Linked Data

Evolving the Web into a Global Data Space

Synthesis Lectures on the Semantic Web: Theory and Technology

Editors
James Hendler, ***Rensselaer Polytechnic Institute***
Frank van Harmelen, ***Vrije Universiteit Amsterdam***

Whether you call it the Semantic Web, Linked Data, or Web 3.0, a new generation of Web technologies is offering major advances in the evolution of the World Wide Web. As the first generation of this technology transitions out of the laboratory, new research is exploring how the growing Web of Data will change our world. While topics such as ontology-building and logics remain vital, new areas such as the use of semantics in Web search, the linking and use of open data on the Web, and future applications that will be supported by these technologies are becoming important research areas in their own right. Whether they be scientists, engineers or practitioners, Web users increasingly need to understand not just the new technologies of the Semantic Web, but to understand the principles by which those technologies work, and the best practices for assembling systems that integrate the different languages, resources, and functionalities that will be important in keeping the Web the rapidly expanding, and constantly changing, information space that has changed our lives. Topics to be covered:

- Semantic Web Principles from linked-data to ontology design
- Key Semantic Web technologies and algorithms
- Semantic Search and language technologies
- The Emerging "Web of Data" and its use in industry, government and university applications
- Trust, Social networking and collaboration technologies for the Semantic Web
- The economics of Semantic Web application adoption and use
- Publishing and Science on the Semantic Web
- Semantic Web in health care and life sciences

Linked Data: Evolving the Web into a Global Data Space

Tom Heath and Christian Bizer

2011

© Springer Nature Switzerland AG 2022
Reprint of original edition © Morgan & Claypool 2011

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Linked Data: Evolving the Web into a Global Data Space
Tom Heath and Christian Bizer

ISBN: 978-3-031-79431-5 paperback
ISBN: 978-3-031-79432-2 ebook

DOI 10.1007/978-3-031-79432-2

A Publication in the Springer series
SYNTHESIS LECTURES ON THE SEMANTIC WEB: THEORY AND TECHNOLOGY

Lecture #1
Series Editors: James Hendler, *Rensselaer Polytechnic Institute*
Frank van Harmelen, *Vrije Universiteit Amsterdam*
First Edition
10 9 8 7 6 5 4 3 2 1

Series ISSN
Synthesis Lectures on the Semantic Web: Theory and Technology
ISSN pending.

Photo credits:

Figure 2.1 © Till Krech, http://www.flickr.com/photos/extranoise/155085339/, reused under Creative Commons Attribution 2.0 Generic (CC BY 2.0) License.
Photo of Dr. Tom Heath on page 121 © Gregory Todd Williams, reused under Creative Commons Attribution 2.0 Generic (CC BY 2.0) License.

Linked Data

Evolving the Web into a Global Data Space

Tom Heath
Talis

Christian Bizer
Freie Universität Berlin

SYNTHESIS LECTURES ON THE SEMANTIC WEB: THEORY AND TECHNOLOGY #1

ABSTRACT

The World Wide Web has enabled the creation of a global information space comprising linked documents. As the Web becomes ever more enmeshed with our daily lives, there is a growing desire for direct access to raw data not currently available on the Web or bound up in hypertext documents. Linked Data provides a publishing paradigm in which not only documents, but also data, can be a first class citizen of the Web, thereby enabling the extension of the Web with a global data space based on open standards - the Web of Data. In this Synthesis lecture we provide readers with a detailed technical introduction to Linked Data. We begin by outlining the basic principles of Linked Data, including coverage of relevant aspects of Web architecture. The remainder of the text is based around two main themes - the publication and consumption of Linked Data. Drawing on a practical Linked Data scenario, we provide guidance and best practices on: architectural approaches to publishing Linked Data; choosing URIs and vocabularies to identify and describe resources; deciding what data to return in a description of a resource on the Web; methods and frameworks for automated linking of data sets; and testing and debugging approaches for Linked Data deployments. We give an overview of existing Linked Data applications and then examine the architectures that are used to consume Linked Data from the Web, alongside existing tools and frameworks that enable these. Readers can expect to gain a rich technical understanding of Linked Data fundamentals, as the basis for application development, research or further study.

KEYWORDS

web technology, databases, linked data, web of data, semantic web, world wide web, dataspaces, data integration, data management, web engineering, resource description framework

Contents

List of Figures ..... **xi**

Preface ..... **xiii**

1 Introduction **1**

1.1 The Data Deluge 1
1.2 The Rationale for Linked Data 2
1.2.1 Structure Enables Sophisticated Processing 2
1.2.2 Hyperlinks Connect Distributed Data 3
1.3 From Data Islands to a Global Data Space 4
1.4 Introducing Big Lynx Productions 5

2 Principles of Linked Data **7**

2.1 The Principles in a Nutshell 7
2.2 Naming Things with URIs 9
2.3 Making URIs Defererenceable 10
2.3.1 303 URIs 11
2.3.2 Hash URIs 13
2.3.3 Hash versus 303 14
2.4 Providing Useful RDF Information 15
2.4.1 The RDF Data Model 15
2.4.2 RDF Serialization Formats 18
2.5 Including Links to other Things 20
2.5.1 Relationship Links 21
2.5.2 Identity Links 22
2.5.3 Vocabulary Links 24
2.6 Conclusions 25

3 The Web of Data **29**

3.1 Bootstrapping the Web of Data 30
3.2 Topology of the Web of Data 30
3.2.1 Cross-Domain Data 33

3.2.2 Geographic Data . . . 34
3.2.3 Media Data . . . 34
3.2.4 Government Data . . . 35
3.2.5 Libraries and Education . . . 36
3.2.6 Life Sciences Data . . . 37
3.2.7 Retail and Commerce . . . 38
3.2.8 User Generated Content and Social Media . . . 38
3.3 Conclusions . . . 39

4 Linked Data Design Considerations . . . 41

4.1 Using URIs as Names for Things . . . 41
4.1.1 Minting HTTP URIs . . . 41
4.1.2 Guidelines for Creating Cool URIs . . . 42
4.1.3 Example URIs . . . 43
4.2 Describing Things with RDF . . . 44
4.2.1 Literal Triples and Outgoing Links . . . 45
4.2.2 Incoming Links . . . 46
4.2.3 Triples that Describe Related Resources . . . 46
4.2.4 Triples that Describe the Description . . . 47
4.3 Publishing Data about Data . . . 48
4.3.1 Describing a Data Set . . . 48
4.3.2 Provenance Metadata . . . 52
4.3.3 Licenses, Waivers and Norms for Data . . . 52
4.4 Choosing and Using Vocabularies to Describe Data . . . 56
4.4.1 SKOS, RDFS and OWL . . . 56
4.4.2 RDFS Basics . . . 57
4.4.3 A Little OWL . . . 60
4.4.4 Reusing Existing Terms . . . 61
4.4.5 Selecting Vocabularies . . . 62
4.4.6 Defining Terms . . . 63
4.5 Making Links with RDF . . . 64
4.5.1 Making Links within a Data Set . . . 64
4.5.2 Making Links with External Data Sources . . . 64
4.5.3 Setting RDF Links Manually . . . 66
4.5.4 Auto-generating RDF Links . . . 66

5 Recipes for Publishing Linked Data 69

5.1 Linked Data Publishing Patterns 69
5.1.1 Patterns in a Nutshell 69
5.1.2 Additional Considerations 71
5.2 The Recipes 72
5.2.1 Serving Linked Data as Static RDF/XML Files 72
5.2.2 Serving Linked Data as RDF Embedded in HTML Files 74
5.2.3 Serving RDF and HTML with Custom Server-Side Scripts 77
5.2.4 Serving Linked Data from Relational Databases 77
5.2.5 Serving Linked Data from RDF Triple Stores 79
5.2.6 Serving Linked Data by Wrapping Existing Application or Web APIs . . 79
5.3 Additional Approaches to Publishing Linked Data 80
5.4 Testing and Debugging Linked Data 80
5.5 Linked Data Publishing Checklist 82

6 Consuming Linked Data 85

6.1 Deployed Linked Data Applications 85
6.1.1 Generic Applications 86
6.1.2 Domain-specific Applications 90
6.2 Developing a Linked Data Mashup 93
6.2.1 Software Requirements 94
6.2.2 Accessing Linked Data URIs 94
6.2.3 Representing Data Locally using Named Graphs 95
6.2.4 Querying Local Data with SPARQL 96
6.3 Architecture of Linked Data Applications 97
6.3.1 Accessing the Web of Data 100
6.3.2 Vocabulary Mapping 101
6.3.3 Identity Resolution 102
6.3.4 Provenance Tracking 103
6.3.5 Data Quality Assessment 103
6.3.6 Caching Web Data Locally 105
6.3.7 Using Web Data in the Application Context 105
6.4 Effort Distribution between Publishers, Consumers and Third Parties 105

7 Summary and Outlook 109

Bibliography 111

Authors' Biographies 121

List of Figures

2.1 URIs are used to identify people and the relationships between them. 9

3.1 Growth in the number of data sets published on the Web as Linked Data. 31

3.2 Linking Open Data cloud as of November 2010. The colors classify data sets by topical domain. 32

5.1 Linked Data Publishing Options and Workflows. 70

5.2 Architecture Diagram of D2R Server 78

6.1 The Marbles Linked Data browser displaying data about Tim Berners-Lee. The colored dots indicate the data sources from which data was merged. 87

6.2 Sig.ma Linked Data search engine displaying data about Richard Cyganiak. 88

6.3 Google search results containing structured data in the form of Rich Snippets. . . 89

6.4 Google result answering a query about the birth date of Catherine Zeta-Jones. . 90

6.5 US Global Foreign Aid Mashup combining and visualizing data from different branches of the US government. 91

6.6 The HTML view of a Talis Aspire List generated from the underlying RDF representation of the data. 92

6.7 Architecture of a Linked Data application that implements the crawling pattern. 99

Preface

This book provides a conceptual and technical introduction to the field of Linked Data. It is intended for anyone who cares about data – using it, managing it, sharing it, interacting with it – and is passionate about the Web. We think this will include data geeks, managers and owners of data sets, system implementors and Web developers. We hope that students and teachers of information management and computer science will find the book a suitable reference point for courses that explore topics in Web development and data management. Established practitioners of Linked Data will find in this book a distillation of much of their knowledge and experience, and a reference work that can bring this to all those who follow in their footsteps.

Chapter 2 introduces the basic principles and terminology of Linked Data. Chapter 3 provides a 30,000 ft view of the Web of Data that has arisen from the publication of large volumes of Linked Data on the Web. Chapter 4 discusses the primary design considerations that must be taken into account when preparing to publish Linked Data, covering topics such as choosing and using URIs, describing things using RDF, data licensing and waivers, and linking data to external data sets. Chapter 5 introduces a number of recipes that highlight the wide variety of approaches that can be adopted to publish Linked Data, while Chapter 6 describes deployed Linked Data applications and examines their architecture. The book concludes in Chapter 7 with a summary and discussion of the outlook for Linked Data.

We would like to thank the series editors Jim Hendler and Frank van Harmelen for giving us the opportunity and the impetus to write this book. Summarizing the state of the art in Linked Data was a job that needed doing – we are glad they asked us. It has been a long process, throughout which Mike Morgan of Morgan & Claypool has shown the patience of a saint, for which we are extremely grateful. Richard Cyganiak wrote a significant portion of the 2007 tutorial "How to Publish Linked Data on the Web" which inspired a number of sections of this book – thank you Richard. Mike Bergman, Dan Brickley, Fabio Ciravegna, Ian Dickinson, John Goodwin, Harry Halpin, Frank van Harmelen, Olaf Hartig, Andreas Harth, Michael Hausenblas, Jim Hendler, Bernadette Hyland, Toby Inkster, Anja Jentzsch, Libby Miller, Yves Raimond, Matthew Rowe, Daniel Schwabe, Denny Vrandecic, and David Wood reviewed drafts of the book and provided valuable feedback when we needed fresh pairs of eyes – they deserve our gratitude. We also thank the European Commission for supporting the creation of this book by funding the LATC – LOD Around The Clock project (Ref. No. 256975). Lastly, we would like to thank the developers of LaTeX and Subversion, without which this exercise in remote, collaborative authoring would not have been possible.

Tom Heath and Christian Bizer
February 2011