Validating RDF Data

Synthesis Lectures on Semantic Web: Theory and Technology

Synthesis Lectures on the Semantic Web: Theory and Technology is edited by Ying Ding of Indiana University and Paul Groth of Elsevier Labs. 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 included:

- 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

Validating RDF Data
Jose Emilio Labra Gayo, Eric Prud'hommeaux, Iovka Boneva, and Dimitris Kontokostas
2017

Natural Language Processing for the Semantic Web
Diana Maynard, Kalina Bontcheva, and Isabelle Augenstein
2016

The Epistemology of Intelligent Semantic Web Systems
Mathieu d'Aquin and Enrico Motta
2016

Entity Resolution in the Web of Data
Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis
2015

Library Linked Data in the Cloud: OCLC's Experiments with New Models of Resource Description
Carol Jean Godby, Shenghui Wang, and Jeffrey K. Mixter
2015

Semantic Mining of Social Networks
Jie Tang and Juanzi Li
2015

Social Semantic Web Mining
Tope Omitola, Sebastián A. Ríos, and John G. Breslin
2015

Semantic Breakthrough in Drug Discovery
Bin Chen, Huijun Wang, Ying Ding, and David Wild
2014

Semantics in Mobile Sensing
Zhixian Yan and Dipanjan Chakraborty
2014

Provenance: An Introduction to PROV
Luc Moreau and Paul Groth
2013

Resource-Oriented Architecture Patterns for Webs of Data
Brian Sletten
2013

Aaron Swartz's A Programmable Web: An Unfinished Work
Aaron Swartz
2013

Incentive-Centric Semantic Web Application Engineering
Elena Simperl, Roberta Cuel, and Martin Stein
2013

Publishing and Using Cultural Heritage Linked Data on the Semantic Web
Eero Hyvönen
2012

VIVO: A Semantic Approach to Scholarly Networking and Discovery
Katy Börner, Michael Conlon, Jon Corson-Rikert, and Ying Ding
2012

Linked Data: Evolving the Web into a Global Data Space
Tom Heath and Christian Bizer
2011

Validating RDF Data

Jose Emilio Labra Gayo, Eric Prud'hommeaux, Iovka Boneva, and Dimitris Kontokostas

ISBN: 978-3-031-79477-3 paperback
ISBN: 978-3-031-79478-0 ebook
ISBN: 978-3-031-79478-0 e-pub

DOI 10.1007/978-3-031-79478-0

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

Lecture #16
Series Editors: Ying Ding, *Indiana University*
Paul Groth, *Elsevier Labs*
Founding Editor Emeritus: James Hendler, *Rensselaer Polytechnic Institute*
Series ISSN
Print 2160-4711 Electronic 2160-472X

Cover art: Amy van der Hiel

Validating RDF Data

Jose Emilio Labra Gayo
University of Oviedo

Eric Prud'hommeaux
W3C/MIT and Micelio

Iovka Boneva
University of Lille

Dimitris Kontokostas
University of Leipzig

SYNTHESIS LECTURES ON SEMANTIC WEB: THEORY AND TECHNOLOGY #16

ABSTRACT

RDF and Linked Data have broad applicability across many fields, from aircraft manufacturing to zoology. Requirements for detecting bad data differ across communities, fields, and tasks, but nearly all involve some form of data validation. This book introduces data validation and describes its practical use in day-to-day data exchange.

The Semantic Web offers a bold, new take on how to organize, distribute, index, and share data. Using Web addresses (URIs) as identifiers for data elements enables the construction of distributed databases on a global scale. Like the Web, the Semantic Web is heralded as an information revolution, and also like the Web, it is encumbered by data quality issues. The quality of Semantic Web data is compromised by the lack of resources for data curation, for maintenance, and for developing globally applicable data models.

At the enterprise scale, these problems have conventional solutions. Master data management provides an enterprise-wide vocabulary, while constraint languages capture and enforce data structures. Filling a need long recognized by Semantic Web users, shapes languages provide models and vocabularies for expressing such structural constraints.

This book describes two technologies for RDF validation: Shape Expressions (ShEx) and Shapes Constraint Language (SHACL), the rationales for their designs, a comparison of the two, and some example applications.

KEYWORDS

RDF, ShEx, SHACL, shape expressions, shapes constraint language, data quality, web of data, Semantic Web, linked data

Contents

Preface

This book describes two languages for implementing constraints on RDF data, describing the main features of both Shape Expressions (ShEx) and Shapes Constraint Language (SHACL) from a user perspective, and also offering a comparison of the technologies. Throughout this book, we develop a small number of examples that typify validation requirements and demonstrate how they can be met with ShEx and SHACL. The book is not intended to be a formal specification of the languages, for which the interested reader can consult the corresponding reference documents, but rather, it is meant to serve as an introduction to the technologies with some background about the rationale of their design and some points of comparison.

Chapter 1 provides a brief introduction to the topic. Chapter 2 presents a short overview of the RDF data model and RDF-related technologies; this chapter could be skipped by any reader who already knows RDF or Turtle. Chapter 3 helps the reader to understand what to expect from data validation. It describes the problem of RDF validation and some approaches that have been proposed. This book specifically reviews two of these approaches in further detail: ShEx (Chapter 4) and SHACL (Chapter 5). These chapters describe each language and provide a practical introduction using examples. Following the discussion of both languages, Chapter 6 presents some applications using either ShEx, SHACL, or both. Finally, Chapter 7 compares ShEx and SHACL and offers some conclusions.

The goal of this book is to serve as a practical introduction to ShEx and SHACL using examples. While we omitted formal definitions or specifications, references for further reading can be found at the end of each chapter. We give a quick overview of some background and related technologies so that readers without RDF knowledge can follow the book's contents. Also, it is not necessary to have any prior knowledge of programming or ontologies to understand RDF validation technologies. The intended audience is anyone interested in data representation and quality.

Jose Emilio Labra Gayo, Eric Prud'hommeaux, Iovka Boneva, and Dimitris Kontokostas
July 2017

Foreword by Phil Archer

"Anyone can say anything about anything," says the mantra for the Semantic Web. More formally, the Semantic Web adopts the Open World Assumption: just because your data encodes a set of facts, that doesn't mean there aren't other facts stated elsewhere about the same thing. All of which is fine and part of the design of RDF which supports the creation of a graph at Web scale, but in a lot of practical applications you just need to know whether the triples you've ingested match what you were expecting; you need validation. You might think of it as a defined subset of the whole graph, or maybe a profile, providing a huge boost to interoperability between disparate systems. If you can validate the data you've received then you can process it with confidence, using more terse code, perhaps with more performant queries. I don't accept that RDF is hard, certainly no harder than any other Web technology; what is hard is thinking in graphs. Keeping in your head that this node supports these properties and has relationships with those other nodes becomes complex for anything other than trivial datasets. The validation techniques set out in this book provide a means to tame that complexity, to set out for humans and machines exactly what the structure of the data is or should be. That's got to be helpful and, incidentally, ties in with new work now under way at W3C on dataset exchange. In my role at W3C I watched as the SHACL and ShEx camps tried hard to converge on a single method: they couldn't, hence the two different approaches. Both are described in detail here with copious examples, which is just what you need to get started. How can you choose between the two methods? Chapter 7 gives a detailed comparison and allows you to make your own choice. Whichever you choose, this is the book you need to make sense of RDF validation.

Phil Archer, Former W3C Data Strategist
July 2017

Foreword by Tom Baker

The technologies described here meet a need first recognized, albeit dimly, two decades ago. Rewind to circa 2000, when the parallel development of two W3C specifications, XML Schema and RDF Schema, both called "schema languages" but with radically different uses, caused some confusion.

This confusion permeated our early discussions about the Dublin Core. Was it an XML format, an RDF vocabulary, or somehow both? Could metadata just follow arbitrary formats or did it need a data model? In 2000, the Dublin Core community turned to "application profiles" as a way to mix and match multiple vocabularies to meet specific needs, and the idea was an instant hit even if people disagreed about their use. Were they more for validating data, or more about finding rough consensus on a metadata model within a community of practice? Attempts to bridge the XML and RDF mindsets in the DCMI community, notably with a Description Set Profile constraint language for validating RDF-based metadata (2008), never quite caught on. Perhaps the idea needed a bigger push?

Fast-forward to 2013, when W3C convened a workshop on RDF validation which revealed that many communities had been circling around the same issues, and which ultimately led to the results described here [82]. This book focuses on data validation, an addition to the Semantic Web stack that is long overdue. But from a DCMI perspective, the ideas for future work outlined in its Conclusion are just as exciting: the prospect of using ShEx- or SHACL-based application profiles to map and convert between data models, size up aggregated datasets, or provide nuanced feedback to data providers on quality issues. ShEx and SHACL, finally production-ready, are full of potential for further evolution.

Tom Baker, Dublin Core Metadata Initiative
July 2017

Foreword by Dan Brickley and Libby Miller

People think RDF is a pain because it is complicated. The truth is even worse. RDF is painfully simplistic, but it allows you to work with real-world data and problems that are horribly complicated. While you can avoid RDF, it is harder to avoid complicated data and complicated computer problems. RDF brings together data across application boundaries and imposes no discipline on mandatory or expected structures. This can make working with RDF data frustrating. Its schema and ontology languages can help define the meaning of RDF content but, again, can't express rules about how actual data records should look. The contents of this book are nearly 20 years too late, but better now than never. Recent developments around RDF validation have finally made it easier to record, exchange, and understand rules about validating and otherwise checking RDF data. Who knows what wonders await us in another 20 years.

Dan Brickley, Schema.org and Google
Libby Miller, BBC
July 2017

Acknowledgments

This book started as an RDF validation tutorial that we prepared for the International Semantic Web Conference (ISWC) 2016. The slides are still available at http://weso.github.io/RDFValidation_ESWC16/. After the tutorial, the editors invited us to extend the slides to write a full book on the subject.

We want to thank Harold Solbrig for his collaboration with the tutorial and his support and participation in the development of ShEx.

Although the task of extending the tutorial seemed easy at first, writing a book on this subject was much more daunting than expected. One challenge was to update the material to the changes that ShEx and SHACL were experimenting with during the W3C process.

There were different points of view on how to tackle the RDF validation problem that led to the appearance of two camps, which at the time of this writing are represented in two W3C groups: the Shape Expressions community group and the SHACL community group. We would like to acknowledge the people participating in those groups as well as the people that were part of the W3C Data Shapes Working group.

During the development of the book we were also updating our implementations. We also want to thank all the people that have been using our implementations and online demos for their patience when something was not working as expected and for submitting issues or suggesting improvements to the tools.

We want to thank Amy van der Hiel for the design of the cover art, as well as Dan Brickley and Tom Baker for their comments and suggestions.

We are also grateful to Vladimir Alexiev for his extensive review and proofreading of the book.

Jose Emilio Labra Gayo wants to thank his colleagues from Oviedo3 and the WESO research group. Special thanks to Edita for her support and patience. He dedicates the book to Sergio, Nuria, and Alex.

Dimitris wants to thank Sebastian Hellmann and the University of Leipzig for supporting his participation in the Shape Expressions and Data Shapes groups as well as his family for their continuous support.

Eric wants to dedicate this book to his family, who put up with him writing when he should be playing. He also has great gratitude to Gregg Kellogg, who always takes time to hack with him, and Tom Baker, who always takes time to write with him.

Jose Emilio Labra Gayo, Eric Prud'hommeaux, Iovka Boneva, and Dimitris Kontokostas
July 2017