CHAPTER 1

Introduction to the Semantic Web

The content of conventional web sites is human-readable only, which is unsuitable for automatic processing and inefficient when searching for related information. Web datasets can be considered as isolated data silos that are not linked to each other. This limitation can be addressed by organizing and publishing data, using powerful formats that add structure and meaning to the content of web pages and link related data to one another. Computers can "understand" such data better, which can be useful for task automation.

The Semantic Web

While binary files often contain machine-readable metadata, such as the shutter speed in a JPEG image[1] or the album title in an MP3 music file, the textual content of traditional web sites cannot be interpreted (that is, not understood) by automated software agents. The web sites that provide semantics (meaning) to software agents form the *Semantic Web*, an extension of the conventional Web [1] introduced in the early 2000s [2]. The Semantic Web is a major aspect of Web 2.0 [3] and Web 3.0 [4]. *Web 2.0* is an umbrella term used for a collection of technologies behind instant messaging, Voice over IP, wikis, blogs, forums, social media portals, and web syndication. The next generation of the Web is denoted as *Web 3.0*, which is an umbrella term for customization, semantic contents, and more sophisticated web applications toward artificial intelligence, including computer-generated contents (see Figure 1-1) .

Caution The word *semantic* is used on the Web in other contexts as well. For example, in HTML5 there are semantic (in other words, meaningful) structuring elements, but this expression refers to the "meaning" of elements. In this context, the word *semantic* contrasts the "meaning" of elements, such as that of `section` (a thematic grouping), with the generic elements of older HTML versions, such as the "meaningless" `div`. The semantics of markup elements should not be confused with the semantics (in other words, machine-processability) of metadata annotations and web ontologies used on the Semantic Web. The latter can provide far more sophisticated data than the meaning of a markup element.

[1] Exif or XMP. For more information, see Leslie Sikos: *Web Standards: Mastering HTML5, CSS3, and XML* (New York, Apress, 2014).

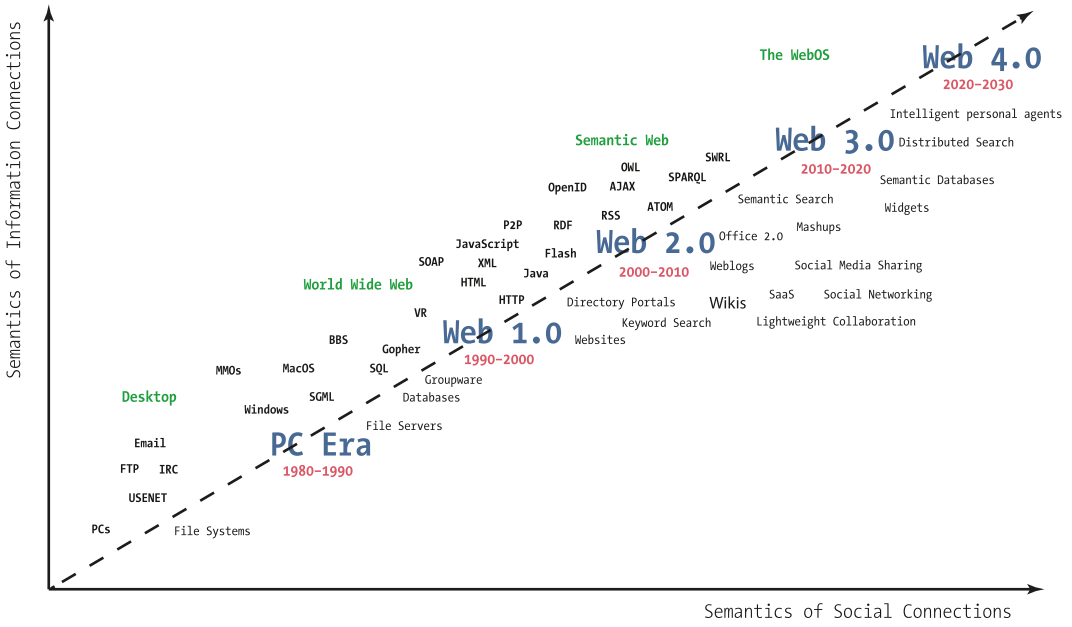

Figure 1-1. *The evolution of the Web [5]*

In contrast to the conventional Web (the "Web of documents"), the Semantic Web includes the "Web of Data" [6], which connects "things"[2] (representing real-world humans and objects) rather than documents meaningless to computers. The machine-readable datasets of the Semantic Web are used in a variety of web services [7], such as search engines, data integration, resource discovery and classification, cataloging, intelligent software agents, content rating, and intellectual property right descriptions [8], museum portals [9], community sites [10], podcasting [11], Big Data processing [12], business process modeling [13], and medical research. On the Semantic Web, data can be retrieved from seemingly unrelated fields automatically, in order to combine them, find relations, and make discoveries [14].

Structured Data

Conventional web sites rely on markup languages for document structure, style sheets for appearance, and scripts for behavior, but the content is human-readable only. When searching for "Jaguar" on the Web, for example, traditional search engine algorithms cannot always tell the difference between the British luxury car and the South American predator (Figure 1-2).

[2] The concept of "thing" is used in other contexts as well, such as in the "Internet of Things" (IoT), which is the network of physical objects embedded with electronics, software, and sensors, including smart objects such as wearable computers, all of which are connected to the manufacturer and/or the operator, and/or other devices.

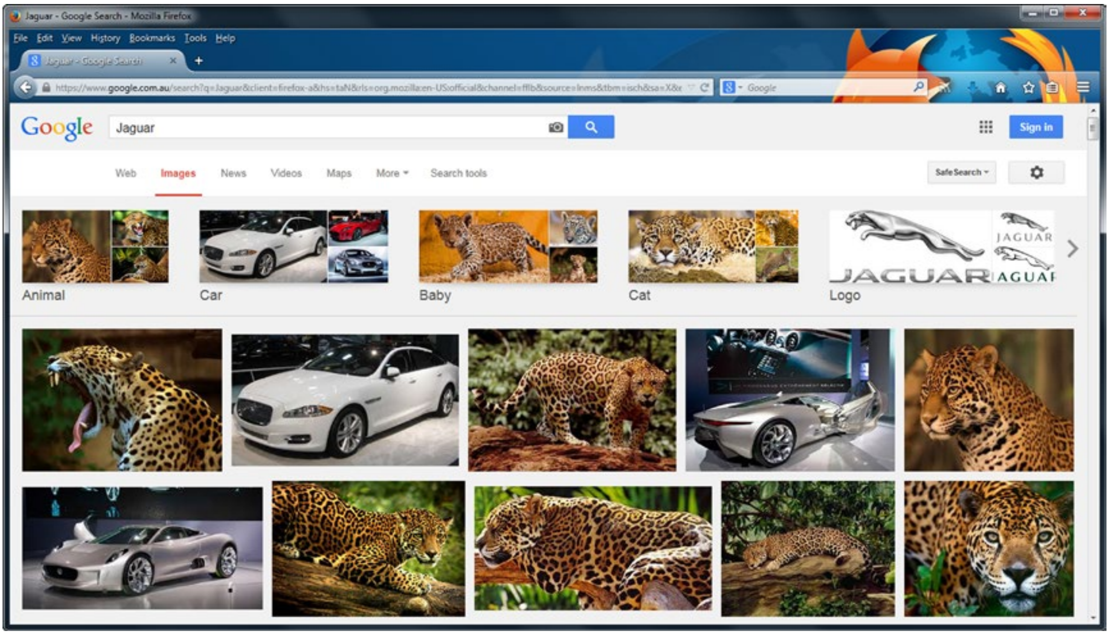
Jaguar - Google Search - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Jaguar - Google Search
https://www.google.com.au/search?q=Jaguar&client=firefox-a&hs=taN&rls=org.mozilla:en-US:official&channel=fflb&source=lnms&tbm=isch&sa=X&
Google
Google
Jaguar
Sign in
Web
Images
News
Videos
Maps
More
Search tools
SafeSearch
Animal
Car
Baby
Cat
Logo
JAGUAR

Figure 1-2. *Traditional web search algorithms rely heavily on context and file names*

A typical web page contains structuring elements, formatted text, and some even multimedia objects. By default, the headings, texts, links, and other web site components created by the web designer are meaningless to computers. While browsers can display web documents based on the markup, only the human mind can interpret the meaning of information, so there is a huge gap between what computers and humans understand (see Figure 1-3). Even if alternate text is specified for images (`alt` attribute with descriptive value on the `img` or `figure`[3] elements), the data is not structured or linked to related data, and human-readable words of conventional web page paragraphs are not associated with any particular software syntax or structure. Without context, the information provided by web sites can be ambiguous to search engines.

[3] This is only supported in (X)HTML5.

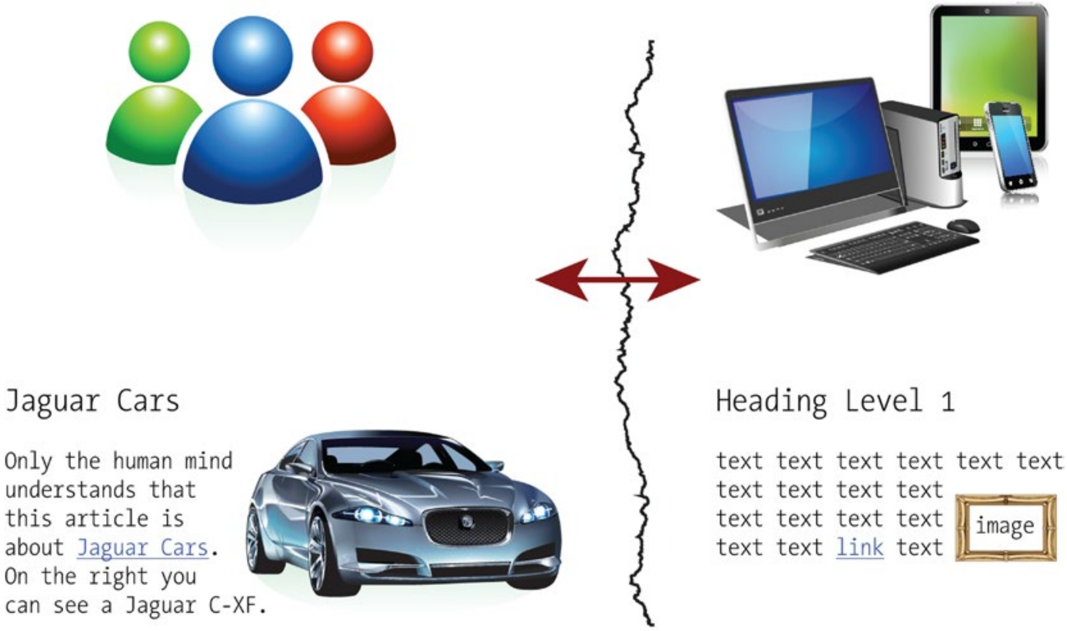
Jaguar Cars
Only the human mind understands that this article is about Jaguar Cars. On the right you can see a Jaguar C-XF.
Heading Level 1
text text text text text text
text text text text
text text text text
text text link text
image

Figure 1-3. *Traditional web site contents are meaningless to computers*

The concept of machine-readable data is not new, and it is not limited to the Web. Think of the credit cards or barcodes, both of which contain human-readable and machine-readable data (Figure 1-4). One person or product, however, has more than one identifier, which can cause ambiguity.

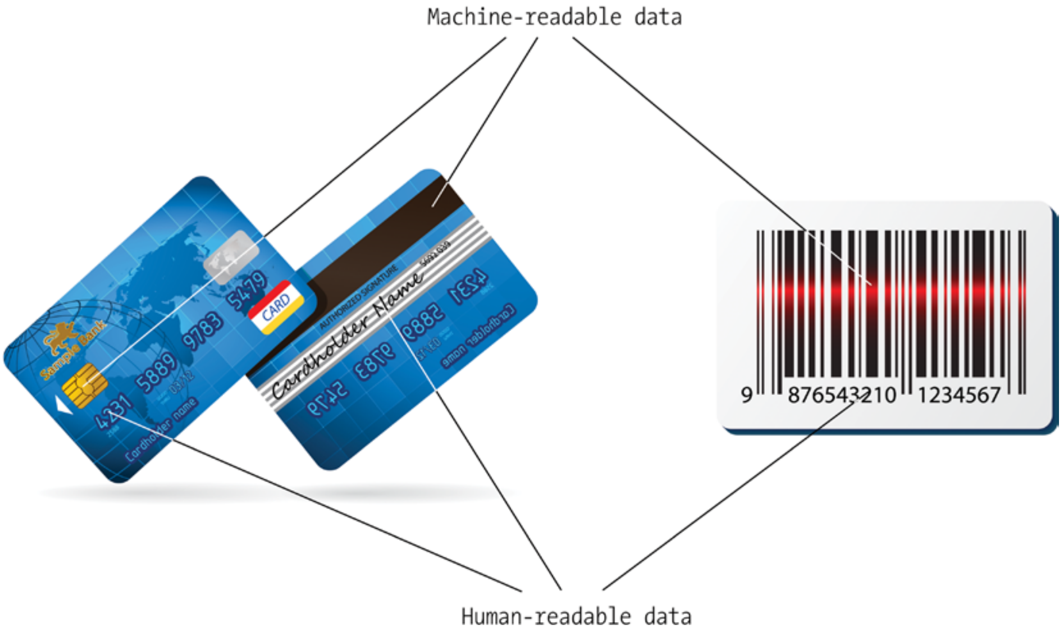

Figure 1-4. *Human-readable and machine-readable data*

Even the *well-formed XML* documents, which follow rigorous syntax rules, have serious limitations when it comes to machine-processability. For instance, if an XML entity is defined between `<SLR>` and `</SLR>`, what does *SLR* stand for? It can refer to a single-lens reflex camera, a self-loading rifle, a service-level report, system-level requirements, the Sri Lankan rupee, and so on.

Contents can be made machine-processable and unambiguous by adding organized (structured) data to the web sites, as markup annotations or as dedicated external metadata files, and linking them to other, related structured datasets. Among other benefits, structured data files support a much wider range of tasks than conventional web sites and are far more efficient to process. Structured data formats have been used for decades in computing, especially in Access and SQL relational databases, where queries can be performed to retrieve information efficiently. Because there are standards for direct mapping of relational databases to core Semantic Web technologies, databases that were publicly unavailable can now be shared on the Semantic Web [15]. Commercial database software packages powered by Semantic Web standards are also available on the market (5Store, AllegroGraph, BigData, Oracle, OWLIM, Talis Platform, Virtuoso, and so on) [16].

In contrast to relational databases, most data on the Web is stored in (X)HTML documents that contain *unstructured data* to be rendered in web browsers as formatted text, images, and multimedia objects. Publishing unstructured data works satisfactorily for general purposes; however, a large amount of data stored in, or associated with, traditional web documents cannot be processed this way. The data used to describe social connections between people is a good example, which should include the relationship type and multiple relationship directions inexpressible with the hyperlinks of the conventional Web [17].

The real benefit of semantic annotations is that humans can browse the conventional web documents, while Semantic Web crawlers can process the machine-readable annotations to classify data entities, discover logical links between entities, build indices, and create navigation and search pages.

Semantic Web Components

Structured data processing relies on technologies that provide a formal description of concepts, terms, and relationships within a knowledge domain (field of interest, discipline). *Knowledge Representation and Reasoning* is the field of Artificial Intelligence (AI) used to represent information in a machine-readable form that computer systems can utilize to solve complex tasks. *Taxonomies* or *controlled vocabularies* are structured collections of terms that can be used as metadata element values. For example, an events vocabulary can be used to describe concerts, lectures, and festivals in a machine-readable format, while an organization vocabulary is suitable for publishing machine-readable metadata about a school, a corporation, or a club. The controlled vocabularies are parts of conceptual data *schemas* (data models) that map concepts and their relationships.

The most widely adopted knowledge-management standards are the Resource Description Framework (RDF), the Web Ontology Language (OWL), and the Simple Knowledge Organization System (SKOS). *Knowledge Organization Systems* (*KOS*) are used for processing authority lists, classifications, thesauri, topic maps, ontologies, and so on. Web *ontologies* are formalized conceptual structures, in other words, complex knowledge representations of sets of concepts in a domain and the relationships between them. The *namespace* mechanism is used to reveal the meaning of tags and attributes by pointing to an external vocabulary that describes the concepts of the discipline in a machine-processable format, extending the vocabulary (set of elements and attributes) of markup languages. For example, a smartphone ontology defines all features of smartphones and the relationships between those features in a machine-processable format, so that software agents can “understand” the meaning of any of these features used to annotate a word on a web page by pointing to the ontology file. Web ontologies make it possible to describe complex statements in any topic in a machine-readable format. The architecture of the Semantic Web is illustrated by the “Semantic Web Stack,” which shows the hierarchy of standards in which each layer relies on the layers below (see Figure 1-5).

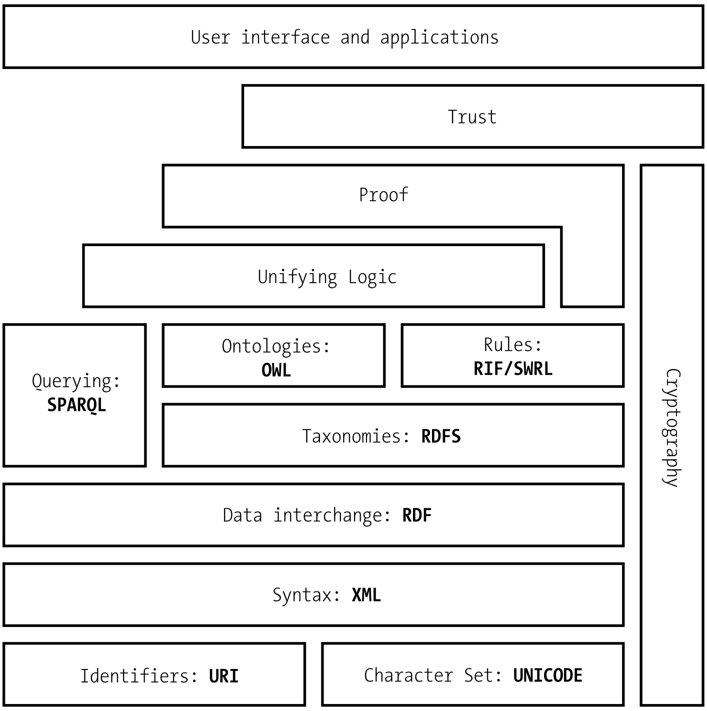

Figure 1-5. *The Semantic Web Stack*

While the preceding data formats are primarily machine-readable, they can be linked from human-readable web pages or integrated into human-readable web pages as *semantic annotations* such as microformats, RDFa, or HTML5 microdata.

Ontologies

The word *ontology* was originally introduced in philosophy to study the nature of existence. In computer science, ontology refers to conceptualization through a data model for describing a piece of our world, such as an organization, a research project, a historical event, our colleagues, friends, etc., in a machine-readable manner, by formally defining a set of classes (concepts), properties (attributes), relationship types, and entities (individuals, instances). The most advanced ontology languages (such as OWL) support the following components:

- *Classes*: Abstract groups, sets or collections of objects, object types. Classes usually represent groups or classes whose members share common properties. The hierarchy of classes is expressed as higher-level (superclass or parent class) and lower-level classes (subclass or child class). For example, a company can be represented as a class with subclasses such as departments and employees.
- *Attributes*: Aspects, properties, characteristics, or parameters that feature objects and classes
- *Individuals*: Instances or objects. For example, if our domain covers companies, each employee is an individual.

- *Relations*: The logical bond between classes, between individuals, between an individual and a class, between a single object and a collection, or between collections
- *Function terms*: Complex structures formed from certain relations that can be used in place of an individual term in a statement
- *Restrictions*: Formally defined limitations or ranges of valid values
- *Rules*: If-then statements (antecedent-consequent sentence) defining the logical inferences
- *Axioms*: Assertions in a logical form that, together with rules, form the overall theory the ontology describes. Unlike the definition of axiom in generative grammar or formal logic, where axioms include only statements defined as a priori knowledge, the axioms of Semantic Web ontologies also include the theory derived from axiomatic statements. Axioms are used to impose constraints on the values of classes or instances, so axioms are generally expressed using logic-based languages. Axioms are suitable for verifying the consistency of the ontology.
- *Events*: Attribute or relationship changes

Ontology Engineering

Ontology engineering is a field of computer science that covers the methods and methodologies for building ontologies. The purpose of Semantic Web ontologies is to achieve a common and shared knowledge ready to be transmitted between applications, providing interoperability across organizations of different areas or different views of the same area. *Ontology transformation* is the development of a new ontology to deal with the new requirements of an existing ontology for a new purpose. Creating a new single coherent ontology from two or more ontologies of the same knowledge domain is known as *ontology merging. Ontology integration* is the creation of a new ontology from two or more source ontologies from different knowledge domains. *Ontology mapping* is a formal expression to define the semantic relationships between entities from different ontologies. *Ontology alignment* is the process of creating a consistent and coherent link between two or more ontologies where statements of the first ontology confirm the statements of the second ontology.

Inference

The automatic discovery of relationships between seemingly unrelated structured data is called *inference*. The automatically generated new relationships are based on the structured data and additional information defined in the vocabulary, as, for example, a set of rules. The new relationships are either explicitly declared or returned through queries. As an example, if we make the statement "Unforgiven is a western," while an ontology declares that "every western is a movie," Semantic Web agents can automatically infer the statement "Unforgiven is a movie," which was originally not declared.

Semantic Web Features

The Semantic Web has many distinctive features that are rarely seen or not used at all on traditional web sites. For example, a large share of the data is published with explicitly declared open license, allowing data sharing and distribution that is truly free. The formally defined data connections make automatic knowledge discovery possible, along with accurate statement verification. Each and every object and feature is associated with a web address, so that you can refer to virtually everything, from a table cell to an image or a friendship of two people.

Free, Open Access Data Repositories

> *Open data and content can be freely used, modified, and shared by anyone for any purpose*
>
> —OpenDefinition.org

Automation and data processing rely on data access, so "open data" is a fundamental feature of the Semantic Web. There are already hundreds of government organizations, enterprises, and individuals publishing machine-readable, structured data as open data (`https://data.cityofchicago.org`, `http://data.alberta.ca`, `http://data.gov.uk`, etc.), although not all of them provide an open license explicitly that makes data truly "open." Semantic Web applications also benefit from open APIs, open protocols, open data formats, and open source software tools to reuse, remix, and republish data. On the Semantic Web, the machine-readable data of persons, products, services, and objects of the world are open and accessible, without registering and paying membership or subscription fees, and software agents can access these data automatically on your behalf.

Adaptive Information

While on the conventional Web, each web document is edited by a single person or a team, Semantic Web documents are edited by groups of people through related documents as a "global database." As a result, datasets are more accurate, more connected, and more descriptive. Information is not just about pages, but, rather, about connected and reusable data. The classification, reorganization, and reuse of data is a fundamental concept of the Semantic Web.

Unique Web Resource Identifiers

Web resources can be located by unique IP addresses. However, they are hard to remember, and their number is limited. This is why *domain names* are used in most cases. Figure 1-6 shows the relationship between a domain name and a URL: `www.masteringhtml5css3.com` is a subdomain of the node `masteringhtml5css3.com`, which is the subdomain of the `com` (which stands for *commercial*) domain. The domain name syntax rules are defined by RFC 1035 [18], RFC 1123 [19], and RFC 2181 [20].

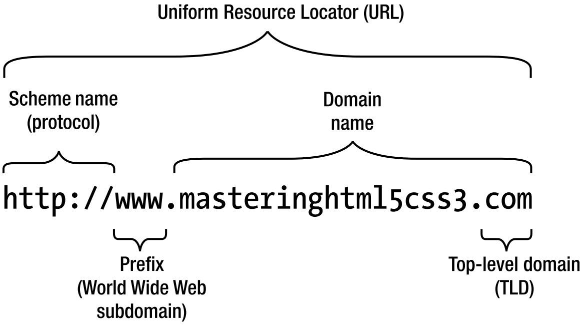
Uniform Resource Locator (URL)
Scheme name (protocol)
Domain name
http://www.masteringhtml5css3.com
Prefix (World Wide Web subdomain)
Top-level domain (TLD)

Figure 1-6. *The domain within the URL*

The tree of subdomains can contain a maximum of 127 levels. Each label may contain up to 63 characters. According to RFC 2181, the full length of a domain name is 253 characters. Conventional domain names cannot contain Latin alphabet–based characters with diacritics, non-Latin characters, or scripts. With the introduction of Internationalized Domain Names (IDN), it is possible to represent names and words in several languages in native alphabets and scripts.

A *Uniform Resource Identifier* (*URI*) is a character string that identifies a name or a resource on the Internet (RFC 2396 [21]). URIs can be classified as *Uniform Resource Locators* (*URLs*; RFC 1738 [22]), *Uniform Resource Names* (*URNs*), or both. A URN defines the identity of a resource, while the URL provides a method for finding it (including protocol and path). URIs are often used incorrectly as a synonym for URLs, although URI is a broader term (RFC 3305 [23]). Both the URN and the URL are subsets of the URI, but they are generally disjoint sets. The best-known examples of URLs are the web site addresses on the World Wide Web. Listing 1-1 shows the general URL syntax.

Listing 1-1. URL Syntax

```
protocol://domain:port/path?query_string#fragment_identifier
```

The *protocol* is followed by a colon. The other parts of URLs depend on the scheme being used. Usually, there is a domain name or an IP address, an optional port number, and an optional path to the resource or script. Programs such as PHP or CGI scripts might have a query string. The end of the URL can be an optional *fragment identifier*, which starts with the number sign (#), and in markup languages, it points to a section of the document available through the provided path. Fragment identifiers are widely used on the Semantic Web for distinguishing a file from the entity it represents (e.g., a person vs. the file that describes him/her or a book vs. the file that describes it). This use of fragment identifiers provides an unlimited number of unique identifiers, regardless of the restricted choice of domain names. Listing 1-2 shows an example, in which `http` is the protocol, `www.masteringhtml5css3.com` is the domain, and the URL identifies a book rather than the file `webstandardsbook.rdf` stored in the `metadata` directory.

Listing 1-2. A Typical URL with Fragment Identifier

```
http://www.masteringhtml5css3.com/metadata/webstandardsbook.rdf#book
```

Because many of the URL components are optional, one or more of them are often omitted.

To avoid the inconvenience of registering and renewing domain names and address the issue that not every user has his/her own domain name, web addresses can be redirected (typically using 302 HTTP redirection) to an intermediate and more persistent location on the Web, instead of the actual physical location of the file or directory. Such URLs are called *Persistent Uniform Resource Locators* (*PURLs*). Official PURLs are registered on `https://purl.org`. Many semantic vocabularies and ontologies use PURLs. Ontologies are listed under the `ontology` directory such as `http://purl.org/ontology/vidont/`.

Summary

In this chapter, you learned that a major limitation of conventional web sites is their unorganized and isolated contents, which is created mainly for human interpretation. You became familiar with the fundamental concepts of the Semantic Web, the main application areas, and the efficiency of structured data processing. You know the core Semantic Web components and understand that many standards rely on other technologies. You are familiar with distinctive Semantic Web features, such as open licensing, decentralized data storage with automatically inferred statements and knowledge discovery, and unique URI indentifiers associated with every bit of represented knowledge.

The next chapter will introduce you to knowledge representation and reasoning used to represent information with machine-processable standards.

References

1. Hausenblas, M., Adida, B., Herman, I. (2008) RDFa—Bridging the Web of Documents and the Web of Data. Joanneum Research, Creative Commons, World Wide Web Consortium. www.w3.org/2008/Talks/1026-ISCW-RDFa/. Accessed 18 January 2015.

2. Berners-Lee, T. (2001) Business Model for the Semantic Web. World Wide Web Consortium. www.w3.org/DesignIssues/Business. Accessed 18 January 2015.

3. Ankolekar, A., Krötzsch, M., Tran, T., Vrandečić, D. The Two Cultures: Mashing Up Web 2.0 and the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web* 2008, 6(1):70–75.

4. Shannon, V. (2006) A "more revolutionary" Web. *International Herald Tribune*. The New York Times Company. www.nytimes.com/2006/05/23/technology/23iht-web.html?scp=1&sq=A+%27more+revolutionary%27+Web&st=nyt. Accessed 18 January 2015.

5. Spivack, N. (2015) Web 3.0: The Third Generation Web Is Coming. http://lifeboat.com/ex/web.3.0. Accessed 16 March 2015.

6. Herman, I. (ed.) (2009) How would you define the main goals of the Semantic Web? In: W3C Semantic Web FAQ. World Wide Web Consortium. www.w3.org/2001/sw/SW-FAQ#swgoals. Accessed 18 January 2015.

7. Sbodio, L. M., Martin, D., Moulin, C. Discovering Semantic Web services using SPARQL and intelligent agents. Web Semantics: Science, Services and Agents on the World Wide Web 2010, 8(4): 310–328.

8. Herman, I. (2009) W3C Semantic Web Frequently Asked Questions. World Wide Web Consortium. www.w3.org/RDF/FAQ. Accessed 18 January 2015.

9. Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S. MuseumFinland—Finnish museums on the Semantic Web. Web Semantics: Science, Services and Agents on the World Wide Web 2005, 3(2–3): 224–241.

10. Bojārs, U., Breslin, J. G., Finn, A., Decker, S. Using the Semantic Web for linking and reusing data across Web 2.0 communities. Web Semantics: Science, Services and Agents on the World Wide Web 2008, 6(1): 21–28.

11. Celma, Ò., Raimond, Y. ZemPod: A Semantic Web approach to podcasting. Web Semantics: Science, Services and Agents on the World Wide Web 2008, 6(2): 162–169.

12. Saleem, M., Kamdar, M. R., Iqbal, A., Sampath, S., Deus, H. F., Ngomo, A.-C. Big linked cancer data: Integrating linked TCGA and PubMed. Web Semantics: Science, Services and Agents on the World Wide Web 2014, http://dx.doi.org/10.1016/j.websem.2014.07.004.

13. Oinonen, K. (2005) On the road to business application of Semantic Web technology. Semantic Web in Business—How to proceed. In: *Industrial Applications of Semantic Web: Proceedings of the 1st IFIP WG12.5 Working Conference on Industrial Applications of Semantic Web*. International Federation for Information Processing. Springer Science+Business Media, Inc., New York.

14. Murphy, T. (2010) Lin Clark On Why Drupal Matters. Socialmedia. http://socialmedia.net/2010/09/07/lin-clark-on-why-drupal-matters. Accessed 9 September 2010.

15. Arenas, M., Bertails, A., Prud'hommeaux, E., Sequeda, J. (eds.) (2012) A Direct Mapping of Relational Data to RDF. www.w3.org/TR/rdb-direct-mapping/. Accessed 18 January 2015.

16. Clark, K. (2010) The RDF Database Market. Clark & Parsia, LLC. http://weblog.clarkparsia.com/2010/09/23/the-rdf-database-market/. Accessed 18 January 2015.

17. Dertouzos, L. M., Berners-Lee, T., Fischetti, M. (1999) *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. Harper San Francisco, San Francisco.

18. Mockapetris, P. (1987) Domain names—Implementation and specification. RFC 1035. The Internet Engineering Task Force. http://tools.ietf.org/html/rfc1035. Accessed 18 January 2015.

19. Braden, R. (ed.) (1989) Requirements for Internet Hosts—Application and Support. RFC 1123. The Internet Engineering Task Force. http://tools.ietf.org/html/rfc1123. Accessed 18 January 2015.

20. Elz, R., Bush, R. (1997) Clarifications to the DNS Specification. RFC 2181. The Internet Engineering Task Force. http://tools.ietf.org/html/rfc2181. Accessed 18 January 2015.

21. Berners-Lee, T., Fielding, R., Masinter, L. (1998) Uniform Resource Identifiers (URI): Generic Syntax. RFC 2396. The Internet Society. http://tools.ietf.org/html/rfc2396. Accessed 18 January 2015.

22. Berners-Lee, T., Masinter, L., McCahill, M. (eds.) (1994) Uniform Resource Locators (URL). RFC 1738. The Internet Engineering Task Force. http://tools.ietf.org/html/rfc1738. Accessed 18 January 2015.

23. Mealling, M., Denenberg, R. (eds.) (2002) Uniform Resource Identifiers (URIs), URLs, and Uniform Resource Names (URNs): Clarifications and Recommendations. RFC 3305. The Internet Society. http://tools.ietf.org/html/rfc3305. Accessed 18 January 2015.