THE EXPERT'S VOICE® IN WEB DEVELOPMENT

Mastering Structured Data on the Semantic Web

From HTML5 Microdata to Linked Open Data

MAKE YOUR WEB CONTENTS MACHINE-INTERPRETABLE USING ARTIFICIAL INTELLIGENCE

lie F. Sikos, Ph.D.

Apress®

Mastering Structured Data on the Semantic Web

From HTML5 Microdata to Linked Open Data

Leslie F. Sikos, Ph.D.

Apress®

Mastering Structured Data on the Semantic Web: From HTML5 Microdata to Linked Open Data

ISBN-13 (pbk): 978-1-4842-1050-5

ISBN-13 (electronic): 978-1-4842-1049-9

Managing Director: Welmoed Spahr
Lead Editor: Ben Renow-Clarke
Technical Reviewer: Maria Maleshkova
Editorial Board: Steve Anglin, Mark Beckner, Gary Cornell, Louise Corrigan, Jim DeWolf, Jonathan Gennick, Robert Hutchinson, Michelle Lowman, James Markham, Susan McDermott, Matthew Moodie, Jeffrey Pepper, Douglas Pundick, Ben Renow-Clarke, Gwenan Spearing, Matt Wade, Steve Weiss
Coordinating Editors: Melissa Maldonado and Christine Ricketts
Copy Editor: Michael G. Laraque
Compositor: SPi Global
Indexer: SPi Global
Artist: SPi Global

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science+Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a Delaware corporation.

For information on translations, please e-mail rights@apress.com, or visit www.apress.com.

Apress and friends of ED books may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Special Bulk Sales–eBook Licensing web page at www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this text is available to readers at www.apress.com and on the author's site at www.lesliesikos.com. For detailed information about how to locate your book's source code, go to www.apress.com/source-code/.

About the Author

Leslie F. Sikos, Ph.D., is a Semantic Web researcher at Flinders University, South Australia, specializing in semantic video annotations, ontology engineering, and natural language processing using Linguistic Linked Open Data. On the cutting edge of Internet technologies, he is a member of industry-leading organizations, such as the World Wide Web Consortium, the Internet Engineering Task Force, and the Internet Society. He is an invited editor and journal reviewer actively contributing to the development of open standards. Dr. Sikos is the author of 15 textbooks covering a wide range of topics from computer networks to software engineering and web design. Devoted to lifelong learning, he holds multiple degrees in computer science and information technology, as well as professional certificates from the industry. Thanks to his hands-on skills, coupled with a pedagogical background, he can introduce technical terms and explain complex issues in plain English.

Dr. Sikos creates fully standard-compliant, mobile-friendly web sites with responsive web design—complemented by machine-readable annotations—and develops multimedia applications leveraging Semantic Web technologies. He works on the standardization of Linked Data implementations for the precise identification, description, and classification of multimedia fragments, advancing the traditional video annotation techniques. To solve syntactic interoperability, conceptual ambiguity, and implementation complexity problems of RDFS and OWL multimedia ontologies mapped from general-purpose XML Schema vocabularies, Dr. Sikos introduced a global video production and broadcasting ontology. Inspired by the creation and exploitation of rich Linked Open Data datasets, he proudly contributes to the Open Data and Open Knowledge initiatives. When he is not working, he enjoys reading, playing the organ, and cycling. For more information, visit `www.lesliesikos.com`.

About the Technical Reviewer

Maria Maleshkova, Ph.D., is a senior researcher at the Karlsruhe Institute of Technology (KIT), Germany. Her areas of expertise include Semantic Web services and web architectures, in particular focusing on the semantic description of Web APIs, RESTful services, and their joined use with Linked Data. In addition, she works on data integration with semantic technologies, domain modeling, and data annotation. She received a Ph.D. in computer science from the Knowledge Media Institute (KMi) at the Open University in Milton Keynes, England, where she worked on projects in the domain of service-oriented architecture and web services. Dr. Maleshkova has published more than 50 papers in conferences and international journals related to Semantic Web services, knowledge-based systems, knowledge engineering, and business process analysis. She is an active member of the Semantic Web community and has worked on several national and international research projects, involving both research and industrial partners.

Preface

With the evolution of the World Wide Web, more and more sophisticated web applications and services appear, making it possible not only to publish and search for information but also to share photos and videos, buy products online, chat with friends, book a hotel room, play games, and more. Many of these applications rely on file characteristics, metadata, tracking cookies, and data from user registrations, which makes it possible to provide customized services and to make offers of products or services the users might be interested in. However, because there is a huge gap between what the human mind understands and what computers can interpret, a large amount of data on the Internet cannot be processed efficiently with computer software. For example, a scanned table in an image file is unstructured and cannot be interpreted by computers. While optical character recognition programs can be used to convert images of printed text into machine-encoded text, such conversions cannot be done in real time and with 100% accuracy, rely on a relatively clear image in high resolution, and require different processing algorithms, depending on the image file format. More important, table headings and table data cells will all become plain text, with no correlation whatsoever. In other words, you lose the relationships between the data cells, including the table columns and rows, making data reusability very limited.

Beyond the lack of structure, much data are locked down in proprietary file formats that can be opened only in the commercial software they were created in. As well, related data are often stored in isolated data silos, making it impossible to automatically find new relations between seemingly unrelated data or make new discoveries based on the relation between data. Even if data is provided in a standardized, open-access format, software agents often cannot interpret the meaning of the represented information. HTML documents, which implement the core markup language of the Web, are semistructured but have limitations when it comes to machine-processability, because they are written primarily for humans. While they can be used to display data on a web site, software tools cannot "understand" the description of real-world persons and objects described in HTML, nor the relationships between them. For instance, on conventional web sites, a character sequence containing numbers can be a phone number, a date, an ISBN number of a book, or the age of a person, and there is no way for a computer program to interpret the meaning of such data. The situation can be improved by adding metadata annotations to the markup documents, but this won't make the entire document machine-interpretable, only small portions of it. Moreover, HTML documents use hyperlinks to link related web resources or parts of web documents to one another; however, there is no information about the type of these links. As a result, machines cannot interpret the relationships represented by hyperlinks, such as whether a link relates to additional information about a topic or the friendship between two people. Another popular format, XML, is used in structured documents that are both human- and machine-readable. XML is widely deployed from web site markup to configuration settings to web news feeds to office software tools; however, XML files are not machine-interpretable either.

Owing to the fact that the content of conventional web resources is primarily human-readable only, automatic processing is infeasible, and searching for related information is inefficient. 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 more easily and better, which can be used for task automation. The web sites and structured datasets that provide semantics (meaning) to software agents form the Semantic Web, the Artificial Intelligence extension of the conventional Web. On the Semantic Web, knowledge is represented in formal languages based on strict grammar, describing every resource and link in a machine-interpretable manner, most of the time with

truly open access. This book is an example-driven tutorial for Semantic Web developers, Internet marketers, and researchers who want to unleash the potential of the Semantic Web. By bridging the gap between academia and the web design industry, this book will explain the core concepts as well as the mathematical background of the Semantic Web, based on graph theory and knowledge representation. You will learn how to annotate your web site markup with machine-readable metadata to boost your site's performance on next-generation Search Engine Result Pages. You will also understand how to reuse machine-readable definitions and how to describe your own concepts. By implementing best practices, you will be able to create typed links, so that computers can interpret a link, say, between two people who know each other, or a link between your web site and the machine-readable definition of topics you are interested in. Step-by-step guides will demonstrate the installation of integrated software development environments and the development of Semantic Web applications in Java.

These interlinked, machine-interpretable data can be used in task automation for web services, as well as for automatic knowledge discovery in research. The benefits of Semantic Web technologies have already been recognized by industrial giants such as Amazon.com, the BBC, Facebook, Flickr, Google, IBM, Thomson Reuters, New York Times, and Yahoo!, and the list is constantly growing. By implementing Semantic Web technologies to represent and connect structured data, you will reach a wider audience, encourage data reuse, and provide content that can be automatically processed with full certainty. As a result, your web sites will be integral parts of the next revolution of the Web.

Contents at a Glance

Contents

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).

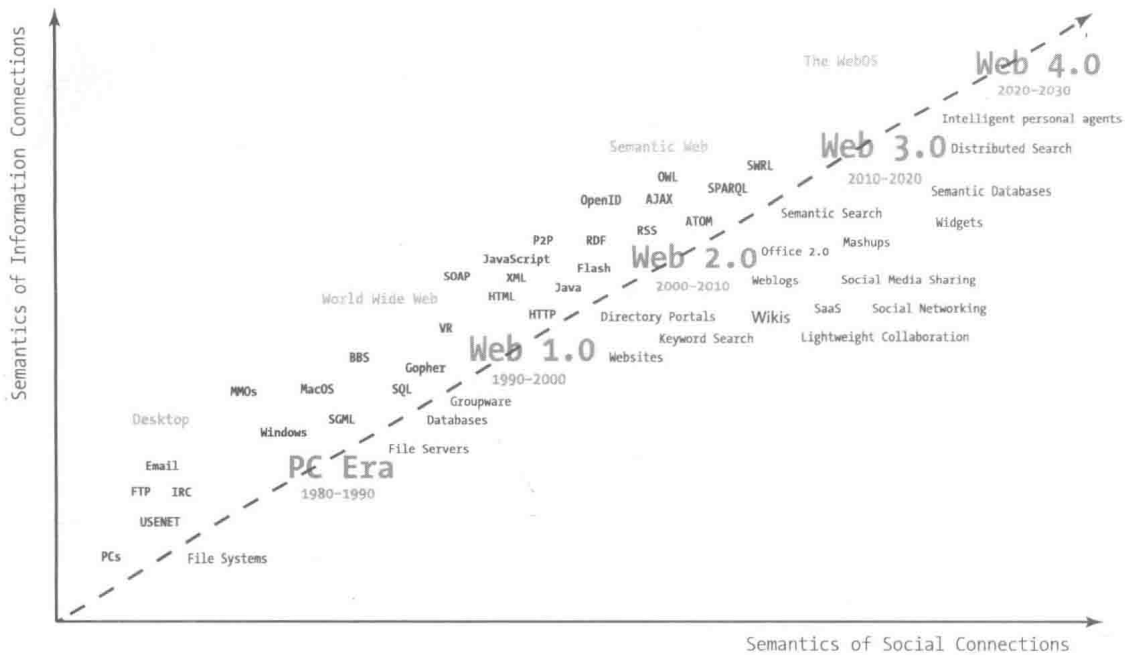

Figure 1-1. *The evolution of the Web [5]*

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.