

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

Copyright © 2015 by Leslie F. Sikos, Ph.D.

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

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

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

Trademarked names, logos, and images may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, logo, or image, we use the names, logos, and images only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

All names, companies, and domains provided in the listings are fictitious. No identification with actual persons, companies, and web sites is intended or should be inferred.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the author nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

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

About the Author xiii

About the Technical Reviewerxv

Prefacexvii

■Chapter 1: Introduction to the Semantic Web..... 1

■Chapter 2: Knowledge Representation 13

■Chapter 3: Linked Open Data 59

■Chapter 4: Semantic Web Development Tools 79

■Chapter 5: Semantic Web Services..... 121

■Chapter 6: Graph Databases 145

■Chapter 7: Querying..... 173

■Chapter 8: Big Data Applications..... 199

■Chapter 9: Use Cases..... 217

Index..... 227

Contents

- About the Author xiii
- About the Technical Reviewerxv
- Prefacexvii
- Chapter 1: Introduction to the Semantic Web 1
 - The Semantic Web..... 1
 - Structured Data2
 - Semantic Web Components 5
 - Ontologies.....6
 - Inference.....7
 - Semantic Web Features 7
 - Free, Open Access Data Repositories8
 - Adaptive Information8
 - Unique Web Resource Identifiers.....8
 - Summary..... 9
 - References 10
- Chapter 2: Knowledge Representation 13
 - Vocabularies and Ontologies 13
 - The schema.org Vocabulary Collection..... 14
 - General, Access, and Structural Metadata..... 15
 - Person Vocabularies 15
 - Book Vocabularies 16
 - PRISM: A Publishing Vocabulary 16
 - GoodRelations: An E-commerce Ontology 16

Publication Ontologies	16
DOAP: A Project Management Vocabulary	17
Licensing Vocabularies	17
Media Ontologies	18
Vocabularies for Online Communities	18
Knowledge Management Standards	18
Resource Description Framework (RDF)	18
Machine-Readable Annotations	23
GRDDL: XML Documents to RDF	39
R2RML: Relational Databases to RDF	40
RDFS	41
Web Ontology Language (OWL)	45
Simple Knowledge Organization System (SKOS)	53
Rule Interchange Format (RIF)	53
Reasoning	54
Parsers	54
Summary	54
References	55
■ Chapter 3: Linked Open Data	59
Linked Data Principles	59
The Five-Star Deployment Scheme for Linked Data	60
LOD Datasets	62
RDF Crawling	62
RDF Dumps	62
SPARQL Endpoints	62
Frequently Used Linked Datasets	63
LOD Dataset Collections	67
The LOD Cloud Diagram	67
Creating LOD Datasets	70
RDF Structure	70
Licensing	71

RDF Statements	72
Interlinking	72
Registering Your Dataset	74
Linked Data Visualization	75
Summary	76
References	77
■ Chapter 4: Semantic Web Development Tools	79
Advanced Text Editors	79
Semantic Annotators and Converters	81
RDFa Play	82
RDFa 1.1 Distiller and Parser	82
RDF Distiller	83
DBpedia Spotlight	84
Google Structured Data Testing Tool	84
RDFizers	85
Apache Any23	85
General Architecture for Text Engineering (GATE)	86
OpenRefine	86
Ontology Editors	86
Protégé	86
SemanticWorks	89
TopBraid Composer	90
Apache Stanbol	91
Fluent Editor	91
Ontology Analysis Tools	91
ZOOMA	91
Semantic Measures Library	92
Reasoners	92
HermiT	92
Pellet	93

FaCT++.....	94
RACER.....	94
Application Development Frameworks.....	94
Jena.....	94
Sesame.....	96
Integrated Development Environments	98
Eclipse	98
NetBeans	108
CubicWeb.....	109
Linked Data Software.....	110
Sindice.....	110
Apache Marmotta.....	111
sameAs.org.....	112
Callimachus	112
Neologism.....	112
LODStats.....	113
Semantic Web Browsers	113
Tabulator.....	113
Marbles.....	114
OpenLink Data Explorer (ODE)	114
DBpedia Mobile	116
IsaViz	116
RelFinder	117
Summary.....	117
References	117
■ Chapter 5: Semantic Web Services.....	121
Semantic Web Service Modeling.....	121
Communication with XML Messages: SOAP.....	122
Web Services Description Language (WSDL).....	124
Web Ontology Language for Services (OWL-S).....	129
Web Service Modeling Ontology (WSMO)	133

Web Service Modeling Language (WSML)	138
Web Services Business Process Execution Language (WS-BPEL)	140
Semantic Web Service Software	141
Web Service Modeling eXecution environment (WSMX).....	141
Internet Reasoning Service (IRS-III).....	141
Web Services Modeling Toolkit (WSMT).....	141
Semantic Automated Discovery and Integration (SADI).....	142
UDDI Semantic Web Service Listings	142
Summary	142
References	143
■ Chapter 6: Graph Databases	145
Graph Databases	145
Triplestores	149
Quadstores	149
The Most Popular Graph Databases	150
AllegroGraph.....	151
Neo4j	161
4Store	169
Oracle	171
Summary	172
References	172
■ Chapter 7: Querying.....	173
SPARQL: The Query Language for RDF	173
Structure and Syntax	173
SPARQL 1.0 and SPARQL 1.1	175
Query Types	176
Pattern Matching	176
Solution Modifiers.....	178
SELECT Queries	178
ASK Queries	179

CONSTRUCT Queries	180
DESCRIBE Queries	180
Federated Queries	181
REASON Queries	181
URL Encoding of SPARQL Queries.....	182
Graph Update Operations.....	182
Graph Management Operations.....	183
Proprietary Query Engines and Query Languages.....	186
SeRQL: The Sesame RDF Query Language	186
CQL: Neo4j's Query Language.....	188
Identify Datasets to Query.....	189
Public SPARQL Endpoints.....	189
Setting Up Your Own SPARQL Endpoint.....	190
OpenLink Virtuoso.....	190
Fuseki	192
D2R	193
4store SPARQL Server	195
PublishMyData.....	195
Summary.....	197
References	197
■ Chapter 8: Big Data Applications.....	199
Big Semantic Data: Big Data on the Semantic Web	199
Google Knowledge Graph and Knowledge Vault.....	200
Get Your Company, Products, and Events into the Knowledge Graph	202
Social Media Applications	205
Facebook Social Graph.....	206
Twitter Cards	211
IBM Watson	212
BBC's Dynamic Semantic Publishing	212

The Library of Congress Linked Data Service	213
High-Performance Storage: The One Trillion Triples Mark.....	213
Summary	214
References	215
■ Chapter 9: Use Cases	217
RDB to RDF Direct Mapping.....	217
A Semantic Web Service Process in OWL-S to Charge a Credit Card.....	221
Modeling a Travel Agency Web Service with WSMO.....	223
Querying DBpedia Using the RDF API of Jena	224
Summary	225
References	226
Index.....	227

CHAPTER 1

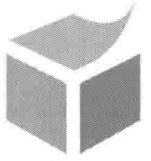


Introduction to the Semantic Web

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 image¹ 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.

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

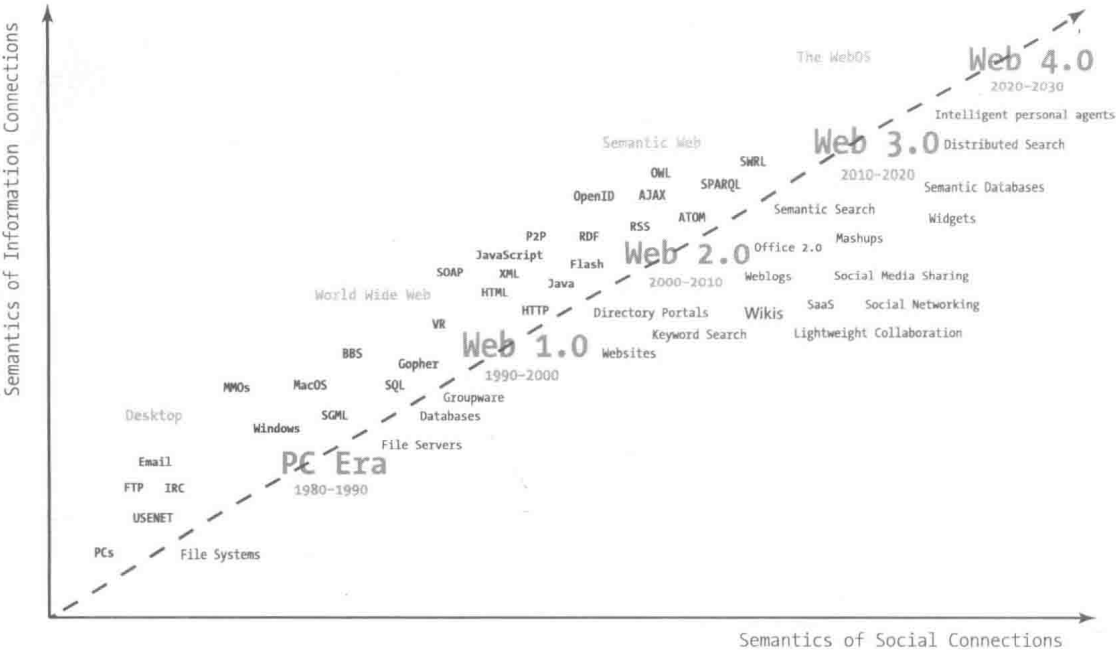


Figure 1-1. The evolution of the Web [5]

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).

²The 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.