

Linking Enterprise Data

David Wood
Editor

Linking Enterprise Data

 Springer

Editor

David Wood
3 Round Stones LLC
22408 Fredericksburg Virginia
USA
david@3roundstones.com

ISBN 978-1-4419-7664-2 e-ISBN 978-1-4419-7665-9
DOI 10.1007/978-1-4419-7665-9
Springer New York Dordrecht Heidelberg London

© Springer Science+Business Media, LLC 2010

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. 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.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

"All problems in computer science can be solved by another layer of indirection, but that will usually create another problem."

David John Wheeler (1927 - 2004)

Preface

Linking Enterprise Data is a new concept, based on an idea more than twenty years old. Tim Berners-Lee's original proposal for the World Wide Web in March 1989 was based on a system of linked information systems. The early Web was intended to interlink information from various systems to solve organizational problems, such as the high turnover of people and the restriction of information to data silos. The hope was to create a distributed information system that would allow "a pool of information to develop which could grow and evolve with the organisation and the projects it describes."

The Web has grown into the world's largest information system. By 2000, Web architecture had been dissected and described by Roy Fielding. Representational State Transfer (REST) was Roy's answer to why the Web worked so well. In a world plagued by software problems, machine crashes, and network outages, the Web never fails. The Web is robust and resilient to change. The Web survives changing machinery, operating system updates, changes in the way we structure index and find information. No other software system provides the features and functions of World Wide Web.

Linked Data techniques have become interesting to organizations of every shape and size. The Linked Open Data (LOD) project began as a community effort of the World Wide Web Consortium's Semantic Web Education and Outreach Group. The project has begun to turn the document-oriented Web into a database of global proportions. The ability of the modern Web to deal with both documents and data have shaped a general solution for information dissemination and integration. The time for linking enterprise data has come.

This book records some of the earliest production applications of linking enterprise data. Parts of it serve as a roadmap for those seeking to replicate their successes. Part I of this book attempts to answer the question why enterprise data should be linked. The chapters in Part I provide valuable guidance to those writing business cases, for those needing to justify internal development efforts, or for those writing requests for proposals to external vendors. Dean Allemang discusses why enterprises must adopt Web techniques for data integration and provides such techniques fit into enterprise systems. Dean makes a strong case that enterprises

must change the way they approach information technology systems. Indeed, since information systems have such a profound impact on the operational aspects of a business, he makes the case that enterprises need to change the way they approach their operations.

Edward Curry, Andre Freitas, and Sean O’Riain discuss the role of community-based data curation. Enterprises have become more distributed, less centrally managed and less integrated in their systems. The lessons Ed and his colleagues have captured from real-world attempts to curate distributed data for the purposes of ensuring data quality will apply to many enterprises. They provide some important best practices extracted from early adopters of Linked Data techniques.

Part II is short, but critically important. Part II provides material assistance for business managers seeking to propose Linked Data projects. Bernadette Hyland discusses the characteristics of enterprises ready to take on Linked Data projects and provides useful fodder for business cases. Her simple guidelines for getting a Linked Data project started have generally been lacking in the public discussion to date. Kristen Harris’ real-world experiences creating and managing the *sworDFish* project at Sun Microsystems demonstrated the potential of linked enterprise data to integrate disparate systems in large enterprises. She provides guidance for the navigation of corporate management to approve and support projects with far-reaching infrastructural ramifications.

The techniques of Linked Data can be subtle and technical, although not out of reach for those with traditional enterprise skills. Part III provides three explanatory chapters that address different technical aspects of linking data. Alexandre Passant, Philippe Laublet, John G. Breslin and Stefan Decker present ways to integrate enterprise social networking solutions such as wikis and blogs. Initial enterprise adoption of new technologies can sometimes create new problems. Alexandre and his co-authors offer both insight and solutions to the integration of Web 2.0 and Web 3.0 techniques.

Roberto Garcia and Rosa Gil demonstrate how the translation of existing data sources may be brought to the Web of data. Reza B’Far and I offer technical approaches to enterprise problems of scale. Reza addresses logical reasoning techniques for enterprise-scale data and I present ways to ensure the long-term viability of Linked Data identifiers.

Part IV provides five success stories from the front line of enterprise adoption. Each story highlights a different aspect of Linked Data in an enterprise context. Thomas Baker and Johannes Keizer address standards for highly distributed operations developed for the Food and Agriculture Organization of the United Nations. Steve Harris, Tom Ilube and Mischa Tuffield show how Linked Data techniques are used as the basis for their Web-scale company, Garlik. Constantine Hondros of publisher Wolters Kluwer illustrates and presents several approaches for solving integration problems of textual content. Chimezie Ogbuji develops a new enterprise system using a Linked Data approach and Yves Raimond, Tom Scott, Silver Oliver, Patrick Sinclair and Michael Smethurst of the British Broadcasting Corporation present their innovative corporate treatment of the Web itself as their content management system.

We have been able to draw some tentative conclusions regarding success criteria for Linked Data projects in an enterprise. First and foremost may be from Jeff Pollock of Oracle Corporation when he said, "If information systems are to keep up with business, we need to change more than technology - we need to change how people deal with technology." Linked Data techniques offer us a means to do just that; we can radically change the interfaces to our existing systems while we build upon them. We can wrap and expose our silos in order to layer a Web-like distributed system over them.

Secondly, the lessons of the Web clearly apply to enterprises. The Web works for some very good, and very explainable, reasons. Those reasons transcend Representational State Transfer (REST), the architectural principles that underlay the idealized Web, and add the techniques of the Semantic Web, especially that subset being used by the Linked Data community. Individual technologies, though, are clearly less important than techniques that have proven their worth. Technologies continue to evolve; good techniques are more resilient and worth building upon.

Note how different the organizations in the success stories are from one another! A broadcasting company, a publishing company, a healthcare provider, a data security firm, an international policy organization. Other chapters referenced other types of organizations, including a utility company and library organizations. If Linked Data techniques work for all of them, those techniques are very likely to apply to others.

All of our success stories have some interesting commonalities: At least one expert in Semantic Web techniques was used by each organization. Each attacked a significant business problem instead of relying on the technologies to "sell themselves". Each leveraged significant existing investments, especially those with captured or implied semantics. Every success relied upon universal addressing of resources via the Web's Uniform Resource Indicator (URI) scheme.

There were also some major differences between the success stories. Those differences define tools and techniques that are more situationally dependent. The most noteworthy is that very different degrees of data modeling were employed. Complete, top-down data modeling is expensive, difficult and should be undertaken only where it provides value. Specific technologies to describe data (OWL, SKOS, RDF serialization formats) varied widely, as did the use of the SPARQL query language.

Trust may be a larger issue in intra-business data than it is on the general Web if business decisions are being made based on the information. Issues of trust in large organizations may be facilitated by social considerations, e.g. via signing of work, taking credit for additions or edits, tying comments to logins. Many of today's enterprises are large and distributed enough to make use of Web techniques for building and maintaining trust socially over a technical framework.

The four parts of this book are presented hierarchically, like most books in the last 2300 years of Western tradition. The material in this book should not be thought of as a hierarchy, but rather like a graph, like the Web itself. All of the chapters in this book contain nuggets of information useful to enterprise professionals looking to apply Linked Data techniques. The opening chapters do address technology and success stories as well as laying and conceptual foundation. The technique chapters

reference success stories of their own. The chapters addressing success stories are chock full of lessons learned in relation to management, approach and style. It is no more possible to fit these chapters into a strict hierarchy than it is to do so with content on the Web. Readers are encouraged to troll the index and review the notes at the beginning of each Part to find the information most relevant to themselves.

The enterprise application of Web architecture to business problems is in its infancy. We hope that this book can be used to assist those managers, data professionals and developers at the forefront of solving today's formidable enterprise data challenges.

Observant readers may notice that any given chapter may use either American and British spelling. The use of mixed spelling systems represents the international nature of the contributing authors and, indeed, the international range of Linked Data research and deployment. We consider such diversity to be a feature, not a bug.

Nigam Shah of the National Center for Biomedical Ontology provided reviews and commentary on this book's contents, as did most of the individual chapter authors. Ivan Herman of the W3C and Eric Miller and Uche Ogbuji of Zepheira provided introductions to prospective authors and suggested content. Our editors at Springer, Susan Lagerstrom-Fife and Jennifer Maurer helped to make the creation of this book much easier than it could have been. Thank you to all.

Fredericksburg, Virginia, USA,

David Wood
June 2010

Contents

Part I Why Link Enterprise Data?

Semantic Web and the Linked Data Enterprise	3
Dean Allemang	
1 Social Data in the Enterprise	3
1.1 Causes	5
1.2 Technology Solutions	6
1.3 Localization and Globalization	9
2 The Linked Data Enterprise	10
2.1 Controlled Vocabularies	11
2.2 Prerequisites for Linked Data Vocabularies	18
3 Examples	20
3.1 Publishing	20
3.2 Government	21
4 Conclusions	22
References	22
The Role of Community-Driven Data Curation for Enterprises	25
Edward Curry, Andre Freitas, and Sean O’Riáin	
1 Introduction	25
2 The Business Need for Curated Data	26
3 Data Curation	28
3.1 How to Curate Data	28
4 Community-based Curated Enterprise Data	30
4.1 Internal Corporate Community	30
4.2 External Pre-competitive Communities	31
5 Case Study: Wikipedia - The World Largest Open Digital Curation Community	32
5.1 Social Organization	33
5.2 Artifacts, Tools and Processes	34
5.3 DBPedia - Community Curated Linked Open Data	35

6	Case Study: The New York Times - 100 Years of Expert Data Curation	36
6.1	Data Curation	36
6.2	Publishing Curated Linked Data	37
7	Case Study: Thomson Reuters - Data Curation, a Core Business Competency	38
7.1	Data Curation	39
8	Case Study: ChemSpider - Open Data Curation in the Global Chemistry Community	40
8.1	Community Objectives	41
8.2	Curation Approach & Types	41
9	Case Study: Protein Data Bank, Pre-competitive Bioinformatics ..	42
9.1	Serving the Community	42
9.2	Curation Approaches & Types	42
9.3	Observations	43
10	Case Study Learnings	44
10.1	Social Best Practices	44
10.2	Technical Best Practices	45
11	Conclusion	46
	References	46

Part II Approval and Support of Linked Data Projects

Preparing for a Linked Data Enterprise 51

Bernadette Hyland

1	Introduction	52
2	The Cost of Linked Data	52
2.1	The Cost of Services and Support	53
2.2	Education and Training	53
2.3	Infrastructure	54
3	Is your Organization Ready for Linked Data?	54
4	The Linked Data Initiative	57
5	A Decentralized Approach to Data Management	58
6	Being On the Web vs. In the Web	59
7	Leverage Vocabularies	60
8	A Simple Approach to Linked Data	61
9	Conclusions	62
9.1	Prepare for a Linked Data Enterprise	62
	References	63

Selling and Building Linked Data: Drive Value and Gain Momentum 65

Kristen Harris

1	The Data Burden	66
2	Driving Value Principles	67
3	Building a Team	70
4	Committing to Something Bigger	72

- 5 Putting it together 73
- 6 Conclusions 74
- References 76

Part III Techniques for Linking Enterprise Data

Enhancing Enterprise 2.0 Ecosystems Using Semantic Web and Linked Data Technologies: The SemSLATES Approach 79

Alexandre Passant, Philippe Laublet, John G. Breslin and Stefan Decker

- 1 Introduction 80
- 2 Issues with Current Enterprise 2.0 Ecosystems 81
 - 2.1 Information Fragmentation and Heterogeneity of Data Formats 82
 - 2.2 Knowledge Capture and Re-use 83
 - 2.3 Tagging and Information Retrieval 83
- 3 SemSLATES: A Social and Semantic Middleware Approach for Enterprise 2.0 84
 - 3.1 The SemSLATES Architecture 85
 - 3.2 Ontologies for Enterprise 2.0 88
 - 3.3 Generating Semantic Annotations Through Software Add-ons 89
 - 3.4 Deploying Additional Services 90
- 4 Case-study: Enabling SemSLATES at EDF R&D 91
 - 4.1 Background 91
 - 4.2 Extending Popular Ontologies 92
 - 4.3 Automated SIOC-based Annotations 93
 - 4.4 Knowledge Capture Using UfoWiki 93
 - 4.5 Semantic Tagging Add-ons 95
 - 4.6 Additional Features of the Platform 96
- 5 Conclusion 99
- References 100

Linking XBRL Financial Data 103

Roberto García and Rosa Gil

- 1 Introduction 103
 - 1.1 XBRL 106
 - 1.2 Related Work 108
- 2 Approach 109
 - 2.1 XSD2OWL Mapping 110
 - 2.2 XML2RDF Mapping 112
 - 2.3 Algorithm 113
- 3 Results 113
 - 3.1 Links to External Data 115
 - 3.2 Semantic Integration 118
- 4 Evaluation 119
 - 4.1 Use Case 122

5 Conclusions and Future Work 122
 References 124

Scalable Reasoning Techniques for Semantic Enterprise Data 127

Reza B’Far

1 Introduction 127
 2 Survey of Reasoning Techniques 128
 2.1 Traditional Rule Engines 130
 2.2 Forward Chaining and the RETE algorithm 131
 2.3 Backward Chaining 132
 3 Bayesian Networks 133
 3.1 Representing Probabilities within the Ontological Model 135
 4 Unsupervised Reasoning 136
 5 Semantic Reasoning 137
 5.1 Performance and Reasoning 139
 5.2 Applying Best-First Search (A* Search) to Semantic Reasoning 140
 5.3 High-level View of Distributed Reasoning 140
 5.4 Map-Reduce and Similar Techniques 141
 5.5 Performance and Ontology Engineering 143
 6 Semantic Reasoning vs. Business Intelligence 143
 7 Best Practices for Application Developers and System Integrators 144
 8 Summary 146
 References 146

Reliable and Persistent Identification of Linked Data Elements 149

David Wood

1 Introduction 150
 2 Metadata Before the World Wide Web 150
 3 Metadata on the World Wide Web 154
 4 Persistent URLs 159
 5 Extending Persistent URLs for Web Resource Curation 160
 6 Redirection of URL Fragments 164
 7 Using Persistent URLs and Retrieved Metadata 164
 8 Federations of PURL Servers 166
 9 Conclusions and Further Work 170
 References 171

Part IV Success Stories

Linked Data for Fighting Global Hunger: Experiences in setting standards for Agricultural Information Management 177

Thomas Baker and Johannes Keizer

1 Agricultural information and Semantic Web 177
 2 Integrating access using Dublin Core metadata 180
 3 AGROVOC and specialized domain ontologies 186

- 4 Networking, capacity development, and outreach 197
- References 201
- Enterprise Linked Data as Core Business Infrastructure 203**
- Steve Harris and Tom Ilube and Mischa Tuffield
- 1 Introduction 203
- 2 Motivations 204
- 3 Garlik’s System Architectures 206
 - 3.1 DataPatrol 207
 - 3.2 QDOS 211
- 4 Schema Driven Software Deployment 215
- 5 Technology and the Need to Scale 216
 - 5.1 4store 216
 - 5.2 5store 217
- 6 Conclusions 218
- 7 Future Work 219
- References 219
- Standardizing Legal Content with OWL and RDF 221**
- Constantine Hondros
- 1 Introduction 221
 - 1.1 The problem domain 221
 - 1.2 Application of Semantic Web technologies 222
- 2 Toward a Common Legal Content Format 223
- 3 OWL Ontology 224
 - 3.1 Creating the ontology 224
 - 3.2 Domain Ontology Mapping 226
- 4 Content Architecture 227
 - 4.1 Modularized XHTML + RDFa for Textual Content 227
 - 4.2 RDF for Metadata, Relations and Classifications 228
- 5 Working with RDF in a Content Supply Chain 229
 - 5.1 The Open World Enigma 230
 - 5.2 Ensuring RDF Data Integrity 230
 - 5.3 Managing Fragmented Ontologies 232
 - 5.4 Managing Performance 232
 - 5.5 Using RDF with XSLT 233
- 6 Enabling Large-Scale Triple Production 234
 - 6.1 Experimental XSD generation 235
 - 6.2 RDFBeans 236
- 7 Conclusions 238
- References 239

A Role for Semantic Web Technologies in Patient Record Data Collection 241

Chimezie Ogbuji

- 1 Introduction 242
- 2 Architectural Styles 243
 - 2.1 REST Architectural Style 243
 - 2.2 Service Oriented Architecture 244
- 3 Semantic Web Technologies 246
- 4 SemanticDB Concurrent Data Collection Workflow 247
 - 4.1 Requirements 248
 - 4.2 XML and RDF Content Management 249
 - 4.3 RESTful XSLT Services 250
 - 4.4 Declarative AJAX Framework 250
 - 4.5 Implementation 251
- 5 General Architectural Observations 257
- 6 Review of Service-oriented Metrics 257
- 7 Conclusions 260
- References 260

Use of Semantic Web technologies on the BBC Web Sites 263

Yves Raimond, Tom Scott, Silver Oliver, Patrick Sinclair and Michael Smethurst

- 1 Introduction 263
 - 1.1 Linking microsites for cross-domain navigation 264
 - 1.2 Making data available to developers 264
 - 1.3 Making use of the wider Web 265
- 2 Programme support on the Web 265
 - 2.1 BBC Programmes 266
 - 2.2 The Programmes Ontology 266
 - 2.3 Web identifiers for broadcast radio and television sites ... 269
- 3 BBC Music 270
 - 3.1 BBC Music as Linked Data 271
 - 3.2 Web identifiers for BBC Music 271
 - 3.3 The Web as a content management system 272
 - 3.4 Using the BBC Programmes and the BBC Music
Linked Data 272
- 4 BBC Wildlife Finder 274
 - 4.1 The Wildlife Ontology 274
 - 4.2 Web identifiers 276
 - 4.3 The Web as a Content Management System 278
 - 4.4 The importance of curation 278
- 5 Journalism 279
 - 5.1 Populating and using the ontology 280
 - 5.2 Future developments 281
- 6 Conclusion 282
- References 283

Contents	xvii
Glossary	285
Index	289

List of Contributors

Dean Allemang

TopQuadrant, Inc., 330 John Carlyle Street, Suite 180, Alexandria, VA 22314-5760, USA, e-mail: dallemang@topquadrant.com

Thomas Baker

Washington DC, USA, e-mail: tbaker@tbaker.de

Reza B'Far

Oracle Corporation, 500 Oracle Parkway, Redwood Shores, CA 94065, USA, e-mail: reza.bfar@oracle.com

John G. Breslin

Digital Enterprise Research Institute, National University of Ireland, Galway, IDA Business Park, Lower Dangan, Galway, Ireland, e-mail: john.breslin@deri.org and Department of Electronic Engineering, National University of Ireland, Galway, Galway, Ireland, e-mail: john.breslin@nuigalway.ie

Edward Curry

Digital Enterprise Research Institute, National University of Ireland, Galway, IDA Business Park, Lower Dangan, Galway, Ireland, e-mail: ed.curry@deri.org

Stefan Decker

Digital Enterprise Research Institute, National University of Ireland, Galway, IDA Business Park, Lower Dangan, Galway, Ireland, e-mail: stefan.decker@deri.org

Andre Freitas

Digital Enterprise Research Institute, National University of Ireland, Galway, IDA Business Park, Lower Dangan, Galway, Ireland, e-mail: andre.freitas@deri.org

Roberto Garcia

Universitat de Lleida. Jaume II, 69. 25001 Lleida, Spain, e-mail:

rgarcia@diei.udl.cat

Rosa Gil

Universitat de Lleida, Jaume II, 69. 25001 Lleida, Spain, e-mail:

rgil@diei.udl.cat

Kristen Harris

Oracle Corporation, 500 Oracle Parkway, Redwood Shores, CA 94065, USA,

e-mail: kristen.harris@oracle.com

Steve Harris

Garlik Ltd, 1-3 Halford Road, London TW10 6AW, United Kingdom, e-mail:

steve.harris@garlik.com

Constantine Hondros

Wolters Kluwer, Zuidpoolsingel 2, 2408 ZE Alphen aan den Rijn, The Netherlands,

e-mail: Constantine.Hondros@wolterskluwer.com

Bernadette Hyland

3 Round Stones Inc., Fredericksburg, VA 22408, USA, e-mail:

bernadette.hyland@3roundstones.com

Tom Ilube

Garlik Ltd, 1-3 Halford Road, London TW10 6AW, United Kingdom, e-mail:

tom.ilube@garlik.com

Johannes Keizer

FAO, Viale delle Terme di Caracalla, 00153 Rome, Italy, e-mail:

Johannes.Keizer@fao.org

Philippe Laublet

STIH (Sens - Texte - Informatique - Histoire), Universit Paris-Sorbonne, 28 rue Serpente, 75006 Paris, France, e-mail: philippelaublet@paris-sorbonne.fr

Chimezie Ogbuji

Cleveland Clinic, 9500 Euclid Ave. Cleveland OH 44195, USA, e-mail:

ogbujic@ccf.org

Silver Oliver

British Broadcasting Corporation, Broadcasting House, Portland Place, London,

United Kingdom, e-mail: silver.oliver@bbc.co.uk

Sean O'Riain

Digital Enterprise Research Institute, National University of Ireland,

Galway, IDA Business Park, Lower Dangan, Galway, Ireland, e-mail:

sean.oriain@deri.org

Alexandre Passant

Digital Enterprise Research Institute, National University of Ireland,

Galway, IDA Business Park, Lower Dangan, Galway, Ireland, e-mail:

alexandre.passant@deri.org

Yves Raimond

British Broadcasting Corporation, Broadcasting House, Portland Place, London,
United Kingdom, e-mail: yves.raimond@bbc.co.uk

Tom Scott

British Broadcasting Corporation, Broadcasting House, Portland Place, London,
United Kingdom, e-mail: tom.scott@bbc.co.uk

Patrick Sinclair

British Broadcasting Corporation, Broadcasting House, Portland Place, London,
United Kingdom, e-mail: patrick.sinclair@bbc.co.uk

Michael Smethurst

British Broadcasting Corporation, Broadcasting House, Portland Place, London,
United Kingdom, e-mail: michael.smethurst@bbc.co.uk

Mischa Tuffield

Garlik Ltd, 1-3 Halford Road, London TW10 6AW, United Kingdom, e-mail:
mischa.tuffield@garlik.com

David Wood

3 Round Stones Inc., Fredericksburg, VA 22408, USA, e-mail:
david.wood@3roundstones.com

Acronyms

- API Application Programmer Interface: An abstraction implemented in software that defines how others should make use of a software package such as a library or other reusable program.
- BPML Web Services Business Process Execution Language: An executable software language for defining interactions with Web Services. A standard of the Organization for the Advancement of Structured Information Standards (OASIS).
- BPMS Business Process Management System or Suite: Enterprise software purporting to assist a business to align with its customers' needs. BPMS systems may be composed of a rules engine to encode business processes, analytics, content management and collaboration tools.
- BI Business Intelligence: Enterprise software approaches to finding and analyzing critical information in information silos, especially information related to important business functions such as sales figures.
- D2RQ Database to RDF Querying: A mechanism to query information in traditional management systems such as relational databases via the SPARQL query language. D2RQ may refer to the language definition or the Open Source Software project.
- DAG Directed Acyclic Graph: A directed graph (like RDF) with the additional restriction that no loops or cycles are permitted. A cycle is a path from a given node that would allow one to find their way back to the starting node.
- DC Dublin Core Element Set: A vocabulary of fifteen properties for use in resource descriptions, such as may be found in a library card catalog (author, publisher, etc). The most commonly used vocabulary for Semantic Web applications.
- DCMI Dublin Core Metadata Initiative: An open international organization engaged in the development of interoperable metadata standards, including the Dublin Core Element Set.

- DNS** Domain Name System: The Internet's mechanism for mapping between a human-readable host name (e.g. www.example.com) and an Internet Protocol (IP) Address (e.g. 203.20.51.10).
- DOI** Digital Object Identifier. A persistent identifier scheme used mostly in the publishing market. Compare with PURLs.
- DTD** Document Type Definition: A type of schema for defining a markup language, such as in XML or HTML (or their predecessor SGML).
- EDGAR** The Electronic Data-Gathering, Analysis, and Retrieval system of the the U.S. Securities and Exchange Commission: An online service providing access to filings by public corporations, such as company registration, sales figures or annual reports.
- ERP** Enterprise Resource Planning (system): An integrated suite of software products that serve many or all enterprise departments.
- FLOSS** Free/Libre/Open Source Software. A generic and internationalized term for software released under an Open Source license.
- FOAF** Friend of a Friend: A Semantic Web vocabulary describing people and their relationships for use in resource descriptions.
- GRDDL** Gleaning Resource Descriptions from Dialects of Languages: A mechanism for extracting Semantic Web data in RDF from XML formats using transformations identified by URIs and typically expressed in XSLT.
- HTML** Hypertext Markup Language: The predominant markup language for hypertext pages on the Web. HTML defines the structure of Web pages. A family of W3C standards.
- HTTP** Hypertext Transfer Protocol: The standard transmission protocol used on the World Wide Web to transfer hypertext requests and information between Web servers and Web clients (such as browsers). An IETF standard.
- IETF** Internet Engineering Task Force: An open international community concerned with the evolution of Internet architecture and the operation of the Internet. Defines standards such as HTTP and DNS.
- ISO** International Standards Organization: A network of the national standards institutes of 162 countries that cooperate to define international standards. Defines many standards including in the context of this book formats for dates and currency.
- LED** Linking Enterprise Data: The use of tools and techniques of the Semantic Web to connect, expose and use data from enterprise systems.
- LOD** Linked Open Data: An open community project to interlink data on the Semantic Web using URIs and RDF.
- LSID** Life Sciences Identifier. A persistent identifier scheme for the life sciences, mostly overtaken by PURLs.
- MDM** Master Data Management: A set of processes and tools that attempts to consistently define and manage an enterprise's non-transactional data entities.
- N3** Notation 3: An RDF syntax intended to be readable by humans.
- ODP** Open Directory Project, a community effort to collect, tag and organize information on the World Wide Web.

- OLAP** Online Analytical Processing: An approach to answering multi-dimensional analytical queries using specialized databases. OLAP is considered part of BI.
- OWL** Web Ontology Language: A family of knowledge representation and vocabulary description languages for authoring ontologies, based on RDF and standardized by the W3C. Standardized variants include OWL Full, OWL DL (for "description logic") and OWL Lite.
- PICS** Platform for Internet Content Selection: An older W3C standard for associating metadata with Web resources. PICS has been superseded by the Protocol for Web Description Resources (POWDER) that is based on RDF.
- PURL** Persistent Uniform Resource Locator: A persistent identifier for Web-based information resources that is protected from change with time. PURLs are URLs and generally use HTTP redirection to resolve a persistent address to a currently valid one.
- RDF** Resource Description Framework: An international standard for data interchange on the Web. A W3C standard.
- RDFa** Resource Description Framework Attributes: An RDF syntax encoded in HTML documents. A W3C standard.
- RDFS** Resource Description Framework Schema: The simplest RDF vocabulary description language that provides much less descriptive capability than SKOS or OWL. A W3C standard.
- RDF/XML** Resource Description Framework eXtensible Markup Language serialization format: An RDF syntax encoded in XML. A W3C standard.
- RELAX NG** Regular Language Description for XML, Next Generation: A simple schema language for XML. An ISO standard.
- REST** Representational State Transfer: An architectural style for information systems used to greater or lesser degree on the Web and explains some of the Web's key features, such as extreme scalability and robustness to change.
- RFC** Request for Comments: A document submitted to the IETF. Internet standards started as RFCs and are often referenced by their RFC numbers.
- SKOS** Simple Knowledge Organisation System: A vocabulary description language for RDF designed for representing traditional knowledge organization systems such as enterprise taxonomies in RDF. A W3C standard.
- SOA** Service Oriented Architecture: A set of architectural design guidelines used to expose services, often as Web Services.
- SOAP** Simple Object Access Protocol: A protocol over HTTP for exchanging structured information in XML to and from Web Services.
- SPARQL** SPARQL Protocol and RDF Query Language: A query language for RDF data on the Semantic Web; analogous to the Structured Query Language (SQL) for relational databases. A W3C standard.
- TAG** The Technical Architecture Group of the World Wide Web Consortium.

- UDEF** Universal Data Element Framework: A mechanism for building controlled vocabularies describing enterprise data. A project of The Open Group, a standards body.
- UPDM** The Unified Profile for DoDAF/MODAF: A modeling standard that supports the USA Department of Defense Architecture Framework (DoDAF) and the UK Ministry of Defence Architecture Framework (MODAF).
- URI** Uniform Resource Indicator: A global identifier for the Web standardized by joint action of the W3C and IETF. A URI may or may not be resolvable on the Web (see URL).
- URL** Uniform Resource Locator: A global identifier for Web resources standardized by joint action of the W3C and IETF. A URL is resolvable on the Web and is commonly called a "Web address".
- UUID** Universally Unique Identifier: A large hexadecimal number that may be calculated by anyone without significant central coordination and used to uniquely identify a resource. A standard of the Open Software Foundation.
- W3C** World Wide Web Consortium: An international community that develops standards for the World Wide Web. Defines standards such as HTML, XML and RDF.
- WFMS** Workflow Management System: Information systems that define and manage the execution of workflows through the use of a workflow engine.
- XBRL** Extensible Business Reporting Language: A mechanism for exchanging business information in XML. A standard of XBRL International.
- XHTML** Extensible Hypertext Markup Language: A family of versions of HTML based on XML and standardized by the W3C.
- XLINK** XML Linking Language: An extension to XML that provides hyperlinks for XML documents. A W3C standard.
- XML** Extensible Markup Language: A specification for creating structured textual computer documents. Many thousands of XML formats exist, including XHTML. A family of standards from the W3C.
- XSD** XML Schema: Limitations on the content of an XML document that defines what structural elements are allowed.
- XSLT** Extensible Stylesheet Language Transformations: Declarative programs to transform one XML document into another XML document.

Part I
Why Link Enterprise Data?

According to studies by Robert Steele at the UC Berkeley School of Information Management and Systems¹ and Roger E. Bohn and James E. Short of the Global Information Industry Center at the University of California, San Diego², the amount of digital information being produced has been growing exponentially for the past two decades. Enterprise information systems have become stressed and have surpassed the point where they are able to scale effectively. Information critical to business success has thus become harder to find, integrate and use. How can industry regain control over business critical information? We know of only one proven approach: The application of Web architecture to the integration of enterprise systems.

Chapters in this part provide a discussion of the information flood and guidance on how to handle it. Dean Allemang tells us that enterprise technology simply must change. He introduces the concept of the Linked Data enterprise and discusses the foundational approaches that define it. Ed Curry, Andre Freitas and Sean O’Riain provide us with a guide for some necessary social changes within enterprises, specifically in relation to curation of information.

Both chapters provide examples of Linked Data techniques being applied now. The use cases given in these chapters supplement the success stories of Part IV and help to broaden our understanding of where and why Linked Data techniques apply.

¹ Robert David Steele, Information Operations: Putting the "I" Back into DIME, DIANE Publishing, 2006

² http://hmi.ucsd.edu/pdf/HMI_2009.ConsumerReport.Dec9_2009.pdf

Semantic Web and the Linked Data Enterprise

Dean Allemang

Abstract Enterprise agility is more important now than ever. An agile enterprise needs to involve a wide variety of stakeholders in its information gathering and management efforts. This results in a number of disconnected data silos. A number of technologies have been applied to this problem, but none have been fully successful in resolving the fundamental enterprise-level issues. The World Wide Web is the only technology that has been proven to be able to scale to an appropriate size to resolve the fundamental enterprise issues. This paper describes the *Linked Data enterprise*, in which Semantic Web technology is used to address the fundamental issues that prevent enterprises from achieving the agility they require.

1 Social Data in the Enterprise

Agility is the name of the game in modern business. While it has always been true that the ability to bring a product to market more quickly than a competitor provides a clear advantage, today's world expects unprecedented high volume of novel products, relying upon complex supply chains, delivered through elaborate distribution channels. These products are delivered to an ever-changing international market, in which regulations somewhere in the world change on a regular basis. New product categories have become the rule rather than the exception. Delicate economic times result in corporate mergers, with a concomitant challenge to make the whole more viable than the sum of the parts. It is no wonder that agility, or the ability for an organization to cope with organizational change, has become the key competitive edge in today's business economy.

At the same time, information has become central to the execution of any business model. Supply chains run on accurate and timely information about availability

Dean Allemang
TopQuadrant Inc., 330 John Carlyle Street, Suite 180, Alexandria, VA 22314-5760, USA e-mail:
dallemang@topquadrant.com

and stock. Business development, marketing and sales are all information intensive activities. For many industries, information forms a large part of the business itself. Pharmaceutical research consists to a large extent of information management, ranging from bioinformatics to the statistics of clinical trials. Finance instruments are largely information-based. Publishing has always been an information-based industry, but with the ascendance of electronic media over hardcopy, the business of publishing is even more one of information management. Agility in business often means agility in information systems.

A few decades ago, this would seem to be good news. Software was the most flexible, agile part of any business. Unlike hardware systems, it was easy to change a software system to adapt to new business models or new production processes. But the information load on business has taken its toll; this is no longer the case. It typically takes several months to a year to bring a new data backed system online. With new products coming out every few months, the ability to build the software system to manage a product has become a severe bottleneck.

In 2008, industry leaders in informatics met at the Claremont resort in Berkeley for the 7th conference to evaluate the state of database research and its impact on industry. This is a well-established, three decade old industry that is well entrenched in just about every other industry. The resulting report identifies a number of challenges to enterprise information management, including distribution and volatility. The report even goes so far as to point out that relational database technology, which has been such a mainstay of corporate information management, will require an overhaul to be up to the task.

When relational databases took a central place in information management, it was possible to think of the database as a place where one would go to find all information about the business. The database was a destination, where questions could be answered. Even the name of one of the major database vendors echoes the metaphor of the Oracle of Delphi, who sat high upon a mountain, and to whom Kings and Heroes, CEOs and Entrepreneurs, would make a pilgrimage, seeking answers to important questions.

In today's distributed information landscape, this metaphor no longer holds. We expect information to come to us; to be available on our desktops, on our phones, to travel with us over land and in the air. The Web has made us grow accustomed to having a wide variety of sources at our fingertips. Search engines like Google and Yahoo! let us pick and choose from information sources, each vying for our eyeballs. The days of a single, definitive information destination are over. We expect a web of interconnected information.

The structure of that information is no longer static, no longer under the control of a single information architect. Valuable enterprise assets live on desktops, as spreadsheets, presentations or documents. It has become more and more difficult for an enterprise to even know what information assets it holds, giving rise to a different viewpoint on enterprise information, that goes by the name of Enterprise Architecture.

Enterprise Architecture refers to a description of all the data owned by an enterprise; what role it plays in business process, who owns it, how it is maintained and