MORGAN & CLAYPOOL PUBLISHERS

Linked Data

Evolving the Web into a Global Data Space

Tom Heath
Christian Bizer

Synthesis Lectures on The Semantic Web: Theory and Technology
James Hendler, *Series Editor*

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

Linked Data: Evolving the Web into a Global Data Space

Tom Heath and Christian Bizer

www.morganclaypool.com

ISBN: 9781608454303 paperback
ISBN: 9781608454310 ebook

DOI 10.2200/S00334ED1V01Y201102WBE001

A Publication in the Morgan & Claypool Publishers 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.

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

MORGAN & CLAYPOOL PUBLISHERS

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

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

CHAPTER 1

Introduction

1.1 THE DATA DELUGE

We are surrounded by data – data about the performance of our locals schools, the fuel efficiency of our cars, a multitude of products from different vendors, or the way our taxes are spent. By helping us make better decisions, this data is playing an increasingly central role in our lives and driving the emergence of a data economy [47]. Increasing numbers of individuals and organizations are contributing to this deluge by choosing to share their data with others, including *Web-native* companies such as *Amazon* and *Yahoo!*, newspapers such as *The Guardian* and *The New York Times*, public bodies such as the UK and US governments, and research initiatives within various scientific disciplines.

Third parties, in turn, are consuming this data to build new businesses, streamline online commerce, accelerate scientific progress, and enhance the democratic process. For example:

- The online retailer Amazon makes their product data available to third parties via a Web *API*[1]. In doing so they have created a highly successful ecosystem of affiliates[2] who build micro-businesses, based on driving transactions to Amazon sites.

- Search engines such as Google and Yahoo! consume structured data from the Web sites of various online stores, and use this to enhance the search listings of items from these stores. Users and online retailers benefit through enhanced user experience and higher transaction rates, while the search engines need expend fewer resources on extracting structured data from plain HTML pages.

- Innovation in disciplines such as Life Sciences requires the world-wide exchange of research data between scientists, as demonstrated by the progress resulting from cooperative initiatives such as the Human Genome Project.

- The availability of data about the political process, such as members of parliament, voting records, and transcripts of debates, has enabled the organisation *mySociety*[3] to create services such as *TheyWorkForYou*[4], through which voters can readily assess the performance of elected representatives.

[1] API stands for *Application Programming Interface* - a mechanism for enabling interaction between different software programs.
[2] `https://affiliate-program.amazon.co.uk/`
[3] `http://www.mysociety.org/`
[4] `http://www.theyworkforyou.com/`

The strength and diversity of the ecosystems that have evolved in these cases demonstrates a previously unrecognised, and certainly unfulfilled, demand for access to data, and that those organizations and individuals who choose to share data stand to benefit from the emergence of these ecosystems. This raises three key questions:

- How best to provide access to data so it can be most easily reused?
- How to enable the discovery of relevant data within the multitude of available data sets?
- How to enable applications to integrate data from large numbers of formerly unknown data sources?

Just as the World Wide Web has revolutionized the way we connect and consume documents, so can it revolutionize the way we *discover*, *access*, *integrate* and *use* data. The Web is the ideal medium to enable these processes, due to its ubiquity, its distributed and scalable nature, and its mature, well-understood technology stack.

The topic of this book is on how a set of principles and technologies, known as *Linked Data*, harnesses the ethos and infrastructure of the Web to enable data sharing and reuse on a massive scale.

1.2 THE RATIONALE FOR LINKED DATA

In order to understand the concept and value of Linked Data, it is important to consider contemporary mechanisms for sharing and reusing data on the Web.

1.2.1 STRUCTURE ENABLES SOPHISTICATED PROCESSING

A key factor in the re-usability of data is the extent to which it is well *structured*. The more regular and well-defined the structure of the data the more easily people can create tools to reliably process it for reuse.

While most Web sites have some degree of structure, the language in which they are created, HTML, is oriented towards structuring textual documents rather than data. As data is intermingled into the surrounding text, it is hard for software applications to extract snippets of structured data from HTML pages.

To address this issue, a variety of *microformats*[5] have been invented. Microformats can be used to published structured data describing specific types of entities, such as people and organizations, events, reviews and ratings, through embedding of data in HTML pages. As microformats tightly specify how to embed data, applications can unambiguously extract the data from the pages. Weak points of microformats are that they are restricted to representing data about a small set of different types of entities; they only provide a small set of attributes that may used to describe these entities; and that it is often not possible to express relationships between entities, such as, for example, that

[5] http://microformats.org/

a person is the speaker of an event, rather than being just an attendee or the organizer of the event. Therefore, microformats are not suitable for sharing arbitrary data on the Web.

A more generic approach to making structured data available on the Web are *Web APIs*. Web APIs provide simple query access to structured data over the HTTP protocol. High profile examples of these APIs include the *Amazon Product Advertising API*[6] and the *Flickr API*[7]. The site *ProgrammableWeb*[8] maintains a directory containing several thousand Web APIs.

The advent of Web APIs has led to an explosion in small, specialized applications (or *mashups*) that combine data from several sources, each of which is accessed through an API specific to the data provider. While the benefits of programmatic access to structured data are indisputable, the existence of a specialized API for each data set creates a landscape where significant effort is required to integrate each novel data set into an application. Every programmer must understand the methods available to retrieve data from each API, and write custom code for accessing data from each data source.

1.2.2 HYPERLINKS CONNECT DISTRIBUTED DATA

It is common for Web APIs to provide results in structured data formats such as XML and JSON[9], which have extensive support in a wide range of programming languages. However, from a Web perspective, they have some limitations, which are best explained by comparison with HTML. The HTML specification defines the *anchor* element, `a`, one of the valid attributes of which is the `href`. When used together, the anchor tag and `href` attribute indicate an outgoing link from the current document. Web *user agents*, such as browsers and search engine crawlers, are programmed to recognize the significance of this combination, and either render a clickable link that a human user can follow, or to traverse the link directly in order to retrieve and process the referenced document. It is this connectivity between documents, supported by a standard syntax for indicating links, that has enabled the Web of documents. By contrast, the data returned from the majority of Web APIs does not have the equivalent of the HTML *anchor* tag and `href` attribute, to indicate links that should be followed to find related data.

Furthermore, many Web APIs refer to items of interest using identifiers that have only local scope – e.g., a product identifier *123456* that is meaningless when taken out of the context of that specific API. In such cases, there is no standard mechanism to refer to items described by one API in data returned by another.

Consequently, data returned from Web APIs typically exists as isolated fragments, lacking reliable onward links signposting the way to related data. Therefore, while Web APIs make data accessible *on the Web*, they do not place it truly *in the Web*, making it linkable and therefore discoverable.

[6] http://docs.amazonwebservices.com/AWSECommerceService/latest/DG/
[7] http://www.flickr.com/services/api/
[8] http://www.programmableweb.com/
[9] http://www.json.org/

To return to the comparison with HTML, the analogous situation would be a search engine that required *a priori* knowledge of all Web documents before it could assemble its index. To provide this *a priori* knowledge, every Web publisher would need to register each Web page with each search engine. The ability for anyone to add new documents to the Web at will, and for these documents to be automatically discovered by search engines and humans with browsers, have historically been key drivers of the Web's explosive growth. The same principles of linking, and therefore ease of discovery, can be applied to data on the Web, and Linked Data provides a technical solution to realize such linkage.

1.3 FROM DATA ISLANDS TO A GLOBAL DATA SPACE

Linking data distributed across the Web requires a standard mechanism for specifying the existence and meaning of connections between items described in this data.

This mechanism is provided by the Resource Description Framework (RDF), which is examined in detail in Chapter 2. The key things to note at this stage are that RDF provides a flexible way to describe things in the world – such as people, locations, or abstract concepts – and how they relate to other things. These statements of relationships between things are, in essence, links connecting things in the world. Therefore, if we wish to say that a book described in data from one API is for sale at a (physical) bookshop described in data from a second API, and that bookshop is located in a city described by data from a third, RDF enables us to do this, and publish this information on the Web in a form that others can discover and reuse.

To conclude the comparison with HTML documents and conventional Web APIs, the key features of RDF worth noting in this context are the following:

- **RDF links things, not just documents**: therefore, in the book selling example above, RDF links would not simply connect the data fragments from each API, but assert connections between the entities described in the data fragments – in this case the book, the bookshop and the city.

- **RDF links are typed**: HTML links typically indicate that two documents are related in some way, but mostly leave the user to infer the nature of the relationship. In contrast, RDF enables the data publisher to state explicitly the nature of the connection. Therefore, in practice, the links in the book selling example above would read something like: *mybook* `forSaleIn` *thatbookshop*, *thatbookshop* `locatedIn` *mycity*.

While these sorts of connections between things in the world may be implicit in XML or JSON data returned from Web APIs, RDF enables Web publishers to make these links explicit, and in such a way that RDF-aware applications can follow them to discover more data. Therefore, a Web in which data is both published and linked using RDF is a Web where data is significantly more discoverable, and therefore more usable.

Just as hyperlinks in the classic Web connect documents into a single global information space, Linked Data enables links to be set between items in different data sources and therefore connect these sources into a single global data space. The use of Web standards and a common data model make it possible to implement generic applications that operate over the complete data space. This is the essence of *Linked Data*.

Increasing numbers of data providers and application developers have adopted Linked Data. In doing so they have created this global, interconnected data space - *the Web of Data*. Echoing the diversity of the classic document Web, the Web of Data spans numerous topical domains, such as people, companies, films, music, locations, books and other publications, online communities, as well as an increasing volume of scientific and government data.

This *Web of Data* [30], also referred to as *Semantic Web* [21], presents a revolutionary opportunity for deriving insight and value from data. By enabling seamless connections between data sets, we can transform the way drugs are discovered, create rich pathways through diverse learning resources, spot previously unseen factors in road traffic accidents, and scrutinise more effectively the operation of our democratic systems.

The focus of this book is data sharing in the context of the public Web. However, the principles and techniques described can be equally well applied to data that exists behind a personal or corporate firewall, or that straddles the public and the private. For example, many aspects of Linked Data have been implemented in desktop computing environments through the *Semantic Desktop* initiative[10]. Similarly, these principles can be employed entirely behind the corporate firewall, to help ease the pain of data integration in enterprise environments [114]. The *Linking Open Drug Data* [68] initiative represents a hybrid scenario, where Linked Data is enabling commercial organizations to connect and integrate data they are willing to share with each other for the purposes of collaboration.

1.4 INTRODUCING BIG LYNX PRODUCTIONS

Throughout this book we will illustrate the principles and technical aspects of Linked Data with examples from a scenario involving *Big Lynx Productions*. *Big Lynx* is a (fictional) independent television production company specialising in wildlife documentaries, primarily produced under contract for major television networks in the UK. The company employs around 30 permanent staff, such as *Managing Director* Dave Smith, *Lead Cameraman* Matt Briggs, and *Webmaster* Nelly Jones, plus a large team of freelancers that evolves according to the needs of current contracts.

Big Lynx maintains its own Web site at http://biglynx.co.uk/ that contains:

- information about the company's goals and structure
- profiles of the permanent staff and of freelancers
- listings of vacancies for freelancers to work on specific contracts
- listings of productions that have been broadcast by the commissioning network
- a blog where staff post news items of interest to the television networks and/or freelancers

[10] http://www.semanticdesktop.org/

Information that changes rarely (such as the company overview) is published on the site as static HTML documents. Frequently changing information (such as listing of productions) is stored in a relational database and published to the Web site as HTML by a series of PHP scripts developed for the company. The company blog is based on a blogging platform developed in-house and forms part of the main *Big Lynx* site.

In the remainder of this book we will explore how Linked Data can be integrated into the workflows and technical architectures of *Big Lynx*, thereby maximising the discoverability of the *Big Lynx* data and making it easy for search engines as well as specialized Web sites, such as film and TV sites, freelancer directories or online job markets, to pick up and integrate *Big Lynx* data with data from other companies.