{[5]} analyze the special

characteristics of using Semantic Web technologies in corporations. They

argue, that it offers local and private gains indeed for individuals and or-

ganizations that link their data and information. Some of the factors to

\textsuperscript{5}\url{http://swoogle.umbc.edu}

\textsuperscript{6}\url{http://sindice.com}
make the deployment of Semantic Web technologies attractive are: Minimize disruption to existing infrastructure, e.g. gradually convert existing data to Semantic Web formats with simple scripts. Use small, well-focused ontologies for individual information assets to keep efforts of ontology development low. Show the added value gained by integration and shared access, for example consistency checking, and provide relative ease of integration and efficient data exchange and merging.

Already in 2006 van Harmelen [185] observed a shift in company profiles that are active in the Semantic Web field from small start-ups to big corporations. He lists the following areas where respective technologies begin to take shape: knowledge management, mostly for intranets of big corporations; data-integration (e.g. at Boeing); e-Science, esp. life sciences; convergence of the Semantic Grid.

This overview of some of the aspects of Semantic Web applications concludes with a few examples of current Semantic Web applications. Siri\(^7\) is a personal assistant for the mobile phone capable of doing simple assistance jobs and answer questions such as “Where is the nearest shop?”, or to execute commands like “I need a cab”. Siri is born out of SRI’s CALO Project, the largest Artificial Intelligence project in U.S. history (according to the Siri Web site). The ambitious vision is that in the next five years almost everyone with a connected lifestyle will delegate details of day-to-day tasks to intelligent assistants, which coordinate and simplify the details of their lives. True Knowledge\(^8\) provides a question-answering system to respond to questions in any domain. It has a search engine-like natural language user interface. The application aims at giving instant and precise answers to questions – as opposed to current Web search engines, which just return a long list of possibly related documents. True Knowledge relies on Semantic Web technologies to answer complex questions by drawing inferences and conclusions on that data. Wolfram\(|\)Alpha\(^9\) is another question-answering system. It relies on a formal Mathematica representation at its heart. Wolfram\(|\)Alpha mostly depends on its own data and does not apply Semantic Web technologies or data extensively, and therefore is no Semantic Web application in the narrow sense. TripIt\(^10\) automatically organizes all of a user’s travel information into a master travel itinerary that is easy to share and access. The master itinerary aggregates a lot of travel-related information in one place. Twine\(^11\) is a service to track, find and share content. Twine uses

\(^{7}\)http://www.siri.com
\(^{8}\)http://www.trueknowledge.com
\(^{9}\)http://www.wolframalpha.com
\(^{10}\)http://www.tripit.com
\(^{11}\)http://www.twine.com
Semantic Web technology to help people organize, disseminate and discover information related to their interests. The application stores information as RDF triples and makes them accessible via the Twine APIs. TopQuadrant\textsuperscript{12} supports companies in moving from disparate data into integrated, actionable and reusable knowledge, using the product TopBraid Suite, which is a set of components for semantic solutions. The SemanticMiner is one of the products created by ontoprise.\textsuperscript{13} This application provides semantic search capabilities for companies. Leveraging the power of ontologies, this product supports moderated search, the optimization of search results and also gives an integrated view on heterogeneous sources of data and information.

\textsuperscript{12}http://www.topquadrant.com
\textsuperscript{13}http://www.ontoprise.de