Creating, Interrelating and Consuming Linked Data on the Web

Wolfgang Halb

Creating, Interrelating and Consuming Linked Data on the Web

Dissertation

at

Graz University of Technology

submitted by

Wolfgang Halb

Knowledge Management Institute,
Graz University of Technology
8010 Graz, Austria

and

DIGITAL - Institute for Information and Communication Technologies,
JOANNEUM RESEARCH Forschungsgesellschaft mbH
8010 Graz, Austria

August 2012

First Assessor: Univ.-Prof. Dr.rer.nat. Klaus Tochtermann
Second Assessor: Prof. Dr.techn. Michael Granitzer

Abstract

The World Wide Web has radically changed the way people communicate and share knowledge. An ever increasing amount of digital content is being produced and made available on the Web. The ideas of the Semantic Web and the related Linking Open Data (LOD) project contribute to making the access to information on the Web more efficient. Through such approaches machines become able to assist humans in their information needs and daily tasks. This thesis addresses how the generation and utilization of Linked Data can be optimized.

The research has been structured along three major research themes that are significant for the optimization of Linked Data use: creating linkable data, interrelating Linked Data, and consuming Linked Data. Two major use cases have guided the thesis. The first use case riese is in the application area of government data and makes EuroStat statistics available as Linked Data. The second use case *Link2WoD* targets the media industry and has been developed with the intention to support editors at online content providers but can also be used as a general-purpose tool for enriching unstructured data with Linked Data.

The thesis presents approaches for creating linkable data based on structured data that is available as relational data in many cases. Motivated by one of our use cases also the Statistical Core Vocabulary (SCOVO) is introduced which can be used to represent statistical data as Linked Data. Approaches for extracting linkable data from unstructured data such as plain text are also briefly addressed.

Regarding the interrelation of Linked Data both user based and automated approaches are presented. With the “User Contributed Interlinking” (UCI) a Wiki-style approach is introduced that enables users to easily contribute links to datasets. In addition, further applications of user based approaches as well as automated approaches based on mapping specifications and conceptual relatedness are presented.

The consumption of Linked Data is addressed where general approaches for providing Linked Data for human and machine access are shown. For the application areas of government data and media industry our demonstrators and prototypes are discussed. Finally, also general trends and ideas for future work to exploit the full potential of the Web are presented.

Kurzfassung

Das World Wide Web hat das Kommunikationsverhalten von Menschen sowie den Austausch von Informationen grundlegend verändert. Eine ständig wachsende Menge an digitalen Inhalten wird produziert und am Web verfügbar gemacht. Die Ideen des Semantic Webs und des Linking Open Data (LOD) Projekts tragen dazu bei, um den Zugriff auf Informationen am Web effizient zu ermöglichen. Durch derartige Ansätze wird es möglich, dass automatisierte Anwendungen Menschen bei deren Informationsbedürfnissen und täglichen Aufgaben unterstützen. Im Fokus dieser Dissertation stehen Herangehensweisen, um die Generierung und Nutzung von Linked Data zu optimieren.

Die Forschungsarbeiten wurden in drei Themenbereiche gegliedert, welche besonders zur Optimierung von Linked Data beitragen: das Erstellen von vernetzbaren Daten, das Vernetzen von Linked Data und das Konsumieren von Linked Data. Zwei Hauptanwendungsfälle begleiten die Arbeit. Der erste Anwendungsfall riese befindet sich im Bereich öffentlicher Daten und stellt EuroStat Statistiken als Linked Data zur Verfügung. Der zweite Anwendungsfall *Link2WoD* adressiert die Medienindustrie und wurde entwickelt, um Online-Redakteure zu unterstützen. Der Demonstrator kann jedoch auch allgemein als Werkzeug für die Anreicherung von unstrukturierten Daten mit Linked Data eingesetzt werden.

In der Arbeit werden Möglichkeiten für das Erstellen von vernetzbaren Daten aus strukturierten Daten, welche meist als relationale Daten vorliegen, gezeigt. Motiviert durch einen unserer Anwendungsfälle wird das Statistical Core Vocabulary (SCOVO) vorgestellt, welches der Repräsentation von statistischen Daten als Linked Data dient. Es erfolgt auch eine kurze Darstellung von Herangehensweisen, um vernetzbare Daten aus unstrukturierten Datenquellen zu extrahieren.

In Bezug auf das Vernetzen von Linked Data werden sowohl benutzerbasierte wie auch automatische Methoden vorgestellt. Mit dem "User Contributed Interlinking" (UCI) haben wir eine auf Prinzipien von Wikis basierende Herangehensweise präsentiert, welche es Benutzern ermöglicht, einfach Links zu Datenbeständen hinzuzufügen. Darüber hinaus werden weitere Anwendungsbeispiele dieser Methodik gezeigt sowie automatisierte Ansätze, welche auf speziellen Spezifikationen und konzeptuellen Beziehungen basieren.

Für das Konsumieren von Linked Data werden allgemeine Ansätze diskutiert, um die Daten sowohl für Menschen als auch für eine maschinelle Verarbeitung nutzbar zu machen. Dies erfolgt auch anhand einer Darstellung unserer Anwendungsfälle und Demonstratoren in den Anwendungsgebieten von öffentlichen Daten und in der Medienindustrie. Schließlich werden allgemeine Trends und Ideen für zukünftige Arbeiten präsentiert, um das volle Potenzial des Webs auszuschöpfen.

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Ort Datum Unterschrift

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Place Date Signature

Contents

List of Figures

List of Tables

Listings

Acknowledgments

My uttermost thanks go to my supervisor, Klaus Tochtermann, for his support and for letting me benefit from his rich experience. I would also like to thank Michael Granitzer for being part of my dissertation evaluation committee and for the helpful discussions.

Special thanks go to my colleagues at JOANNEUM RESEARCH and at other organizations for taking the time to discuss, implement, and publish together. I would like to express my gratitude to Michael Hausenblas for introducing me to the topic and doing research jointly. Thanks go to Werner Haas and Harald Mayer for the great support at the institute. Without any particular order I would further like to thank Yves Raimond, Tobias Bürger, Selver Softic, Wolfgang Weiss, Tom Heath, Bernhard Schandl, Bernhard Haslhofer, Andreas Langegger, Alexander Stocker, Helmut Mülner, Ilir Ademi, Claudia Wagner, Peter Scheir, Lee Feigenbaum, Danny Ayers, Behnam Taraghi, Werner Bailer, Georg Thallinger, Herwig Zeiner, Christian Derler, Harald Lernbeiß, Bernhard Jandl, Robert Villa, Punitha Swamy, Richard Cyganiak, Simon Field, Arofan Gregory, Jeni Tennison, Neil Briscombe, Stuart Bertram, Mike Kirton, Satya Sahoo, Sebastian Hellmann, Kingsley Idehen, Ted Thibodeau, Sören Auer, Ahmed Ezzat, Juan Sequeda, and everybody else I have had the pleasure to work with.

Many thanks also go to the funding agencies that partly supported this research: the Austrian Federal Ministry for Transport, Innovation and Technology (bmvit), the Austrian FIT-IT Programme, the European Commission under contracts FP6-027122 SALERO, FP6-027026 K-SPACE, and FP7-225336 SERSCIS.

I also wish to thank my family and my close friends, whose enthusiasm, interest, and support have given me the motivation to realize this achievement.

Wolfgang Halb
Graz, Austria, August 2012

Part I

Introduction and Foundations

Chapter 1

Introduction

The World Wide Web has radically changed the way people communicate and share knowledge. An ever increasing amount of digital content is being produced and made available on the Web. According to a recent study [61] the amount of digital information created and replicated per year has grown by the factor of 9 in only five years. The continuous rapid growth of information on the Web needs to be accompanied with new ways of accessing this information in order to gain valuable insights and have the right information available in the right place at the right time. The traditional approach of simply publishing documents on the Web needs to be improved to enable intelligent information consumption by humans and machines. Semantic technologies and the Linked Data principles are one of the possible solutions for tackling this problem.

1.1 Motivation

Already in the year 1998 Sir Tim Berners-Lee has outlined a visionary road map for future Web design called the *Semantic Web*:

> The Web was designed as an information space, with the goal that it should be useful not only for human-human communication, but also that machines would be able to participate and help. One of the major obstacles to this has been the fact that most information on the Web is designed for human consumption, and even if it was derived from a database with well defined meanings (in at least some terms) for its columns, that the structure of the data is not evident to a robot browsing the web. Leaving aside the artificial intelligence problem of training machines to behave like people, the Semantic

> Web approach instead develops languages for expressing information in a machine processable form. [13]

The goal of such an approach is to improve the usage of Web information by machines where in the end humans benefit from being able to easier access relevant information. In order to make the machine interpretable use of the Web possible the structure and content of the Web needs to be transformed. This approach then even allows intelligent software agents to create additional benefits for humans.

A more detailed description of the Semantic Web vision and its possibilities has been published in a 2001 *Scientific American* article [18]. New applications are presented that could be realized through the Semantic Web. It shall, for instance, be possible to create intelligent Semantic Web agents that can automatically accomplish tasks for human users. To aid the realization of the Semantic Web the World Wide Web Consortium (W3C) has started several standardization initiatives to create recommendations and the surrounding technology landscape. Among the most notable recommendations for providing machine interpretable information on the Web the *Web Ontology Language* (OWL) and *Resource Description Framework* (RDF) have been defined.

In 2006 Berners-Lee and colleagues state that the initial idea of the Semantic Web "remains largely unrealized" [152] but reinforce their believe in the Semantic Web. They argue that well established standards are needed and "that the Web standards for expressing shared meaning have progressed steadily" [152]. In the same year Berners-Lee also announced a set of principles to foster the creation of semantic data on the Web. The four simple rules of the Linked Data principles [14] provide guidance for data publishers to become part of this distributed global database for machine interpretable data, also called the *Web of Data*. The following year 2007 is of special importance in this movement as the *Linking Open Data* (LOD) community project has been started by W3C's Semantic Web Education and Outreach (SWEO) Interest Group. The goal of this community project is to bootstrap the Semantic Web and extend the Web with a data commons by publishing open data as RDF on the Web and interlinking the different datasets using typed RDF links. The LOD project has soon gained considerable attention in the scientific community and has also proceeded to other communities. Driven by the growing interest in the LOD project a steadily increasing amount of Linked Data has been made available. The main principle and goal of Linked Data is to use the Web for creating typed links between data from different sources. This global data space connects data from a variety of different domains such as encyclopedias, geographical information, statistics, movies, TV and radio programs, books, scientific publications, companies and much more.

This new ecosystem of data on the Web enables a plethora of innovative applications. Tremendous research efforts were necessary in the evolution from the first initially published Linked Data sources in 2007 to the currently more than 50 billion facts (triples) connected via the LOD cloud in the year 2012. Several open issues related to the realization of the Linked Data vision have been resolved in the recent years, with some open research issues still remaining to unleash the full potential of Linked Data.

The possibilities and potential benefits of Linked Data motivated the author of this thesis to research on open issues in relation with Linked Data. In an initial study [85] that we conducted in 2008 we analyzed the state of the Semantic Web and Linked Data at that time. We identified several important factors that needed to be addressed to optimize the benefits of Linked Data. Besides the sheer amount of Linked Data also the number, type, and quality of links between the datasets presents a crucial measure. Motivated by these findings we focused our research on improving the generation and use of Linked Data with high-quality interlinks. The achieved research results are presented in this thesis and the Link2WoD framework combines several of the sub-results in a demonstration prototype where Linked Data can be generated from unstructured data sources, interrelated, and consumed.

1.2 Research Questions

Based on the initial idea of the Semantic Web and the Linked Data principles several issues arise when trying to use the full potential of Linked Data on the Web. Two higher level research questions are addressed in this thesis:

- *What is the size of the Semantic Web?* The rationale for this question is to get a basic understanding of the characteristics of the Semantic Web and Linked Data. This also lays the foundation for the second research question:
- *How can Linked Data generation and utilization be optimized?*

The research presented in this thesis has been structured along three major research themes that have been identified as being significant for the optimization of Linked Data use:

- creating linkable data,
- interrelating Linked Data, and

- consuming Linked Data.

Research in these three areas has been aligned to two main application areas: media industry and government data. Companies and organizations that are active in the media industry usually take the role of a content provider, i.e. they produce text or multimedia content which usually represents unstructured content that can benefit from augmentation as Linked Data. Government data as another application area can also benefit from Linked Data augmentation, which is especially true for open government data. This data, however, is often available as structured data, as it is for instance the case with statistical information. Depending on whether the source data is available as structured or unstructured data the methodologies for creating Linked Data need to be adapted accordingly. The research questions take both cases into account and aim for a flexible approach for the creation, interrelation, and consumption of Linked Data.

1.2.1 Creating Linkable Data

At the very core of the Linked Data vision is the need for data that can be linked and used in various applications. In 2009 Berners-Lee pronounced this need with the request for "raw data now" in his famous TED talk [16]. However, raw data that is available as structured or unstructured data still needs to be processed so that it carries sufficient semantic information to be linkable with other datasets. Data described with the Resource Description Framework (RDF) provides a base for later linking and can be considered linkable data.

Research questions: Which data is linkable? What are the potentials of public data and how can the data be represented? How can structured data be RDFized? How can linkable data be extracted from unstructured data?

Scope: As the creation of linkable data potentially covers all data that is available globally the scope has to be narrowed. Along the two main application areas the focus for structured data is laid on government data that is available in many cases as relational and/or tabular data. The focus for unstructured data is given on content from the media industry.

1.2.2 Interrelating Linked Data

Once linkable data has been created it needs to be interrelated. Links between data from different datasets allow humans and machines to use this data and follow semantically typed links in the Web of Data. As our initial study [85] has shown only few different properties are used to interlink data from different datasets.

Research Questions: Why is there a low variety of different properties used for the interlinking? How can the manual creation of links be supported? How can user based approaches be realized? How can links be created automatically?

Scope: The research on the interrelation of Linked Data focuses again on the two main application areas and the specifics of data in these two areas. The interrelation is also based on the instance level and thus topics of ontology matching are not covered in this thesis.

1.2.3 Consuming Linked Data

When Linked Data has been created and interrelated it is ready to be consumed. Linked Data can be used by machines to automatically retrieve information and present it to humans. For the presentation intelligent user interfaces are needed that make the right information available at the right time. The underlying structure and potential complexity is mainly intended for machine use. Human users should not have to care about technical details and simply use the applications realized with Linked Data with ease.

Research Questions: How can Linked Data be consumed by humans and machines? Which application areas are well suited for Linked Data consumption?

Scope: The consumption of Linked Data is shown for exemplary use cases in the main application areas and takes best practices in user interface design into account. It is not in the scope of this thesis to research on fundamental human-computer interaction patterns.

1.3 Scientific Contributions

During the research and development conducted in the course of this thesis the author of this thesis has made several contributions to various areas: The conceptual framework for Linked Data generation and utilization that guides the thesis has been designed. The analyses for the study in 2008 to assess the state of the Semantic Web (cf. chapter 3) have also been mainly conducted by the author. Further on the major contribution to the position statement discussing the issues of trusting interlinked multimedia data (cf. section 5.3) has been made.

Regarding the creation of linkable data out of structured datasources (cf. chapter 6) the following major contributions have been made by the author:

- For the *RDFizing and Interlinking the EuroStat Data Set Effort* (riese) the core schema for the representation of statistical data has been designed. The implementation of the prototype has been conducted jointly with Michael Hausenblas and Yves Raimond.
- The further development of the riese core schema into the Statistical Core Vocabulary (SCOVO) has been led by the author of this thesis and has received input from Michael Hausenblas, Yves Raimond, Lee Feigenbaum, and Danny Ayers.
- For the further alignment between SDMX and SCOVO the author of this thesis has mainly contributed aspects from SCOVO. The further research has been conducted together with Richard Cyganiak, Simon Field, Arofan Gregory, and Jeni Tennison.
- In the course of the RDB2RDF Incubator Group the author of this thesis has contributed to the survey of current approaches for the mapping of relational databases to RDF together with other members of the group.

The creation of linkable data from unstructured datasources (cf. chapter 7) has been mainly covered by Helmut Mülner who contributed to the development of the term extraction module in Link2WoD. The author of this thesis was the research and development lead of Link2WoD.

Regarding the user based approaches for interrelating Linked Data (cf. chapter 9) the following contributions have been made by the author of this thesis:

- The *User Contributed Interlinking* (UCI) approach has been developed jointly together with Michael Hausenblas and Yves Raimond.
- For the interlinking of social e-learning platforms the main research and development has been conducted by Selver Softic and Behnam Taraghi. The author of this thesis has particularly contributed to Linked Data aspects of the work.
- With the *Catch Me If You Can* (CaMiCatzee) a multimedia data interlinking concept demonstrator has been developed together with Michael Hausenblas.
- The development for the SALERO *Intelligent Media Annotation & Search* (IMAS) system has been led by Wolfgang Weiss and the author of this thesis contributed

to the development especially for Linked Data aspects and with experiences from the implementation of the User Contributed Interlinking approach.

The automated approaches for interrelating Linked Data (cf. chapter 10) have been mainly covered by the author of this thesis which includes

- the automated interlinking approach applied in riese and
- the automated interlinking in Link2WoD.

Regarding the consumption of Linked Data the following contributions have been made by the author of this thesis:

- The implementation of riese (cf. chapter 12) which includes the consumption aspects has been conducted jointly with Michael Hausenblas and Yves Raimond.
- For Link2WoD (cf. chapter 13) the author of this thesis was the research and development lead and contributed most parts of the work. Ilir Ademi contributed to the implementation of the Link2WoD standalone web technology preview demonstrator and Alexander Stocker contributed to the related business aspects.
- Linked Data and semantics related aspects have also been contributed by the author of this thesis to the work related to dynamic business process execution in the course of the SERSCIS project (cf. chapter 14) and most of the other research and development aspects have been carried out by colleagues in the project.

1.4 Published Work

The research carried out for this thesis has made several scientific contributions towards the realization of the full potential of Linked Data. Most parts of this thesis have already been published in several places including international conferences, refereed workshops, and a journal as listed in chronological order below:

W. Halb, Y. Raimond, and M. Hausenblas. *Building Linked Data For Both Humans and Machines.* In C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, editors, *Proceedings of the WWW08 Linked Data on the Web Workshop (LDOW 2008)*, volume 369 of *CEUR Workshop Proceedings.* Beijing, China, 2008. `http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-369/paper06.pdf`. ISSN 1613-0073. [73]

M. Hausenblas, W. Halb, and Y. Raimond. *Scripting User Contributed Interlinking.* In *Proceedings of the 4th Workshop on Scripting for the Semantic Web (SFSW08)*, volume 368 of *CEUR Workshop Proceedings*. Tenerife, Spain, 2008. `http://CEUR-WS.org/Vol-368/paper6.pdf`. [83]

M. Hausenblas and W. Halb. *Interlinking of Resources with Semantics (Poster)*. In *5th European Semantic Web Conference (ESWC2008)*. 2008. [82]

M. Hausenblas, W. Halb, Y. Raimond, and T. Heath. *What is the Size of the Semantic Web?* In *Proceedings of the 4th International Conference on Semantic Systems (I-SEMANTICS 2008)*, pages 9–16. Graz, Austria, September 2008. [85]

M. Hausenblas and W. Halb. *Interlinking Multimedia Data*. In *Linking Open Data Triplification Challenge at the International Conference on Semantic Systems (I-Semantics08)*. 2008. [81]

W. Halb and M. Hausenblas. *select * where { :I :trust :you }: How to Trust Interlinked Multimedia Data*. In S. Auer, S. Dietzold, S. Lohmann, and J. Ziegler, editors, Proceedings of the International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW'08), volume 417 of *CEUR-WS*, pages 59–65. Koblenz, Germany, December 2008. `http://ceur-ws.org/Vol-417/paper6.pdf`. [72]

S. S. Sahoo, W. Halb, S. Hellmann, K. Idehen, T. Thibodeau, S. Auer, J. Sequeda, and A. Ezzat. *A Survey of Current Approaches for Mapping of Relational Databases to RDF*, 2009. W3C RDB2RDF Incubator Group January 08 2009. [144]

M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayers. *SCOVO: Using Statistics on the Web of Data*. In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, editors, *The Semantic Web: Research and Applications*, volume 5554 of *Lecture Notes in Computer Science*, pages 708–722. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-02120-6. doi:10.1007/978-3-642-02121-3_52. [84]

S. Softic, B. Taraghi, and W. Halb. *Weaving Social E-Learning Platforms Into the Web of Linked Data*. In *Proceedings of I-KNOW '09 and I-SEMANTICS '09*, pages 559–567. Graz, Austria, 2009. [156]

W. Weiss, T. Bürger, R. Villa, P. Swamy, and W. Halb. *SALERO Intelligent Media Annotation & Search*. In *Proceedings of I-KNOW '09 and I-SEMANTICS '09*, pages 622–629. Graz, Austria, 2009. [181]

W. Weiss, T. Bürger, R. Villa, P. Punitha, and W. Halb. *Statement-Based Semantic Annotation of Media Resources*. In T.-S. Chua, Y. Kompatsiaris, B. Mérialdo, W. Haas, G. Thallinger, and W. Bailer, editors, *Semantic Multimedia*, volume 5887 of *Lecture Notes in Computer Science*, pages 52–64. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-10542-5. doi:10.1007/978-3-642-10543-2_7. [180]

R. Cyganiak, S. Field, A. Gregory, W. Halb, and J. Tennison. *Semantic Statistics: Bringing Together SDMX and SCOVO*. In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, editors, *Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010*, volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010. `http://ceur-ws.org/Vol-628/ldow2010_paper03.pdf`. [40]

W. Halb, H. Zeiner, B. Jandl, H. Lernbeiß, and C. Derler. *Agile Service Oriented Architecture with Adaptive Processes Using Semantically Annotated Workflow Templates*. In *Proceedings of the 2010 IEEE International Conference on Web Services*, ICWS '10, pages 632–633. IEEE Computer Society, Washington, DC, USA, 2010. ISBN 978-0-7695-4128-0. doi:10.1109/ICWS.2010.42. [75]

W. Halb, A. Stocker, H. Mayer, H. Mülner, and I. Ademi. *Towards a commercial adoption of linked open data for online content providers*. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 16:1–16:8. ACM, New York, NY, USA, 2010. ISBN 978-1-4503-0014-8. `http://doi.acm.org/10.1145/1839707.1839727`. [74]

H. Zeiner, W. Halb, H. Lernbeiß, B. Jandl, and C. Derler. *Making business processes adaptive through semantically enhanced workflow descriptions*. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 27:1–27:3. ACM, New York, NY, USA, 2010. ISBN 978-1-4503-0014-8. `http://doi.acm.org/10.1145/1839707.1839741`. [184]

C. Wagner, P. Scheir, A. Stocker, and W. Halb. *Harnessing semantic web technologies for solving the dilemma of content providers*. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 20:1–20:5. ACM, New York, NY, USA, 2010. ISBN 978-1-4503-0014-8. `http://doi.acm.org/10.1145/1839707.1839733`. [178]

N. Briscombe, H. Zeiner, W. Halb, S. Bertram, M. Kirton, and C. Derler. *Dynamic Service Orchestration Using Human and Machine Interpretable System Knowledge with Associated Graphical Software Tools*. In *Proceedings of the Workshop on Open Knowledge Models (OKM-2010) at the 17th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*. Lisbon, Portugal, 2010. [29]

B. Schandl, B. Haslhofer, T. Bürger, A. Langegger, and W. Halb. *Linked Data and multimedia: the state of affairs*. *Multimedia Tools and Applications*, 59:523–556, 2012. ISSN 1380-7501. doi:10.1007/s11042-011-0762-9. [149]

1.5 Thesis Organization

This thesis is organized as shown in figure 1.1 along the conceptual framework for Linked Data generation and utilization. As a first step linkable data needs to be created which can further on be interrelated. Finally, Linked Data is available which can be consumed in various ways.

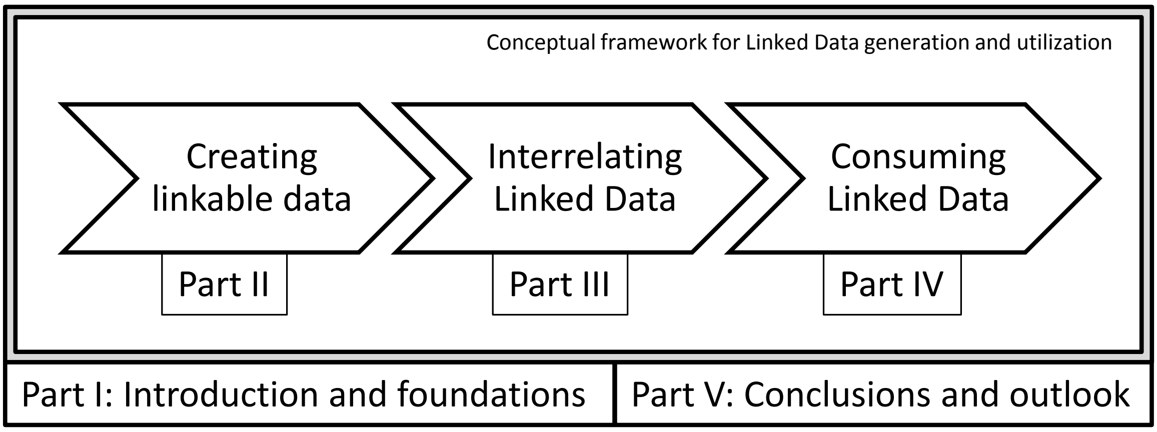

Conceptual framework for Linked Data generation and utilization
Creating linkable data
Part II
Interrelating Linked Data
Part III
Consuming Linked Data
Part IV
Part I: Introduction and foundations
Part V: Conclusions and outlook

Figure 1.1: Structure of the thesis.

Part I - Introduction and foundations The first chapter of the thesis contains an introduction along with a motivation, the research questions addressed and the scientific contributions as well as publications of the author. In chapter 2 we discuss the basic principles of Linked Open Data and related existing work. The underlying research question of *What is the size of the Semantic Web?* is addressed in chapter 3 and is based on our paper [85]. The two major guiding use cases of the thesis (riese and *Link2WoD*) are briefly introduced in chapter 4.

Part II - Creating linkable data The second part of the thesis discusses the initial step of the conceptual framework for Linked Data generation and utilization which is the creation of linkable data. General considerations are discussed in chapter 5 which also includes some additional aspects of provenance, trust, and privacy based on our paper [72]. The creation of linkable data out of structured datasources is covered in chapter 6. Being aligned with our *RDFizing and Interlinking the EuroStat Data Set Effort* (riese) this chapter treats several aspects of the riese demonstrator, statistical data representations and general approaches covered in several of our papers (cf. [73, 84, 40, 144]). The creation of linkable data from unstructured datasources such as plain text on the example of Link2WoD is briefly discussed in chapter 7.

Part III - Interrelating Linked Data The third part of the thesis covers the interrelation of Linked Data, i.e. making interlinks to other datasets based on the linkable data that has been created in the previous step. Some general considerations are presented in chapter 8.

Further we discuss user based approaches for creating interrelations in chapter 9. With User Contributed Interlinking (UCI) we proposed a new way of creating semantic links which we introduced for riese and with "**i**nterlinking of **r**esources with **s**emantics" (IRS) we also created a generalized demonstrator (cf. our papers [73, 82, 83]). In addition we also looked into interlinking e-learning platforms (cf. [156]), developed a multimedia data interlinking concept demonstrator (cf. [81]) and implemented a user based interlinking approach in the *Intelligent Media Annotation & Search* (IMAS) system (cf. [180, 181]).

Automated approaches for interrelating different Linked Data sets are presented in chapter 10 where we also report about our approaches applied in riese (cf. [73]) and in *Link2WoD* (cf. [74]) where we also introduced the *Linked Data Entity Space* concept.

Part IV - Consuming Linked Data The consumption of Linked Data is covered in the fourth part of thesis. Chapter 11 introduces general considerations on the consumption and chapter 12 provides details for the area of government data. In chapter 13 we cover Linked Data consumption in the media industry which is based on our paper [74] and also includes aspects from our paper [178] where we have addressed how Semantic Web technologies can help content providers to open up their content for integration and reuse by third parties. To demonstrate the flexibility of Linked Data and its potential usage and consumption in various domains we also briefly report about its application in process modeling and execution in chapter 14 (cf. our papers [75, 29, 184]).

Part V - Conclusions and outlook The fifth part contains concluding remarks (chapter 15) and an outlook (chapter 16) including general trends and ideas for future work.

Part VI - Appendix Finally the appendix contains an overview of the author's contributions including publications and activities in various fields related to the thesis.

Chapter 2

Linked Open Data

The idea of Linked Open Data (LOD) is closely aligned with the idea of the Semantic Web and many consider LOD as “the Semantic Web vision coming true”. In this chapter we provide a brief introduction to the basic principles of Semantic Web technology and Linked Open Data.

2.1 Semantic Web

The idea of the *Semantic Web* has been described by Sir Tim Berners-Lee in 1998 (cf. [13]). In his - at that time - visionary road map for future Web design he introduced the Semantic Web as an approach to make the Web an information space that is useful for human and machine consumption. As a result, for instance intelligent software agents should also be able to automatically accomplish tasks for human users. At the World Wide Web Consortium (W3C) several initiatives have been started to create the technology landscape for realizing the Semantic Web and promoting it. The related groups and initiatives are organized in the W3C Semantic Web Activity[1].

The *Semantic Web Stack* (also known as *Semantic Web Layer Cake*) shown in figure 2.1 illustrates the hierarchy of languages used for the Semantic Web architecture. Going from the bottom to the top in the stack it can be seen that each layer uses capabilities of the underlying layers. The stack shows how the various technologies are organized and it visualizes that the Semantic Web extends the traditional hypertext Web instead of replacing it. Among the standard technologies that are not specific to Semantic Web are the lower layers such as Uniform Resource Identifiers (URIs), Inter-

[1] `http://www.w3.org/2001/sw/`

nationalized Resource Identifiers (IRIs), and the eXtensible Markup Language (XML). In the following the major components of the Semantic Web stack are briefly described.

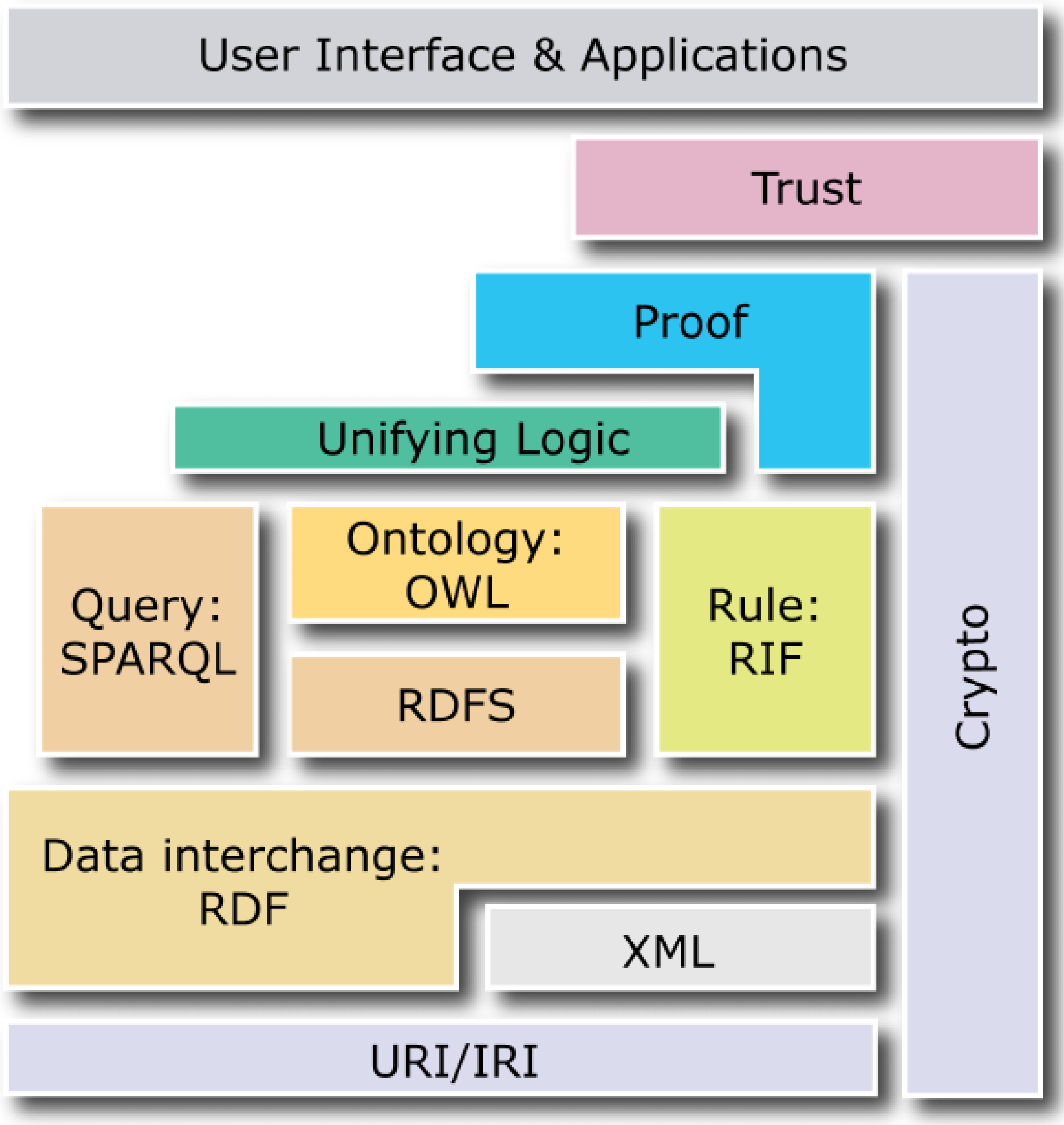


Figure 2.1: Semantic Web Stack.
Source: `http://www.w3.org/2007/03/layerCake.png`

2.1.1 URI/IRI

A Uniform Resource Identifier (URI) is a string that is used to identify a resource. It also represents the basic principle for addressing things on the Web. URIs can further be classified as Uniform Resource Locator (URL) or Uniform Resource Name

(URN). Names simply allow to identify resources whereas Locators also specify how to retrieve them. In many cases, especially technical standards and at the World Wide Web Consortium (W3C), this differentiation is not made and preferably the term URI is used. In the Semantic Web environment the use of HTTP URIs is strongly suggested. The HyperText Transfer Protocol (HTTP) is an application protocol that represents the foundation of communication on the World Wide Web. An example of such a HTTP URI is `http://www.joanneum.at`. In addition also Internationalized Resource Identifiers (IRIs) have been defined that allow the use of Unicode characters (whereas URIs may only use a subset of ASCII characters). A general discussion regarding the use of URIs on the Semantic Web can be found in the note "Cool URIs for the Semantic Web" [147] which also contains references to further technical foundations that are not discussed in detail here.

2.1.2 XML

The eXtensible Markup Language (XML) has been defined by the W3C [170] and is a markup language that defines a set of rules for encoding documents in a format that is supposed to be both human-readable and machine-readable. Several related specifications exist as well and the aim of XML is to provide a textual data format that is simple and generally usable in the Internet. Many applications, protocols and services use XML and it can be considered a widely adopted standard. At the W3C most of the related developments are done within the XML Activity[2].

XML documents have a logical and physical structure. Physically, documents are composed of entities that may refer to other entities. The logical structure is defined through explicit markup that may be in the form of declarations, elements, comments, character references, or processing instructions. In general, a differentiation between markup and content can also be made. Markup or tags are typically enclosed in angle brackets ("<" and ">"). The content is contained in elements between a start-tag and end-tag. By default the semantics of a tag are not defined in the core XML standard but commonly agreed standards exist for various applications areas. An example of a very simple XML document is given in listing 2.1. Further information is referenced for instance from the W3C XML Activity website and is not discussed here in detail.

[2] `http://www.w3.org/XML/Activity`

```
<?xml version=”1.0”?>
<thesis>
  <author>Wolfgang Halb</author>
  <title>Creating, Interrelating and ...</title>
  <year>2012</year>
</thesis>
```

Listing 2.1: XML example snippet.

2.1.3 RDF

The Resource Description Framework (RDF) is a language for representing information about resources in the World Wide Web. Several related recommendations and standards have been produced by the W3C RDF Core Working Group[3] as part of the W3C Semantic Web Activity. RDF is based upon the idea of making statements about a resource in the form of triples, i.e. subject-predicate-object expressions. As an example a RDF triple is shown in listing 2.2 and visualized in figure 2.2. The subject of the triple is `<http://data.semanticweb.org/person/wolfgang-halb>` which represents the resource that the statement is about. Through the predicate `<http://xmlns.com/foaf/0.1/made>` the relation is defined, in the case of the example it relates the subject to something (i.e. the object) it has made. In the example the object is `<http://data.semanticweb.org/workshop/LDOW/2008/paper/10>`.

```
<http://data.semanticweb.org/person/wolfgang-halb>
  <http://xmlns.com/foaf/0.1/made>
    <http://data.semanticweb.org/workshop/LDOW/2008/paper/10>
```

Listing 2.2: RDF triple example.

RDF is based on identifying things on the Web using URIs and describing them via properties (predicates) and property values (objects). In the example shown above only URIs are used in the triple but in addition also the use of literal values and blank nodes is possible. A collection of RDF statements intrinsically represents a labeled, directed multi-graph. Consequently, a RDF graph can also visualize the relations between different resources described in RDF and the use of URIs offers a very useful way of connecting information. An example of such a graph is shown in figure 2.3 that is based on some RDF information about the author of this thesis.

[3] http://www.w3.org/2001/sw/RDFCore/

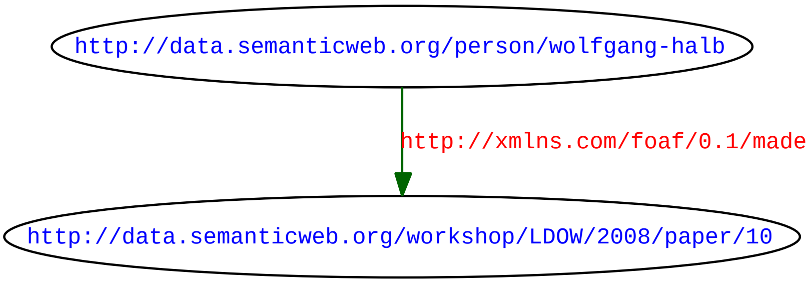

http://data.semanticweb.org/person/wolfgang-halb
http://xmlns.com/foaf/0.1/made
http://data.semanticweb.org/workshop/LDOW/2008/paper/10

Figure 2.2: RDF triple example visualization.

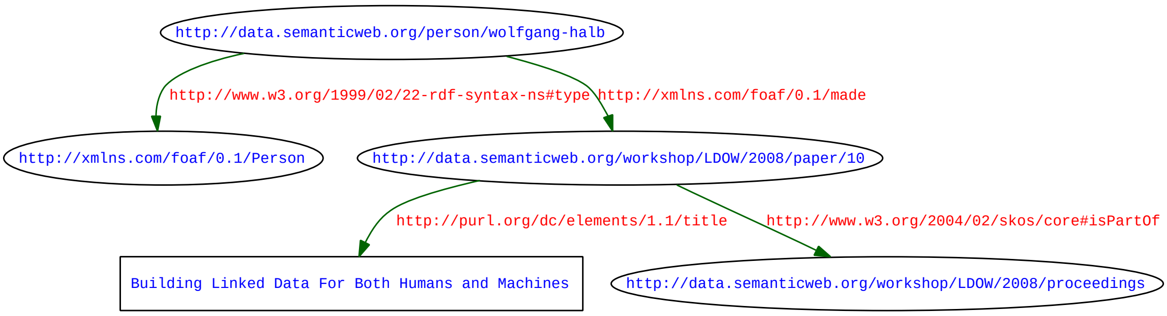

http://data.semanticweb.org/person/wolfgang-halb
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://xmlns.com/foaf/0.1/made
http://xmlns.com/foaf/0.1/Person
http://data.semanticweb.org/workshop/LDOW/2008/paper/10
http://purl.org/dc/elements/1.1/title
http://www.w3.org/2004/02/skos/core#isPartOf
Building Linked Data For Both Humans and Machines
http://data.semanticweb.org/workshop/LDOW/2008/proceedings

Figure 2.3: RDF graph example visualization.

Different serializations of RDF are possible. Representing RDF in XML is common practice but also Notation 3 (N3) is widely used. N3 is more compact and easier to read than XML. It also is a superset of the Turtle syntax which can be used as well. In addition the representation in the N-Triple format is possible. Most software components are able to handle RDF regardless of the serialization format and the different formats have been mainly developed for easier manual handling of the data. RDF can be stored in single serialized files and there also exist dedicated triplestores where RDF can be maintained. Further details about RDF are for instance referenced from the respective group websites at the W3C and are not contained here.

2.1.4 Query, Ontology, Rule

The central parts of the Semantic Web Layer target some mechanisms for handling knowledge that can be exploited by higher layers on top. Through SPARQL query capabilities are provided. Knowledge representation is further supported by RDFS and OWL. With RIF the support of rules is addressed.

SPARQL SPARQL [137] is a recursive acronym for ***S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage* and provides a query language for RDF. It can be used to query different sources of RDF data regardless of their physical storage location. Flexible queries involving graph patterns and various constraints are supported. The results of queries can be represented as a result set or RDF graph and therefore SPARQL also provides efficient means for accessing RDF data. Many software implementations are available that support SPARQL and recently SPARQL 1.1 Update [62] has been proposed as update language for RDF graphs.

RDFS The RDF Schema (RDFS) [27] is a lightweight RDF vocabulary description language. It provides mechanisms for describing groups of related resources and the relationships between these resources. RDFS supports the concept of classes and properties. Classes are groups of resources and properties are instances of the class `rdf:Property` which describe a relation between subject resources and object resources. Simple ontologies and taxonomies can be represented using RDFS. For more complex ontologies richer languages such as OWL need to be used.

OWL The Web Ontology Language (OWL) [123] is used for authoring ontologies. Ontologies have already been a research topic in philosophy and computer science for a long time. In computer science they are formal representations of knowledge as a set of concepts within a domain and the relationships between these concepts. One of the possible definitions is that an “ontology is an explicit specification of a conceptualization.” [69]

Various developments led into OWL and different variants of OWL with different levels of expressiveness exist. *OWL Lite* is intended for classification hierarchies and simple constraints. It has a lower formal complexity than OWL DL and the intention was to make it simpler to provide tool support for OWL Lite. In *OWL DL* the “DL” stands for the corresponding field of description logics which forms the theoretical foundations of OWL. It was designed with the intention to provide maximum expressiveness while retaining computational completeness and decidability. Finally, *OWL Full* offers the maximum expressiveness but no computational guarantees.

As an extension to the first version of OWL from 2004 (cf. [123]) the *OWL 2 Web Ontology Language* [176] has been published as W3C Recommendation in 2009. OWL 2 has a similar structure as its predecessor and is backwards compatible. Some new features that offer more expressivity have been added. A detailed discussion of the changes is provided in [67].

RIF Rules represent one frequently used technique in knowledge representation. Different types of rules exist that allow to reason over data and infer new facts. There also exist several rules languages and the *Rule Interchange Format* (RIF) [104] addresses this aspect. RIF was designed as a family of languages, called dialects, that are also W3C Recommendations. Basically, RIF focuses on two kinds of dialects, logic-based dialects that make use of some kind of logic (e.g., first-order logic) and dialects for rules with actions (e.g., production or reactive rules). The W3C RIF Working Group[4] has produced several documents and W3C Recommendations that are available from the group's website.

2.1.5 Cryptography, Unifying Logic, Proof, Trust

In the Semantic Web Layer cryptography represents a vertical layer that encourages to use public key cryptography and digital signatures to ensure and verify statements from a trusted source. Berners-Lee [13] also envisioned that reasoning engines would have to be tied to signature verification systems so that reasoning can take trust into account. However, this still largely remains to be realized. Another area still under research is the unifying logic which should combine different approaches for representing logic. The proof layer is further on responsible for proofing/verifying statements that have been made in the underlying layers of the stack. This is also related to the topic of provenance about where information originated and how it has been modified. Finally, the trust layer handles the trust of information in the Semantic Web which is influenced by the trust of the information source. Proof as well as cryptography can aid in assessing the trustworthiness.

2.1.6 User Interface & Applications

The final user interface and application layer represents what humans are able to use. Especially in the area of Linked Data several solution examples for this layer have been presented which are also partly discussed in subsequent sections of this thesis. Semantic Web applications can make use of all the underlying layers of the stack to provide the user with the desired information. Intelligent agents can also automatically accomplish tasks for users. Recently several applications and services have been released that can be considered Semantic Web applications even though they might not strictly follow the W3C's Semantic Web Stack. Among the most notable developments are

[4] `http://www.w3.org/2005/rules`

the Google Knowledge Graph [154], the answer engine Wolfram Alpha[5], or Apple's *Speech Interpretation and Recognition Interface* (Siri) as intelligent personal assistant introduced for the iPhone. There also exist several more examples of Semantic Web applications in the wider sense that are gradually becoming widely used. In the following section we discuss the Linking Open Data project that has also contributed much to the success of the Semantic Web.

2.2 LOD Basics

Linked Open Data (LOD) is one of the movements that has attracted considerable academic attention in the recent years and also continues to be widely adopted in various domains. It has its roots in the Semantic Web vision and back in 2006 Berners-Lee and colleagues stated that the initial idea of the Semantic Web "remains largely unrealized" [152] but reinforce their believe in the Semantic Web. They argue that well established standards are needed and "that the Web standards for expressing shared meaning have progressed steadily" [152]. In the same year Berners-Lee also announced a set of principles to foster the creation of semantic data on the Web. The four simple rules of the Linked Data principles [14] provide guidance for data publishers to become part of this distributed global database for machine interpretable data, also called the *Web of Data*.

The following year 2007 is of special importance in this movement as the *Linking Open Data* (LOD) community project has been started by W3C's Semantic Web Education and Outreach (SWEO) Interest Group. The goal of this community project is to bootstrap the Semantic Web and extend the Web with a data commons by publishing open data as RDF on the Web and interlinking the different datasets using typed RDF links. The LOD project has soon gained considerable attention in the scientific community and has also proceeded to other communities. Driven by the growing interest in the LOD project a steadily increasing amount of Linked Data has been made available. The main principle and goal of Linked Data is to use the Web for creating typed links between data from different sources. This global data space connects data from a variety of different domains such as encyclopedias, geographical information, statistics, movies, TV and radio programs, books, scientific publications, companies and much more.

[5] http://www.wolframalpha.com/

2.2.1 Linked Data Principles

At the heart of the LOD vision are the Linked Data principles [14]:

1. Use URIs as names for things

2. Use HTTP URIs so that people can look up those names

3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)

4. Include links to other URIs so that they can discover more things

The first principle simply suggests to use URIs to address resources. These resources can be documents but also real-world entities that are referenced via a URI. This makes it possible to make statements about the resources and also link them with other information. It also resembles the Resource-Oriented Architecture of RESTful Web services (cf. [59]) where a “resource is anything that’s important enough to be referenced as a thing in itself” [142] and “has to have at least one URI” [142].

Through the use of the HyperText Transfer Protocol (HTTP) it is possible to *dereference* URIs and this uniform interface is also one of the foundations of the World Wide Web.

The third principle suggests to provide useful information when the URI is dereferenced. In a revision of the initial principles Berners-Lee also added that standards such as SPARQL or RDF should be used. Through the use of commonly agreed standards the interoperability of different data sources is increased.

The fourth principle of providing links is also referred to as the “follow-your-nose” approach which makes it possible to simply identify additional related information by following the links. Through the use of strongly typed links instead of untyped Web links (`@href`) the contained information can also automatically be processed more precisely.

These four principles lay the foundation of Linked Data and have been followed by several dataset providers as part of the LOD project. Additional guidelines are also available that target more technical aspects. The note “Cool URIs for the Semantic Web” [147] for instance provides a detailed discussion of URI assignment related issues and explains the different implementation approaches.

2.2.2 LOD Evolution

Soon after the start of the LOD project in 2007 it became very popular and attracted interest from the scientific community. The dissemination of the current state of the LOD project is also supported by the famous LOD cloud diagram visualization that shows datasets available as Linked Data and how they are related. The diagram is maintained by Richard Cyganiak and Anja Jentzsch. The latest diagram from September 2011 is shown in figure 2.4. In this visualization each bubble represents a dataset available as Linked Data. The size of the bubble corresponds with the number of triples in each dataset. Arrows between different datasets indicate that at least 50 links exist between the two datasets. The different colors represent the domains of the dataset.

Currently the LOD cloud is reported to contain data from approximately 300 different datasets and on a high-level they have been clustered into the domains of media, geographic, publications, user-generated content, government, cross-domain, and life sciences. According to [25] these datasets provide more than 30 billion triples and more than 500 million interlinks.

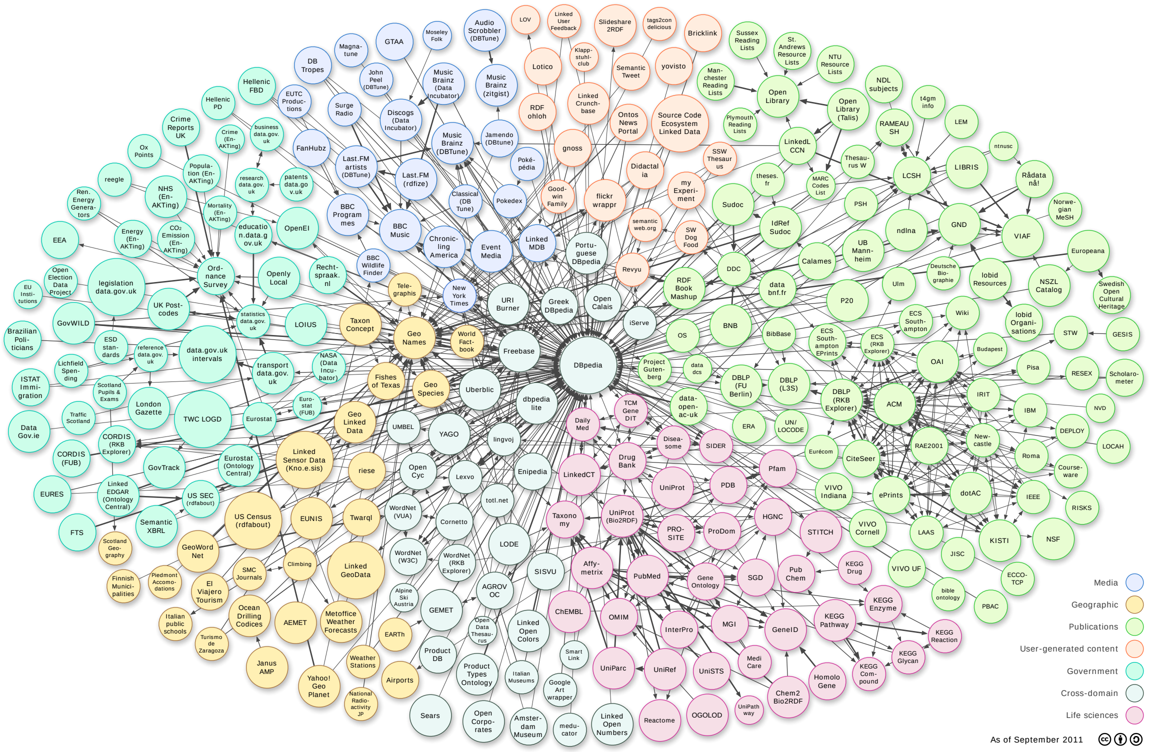
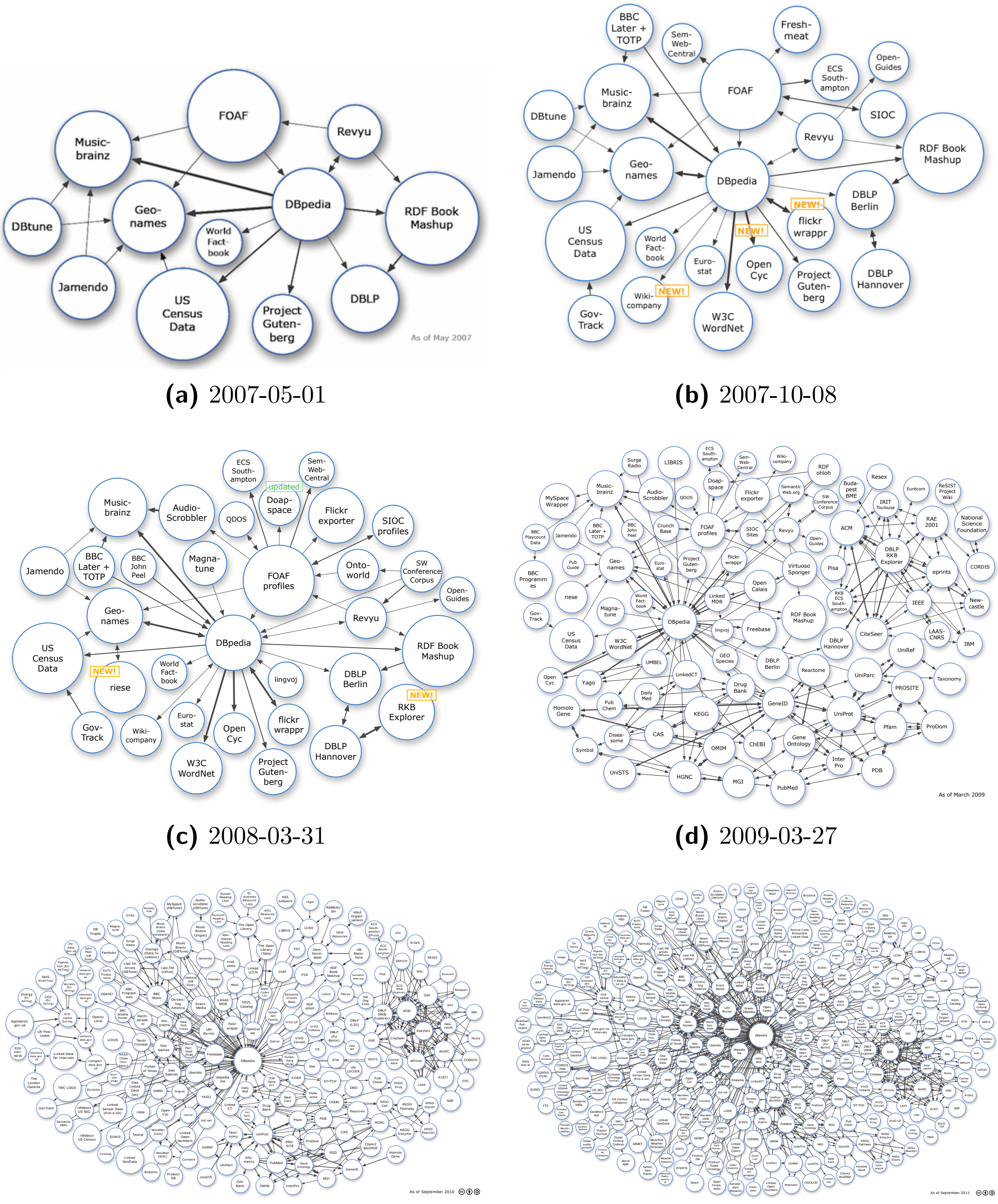


Figure 2.4: Colored LOD cloud diagram, September 2011.
Source: Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/



(a) 2007-05-01

(b) 2007-10-08

(c) 2008-03-31

(d) 2009-03-27

(e) 2010-09-22

(f) 2011-09-19

Figure 2.5: Evolution of the LOD cloud.
Source: Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. `http://lod-cloud.net/`

The success of the LOD project is also impressively illustrated by the evolution of the LOD cloud that is shown in figure 2.5. The first LOD cloud diagram has been produced in May 2007 and is shown in figure 2.5a. At that time only 12 datasets constituted Linked Open Data. The most important dataset DBpedia [8] was already created at the beginning of the LOD project. DBpedia is a semantic version of the data contained in Wikipedia. After just 5 months in October 2007 the number of datasets in the LOD cloud (cf. figure 2.5b) already more than doubled. Figure 2.5c shows the state of the LOD cloud in early 2008 where also riese (cf. section 4.1) joined the LOD cloud. Approximately one year later in 2009 (cf. figure 2.5d) the LOD cloud already contained almost 100 different datasets. In late 2010 (cf. figure 2.5e) more than 200 datasets made up LOD and in late 2011 (cf. figures 2.4 and 2.5f) approximately 300 datasets.

Initially mainly research projects and Semantic Web enthusiasts contributed LOD sets by converting third-party content to Linked Data. Recently also primary data producers have started to provide Linked Data access to their datasets. Through this trend even more original data is made available.

Since September 2010 (cf. [101]) LOD providers are asked to provide information about their datasets on the Comprehensive Knowledge Archive Network (CKAN). CKAN is a web-based data catalog supported by the Open Knowledge Foundation. Information about LOD datasets is consequently available via an API from CKAN and is also used to create the LOD cloud diagrams and high-level statistics. In order for new datasets to be contained in the LOD cloud the Linked Data principles should be followed and in addition the dataset must contain at least 1,000 triples and at least 50 links.

Further details about LOD can be found on `http://linkeddata.org/` and are also partly contained in subsequent sections of this thesis. Along the conceptual framework for Linked Data generation and utilization part II of this thesis discusses the creation of linkable data, part III focuses on interrelating Linked Data and part IV covers the consumption of Linked Data.

2.3 Summary

In this chapter we have introduced the basic principles of Linked Open Data and discussed the foundational elements of the Semantic Web. In section 2.1 we discussed Semantic Web technologies along the Semantic Web Stack. A high-level introduction to the various related layers and components has been given. Further on section 2.2 covered the basics of the Linking Open Data (LOD) community project. The Linked Data principles were presented and we also discussed the evolution of LOD with illustrations of the LOD cloud.

Chapter 3

Size of the Semantic Web

As a starting point for the research presented in this thesis we conducted a study in 2008 to assess the state of the Semantic Web. This chapter is based on the results of this study which have been published as the paper "What is the Size of the Semantic Web?" [85]. The paper is partly reprinted in this chapter and has been written together with Michael Hausenblas, Yves Raimond, and Tom Heath. It has been presented by the author of this thesis together with Michael Hausenblas at the 4th International Conference on Semantic Systems (I-SEMANTICS2008) in Graz, Austria, in 2008. The analyses presented in the paper have mainly been conducted by the author of this thesis.

The motivation for conducting this study was to get a better understanding of the characteristics of the Semantic Web at that time. This helped in identifying potential issues that needed to be addressed in the following research. In order to be able to understand the characteristics of the Semantic Web, we examined Linked Open Data acting as a representative proxy for the Semantic Web at large. In 2008 the LOD project measured its success by the number of triples available as Linked Data and we also investigated in this study if the pure number of triples can serve as the single measure of interest. Our main finding was that regarding the size of the Semantic Web, there is more than the sheer number of triples: the number and type of links is an equally crucial measure.

3.1 Motivation

The start of the Linking Open Data (LOD) community project in 2007 has immediately created large impact in the scientific community and gained great attention. Based on the success a number of prominent members of the Semantic Web community were claiming in 2008 that the Semantic Web had arrived. Initiatives such as LOD have indeed started populating the Web with vast amounts of distributed yet interlinked RDF data. Anyone seeking to implement applications based on this data needs basic information about the system with which they are working. We will argue that regarding the size of the Semantic Web, there is more to find than the sheer numbers of triples currently available. Based on our study we aimed at answering what seems to be a rather a simple question: *What is the size of the Semantic Web*?

3.2 Starting Point

On the Web of Documents, typically the number of users, pages or links are used to gauge its size [30, 71]. However, Web links (`@href`) are untyped, hence leaving its interpretation to the end-user [9]. On the Semantic Web we basically deal with a directed labeled graph where a fair amount of knowledge is captured by the links between its nodes.

From semantic search engines we learn that mainly the documents and triples as such are counted. No special attention is paid to the actual interlinking, i.e. the type of the links [52]. In the development of the semantic search engine *swoogle* [60] it has been reported that "... the size of the Semantic Web is measured by the number of discovered Semantic Web Documents". However, later, they also examined link characteristics [43]. Findings regarding the distribution of URIs over documents are well known in the literature [163, 45]. Unlike other gauging approaches focusing on the schema level [179], we address the interlinking aspect of Semantic Web data represented in RDF, comparable to what Ding et al. [45] did in the FOAF-o-sphere.

3.3 Linked Datasets As A Proxy For The Semantic Web

The *reference test data set* (RTDS) we aim to use should be able to serve as a good proxy for the Semantic Web, hence it (i) must cover a range of different topics (such as people-related data, geo-spatial information, etc.), (ii) must be strongly interlinked, and (iii) must contain a sufficient number of RDF triples (we assume some millions of triples sufficient). As none of the available alternatives in 2008 - such as the Lehigh

University Benchmark dataset[1], Semantic Wikis (such as [166]) or embedded metadata - exhibit the desired characteristics, the Linking Open Data datasets were chosen as the RTDS. We note that embedded metadata (in the form of microformats, RDFa, eRDF and GRDDL) are constituting a large part of the openly published metadata. However, the interlinking of this data is not determinable unambiguously.

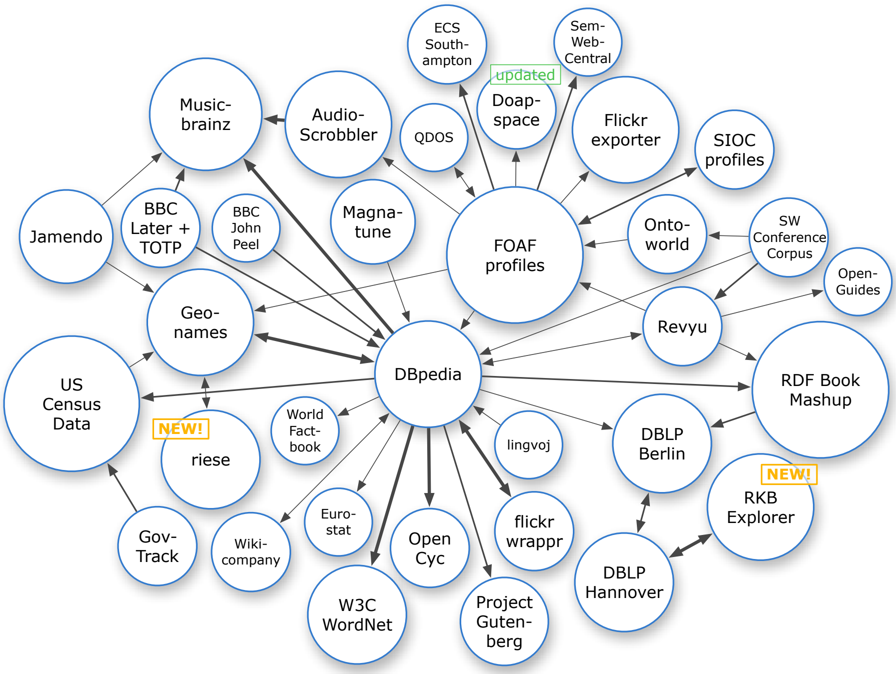


Figure 3.1: The LOD dataset in 2008, at time of conducting the study.

The basic idea of Linked Data has already been introduced in chapter 2. At the time of conducting the study in early 2008 roughly two billion triples and three million interlinks have been reported (cf. figure 3.1[2], ranging from rather centralized ones to those that are very distributed. A detailed description of the datasets contained in the LOD is available in Table 3.1.

3.4 Gauging the Semantic Web

In order to find metrics for the Semantic Web we examined its properties by inducing from the LOD dataset analysis. One possible dimension to assess the size of a system like the Semantic Web is the data dimension. Regarding data on the Semantic Web,

[1] http://swat.cse.lehigh.edu/projects/lubm/

[2] by courtesy of Richard Cyganiak, http://richard.cyganiak.de/2007/10/lod/

Name	**Triples** (millions)	**Interlinks** (thousands)	**Dump** download	**SPARQL** endpoint
BBC John Peel	0.27	2.1		
DBLP	28	0		yes
DBpedia	109.75	2,635	yes	yes
Eurostat	0.01	0.1		yes
flickr wrappr	2.1	2,109		
Geonames	93.9	86.5	yes	
GovTrack	1,012	19.4	yes	yes
Jamendo	0.61	4.9	yes	yes
lingvoj	0.01	1.0	yes	
Magnatune	0.27	0.2	yes	yes
Musicbrainz	50	0		
Ontoworld	0.06	0.1	yes	yes
OpenCyc	0.25	0	yes	
Open-Guides	0.01	0		
Project Gutenberg	0.01	0		yes
Revyu	0.02	0.6	yes	yes
riese	5	0.2	yes	yes
SemWebCentral	0.01	0		
SIOC	N/A	N/A		
SW Conference Corpus	0.01	0.5	yes	yes
W3C Wordnet	0.71	0	yes	
Wikicompany	?	8.4		
World Factbook	0.04	0		yes

Table 3.1: Linking Open Data dataset at a glance in 2008.

we roughly differentiate into: (i) the **schema level**, (ii) the **instance level**, i.e. a concrete occurrence of an item regarding a certain schema (see also [86]), and the actual **interlinking**: the connection between items; represented in URIrefs and interpretable via HTTP. This aspect of the data dimension will be the main topic of our investigations, below.

As stated above, the pure number of triples does not really tell much about the size of the Semantic Web. Analyzing the links between resources exhibits further characteristics. The LOD dataset can roughly be partitioned into two distinct types of datasets, namely (i) **single-point-of-access datasets**, such as DBpedia or GeoNames, and (ii) **distributed datasets** (e.g. the FOAF-o-sphere or SIOC-land). This distinction is significant regarding the access of the data in terms of performance and scalability.

Our initial approach aimed at loading the whole LOD dataset into a relational database (Oracle 11g Spatial) on a single machine. Due to technical limitations this turned out not to be feasible - the overall time to process the data exceeded any sensible time constraints. As not all LOD datasets are available as dumps, it became obvious that additional crawling processes were necessary for the analysis. We finally arrived at a hybrid approach. The available and the self-crawled dumps together were loaded into the relational database, were the analysis took place using SQL. Additionally, we inspected the descriptions provided by the LOD dataset providers in order to identify parts of the dataset which are relevant for interlinking to other datasets. Where feasible, we also used the available SPARQL-endpoints.

3.4.1 Single-point-of-access Datasets

It has to be noted that only a certain subset of the links actually yields desirable results in the strict sense, i.e. return RDF-based information when performing an `HTTP GET` operation.

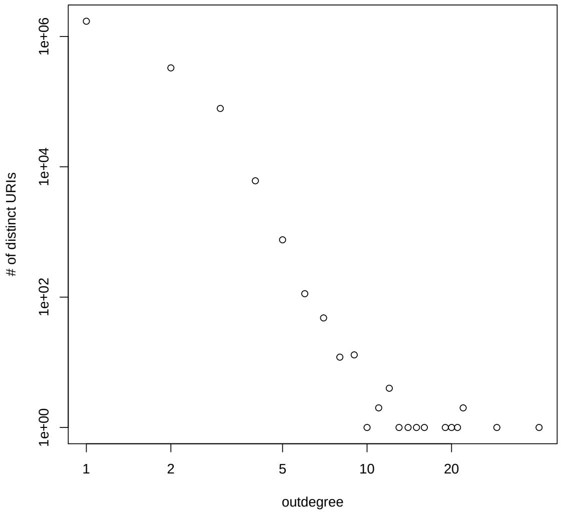


Figure 3.2: Outgoing Links From the DBpedia dataset in 2008.

Taking the DBpedia dataset as an example yields that only half of the properties in this dataset are dereferenceable. Figure 3.2 depicts the distribution of the dereferenceable outgoing links from the DBpedia dataset. We would expect this distribution to be modeled by a power-law distribution considering the degree of DBpedia resources (the number of resources having a given number of links to external datasets). However, figure 3.2 does not clearly suggest this, which may be due to too little data or due to the fact that links from DBpedia to other datasets are created in a supervised way, whereas scale-free networks tend to represent organic and decentralized structures.

Property (Link Type)	**Occurrence**
`http://dbpedia.org/property/hasPhotoCollection`	2.108.962
`http://xmlns.com/foaf/0.1/primaryTopic`	2.108.962
`http://dbpedia.org/property/wordnet_type`	338.061
`http://www.w3.org/2002/07/owl#sameAs`	307.645
`http://xmlns.com/foaf/0.1/based_near`	3.479
`http://www.w3.org/2000/01/rdf-schema#seeAlso`	679

Table 3.2: Overall Occurrence of Link Types in the LOD dataset in 2008.

We found only a limited number of dereferenceable links in the LOD dataset (Table 3.2); this distribution is biased towards the DBpedia dataset and the flickr wrapper, however. In case of the single-point-of-access datasets, we found that mainly one or two interlinking properties are in use (figure 3.3). The reason can be seen in the way these links are usually created. Based on a certain template, the interlinks (such as `owl:sameAs`) are generated automatically.

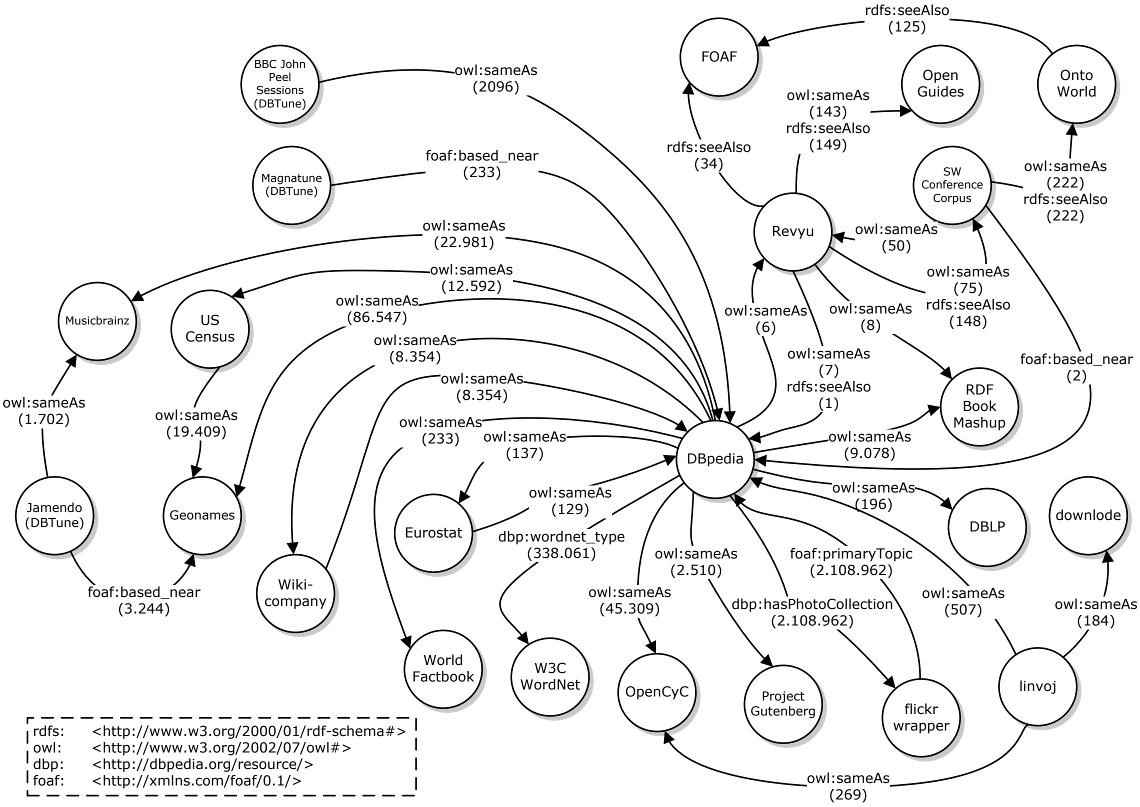


Figure 3.3: Single-point-of-access Partition Interlinking in 2008.

As the data model of the Semantic Web is a graph the question arises if the density of the overall graph can be used to make a statement regarding the system's size. The LOD dataset is a sparse directed acyclic graph; only a few number of links (compared to the overall number of nodes) exist. Introducing links is costly. While manual added, high-quality links mainly stem from user generated metadata, the template-based generated links (cheap but semantically low-level) can be added to a greater extent.

3.4.2 Distributed Datasets

In order to analyze the partition of the LOD covering the distributed dataset, such as the FOAF-o-sphere, we need to sample it. Therefore, from a single seed URI[3], approximately six million RDF triples were crawled. On its way, 97410 HTTP identifiers for persons were gathered. We analyzed the resulting sampled FOAF dataset, yielding the results highlighted in Table 3.3.

To	**Interlinking Property**	**Occurrence**
FOAF	`foaf:knows` (direct)	132.861
FOAF	`foaf:knows+rdfs:seeAlso`	539.759
Geonames	`foaf:based_near`	7
DBLP	`owl:sameAs`	14
ECS Southampton	`rdfs:seeAlso`	21
ECS Southampton	`foaf:knows`	21
DBPedia	`foaf:based_near`	4
DBPedia	`owl:sameAs`	1
RDF Book Mashup	`dc:creator`	12
RDF Book Mashup	`owl:sameAs`	4
OntoWorld	`pim:participant`	3
Revyu	`foaf:made`	142
Other LOD datasets	-	0
Total inter-FOAF links	-	672.620
Total of other links	-	229

Table 3.3: Interlinking from a sampled FOAF dataset to other datasets in 2008.

Although the intra-FOAF interlinking is high (in average, a single person is linked to 7 other persons), the interlinking between FOAF and other datasets is comparably low; some $2 * 10^{-3}$ interlinks per described person have been found. Also, the proportion of *indirect* links from a person to another (using `foaf:knows` and `rdfs:seeAlso`) is higher than *direct* links (through a single `foaf:knows`). These two observations lead to the conclusion that Linked Data principles were not widely used throughout the FOAF

[3] http://kmi.open.ac.uk/people/tom/

community in 2008. Using FOAF does not limit one to link to other resources: one can link to his geographical location on GeoNames, his cultural interests on DBPedia or DBTune, etc.

3.5 Conclusions of the 2008 Study

In this study in 2008 we have attempted to make a step towards answering the question: *What is the size of the Semantic Web*?. Based on a syntactic and semantic analysis of the LOD dataset we believe that answers can be derived for the entire Semantic Web. We have identified two different types of datasets, namely single-point-of-access datasets (such as DBpedia), and distributed datasets (e.g. the FOAF-o-sphere). At least for the single-point-of-access datasets it seems that automatic interlinking yields a high number of semantic links, however of rather shallow quality. Our finding was that not only the number of triples is relevant, but also how the datasets both internally and externally are interlinked. The main conclusions regarding the further development of Linked Open Data were:

- More Linked Data is needed and therefore the creation of linked (or linkable) data needs to be eased
- Interrelations (interlinks) within and between different datasets need to be increased
- The quality of links needs to be improved

3.6 Follow-up Activities of the 2008 Study

The outcomes of this study have been used to align our further research activities. In the follow-up research we have contributed to easing the creation of linkable data and also made contributions to interrelating the various datasets with improved quality. The individual contributions are presented in the following sections of this thesis. We have also investigated the logical final step in the conceptual framework for Linked Data generation and utilization which is the consumption of Linked Data.

Our initial study did not only serve as a guide for our follow-up research but has also attracted interest from the scientific community and other researchers have also investigated ways for characterizing Linked Open Data. In the famous 2009 *International*

Journal on Semantic Web and Information Systems article *Linked Data - The Story So Far* [24] our approach for analyzing the Web of Data has been mentioned as one strategy for estimating LOD statistics and characteristics. The article also mentions an alternative of estimating the size based on statistics provided in W3C's ESW Wiki. In March 2009 this alternative approach has been started to collect statistics about the size of Linked Open Data. Dataset providers have been asked to enter the number of triples available in their datasets on that Wiki page. This Wiki-based approach has been used until late 2010. In September 2010 an improved method for capturing dataset statistics has been announced [101]: Since then dataset providers were asked to provide statistics about their LOD data on the Comprehensive Knowledge Archive Network (CKAN). CKAN is a web-based data catalog supported by the Open Knowledge Foundation. Information about LOD datasets has consequently been used from CKAN to create the well-known LOD cloud diagrams and also create high-level statistics about LOD. The advantage of this approach is that information about LOD datasets can be collaboratively maintained on CKAN and it also supports easier retrieval of relevant LOD datasets. However, regarding the accuracy of the statistics it has to be noted that all information on CKAN about the size of the datasets as well as the number of interlinks are only provided manually to the catalog. Therefore the figures presented there are just estimates and the actual figures could be higher or lower. An automated verification of the information is not done. In comparison, our initial study approach from 2008 relied on investigating the entire data and was thus able to provide detailed statistics of the actual data.

Another approach that is able to capture LOD statistics was presented with the *Vocabulary of Interlinked Datasets (VoID)* [2]. It can be used to describe metadata about RDF datasets and its development started in 2008, also taking findings from our study into account. The vocabulary can be used to describe general metadata, access metadata, structural metadata, and links between datasets. VoID also depends on the dataset provider to make these descriptions available either manually or (if supported by the relevant software) also automatically. The vocabulary has been published as W3C Interest Group Note [3] in 2011 and receives increasing uptake. Characteristics metadata about a dataset provided with VoID can be used to update dataset descriptions on CKAN. In early 2012 approximately one third of the LOD datasets provide descriptions using VoID.

With the *Dynamic Linked Data Observatory (DyLDO)* Käfer et al. [102] have announced in 2012 a further approach for acquiring detailed Linked Data characteristics. As our initial 2008 study of Linked Data characteristics their approach relies on crawling a sample dataset collection and analyzing it. DyLDO is supposed to be suitable for general-purpose studies on a wide range of Linked Data aspects. However, at the time of writing only the announcement was available without concrete results.

In general it can be stated that there is a need for and interest in detailed Linked Data statistics. Our initial study of 2008 already showed some of the Linked Data characteristics and has provided insights into areas with potential for improvement of the Linked Data ecosystem.

3.7 Summary

The research question *What is the size of the Semantic Web?* has been addressed in this chapter. As a starting point for the research presented in this thesis we conducted a study in 2008 to assess the state of the Semantic Web. We have identified two different types of datasets, namely single-point-of-access datasets (such as DBpedia), and distributed datasets (e.g. the FOAF-o-sphere). At least for the single-point-of-access datasets it seems that automatic interlinking yields a high number of semantic links, however of rather shallow quality. Our finding was that not only the number of triples is relevant, but also how the datasets both internally and externally are interlinked. The outcomes of this study have been used to align our further research activities. Follow-up activities of our 2008 study have also been briefly addressed.

Chapter 4

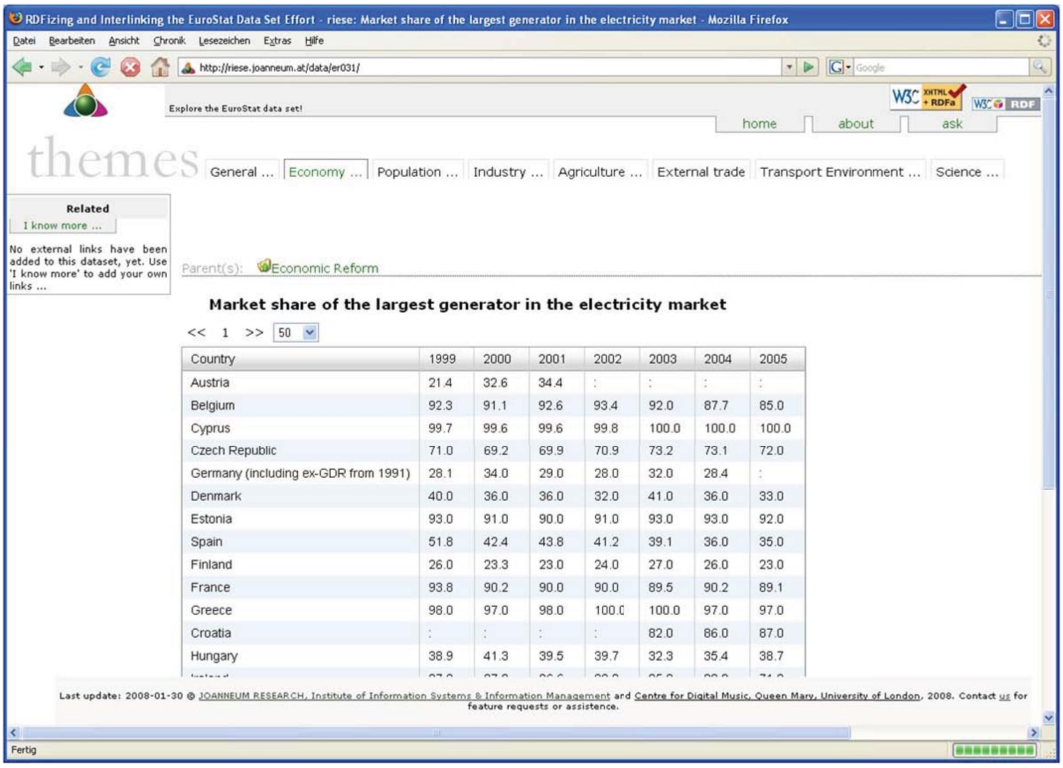
Guiding Use Cases

The research in the course of thesis along the conceptual framework for Linked Data generation and utilization has been aligned with two major use cases: The first use case riese is in the application area of government data and the second use case *Link2WoD* targets the media industry. Both use cases consider the three main steps (creating, interrelating, and consuming) of the conceptual framework. A brief introduction to these two use cases follows below. Additional sub-results have also been achieved in other related settings and use cases that are explained and stated in the relevant sections.

4.1 riese

The *RDFizing and Interlinking the EuroStat Data Set Effort* (riese) has been started by us in 2008 to make EuroStat statistics available as Linked Data. It has been initiated as part of the W3C SWEO Linking Open Data (LOD) project and aims at being useful for both humans and machines. At the time of starting this effort most LOD datasets such as DBpedia [8] were only targeted at machine consumption. With *riese* we realized an approach to satisfy both humans and machines from the same source. An exemplary screenshot of the *riese* Web application's user interface is depicted in figure 4.1.

Eurostat provides detailed statistics for the entire European Union as well as additional statistics for major other countries. The publicly available statistical data contains approximately 350 million data values and serves as a basis for the Linked Data that we made available through riese. Our related research covered the three main steps of the conceptual framework for Linked Data generation and utilization. We have shown an approach to create, interrelate, and consume Linked Data based on



No external links have been added to this dataset, yet. Use 'I know more' to add your own links ...

Market share of the largest generator in the electricity market

Country	1999	2000	2001	2002	2003	2004	2005
Austria	21.4	32.6	34.4	:	:	:	:
Belgium	92.3	91.1	92.6	93.4	92.0	87.7	85.0
Cyprus	99.7	99.6	99.6	99.8	100.0	100.0	100.0
Czech Republic	71.0	69.2	69.9	70.9	73.2	73.1	72.0
Germany (including ex-GDR from 1991)	28.1	34.0	29.0	28.0	32.0	28.4	:
Denmark	40.0	36.0	36.0	32.0	41.0	36.0	33.0
Estonia	93.0	91.0	90.0	91.0	93.0	93.0	92.0
Spain	51.8	42.4	43.8	41.2	39.1	36.0	35.0
Finland	26.0	23.3	23.0	24.0	27.0	26.0	23.0
France	93.8	90.2	90.0	90.0	89.5	90.2	89.1
Greece	98.0	97.0	98.0	100.0	100.0	97.0	97.0
Croatia	:	:	:	:	82.0	86.0	87.0
Hungary	38.9	41.3	39.5	39.7	32.3	35.4	38.7

Last update: 2008-01-30 © JOANNEUM RESEARCH, Institute of Information Systems & Information Management and Centre for Digital Music, Queen Mary, University of London, 2008. Contact us for feature requests or assistence.

Figure 4.1: Screenshot of riese user interface using XHTML+RDFa.

statistical data that is available in a simple tabular format. This also motivated us to contribute to the general issue of generating RDF from relational databases where we have participated in the W3C RDB2RDF Incubator and Working Groups. In addition we have made early contributions to the field of open government data in LOD and the riese core schema evolved into SCOVO, the Statistical Core Vocabulary [84] where a detailed description is presented in section 6.2. SCOVO has also influenced further developments which are detailed in section 6.3.

To realize Linked Data access for both humans and machines the riese Web application supports the following access modes:

- Human users: Users can navigate the dataset provided in XHTML+RDFa
- Semantic Web agents:
  - single item query: XHTML+RDFa per page allows the exploration of the dataset and the query of a single data item
  - global query: to allow an efficient query of the entire dataset, a SPARQL-endpoint is provided

- indexer: to allow semantic search engines (indexers) an effective processing, the entire dataset is offered as a dump along with an according description

More detailed explanations of the individual aspects and the related research follow in further sections of this thesis. Chapter 6 discusses the generation of linkable data from structured sources also on the example of riese. Regarding the interrelation of Linked Data user based approaches (cf. chapter 9) as well as automated approaches (cf. chapter 10) have been developed in the context of *riese*. Finally the consumption of government data is discussed in chapter 12.

4.2 Link2WoD

The second major use case *Link2WoD* (Link to Web of Data) targets the media industry and has been developed with the intention to support editors at online content providers but can also be used as a general-purpose tool for enriching unstructured data with Linked Data. A high-level overview of the Link2Wod operating principles is depicted in figure 4.2. Basically, Link2WoD takes plain text, i.e. unstructured data, as input and processes it. The output is Linked Data that contains interlinked information with enhanced and enriched content from various Linked Data sources.

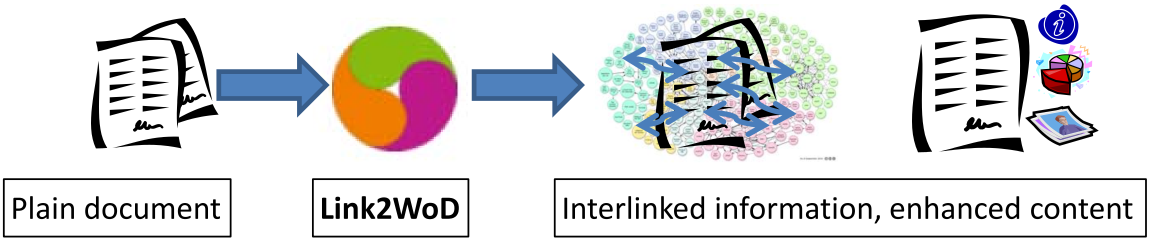


Figure 4.2: High-level overview of Link2Wod operating principles.

Link2WoD consists of three different modules that work together as a unified solution that can be integrated into an existing content management system as a service or can be used as a stand-alone web application. The term extraction and classification module identifies interesting and relevant terms which act as an input for the Linked Data consumption & interlinking module where additional information from Linked Data sources is collected. The Linked Data Provision module makes the discovered information available as Linked Data to the public.

In a use case where Link2WoD is integrated in a content authoring environment for online content providers the editors can continue to use the already established systems. During content creation the text is analyzed and Link2WoD suggests further information from Linked Data and multimedia sources on the Web. The editor can then decide with a single click on which external content to include in the article. This manual decision step has been introduced following feedback from editors and content providers as they prefer to have more control over the published content. It further allows improving content suggestions based on previous selections and preferences. As an added benefit the final article is enriched with more exciting content and provides the reader with further information without having to leave the provider's pages. This increases the attractiveness of the content provider with further potential positive effects on visit durations, returning users, as well as page and ad impressions leading to higher ad revenues for the content provider which is a crucial measure.

A distinguishing feature of the tool is that it is ready to support multilingual content which poses additional challenges in the context of the Web of Data where English is still the predominant language. The term extraction and classification module developed for Link2WoD supports German language text and additional third-party modules can be plugged in for the support of further languages.

For the final outcome we target humans and machines as well again (as we did in the riese use case). Human users can access the original content along with additional information that has been consolidated from other Linked Data sources. The additional information can be made available in separate areas of the web page, e.g., placed in information boxes, or as tooltip when the users moves the cursor over an identified term where supplementary information is available. The data along with its links can be made available via the XHTML+RDFa approach where machine users can easily extract the annotated original data and links to other resources.

A standalone technology preview demonstrator of Link2WoD is available online at `http://link2wod.joanneum.at`. It takes plain text as input, analyzes it and presents all discovered information in the result page. This process functions similar to an integrated version of Link2WoD for editors but it can also be used for general purpose information enrichment.

More detailed explanations of the individual modules and related research follow in the upcoming sections of this thesis. In chapter 7 we briefly discuss how linkable data can be generated from unstructured data sources such as plain text. The specific aspects of interrelating this data in the context of Link2WoD are covered in section 10.3. Finally the consumption of Linked Data for applications in the media industry (which the Link2WoD demonstrator targets) are discussed in chapter 13.

Part II

Creating Linkable Data

Chapter 5

General Considerations on Creating Linkable Data

The idea of Linked Open Data relies on having data available in a machine-readable format that in the best case follows the four Linked Data principles [14]. The creation of linkable data is also the first step in our framework for Linked Data generation and utilization. In addition Berners-Lee has also developed a star rating system (cf. [14]) to classify good Linked (Open) Data. The highest rating of 5 stars can be achieved both for data that is open (Linked *Open* Data) as well as for data that is not open (only Linked Data) which might be desirable for instance for internal use inside organizations. What we consider as the creation of linkable data are all the (preparatory) steps towards 5-star Linked Data except for the final step of the actual interlinking. In our framework the interlinking takes place in the second phase of the framework where data is being interrelated to create true Linked Data. Looking at the star rating the following aspects are covered by creating linkable data:

⋆ **On the Web**: Data should be available on the Web. This also relates to the Linked Data rule of using HTTP URIs to name things that can be looked up. At this low stage a data consumer can look up the data, even if it is just a scanned copy of some printed table. However, the reusability of such data might be very limited. Considerable effort might be required for being able to further process this data. Towards Linked *Open* Data it is also desirable to publish the data under an open license.

⋆⋆ **Machine-readable data**: Data should be available in a structured format which also relates to the Linked Data rule of providing useful information when accessing the resource. When data is machine-readable it can be processed by at least some (potentially proprietary) application. An example for data at this stage is an Excel file which is machine-readable but locked up to processing with some proprietary software.

⋆⋆⋆ **Non-proprietary format**: In addition, data should use a format that is not proprietary. This means that data is not locked in to a specific vendor but can be used with different applications for further processing. An example at this stage is a table represented as comma-separated values (CSV) which is for instance described in RFC 4180 [153]

⋆⋆⋆⋆ **RDF standards**: In the best case the data uses W3C standards such as RDF to identify things. This eases compatibility, integration, and reuse across different dataset providers. This sort of 4-star data is what we are aiming for in the creation of linkable data. Items should have their own URI and can be shared on the Web. This stage is also sometimes referred to as having data *in* the Web.

The final fifth star can be obtained by the interrelation of Linked Data - which is described in the third part of this thesis:

⋆⋆⋆⋆⋆ **Linked RDF**: Data in the final stage can be considered good Linked Data. Data is now *in* the Web and *linked* with other data. This allows both the consumer and the publisher to take advantage of the benefits of Linked Data and its network effects. More related data can be discovered when consuming the data, data can easily be discovered and this also increases the value of the data.

5.1 Sources of Data

The creation of linkable data starts at its very roots. In general some source data needs to be transformed into Linked Data. The question that might arise is: *Which data is linkable?* The short answer to this question is that every sort of source data can become Linked Data. The biggest benefits of Linked Data can be gained with data that is potentially reused. There exist cases and situations where it can be expected that reuse will be high such as with general facts like information about a country, an important person or alike. However, even personal notes and communication bear a potential for reuse and can benefit from being available as (private) Linked Data. In the early days of the LOD project DBpedia started as *a nucleus for a Web of Open Data* (cf. [8]) by providing a Linked Data version of Wikipedia, the well-known free collaborative encyclopedia. Data from several other domains is available as Linked Data by now and there are almost no limitations on the kind of data that could be made available as Linked Data. General facts, medical or biotechnological information, databases, event-based data, news, financial data, social information, multimedia, etc. can all be used as a source for creating linkable (and in the end Linked) data.

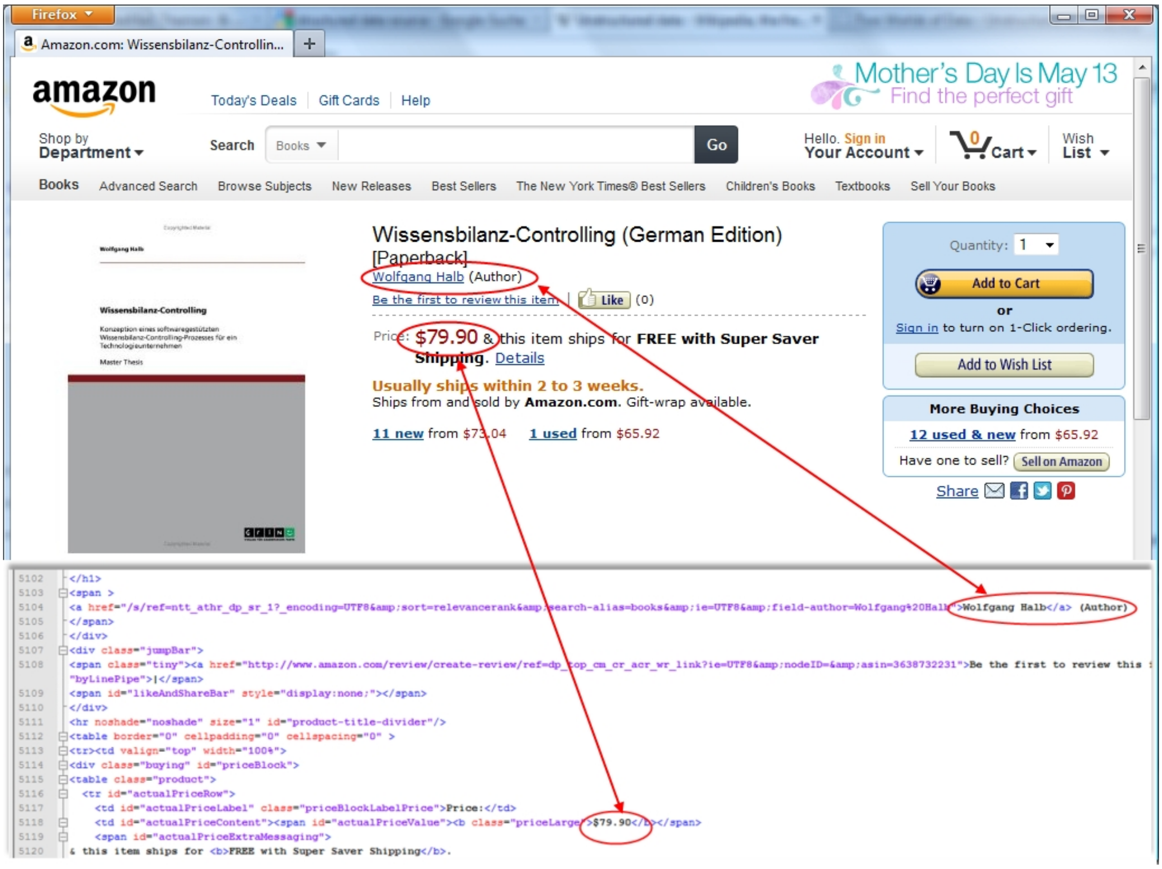
A general distinction can be made between two different major kinds of source data: It can be available as *structured* or *unstructured* data. Structured data is modeled using an abstract model which organizes the data and already carries a certain level of semantics. Common occurrences are in data management systems such as relational databases. Unstructured data refers to information that does not have a pre-defined data model and usually consists of textual or multimedia objects. The unstructured data may also contain structured information that needs to be extracted or enriched. Typical Web pages are also usually considered unstructured data even though HTML uses tags but these tags are mainly used for rendering. The example in figure 5.1 shows a screenshot of a Web page displaying a product and an accompanying HTML source code snippet. Even though the Web page might be dynamically created based on structured content this structure is not conveyed in the wrapping object of this information. Different techniques are required to extract information such as the author or the price of the book. Plain text content is even more difficult to process and various analyses are needed to extract the semantics of unstructured data.

The nature of the source data also plays a role when creating linkable data out of it. In chapter 6 we look at some approaches for creating linkable data out of structured datasources. The handling of unstructured datasources is discussed in chapter 7 along with our approaches.

5.2 Generation Approaches

Following the Linked Data principles [14] everyone is able to contribute information to the Web of Data. In principle it is as easy as publishing RDF data. Simply put, it is possible to manually create RDF data and publish this data on a server. However, some considerations have to be made. Linked Data encourages to use HTTP URIs for naming things because this mechanism provides a simple way to create globally unique names in the Web which can also be accessed (de-referenced) to get information describing the entity. On the classical Web these URIs can be used to identify HTML pages and in the context of Linked Data the URIs identify real-world objects and abstract concepts.

When looking up an URI two different representations can be presented: Usually a RDF data view is intended for machine consumption and a HTML view for humans. The Web provides mechanisms to provide both humans and machines with represen-


Firefox
Amazon.com: Wissensbilanz-Controllin...
amazon
Today's Deals
Gift Cards
Help
Mother's Day Is May 13
Find the perfect gift
Shop by Department
Search
Books
Go
Hello. Sign in Your Account
0 Cart
Wish List
Books
Advanced Search
Browse Subjects
New Releases
Best Sellers
The New York Times® Best Sellers
Children's Books
Textbooks
Sell Your Books
Wissensbilanz-Controlling (German Edition)
[Paperback]
Wolfgang Halb (Author)
Be the first to review this item
Like (0)
Price: $79.90 & this item ships for FREE with Super Saver Shipping. Details
Usually ships within 2 to 3 weeks.
Ships from and sold by Amazon.com. Gift-wrap available.
11 new from $73.04
1 used from $65.92
Quantity: 1
Add to Cart
or
Sign in to turn on 1-Click ordering.
Add to Wish List
More Buying Choices
12 used & new from $65.92
Have one to sell? Sell on Amazon
Share
Wolfgang Halb
Wissensbilanz-Controlling
Master Thesis
GRIN

```
</h1>
<span >
<a href="/s/ref=ntt_athr_dp_sr_1?_encoding=UTF8&amp;sort=relevancerank&amp;search-alias=books&amp;ie=UTF8&amp;field-author=Wolfgang%20Halb">Wolfgang Halb</a> (Author)
</span>
</div>
<div class="jumpBar">
<span class="tiny"><a href="http://www.amazon.com/review/create-review/ref=dp_top_cm_cr_acr_wr_link?ie=UTF8&amp;nodeID=&amp;asin=3638732231">Be the first to review this i
"byLinePipe">|</span>
<span id="likeAndShareBar" style="display:none;"></span>
</div>
<hr noshade="noshade" size="1" id="product-title-divider"/>
<table border="0" cellpadding="0" cellspacing="0" >
<tr><td valign="top" width="100%">
<div class="buying" id="priceBlock">
<table class="product">
  <tr id="actualPriceRow">
    <td id="actualPriceLabel" class="priceBlockLabelPrice">Price:</td>
    <td id="actualPriceContent"><span id="actualPriceValue"><b class="priceLarge">$79.90</b></span>
    <span id="actualPriceExtraMessaging">
& this item ships for <b>FREE with Super Saver Shipping</b>.
```

Figure 5.1: Example screenshot of product Web page and source code snippet.

tations of a resource that meets their needs. This is achieved through the content negotiation feature of HTTP which is "the process of selecting the best representation for a given response when there are multiple representations available." [58]. The client can specify in the HTTP header which type of document it prefers. In the *303 URIs* strategy the server responds to the client with the HTTP response code `303 See Other` and the URI of a Web document which describes the real-world object. Based on the preferences of the client either the URI of a RDF or HTML representation is returned. As an example, the URI `http://example.com/wolfgang-halb` identifies the person Wolfgang Halb. Based on the client preferences a redirect is done either to the URI containing the RDF/XML representation (`http://example.com/wolfgang-halb.rdf`) or to the HTML representation (`http://example.com/wolfgang-halb.htm`). This approach however requires two HTTP requests to retrieve a single entity description and has been criticized therefore. A different possibility uses fragment identifiers and is also called the *hash URI* strategy. It builds on the characteristic that URIs may contain a special part that is separated from the base part of the URI by a hash symbol (#). The client retrieves the entire base document and only further processes the part identified by the fragment identifier. The downside of this approach is that the entire base document is transmitted which can introduce increased bandwidth consumption. A mix of both strategies can be observed in Linked Data sources and both approaches have their advantages and disadvantages. For small datasets or simple vocabulary definitions it might be more comfortable to use the hash URI strategy whereas big datasets tend to prefer the 303 URI strategy.

Apart from choosing URIs for the resources also the vocabulary is important. The Web of Data is open to every vocabulary that one wants to use. To increase interoperability it is considered good practice to use terms from well-known RDF vocabularies such as Friend of a Friend (FOAF) [28] to describe persons, organizations, and the relationships between them, Dublin Core [47] for general metadata, SKOS [151] and alike. Different communities also have specific preferences on the vocabularies and for instance for the domain of statistical data we have developed SCOVO, the statistical core vocabulary (cf. section 6.2).

Regarding the overall strategy for generating linkable and Linked Data there are the two possibilities of entirely manually creating the RDF or using some tool to support in the creation. Manual generation is usually only done for testing purposes or for very small datasets. It is much more comfortable to automate the entire process or at least parts of it. There exist different tools for automated Linked Data generation depending on the nature of the originating source data. In some cases users do not even have to bother with the creation of Linked Data as it is built in into various platforms. Reviews posted at revyu.com for instance are automatically also made available as Linked Data and images posted on flickr also automatically enter the Web of Data through a flickr wrapper. However, in many cases this conversion is not done automatically in the background. If originating data is already available as RDF the most convenient way of publishing this RDF is by using one of the many available triple stores. Triple stores are purpose-built databases for the storage and retrieval of triples and can also be considered one sort of specialized NoSQL solution that follows the standards around RDF. In contrast to relational databases, triple stores are optimized for the handling of triples and usually provide querying via SPARQL.

If the datasource is not yet available as RDF it needs to be processed accordingly. For structured data two basic approaches can be applied. It is possible to create a dynamic RDF view on top of the structured source data which converts to RDF on the fly. The benefit if this approach is that both the relational view as well as the dynamic view are always in sync and updates are propagated instantaneously. It works well when accessing individual resources of a dataset. For analyses that need to access the entire dataset the dynamic view is less well suited than the second approach which is referred to as Extract-Transform-Load (ETL). In the ETL approach the original data is extracted from the structured data source, transformed to RDF, and then loaded into a triplestore. In the case of unstructured source data additional processing is required to produce linkable data. In most cases information extraction techniques have to be applied to the textual content. In the case of multimedia content it will be desirable to add metadata descriptions to the content and either multimedia analyses and/or user-based approaches are needed. More information on the generation follows in the subsequent sections along with descriptions of our examples.

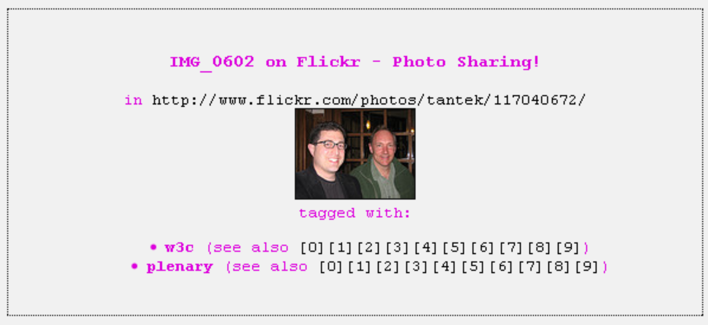
5.3 Further Considerations

There are also further considerations that might need to be taken into account and in a position paper we have discussed the issue of trusting interlinked multimedia data. This section is an adapted reprint of our paper "select * where { :I :trust :you } - How to Trust Interlinked Multimedia Data" [72]. The paper has been written together with Michael Hausenblas and has been presented by the author of this thesis at the International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW'08) in Koblenz, Germany, in 2008.

Finding, accessing and consuming multimedia content on the Web is still a challenge. In this position statement we have discussed three still widely neglected issues arising when one is interacting with multimedia content in social media environments: provenance, trust, and privacy. We will introduce a generic model allowing us to identify potential risks and problems, further discuss this model regarding multimedia content and finally outline how Semantic Web technologies can help.

5.3.1 Motivation

It is a trivial truth that, in order to use any kind of content or service on the Web, one must know how to access it (that is, to know the URI). What is true for the Web of Documents is equally true for the Web of Data. While with the rise of linked data [23, 22] the situation has changed—publishing and consuming data is now possible straight forward—there are still a number of issues in the discovery process. For example, with our multimedia interlinking demonstrator CaMiCatzee [81] we have identified issues around trust and believe of information regarding linked data in general and multimedia content in special. CaMiCatzee allows the FOAF-profile-based search for person depictions on flickr. However, when looking at figure 5.2, how can we find out if one of the depicted persons actually is Sir Tim Berners-Lee?



IMG_0602 on Flickr - Photo Sharing!

in http://www.flickr.com/photos/tantek/117040672/

tagged with:

- w3c (see also [0] [1] [2] [3] [4] [5] [6] [7] [8] [9])
- plenary (see also [0] [1] [2] [3] [4] [5] [6] [7] [8] [9])

Figure 5.2: Exemplary result from CaMiCatzee.

Motivated by this observation we will address—from a practical point of view—widely neglected issues arising when one is interacting with multimedia content in social platforms:

- **Provenance**. Where does content stem from? Who provided annotations?
- **Trust**. Is a person that provided an annotation trustworthy? Is the interlinking eligible?
- **Privacy**. When interacting with content—what are the consequences?

In the following, we will first introduce a generic model addressing the three above listed issues. Based on our experience with interlinked multimedia data we will have a detailed look at the consequences when this model is applied to multimedia content.

5.3.2 The Abstract Provenance-Trust-Privacy Model

In order to identify issues with the usage of content on the Web, we have developed the “Provenance-Trust-Privacy” (PTP) model (figure 5.3). Basically, two aspects are covered by this model: the real life and the online world, the Web. In the PTP model we deal with three orthogonal, nevertheless interdependent dimensions, (i) the **social** dimension, (ii) the **interaction** dimension, and (iii) the **content** dimension.

The Social Dimension. The dotted arrows in figure 5.3 represent social relations between humans, either in the real life or online. While it is straight-forward to deal

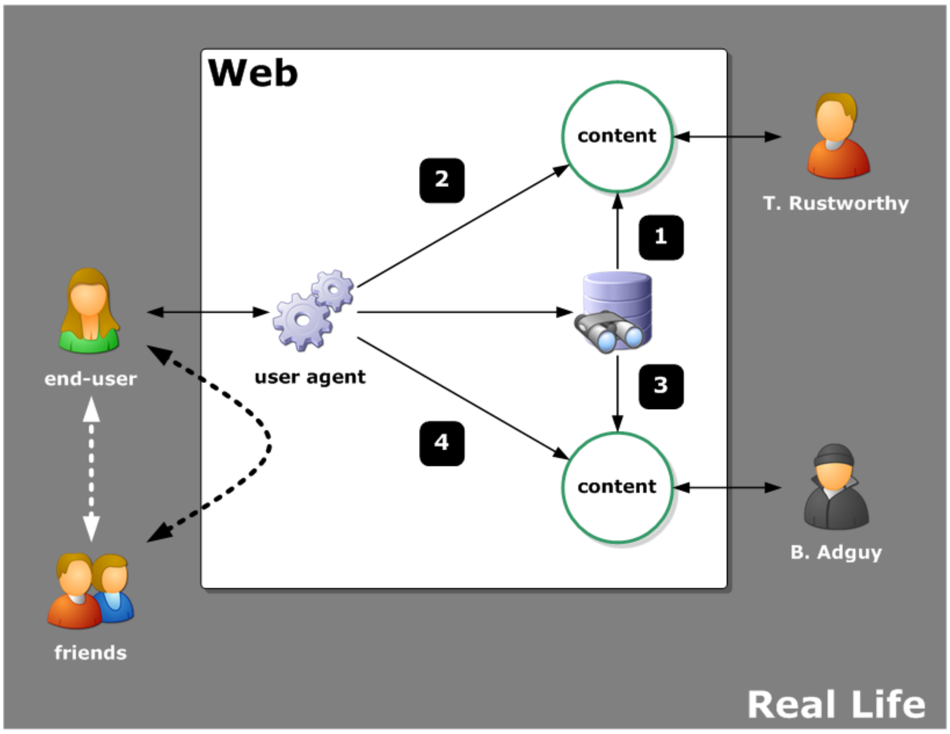


Figure 5.3: The abstract PTP model.

with the case when people know each other in real life (and maybe continue this relation in the online world), the other way round can cause substantial problems. For example, only because I have added someone to my "buddy list" on a social platform such as LinkedIn does not mean that I really know the person and that this person also (wants to) know me.

The Interaction Dimension. In the realm of the PTP model we understand the interaction dimension of taking place online-only. Again, referring to figure 5.3, everything that happens between the user-agent (being instructed by a human) and other participants (server, etc.) on the Web. In general, two interaction patterns for the discovery and access of resources on the Web can be observed:

- A **direct access** of the content. In this case the URI of the content is known in advance by the end-user who instructs her user-agent to access the content. The URI may originally stem from a newspaper advertisement, a friend may have pointed it out in an e-mail, etc.
- An **indirect access** of the content by means of consulting an intermediate such as a search engine, a recommendation system, etc.

Based on these two interaction patterns, four possible paths can be identified:

1. The user-agent, equipped with the end-user's profile and her desire consults an intermediate. For example, a user enters a search string into Google and is presented a list of URIs. The end-user happens to select a trustworthy source.

2. The user-agent, equipped with an URI from the end-user, accesses a trustworthy source.

3. The user-agent, equipped with the end-user's profile and her desire consults an intermediate. This time, the end-user happens to select a troublesome source.

4. The user-agent, equipped with an URI from the end-user, accesses a troublesome source.

Obviously, the first two situations are desirable. The end-user has—for example based on previous experiences or trust in the search results—found some content that she can use and which is not causing her troubles (a virus, Trojan horse, etc.). However, equally, we are after avoiding the latter two cases where the end-user actually finds herself using content and/or services that are harmful and/or violate her privacy.

The Content Dimension. Regarding the content dimension one generally can state that the more is known about the content, the easier it is to assess its usefulness and its capabilities regarding a potential damage. Wherever possible, we are after self-descriptive resources, that is we require a minimum level of metadata being available. In our case, we focus on multimedia content. In the next section we will therefore initially discuss this kind of content and along the metadata in greater detail.

5.3.3 Multimedia Content in the PTP Model

In the position statement at hand we focus on multimedia content. We will in the following discuss characteristics of spatio-temporal multimedia content in the context of the emerging interlinking multimedia effort. Further, we apply the above introduced PTP model to multimedia content and try to derive requirements for it.

Characteristics of Multimedia Content and Multimedia Metadata. Multimedia content has some specific characteristics that allow and/or request special treatment. We have reported on this in great detail elsewhere [31]. A basic observation, however,

Issue	Content	Metadata	Remark
production (prosumers)	++	-	easy to produce high-volume content (e.g., mobile phone)
production (professionals)	++	-	esp. high-level semantic content descriptions expensive
consumption	++	-	easy to consume (also in mobile environments)
search	-	- -	practically, only global descriptions are available
summaries	- -	- -	little automation possible

Table 5.1: Overview on Multimedia Content Characteristics.

with impact on many parts of the interaction process is that with multimedia content we are dealing almost always with spatio-temporal dimensions.

From the prosumers point-of-view, multimedia content is cheap to produce and available in high volumes (mobile phones, etc.). Further, most of the current content in that regard is publicly and freely available (Flickr, youtube). Business models remain vague. On the other hand, for professionally created content for very specific domains such as broadcaster's archives, adult content, etc. the fees are considerably high.

Multimedia content is in general good for consumption in mobile environments (as opposed to reading longish text on a mobile).

In general it is hard and expensive to create good and detailed multimedia content descriptions (for example in MPEG-7, etc.). This leads to a problem regarding the fine-grained search and automated summaries.

In Table 5.1 the above discussed characteristics are summarized, and weighted regarding the content itself on the one hand and the metadata on the other hand.

Applying the PTP Model to Multimedia Content. With the above listed characteristics of multimedia content in mind we claim the following regarding the application of the PTP model:

- Any solution addressing PTP issues must at least avoid accessing "bad" content and should support the discovery and consumption of "good" content.
- Existing and deployed multimedia metadata formats (such as Exif, ID3, etc.)

have to be taken into account.

- The solution at hand needs to scale to the size of the Web.
- It must be practically relevant in terms of availability in widely used platforms such as Drupal, MediaWiki, etc. (for example as a plug-in, etc.; however it needs to be integrated).
- Provider must be able to easily offer and administer it (e.g., "enable" it with little configuration effort).
- Consumer must be able to use it in a non-disruptive way, for example as a part of their everyday tools.

In the next section we will report on how Semantic Web technologies can be used in combination with other, deployed technologies (such as for identification and authentication) in order to address the above listed requirements.

5.3.4 How Semantic Web Technology Can Help

We strongly believe that Semantic Web technologies can help to realize a PTP model for multimedia content. Lots of research is already available[1], however with little practical impact to this end. Apart from avoiding unreliable or even malicious content, the main aim of applying PTP to multimedia content is to help the user in finding trustworthy information sources. In a first step, we consider all content created by a trusted person or authority to be trustworthy. Solving this issue implies that there need to be techniques that can ensure a content's provenance and the content producer's identity respectively. Consequently means have to be made available that can decide which person to trust or not.

In the case of multimedia content it also has to be taken into account that information associated with a single content item can potentially have a multitude of contributors. A photo along with some metadata (title, description) on Flickr for instance might be uploaded by the fictitious trusted user "T. Rustworthy". Another user,"B. Adguy", could add a fake note about who is depicted in the photo. When accessing the photo and the associated metadata it is thus not sufficient to only consider who contributed the image but we would also need to be able to figure out who contributed the metadata about it. Taking this further to video content it might also be relevant

[1] http://www4.wiwiss.fu-berlin.de/bizer/SWTSGuide/

to take into account who contributed which parts of a video (consider, for example, advertisements inserted into a video stream).

In the following we will discuss already available technologies that may be able to address the PTP issues along the three identified dimensions. However, to date only isolated solutions exist and there is still a lack of systems that incorporate all available technologies for the user's benefit. We envision a framework that would allow to combine the below listed technologies and develop plug-ins for widely used platforms (Flickr, Youtube, etc.) and systems (Drupal, etc.).

Social Dimension. For the identification as well as for the authentication several technologies are available. A user can for example provide her OpenID[2] to identify against a service. Further, OAuth[3] can be used for publishing and interacting with protected data. Big players such as Google are already offering support for the above mentioned technologies[4]. It seems advisable to build on this and contemplate what might be missing to align it with the Web of Data, being based on RDF [107].

While FOAF-based white listing and other related approaches [66, 76] are available, there is still a need for an up-front agreed way to deploy it in widely used systems. The same issue can be observed with privacy: there are proposals on the table (for example P3P[5]) but no measurable uptake in terms of users, systems that offer it, etc. can be stated.

Interaction Dimension. Especially for data provenance it seems to us that named graphs [32] offer a solid and scalable solution. With the rise of RDFa[6] one can think of new provenance mechanisms, as the hosting document can actually be understood as the "name" of the graph. Just imagine Flickr (already offering licensing information in RDFa) to include provenance information on both the content and the metadata, based on vocabularies such as the "Semantic Web Publishing Vocabulary" [20, Chapter 6]. Finally, we note that regarding the discovery and usage of linked (multimedia) data issues of provenance and trust have been considered in the work on VoiD, the "Vocabulary of Interlinked Datasets" [3] which followed after this position statement.

[2] http://openid.net/

[3] http://oauth.net/

[4] http://googledataapis.blogspot.com/2008/06/oauth-for-google-data-apis.html

[5] http://www.w3.org/P3P/

[6] http://www.w3.org/TR/xhtml-rdfa-primer/

Content dimension. In our understanding, the content dimension of the PTP model requires the most attention. Basic mechanisms proposed to represent the type of multimedia content in RDF[7] are available. Still, practical ways for creating and consuming rich multimedia content descriptions are missing. Recently, we have proposed ramm.x [79] which allows to use existing multimedia metadata formats such as MPEG-7, Exif, ID3, etc. in the Web of Data. However, we expect a fair amount of further research being required to address provenance and trust issues properly and make tools and applications available in a real-world setup.

5.3.5 Further Activities

After the publication of our position statement [72] (which has been reprinted above) several research activities have taken place. Even though progress has been made there are still open issues related to privacy, trust and provenance of Linked Data. Regarding provenance for instance the vocabulary of interlinked datasets (VoID) [2] allows to express information about the original datasource and it is considered best practice for all Linked Data sets to include such information. However, even if the provenance information is provided by the dataset provider there is still a lack of verification of this information. Different content-, context-, and rating-based techniques can also be used to heuristically assess the relevance, quality and trustworthiness of data as discussed in [21]. Regarding privacy the awareness of users about where to provide which data also plays a crucial role apart from technical or legal aspects.

5.4 Summary

The research question of *Which data is linkable?* has been covered in this chapter where general considerations on creating linkable data have been presented. We found that in principle every sort of source data can become Linked Data and the biggest benefits of Linked Data can be gained with data that is potentially reused. We also discussed general approaches of generating Linked Data. Regarding the overall strategy for generating linkable and Linked Data we stated the two possibilities of entirely manually creating the RDF or using some tool to support in the creation. For tool-supported creation we made a distinction between two different major kinds of source data, namely structured and unstructured data that will be addressed in subsequent chapters. Further considerations that might need to be taken into account such as the

[7] `http://www.w3.org/TR/Content-in-RDF/`

issue of trusting interlinked multimedia data have also been discussed and the abstract “Provenance-Trust-Privacy” (PTP) model has been presented.

Chapter 6

Structured Datasources

The creation of linkable data out of structured datasources involves an enrichment of the underlying relational data so that it can be accessed as RDF via URIs to adhere to the Linked Data principles. Large amounts of structured data are residing in different databases that are used in an almost infinite number of different use cases. In this chapter we look at some specific examples in detail that are aligned with our guiding use case riese which is targeting statistical data. In section 6.1 our approach for generating linkable data out of EuroStat statistical data is presented. Vocabularies are also of importance as already explained. Our contribution to the modeling of statistical data is presented with SCOVO in section 6.2. The further developments beyond SCOVO are highlighted in section 6.3 and finally in section 6.4 a short discussion of the general aspects of generating RDF from relational databases in the context of the RDB2RDF activity is presented.

6.1 riese

This section describes the parts of the *RDFizing and Interlinking the EuroStat Data Set Effort* (riese) related to the generation of linkable data. According to our framework the parts related to interrelation and consumption of Linked Data are discussed in the respective following parts of this thesis. This section contains adapted parts of our paper "Building Linked Data For Both Humans and Machines" [73] that has been written together with Yves Raimond and Michael Hausenblas. The paper has been presented by the author of this thesis at the Linked Data on the Web (LDOW) 2008 workshop in Beijing, China. In this paper we described our experience with building the riese dataset, an interlinked, RDF-based version of the Eurostat data, containing statistical data about the European Union. The riese dataset aims at serving roughly 3 billion RDF triples, along with millions of high-quality interlinks. Our contribution was twofold: Firstly, we suggest using RDFa as the main deployment mechanism, hence serving both humans and machines to effectively and efficiently explore and use the dataset. Secondly, we introduce a new way of enriching the dataset with high-quality links: the User Contributed Interlinking, a Wiki-style way of adding semantic links to data pages.

In this section only the parts relevant to the generation of linkable data are discussed. The actual interlinking is discussed in section 9.1 for the User Contributed Interlinking and in section 10.2 for the automated interlinking in riese.

6.1.1 Motivation

The goal of the “RDFizing and Interlinking the Eurostat Data Set Effort” (riese) is to offer a Semantic Web version of the publicly accessible data provided by the Eurostat data source. riese has been initiated as part of the W3C SWEO Linking Open Data (LOD) project and aims at being useful for both humans and machines.

Prior existing linked datasets such as DBpedia [8] are slanted towards machines as the consumer. Although there are exceptions to this machine-first approach (cf. [87]), we strongly believe that satisfying both humans and machines from a single source is a necessary path to follow.

We subscribe to the view that every LOD dataset can be understood as a Semantic Web application. Every Semantic Web application in turn is a Web application in the sense that it should support a certain task for a human user. Without offering a state-of-the-art Web user interface, potential end-users are scared away. Hence a Semantic Web application needs to have a nice outfit, as well.

6.1.2 Related Work

Statistical data on the (Semantic) Web Looking at related work reveals that there is actual demand for new solutions to disseminate statistical data using semantic technologies. As reported by Assini [7] the European Union funded a research and development project called NESSTAR in 1998, with the aim of bringing the advantages of the Web to the world of statistical data dissemination. Another project that is entirely situated on the Semantic Web is the U.S. Census data [159] where 1 billion

RDF triples containing statistical information about the United States were published in 2007.

An earlier attempt to publish Eurostat is known from the FU Berlin[1], using a very small subset of country and region statistics. Stuckenschmidt [157] has reported on translating and modeling the European fishery statistics in ontologies. Grossenbacher [68] recently pointed out issues with translating the Swiss statistics to an RDF basis. A somehow related approach is the Rswub[2], a package for handling statistical data, based on RDF and capable of handling ontologies.

RDFa As RDFa [1] was turning into a W3C Last Call document at the time of creating riese, we expected the penetration to dramatically increase in the following couple of months which actually also occurred. Although not yet a standard by that time, there existed a number of smaller-sized deployed datasets. We had shortly before also proposed to use RDFa as a base for multimedia metadata deployment [80]. However, to the best of our knowledge there existed no other linked-data set deployed in RDFa by that time.

6.1.3 Requirements and Issues

The Eurostat data

This section provides a short description of the Eurostat data, which served as the primary input for the riese dataset.

Eurostat provides detailed statistics for the entire European Union as well as additional statistics for major non-European countries. The Eurostat data is arranged along the following themes:

- General and regional statistics
- Economy and finance
- Population and social conditions
- Industry, trade and services
- Agriculture and fisheries

[1] http://www4.wiwiss.fu-berlin.de/eurostat/

[2] http://www.biostat.harvard.edu/~carey/hbsfin.html

- External trade
- Transport
- Environment and energy
- Science and technology

Three main data sources are being provided by Eurostat for public download, namely (i) the statistical data itself, (ii) a table of content, and (iii) dictionaries.

The statistical data is provided as dump download of approximately 4,000 single tab-separated values (TSV) documents, having a total size of approximately 5GByte, and containing some 350 million data values. This data is updated twice a day. Only limited semantic exploitable information is contained in these TSV documents, hence it is inevitable to use other available information sources.

A table of content (TOC) provides a hierarchical overview of the datasets—organized in so called themes—allowing to identify the structure and content of a dataset.

Dictionaries are especially valuable as they contain all information for resolving the nearly 100,000 data codes used in the statistical data. These data codes refer to dimensions such as time, location, currency, etc. The data codes also contain an implicit hierarchy, which can be used for further classification. However, various schemas have been used requiring individual processing for extracting classifying features. For example, in order to refer to locations, the Nomenclature of Territorial Units for Statistics (NUTS) is in use. This basically allows to extract information about the structure of administrative divisions of countries. For each of the dictionaries a different terminology is used.

Most of the data is represented in time series with varying granularity, ranging from annual to daily data. Each single data item can be identified using the corresponding dataset and various dimensions as the following example illustrates: The population of the European Union can be seen as one single data item valued at 497,198,740 (contained in the dataset 'Total population'), having as time-dimension the year 2008, as indicator-dimension 'Population on 1. January', and as geo-dimension the 'European Union (27 countries)'. Additionally the data is flagged as provisional and Eurostat estimate.

Requirements

In a first phase we have analyzed the Eurostat data. We have identified the implicit semantics present in the TOC and the dictionaries and gathered a number of issues. Firstly, the Eurostat data set is highly heterogeneous; the data sources formats vary (TSV, HTML) and are not machine-processable per se. Another issue is the modeling of temporal data, more specific how to represent time intervals. Further, the schemas in the dictionaries form a multidimensional space that somehow has to be linearized in order to be represented in a URI format. We have also identified data provenance (and trust) issues, which are currently only handled on a global level.

Based on the analyses given above we state the following requirements for a linked dataset that is designed to serve both humans and machines:

- The system must serve both humans and machines in an adequate way by applying the don't-repeat-yourself (DRY) [91] principle;
- To allow both humans and machines to reveal more information, the follow-your-nose principle [158] must be applied.
- To be a useful (real-world) Semantic Web application, the system must be able to scale to the size of the Web;

Additionally we want to point out that we aim at providing high-quality interlinking. Hence, the sheer template-driven generation of global interlinks is certainly not sufficient.

6.1.4 Linked Data For Humans and Machines

In order to demonstrate how to address the issues raised earlier, we have implemented the riese dataset as a Semantic Web application. This section describes how the mapping—from the available, relational data into RDF form—has been done and introduces the riese system architecture.

Data, Schemas and Mapping

This section explains the schemas utilized in riese and discusses the mapping to RDF.

The data used in riese is a snapshot of the data available for bulk-download taken on 9 Jan 2008. Depending on the type of data, three formats are used by Eurostat:

HTML or plain text for the TOC, and TSV for the dictionary files and the actual data tables.

In figure 6.1 the riese core schema is depicted. Currently the riese core schema is modeled using RDF-Schema [27] rather than OWL [123] based and comprises three main classes: `riese:Dataset`, `riese:Item` and `riese:Dimension`. A dataset is the logical container of either more sub-datasets (related via `skos:narrower`) or data items. An item represents one single data value (like 497,198,740 for the population of the European Union) with all accompanying metadata about the containing dataset and the dimensions used. A dimension semantically describes the value of a data item in terms of, e.g. time, location, unit, etc. In listing 6.1 an exemplary snippet of an item is shown.

```
data:eb040_infl_2006_at a :Item ;
 dc:title ”Inflation rate Austria 2006” ;
 rdf:value ”1.7” ;
 :dimension dim:geo_at ;
 :dimension dim:time_2006 ;
 :dataset data:eb040 .
```

Listing 6.1: A single data item.

Additionally, the following schemas are used or have been extended:

- Dublin Core (DC) Elements [47] and Terms [46]
- GeoNames [64]
- Simple Knowledge Organisation Systems (SKOS) [151]
- Description of a Project (DOAP) [48]
- the event ontology [138]

We decided to model a flat schema for the following reasons:

- Queries can be constructed with very little a-priori knowledge about the structure of the dataset;
- Additional Eurostat datasets can easily be added without changing the schema (and are instantaneously integrated in the hierarchy, hence available to all users regardless of the access method);

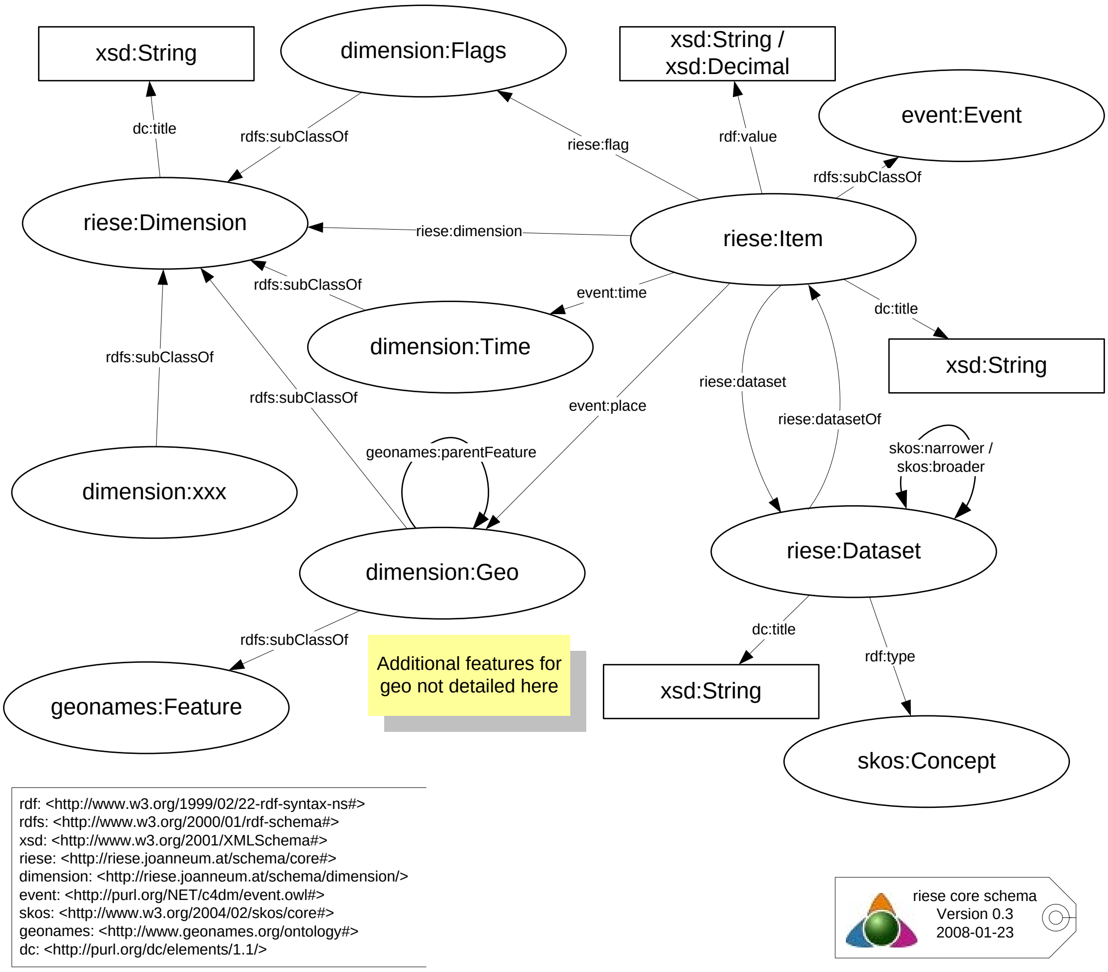

xsd:String
dimension:Flags
xsd:String / xsd:Decimal
event:Event
dc:title
rdfs:subClassOf
riese:flag
rdf:value
rdfs:subClassOf
riese:Dimension
riese:dimension
riese:Item
rdfs:subClassOf
event:time
dc:title
dimension:Time
xsd:String
rdfs:subClassOf
riese:dataset
rdfs:subClassOf
event:place
riese:datasetOf
geonames:parentFeature
skos:narrower / skos:broader
dimension:xxx
riese:Dataset
dimension:Geo
dc:title
rdfs:subClassOf
rdf:type
Additional features for geo not detailed here
xsd:String
geonames:Feature
skos:Concept
rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
rdfs: <http://www.w3.org/2000/01/rdf-schema#>
xsd: <http://www.w3.org/2001/XMLSchema#>
riese: <http://riese.joanneum.at/schema/core#>
dimension: <http://riese.joanneum.at/schema/dimension/>
event: <http://purl.org/NET/c4dm/event.owl#>
skos: <http://www.w3.org/2004/02/skos/core#>
geonames: <http://www.geonames.org/ontology#>
dc: <http://purl.org/dc/elements/1.1/>
riese core schema
Version 0.3
2008-01-23

Figure 6.1: The core schema.

- Dimensions can be added without any changes to the schema;

- Finally, it is possible to formulate very flexible queries.

Other approaches, such as the U.S. Census data [159] use a more complex schema, where for example a new property for every possible description is introduced. This yields properties such as `population15YearsAndOverWithIncomeIn1999`, which do not offer any additional semantic information.

Querying data using these properties can get very cumbersome, as the user would have to know about the exact terms beforehand. We believe that our flat approach, where every value can be identified by the corresponding dataset and dimensions, enables fairly flexible queries.

The example in listing 6.2 demonstrates this. All items for Austria are returned

```
SELECT *
WHERE
{ ?item riese:dimension dim:geo_at.
  ?item riese:dataset ?dataset.
  ?dataset dc:title ?ds_title
  FILTER regex(?ds_title, ”food”,i)}
```

Listing 6.2: A query in riese.

that belong to a dataset with ’food’ in the description (with default namespace `http://riese.joanneum.at/schema/core#`).

System Architecture

Based on the lessons learned from [86] we have developed the riese Web application. It comprises:

1. An (offline) module, being responsible for converting the Eurostat data into an RDF representation and creating the global, pattern-based interlinks (RDFizing & Interlinking; cf. section 10.2 for details), and a
2. Web server including a scripting environment that fills predefined templates with the values from the (static) RDF/XML representation in order to generate an RDFa representation of the themes and the data tables.

The figure 6.2 depicts the riese system architecture and shows as well the interfaces with the environment (in and out ports).

The riese Web application supports the following tasks:

- Human users: Users can navigate the dataset provided in XHTML+RDFa;
- Semantic Web agents:
  - single item query—XHTML+RDFa per page allows the exploration of the dataset and the query of a single data item (FYN);
  - global query—To allow an efficient query of the entire dataset, a SPARQL-endpoint is provided;

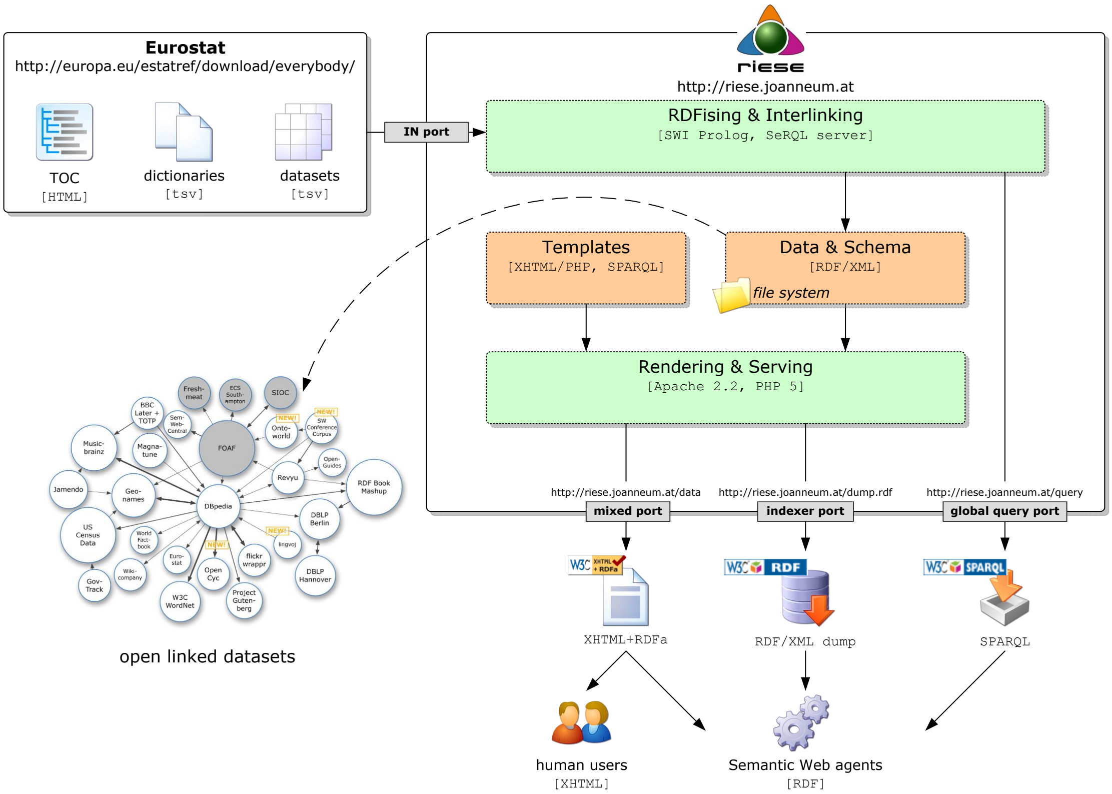


Figure 6.2: The system architecture of riese.

- indexer: to allow semantic search engines (indexer) an effective processing, the entire dataset is offered as a dump and an according description using the semantic crawler sitemap extension protocol[3] is offered.

For creating the RDF representation from the original Eurostat files, SWI-Prolog scripts are used. The SWI-Prolog Semantic Web Library provides an infrastructure for reading, querying and storing Semantic Web documents. Additionally the Prolog-2-RDF (p2r) modules[4] and individually defined mappings are used for translating the input data to RDF.

The resulting RDF can be accessed via a SPARQL endpoint and it is further possible to consume a dump of the entire data. We have created one large dump containing all triples, and also store the triples according to their URI directly into the file system in RDF/XML.

The latter approach is currently used for 'Rendering & Serving' where the PHP scripts looks up the files in the file system and renders a RDFa representation. Beside

[3] http://sw.deri.org/2007/07/sitemapextension/
[4] http://moustaki.org/p2r/

the data that originates from Eurostat (the official statistical data), the UCI module stores the user-contributed triples in a separate document (cf. section 9.1). This physical separation is mainly due to being able to replace parts of the data without too much additional effort.

6.1.5 Conclusion

In the paper that has been partly reprinted above we have presented the riese dataset containing statistical data from Eurostat. We have shown how to RDFize and prepare this data for interlinking, hence making it possible to expose it onto the Semantic Web.

We have also identified some issues and bottlenecks when deploying datasets of such enormous size. Generating a static file-structure with small RDF files requires quite a lot of time. This is due to our current way of storing the data items in the file system. Because in riese several hundred millions of folders and files have to be created, the bottleneck is somehow obvious. Moreover, when accessing datasets (tables) containing thousands of items (cells) in individual files this yields thousands of file access operations for simply parsing them. Regarding the file system we came across another limitation: reserved names on the MS Windows operating systems (as it turned out, it is not possible to create files or folders named 'con', 'aux', etc. [124]). In the following version of riese we have replaced this extract-transform-load (ETL) approach by a dynamic view which allows to access all resources without having to pre-compute each individual file. This means that instead of providing the RDF files via the filesystem the RDF view is generated automatically upon request. As the SPARQL query has already been possible via the SWI Prolog mapping in the first version we are able to provide the same functionality also with this dynamic view approach and therefore in our case we were able to achieve an overall improvement.

When modeling the representation of time related to a certain statistical information we encountered some challenges as the raw data from Eurostat is sometimes ambiguous and can only be resolved by analyzing the corresponding document. For example the statement `time\2007` can stand for the value over a period of time (e.g. entire year) or at the end of the reporting period (e.g. 31 Dec).

In the following section 6.2 we present the further development of the riese core schema into the statistical core vocabulary (SCOVO). The issues of interlinking in the riese use case are discussed in the sections 9.1 and 10.2. Further aspects regarding the consumption of Linked Data in the light of riese are finally discussed in section 12.2.

6.2 SCOVO - the Statistical Core Vocabulary[5]

The riese core schema presented in the preceding section has been further developed to be generically usable for expressing statistical information. SCOVO has been described in the paper "SCOVO: Using Statistics on the Web of Data". The paper has been written together with Michael Hausenblas, Yves Raimond, Lee Feigenbaum, and Danny Ayers. The author of this thesis has designed the original riese core schema and has also led the development of SCOVO. The paper has been presented by the author of this thesis together with Michael Hausenblas at the 6th European Semantic Web Conference (ESWC 2009) in Heraklion, Greece, in 2009. It has been published in Springer's Lecture Notes in Computer Science and this section is an adapted reprint of the original paper with kind permission from Springer Science and Business Media.

6.2.1 Motivation

Statistical data is present everywhere—from governmental bodies to economics, from life-science to industry. With the rise of the Web of Data, the need for sharing, accessing, and using this data has entered a new stage. Available technologies are either not compatible with the Semantic Web or are too complex to be useful for a range of use cases. We have identified the need for a more general and flexible solution to the problem of modeling and publishing statistics on the Web.

Our motivation stems from ongoing work in three distinct efforts, namely riese ("RDFizing and Interlinking the EuroStat Data Set Effort") [73, 83], U.S. Census Bureau's annual Statistical Abstract of the United States, and making the UN data accessible on the Web of Data. In riese, we provide statistical data about European citizens. One of our main use cases of riese was in the context of an advertising analysis application [155] allowing for example a market researcher to better and faster understand a certain market or product. Further, in the U.S. Census Bureau's annual Statistical Abstract of the United States case, one of the authors and representatives of the U.S. Environmental Protection Agency were exploiting the semantics implicit in spreadsheets published yearly since 1878 by the U.S. Census Bureau. This data corpus is a comprehensive

[5] This section is an adapted reprint of "M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayers. *SCOVO: Using Statistics on the Web of Data.* In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, editors, *The Semantic Web: Research and Applications*, volume 5554 of *Lecture Notes in Computer Science*, pages 708–722." © Springer-Verlag Berlin Heidelberg 2009, with kind permission from Springer Science and Business Media under license number 2891941101364.

collection of social, political, and economic statistics compiled from information from over 250 agencies. The goal here was to enable a fast and efficient publishing of the statistics on the Web. Currently, this means making MS Excel documents and PDF documents available for download from the Web.

The following subsections are structured as follows: Firstly, we review related efforts in subsection 6.2.2. Then, in subsection 6.2.3 we discuss issues with representing statistical data and derive requirements for a generic modeling. The core of the work is presented in subsection 6.2.4—where we propose a modeling framework for statistical data—and subsection 6.2.5 in which two reference implementations are discussed.

6.2.2 Related Work

Representing statistical data has a long tradition, hence a plethora of proposals and solutions exists [5]—mainly driven by governmental and international institutions dealing with high volumes of data. In the 1990's the U.S. Bureau of the Census has developed a statistical metadata content standard, which allows to describe all aspects of survey design, processing, analysis, and data sets [114]. Later, the "United Nations Economic Commission for Europe" (UNECE) has developed guidelines [164] covering search, navigation, interpretation, and post-processing of statistical data in their realm. A standard proposed by the International Organization for Standardization (ISO) is the "Statistical data and metadata exchange" (SDMX) [96]. More recently, the OECD has released a report on the management of statistical metadata at the OECD [131]. Another related effort is the Data Documentation Initiative (DDI)[6], which aims at establishing an XML-based standard for the content, presentation, transport, and preservation of documentation for datasets in the social and behavioral sciences. As we have already pointed out in [73], there are known attempts concerning the modeling and use of statistical data on the Web of Data [7, 68, 157, 159]. However, unlike earlier attempts such as [122, 136], we aim at a light-weight solution enabling a quick uptake and wide deployment.

The **Web of Data** is understood as the part of the Web where the Linked Data principles are applied. The basic idea of Linked Data has been outlined already in chapter 2. The LOD community project is an open, collaborative effort applying the linked data principles. It aims at bootstrapping the Web of Data by publishing datasets in RDF on the Web and creating large numbers of links between these datasets [23].

We finally highlight the modeling issue with n-ary relations on the Web of Data.

[6] http://www.ddialliance.org/

The data model of the Web of Data is RDF, hence modeling n-ary relations is a nontrivial task. In 2006 the W3C Semantic Web Best Practices and Deployment Working Group has published a note dealing with this issue [129]; we will use this as a base for our framework.

6.2.3 Requirements and Issues

Issues with Modeling Statistical Data

Independent of the original format of the data (such as a table in an Excel sheet, etc.), the issues discussed in the following need to be addressed properly to ensure a lossless representation of the statistical data.

Handling of Multiple Dimensions It is very often the case that a data item has several dimensions. Roughly, two types of dimensions can be identified, (i) generic dimensions, such as location and time, and (ii) domain specific dimensions. For example we may be interested in the reliability of flights (domain specific) from Cambridge, MA, US to London, UK (both are locations) between 2003 and 2007 (time period). It is crucial that a vocabulary aiming at representing statistics is capable of denoting such dimensions and allows to attach as many as needed to a single data item.

Reusability and Uptake Statistics are no ends in themselves; rather they are **about** something—be it money, flights, death rates or the consumption of YouTube videos. It is therefore essential being able to reuse existing information; both on the schema as on the instance level. Related to reusability is the issue of community uptake: Most statistical metadata formats are rather complex, yielding a small deployment.

Structural vs. Domain Semantics Two kinds of semantics come into mind when modeling statistics:

- structural semantics, stemming from how statistics are presented, such as grouping into time-periods, primary dimensions, etc.;
- domain semantics, stemming from the domain the statistic is about (money, airports, etc.).

Performance and Scalability Issues As discussed elsewhere [86] performance and scalability issues may arise from the way data is encoded and served.

Requirements

Based on the issues listed above we state the following requirements for our framework:

1. The framework *must* be directly usable on the Web of Data. This implies for example that a vocabulary used in the framework must be expressed in RDF. This requirement addresses the issue of structural and domain semantics, as well as ensuring reusability;
2. The framework *must* be extensible both on the schema level and the instance level, enabling reusability;
3. The framework *should* be light-weight, addressing uptake, and performance and scalability issues.

It has to be noted that the first requirement does not stem from the issues discussed earlier—it was rather introduced to benefit from the large deployment base of domain vocabularies as well as available tools and systems.

6.2.4 Statistical Modeling Framework

Driven by the requirements we propose a modeling and publishing framework for statistics on the Web of Data consisting of:

- a core vocabulary for representing statistical data
- a “workflow” to create the statistical data

The framework is depicted at a glance in Figure 6.3. The lower half is the generic part, defined in this work, the upper part is the application-specific part, depending on the technologies used to implement the framework.

Statistical Core Vocabulary (SCOVO)

One of the main contributions of our work at hand is the Statistical Core Vocabulary (SCOVO)[7]. SCOVO (depicted in figure 6.4) defines three basic concepts:

- a dataset, representing the container of some data, such as a table holding some data in its cells;

[7] http://purl.org/NET/scovo

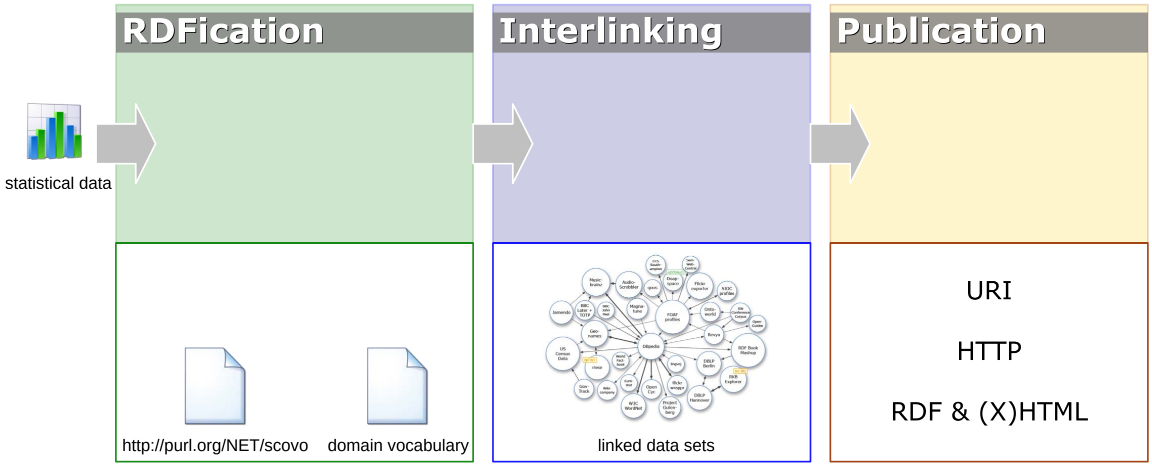


Figure 6.3: The Statistical Modeling Framework.

- a data item, representing a single piece of data (e.g. a cell in a table);
- a dimension, representing some kind of unit of a single piece of data (for example a time period, location, etc.)

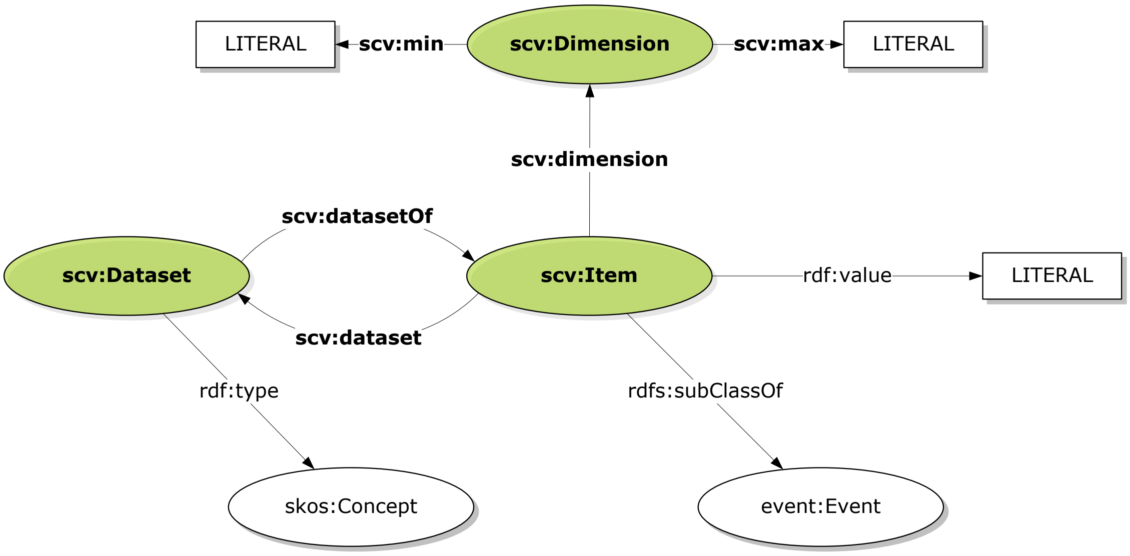


Figure 6.4: The Statistical Core Vocabulary (SCOVO).

A statistical dataset in SCOVO is represented by the class `Dataset`; it is a SKOS concept [151] in order to allow hooking into a categorization scheme. A statistical data item `Item` belongs to a dataset (cf. inverse properties `dataset` and `datasetOf`). An `Item` is subsuming the `Event` concept, as defined in the Event ontology[8]. The Event

[8] `http://purl.org/NET/c4dm/event.owl`

ontology essentially adopts the view from Allen and Fergusson [4]:

> [..] events are primarily linguistic or cognitive in nature. That is, the world does not really contain events. Rather, events are the way by which agents classify certain useful and relevant patterns of change.

An event is then defined in this ontology as the way by which cognitive agents classify arbitrary time/space regions. Our `Item` concept is subsuming this `Event` concept—a statistical item is a particular classification of a time/space region. Dimensions of a statistical item are factors of the corresponding events, attached through the `dimension` property, pointing to an instance of the SCOVO `Dimension` class.

This model is easily extensible by defining new factors and agents pertaining to the actual statistical data. For example, we can relate to a statistical data item the institutional body responsible of it as well as the methodology used. A `Dimension` can have a minimum (and respectively a maximum) range value, captured through the `min` and `max` properties.

The Statistical Core Vocabulary is currently defined in RDF-Schema. It is possible to express SCOVO in OWL-DL, if advanced reasoning is of necessity. Although we have depicted the range of both `:min` and `:max` in figure 6.4 being of literal value, we emphasize that in the RDF-Schema the ranges have not been specified in order to allow an extension for domain-specific purposes. Hence, this can be seen as a kind of recommendation for the default case.

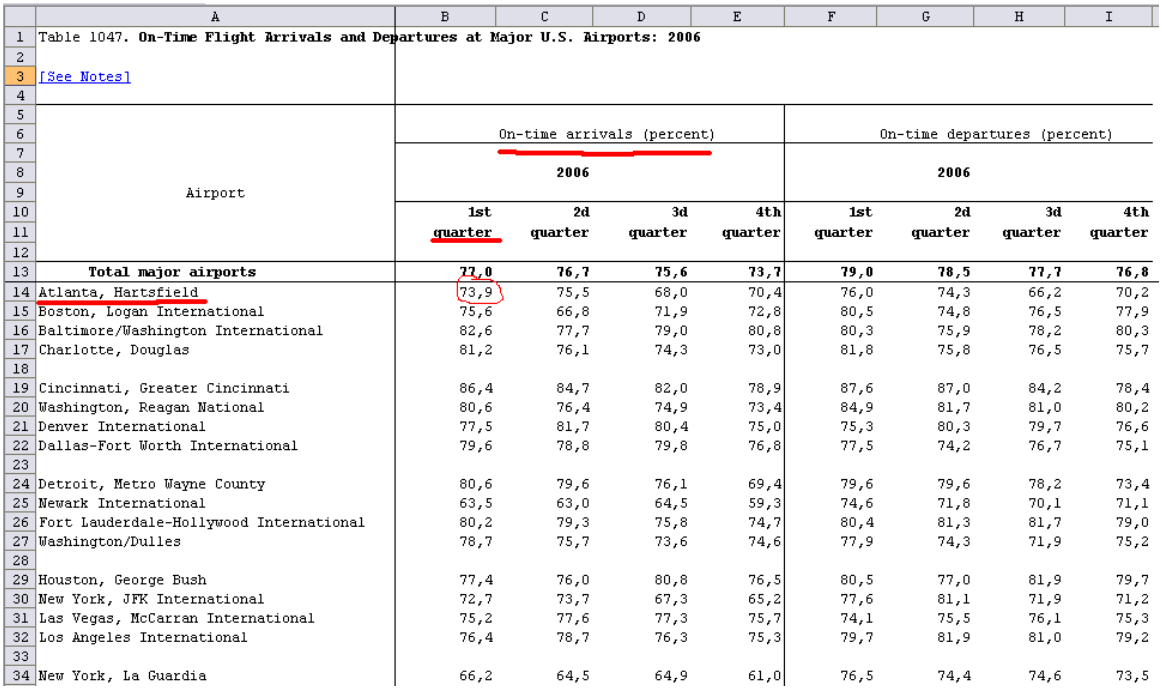
Example To demonstrate the usage of SCOVO, let us assume we want to model airline on-time arrivals and departures. The input in our example is the "Table 1047. On-Time Flight Arrivals and Departures at Major U.S. Airports: 2006"[9] (cf. figure 6.5) from the US Census data set. Every airport, for each time period has an on-time arrival percentage and an on-time departure percentage.

In listing 6.3 an excerpt[10] of the modeling of the airline on-time arrivals and departures is shown[11]. We note that the example has `http://example.org/on-time-flight#` as its base along with the prefix `ex:` (line 1). Lines 4 to 16 define domain-specific entities, such as a `ex:TimePeriod`, an `ex:Airport`, etc. From line 18 to 22 the one and only dataset (`ex:ontime-flights`) is defined, corresponding to the entire Excel table

[9] `http://www.census.gov/compendia/statab/tables/08s1047.xls`

[10] The full example in RDF/XML is available at `http://sw.joanneum.at/scovo/otf-example-full.rdf`

[11] Note, that in all of our examples the well-known prefixes such as rdf:, rdfs:, etc. have been omitted due to readability reasons.



	A	B	C	D	E	F	G	H	I
1	Table 1047. **On-Time Flight Arrivals and Departures at Major U.S. Airports: 2006**								
2									
3	[See Notes]								
4									
5									
6		On-time arrivals (percent)				On-time departures (percent)			
7									
8		**2006**				**2006**			
9	Airport								
10		**1st**	**2d**	**3d**	**4th**	**1st**	**2d**	**3d**	**4th**
11		**quarter**	**quarter**	**quarter**	**quarter**	**quarter**	**quarter**	**quarter**	**quarter**
12									
13	**Total major airports**	**77,0**	**76,7**	**75,6**	**73,7**	**79,0**	**78,5**	**77,7**	**76,8**
14	Atlanta, Hartsfield	73,9	75,5	68,0	70,4	76,0	74,3	66,2	70,2
15	Boston, Logan International	75,6	66,8	71,9	72,8	80,5	74,8	76,5	77,9
16	Baltimore/Washington International	82,6	77,7	79,0	80,8	80,3	75,9	78,2	80,3
17	Charlotte, Douglas	81,2	76,1	74,3	73,0	81,8	75,8	76,5	75,7
18									
19	Cincinnati, Greater Cincinnati	86,4	84,7	82,0	78,9	87,6	87,0	84,2	78,4
20	Washington, Reagan National	80,6	76,4	74,9	73,4	84,9	81,7	81,0	80,2
21	Denver International	77,5	81,7	80,4	75,0	75,3	80,3	79,7	76,6
22	Dallas-Fort Worth International	79,6	78,8	79,8	76,8	77,5	74,2	76,7	75,1
23									
24	Detroit, Metro Wayne County	80,6	79,6	76,1	69,4	79,6	79,6	78,2	73,4
25	Newark International	63,5	63,0	64,5	59,3	74,6	71,8	70,1	71,1
26	Fort Lauderdale-Hollywood International	80,2	79,3	75,8	74,7	80,4	81,3	81,7	79,0
27	Washington/Dulles	78,7	75,7	73,6	74,6	77,9	74,3	71,9	75,2
28									
29	Houston, George Bush	77,4	76,0	80,8	76,5	80,5	77,0	81,9	79,7
30	New York, JFK International	72,7	73,7	67,3	65,2	77,6	81,1	71,9	71,2
31	Las Vegas, McCarran International	75,2	77,6	77,3	75,7	74,1	75,5	76,1	75,3
32	Los Angeles International	76,4	78,7	76,3	75,3	79,7	81,9	81,0	79,2
33									
34	New York, La Guardia	66,2	64,5	64,9	61,0	76,5	74,4	74,6	73,5

Figure 6.5: On-Time Flight Arrivals and Departures at Major U.S. Airports: 2006.

used as an input. Further, from line 24 on an exemplary data item is defined, stating that the on-time arrival of the "Atlanta, Hartsfield" airport in the first quarter of 2006 was round 74%. This corresponds to the highlighted cell in figure 6.5.

The SPARQL query from listing 6.4 can be used to explore high-performing airports. With "high-performing" we define in this context airports with on-time arrivals or departures higher than 85%. Lines 2 to 6 provide the generic pattern for an `Item`. Note how in line 7 and 8 the dimensions are constrained. Line 9 of listing 6.4 basically expresses "give me all kinds of on-timeness", and finally line 10 implements the "high-performance" filter criteria.

The query result—depicted in figure 6.6—shows the list of high-performing airports along with the time period, starting with the best airport in terms of "on-timeness". We note that the complete example, including the exemplary queries in an executable form, is available at `http://purl.org/NET/scovo`.

Workflow—Good Practice Rules

In this section we discuss the overall workflow as shown in figure 6.3. Based on our findings from publishing real-world statistical datasets, the following should be seen as strong advises, helping to avoid failings and to enable a quick adoption.

```
@prefix ex: <http://example.org/on-time-flight#> .
@prefix scv: <http://purl.org/NET/scovo#> .

ex:TimePeriod rdfs:subClassOf scv:Dimension; dc:title ”time period” .

ex:Q12006 rdf:type ex:TimePeriod; dc:title ”2006 Q1”;
          scv:min ”2006-01-01”^^xsd:date ;
          scv:max ”2006-03-31”^^xsd:date .

ex:OnTime rdfs:subClassOf scv:Dimension; dc:title ”on-time ...” .

ex:ota rdf:type ex:OnTime; dc:title ”on-time arrivals” .

ex:Airport rdfs:subClassOf scv:Dimension; dc:title ”airport” .

ex:AtlantaHartsfield rdf:type ex:Airport; dc:title ”Hartsfield-Jackson Atlanta International Airport” .

ex:ontime-flights rdf:type scv:Dataset ;
                  dc:title ”On-time Flight Arrivals ...” ;
                  scv:datasetOf ex:atl-arr-2006q1 .

ex:AtlantaHartsfield-ota-2006-q1 rdf:type scv:Item ;
                                 rdf:value 74 ;
                                 scv:dataset ex:ontime-flights ;
                                 scv:dimension ex:Q12006 ;
                                 scv:dimension ex:ota ;
                                 scv:dimension ex:AtlantaHartsfield .
```

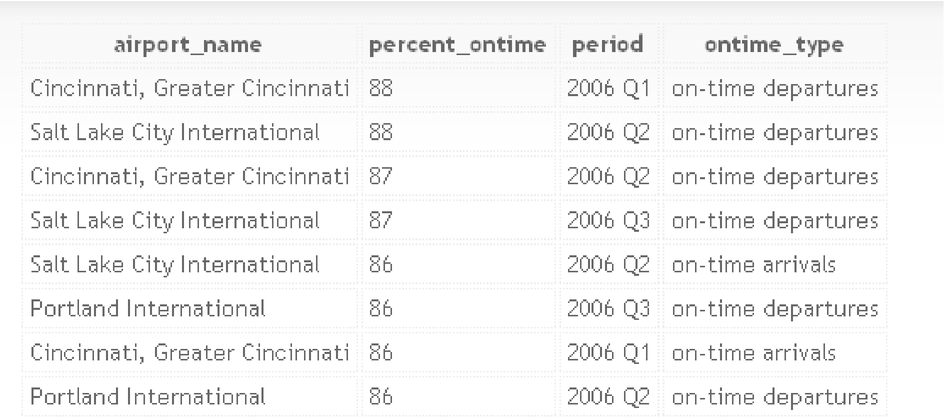
Listing 6.3: Modeling flight on-time arrival statistics.

RDFication In the very first step, the data needs to be converted into an RDF-based form. This is equally true for the schema level as for the instance level. The schema level (e.g. XSD, etc.) is a typical starting point which is followed by the conversion of the actual data in a second step. While creating and populating the ontology with instances several issues arise.

URI Design It has to be ensured that every entity has a URI assigned, which is usually referred to as “URI minting”; cf. [26] for a more detailed discussion on URI design. For example `http://dbpedia.org/resource/Airport` has been minted by DBpedia to represent the concept of an airport.

```
SELECT ?airport_name ?percent_ontime ?period ?ontime_type WHERE {
  ?item rdf:type scv:Item ;
        scv:dimension ?airport ;
        scv:dimension ?time_period ;
        scv:dimension ?ontime ;
        rdf:value ?percent_ontime .
  ?airport rdf:type ex:Airport; dc:title ?airport_name .
  ?time_period rdf:type ex:TimePeriod; dc:title ?period .
  ?ontime rdf:type ex:OnTime; dc:title ?ontime_type .
  FILTER (?percent_ontime > 85)
} ORDER BY DESC (?percent_ontime)
```

Listing 6.4: SPARQL query for high-performing airports.



airport_name	percent_ontime	period	ontime_type
Cincinnati, Greater Cincinnati	88	2006 Q1	on-time departures
Salt Lake City International	88	2006 Q2	on-time departures
Cincinnati, Greater Cincinnati	87	2006 Q2	on-time departures
Salt Lake City International	87	2006 Q3	on-time departures
Salt Lake City International	86	2006 Q2	on-time arrivals
Portland International	86	2006 Q3	on-time departures
Cincinnati, Greater Cincinnati	86	2006 Q1	on-time arrivals
Portland International	86	2006 Q2	on-time departures

Figure 6.6: Results for high-performing airports.

More specifically we recommend using HTTP URIs in order to be compliant with the linked data principles: When dereferencing the aforementioned URI for "Airport" it yields the result presented in listing 6.5, basically telling us that the "concept URI" redirects to an information resource at `http://dbpedia.org/page/Airport`, see also [147].

Domain Ontologies As already mentioned, statistics are always about a certain domain. In order to use domain vocabularies together with SCOVO, several "hooks" can be used:

- Subclassing the SCOVO-`Dimension` class. In most cases it is sufficient to use this technique to incorporate domain-specific concepts, for instance as from the example in listing 6.3:

```
ex:Airport rdfs:subClassOf scv:Dimension
```

```
curl -I http://dbpedia.org/resource/Airport
HTTP/1.1 303 See Other
Server: Virtuoso/05.00.3028 (Solaris) x86_64-sun-solaris2.10-64
   VDB
Content-Type: text/html; charset=UTF-8
Date: Thu, 08 May 2008 10:32:29 GMT
Location: http://dbpedia.org/page/Airport
```

Listing 6.5: Dereferencing URI for “Airport”.

```
ex:AtlantaHartsfield rdf:type ex:Airport
```

- Use the built-in support for `event:Event` and `skos:Concept`. The latter is of particular help if an existing taxonomy or thesaurus is used as a base. The earlier can be used to capture more information pertaining to the creation of a particular statistical item;
- Defining sub-properties of using SCOVO-`min` and `max`. Whenever the need arises to more explicitly declare what kind of range is intended, this technique can be used (e.g. an `xsd:date`).

Interlinking Classes and instances of the domain vocabulary *should* be interlinked to existing LOD entities. The rationale behind is that any dataset can be enriched through this at low costs. For example, to connect the airports to the LOD datasets, one could use the query from listing 6.6 to find according targets in DBpedia (note that this query can be executed at `http://dbpedia.org/snorql/`).

```
SELECT ?airports_state ?airport
WHERE {
 ?airports_state skos:broader
 <http://dbpedia.org/resource/Category:
    Airports_in_the_United_States> .
 ?airport skos:subject ?airports_state ;
         <http://dbpedia.org/property/name> ?name .
 FILTER regex(?name, ”Atlanta”, ”i”)
}
```

Listing 6.6: Interlinking airports to DBpedia.

The result of the query from listing 6.6 may subsequently be used to enrich our example, that is adding for example the triple

```
ex:AtlantaHartsfield owl:sameAs
  <http://dbpedia.org/resource/Hartsfield-Jackson_Atlanta_[...]>
```

in order to express that the two URIs are actually identifying the same thing. With this interlinking we have significantly broadened the possibilities for querying our dataset; for example we could issue a location-based query with the geo-data from DBpedia or could further follow down the path to other LOD datasets containing even more information related to the Hartsfield airport.

Publication When publishing the dataset, one needs to make choices on the formats to be used for the data. While certain circumstances may require the usage of specialized and/or proprietary formats such as PDF or the SPSS file format, there are four basic technologies that we (unsurprisingly) see central to our setup: URIs, HTTP, RDF and (X)HTML; every publishing system on the Web of Data *should* use these as primary technologies. We have discussed URI minting above. Regarding HTTP—beside its basic transport function—we encourage people to use light-weight REST interfaces. One particular issue, however, is how to deploy the metadata. Several options exist, we list some widely used in the following:

- use an RDF standalone format such as RDF/XML, N3, etc. along with 303 redirects or links such as described in [147];
- use XHTML+RDFa[12] for both humans and machines (see also [73]);
- SPARQL-endpoints and RDF dumps [137, 86].

Note that in practice very often the approaches listed above are used in combination. For example offering an RDF dump (in N-Triples) for semantic search engines such as Sindice [163] along a SPARQL-endpoint for cross-site query is a typical pattern.

To allow semantic search engines to efficiently and effectively process the dataset it is advisable to use proper announcement mechanisms such as the semantic crawler sitemap extension protocol [39].

Comparison with other approaches

The following table 6.1 presents a comparison between three different approaches for modeling statistics in RDF. The comparison is based on [57] and highlights some differ-

[12] http://www.w3.org/TR/rdfa-syntax/

ences between the modeling from the D2R Server for Eurostat[13], the 2000 U.S. Census Data [159], and SCOVO itself.

The most distinguishing feature of SCOVO is the ability to express complex statistics over time while still keeping the structural complexity very low. Both other approaches are not capable of representing historical data and only provide statistics for one point-in-time. From the table below we conclude further that SCOVO seems to be the best combination of flexibility and usability, allowing to recreate the data-table structures with a reasonable degree of fidelity in another environment (that is, on the Web). Additionally, in our understanding SCOVO is more aligned with the linked data principles, compared to D2R Eurostat and 2000 U.S. Census.

6.2.5 Usages in the Wild

Eurostat Data—riese

In riese, the "RDFizing and Interlinking the EuroStat Data Set Effort" we have RDFized, interlinked, and published the Eurostat dataset on the Web. A detailed discussion can be found section 6.1 of this thesis. First of all the dimensions and dataset hierarchies defined in Eurostat get translated to RDF and interlinked. The actual data is translated to RDF on-the-fly from the raw Eurostat tables. Both human users and machines can access the data from the same location thanks to embedding RDF on the human-readable pages with XHTML+RDFa. Additional access methods are available as well.

In riese HTTP URIs are used, which are compliant with the Linked Data principles. For instance the currency dimension "Euro" has been minted with the "concept URI" of `http://riese.joanneum.at/dimension/currency/eur` which can be dereferenced for accessing the information resource that describes the concept. Accordingly datasets and individual items have an URI assigned for unambiguous identification of all resources.

Several domain ontologies are being re-used in riese. Geographical dimensions such as countries, etc. for instance make use of the GeoNames ontology as they are subclasses of `geonames:Feature` which allows more expressive descriptions. It is hence easily possible to express further classifications such as the class of the geographical dimension (e.g. country, administrative division, etc.). Furthermore, the riese schema re-uses the Event ontology in the same way as described above. We note that the riese schema can be seen as a predecessor of the SCOVO. However, the schema used in riese is less

[13] `http://www4.wiwiss.fu-berlin.de/eurostat/`

	D2R Eurostat	**2000 U.S. Census**	**Scovo**
Expressivity	Simple, limited	Complex	Complex
Modeling of time	Point-in-time*	Point-in-time	Over time
Historic data	No*	No	Yes
Easy table (re-)generation	No	Yes	Yes
Access to actual statistical data	Using individual predicates	Using individual predicates	Value attached to items
Location of classifying features	Concatenated in the name of the predicate	Concatenated in the name of the predicate and a chain of predicates	Attached to items
Structural complexity	Flat model	Complex, related via individual predicates; sometimes inconsistent	Relatively flat model; all information attached to item; datasets as container
Use of blank nodes	No	Yes	No
Different predicates	Many	Many	Few
Knowledge needed for query	Predicate names of desired dimensions	Predicate names of desired dimensions and nesting chain	Name or URI of desired dimensions
What is modeled?	Real-world, ignoring statistical artefacts such as time, table, sub-tables, etc.	Statistical domain in question	Statistics in general
Use of deref.-able URIS	Very limited	Limited	Each statistical item has an explicit URI

* The use of different named graphs for different points in time is planned in D2R.

Table 6.1: Comparison between three different approaches for modeling statistics.

light-weight (i.e. making more assumptions about the modeling) and not as flexible as SCOVO.

In Other Vocabularies

SCOVO is used in early versions of voiD, the “Vocabulary of Interlinked Datasets” [2] to express information about the number of triples, resources and so forth. Using SCOVO in voiD allows a simple and extendable description of statistical information, however, a shortcoming has been identified: as `scovo:Items` are grouped into `scovo:Datasets`, there is an implicit assumption that all items in such a dataset share the same dimensions. This yields to complex SPARQL expressions, as it will often require a verbose check to make sure that an item has only certain dimensions and no others. An exemplary usage of SCOVO in voiD is given below in listing 6.7.

```
:DBpedia a void:Dataset ;
         void:statItem [
           scovo:dimension void:numberOfTriples ;
           rdf:value 212576239 ;
         ] ;
}
```

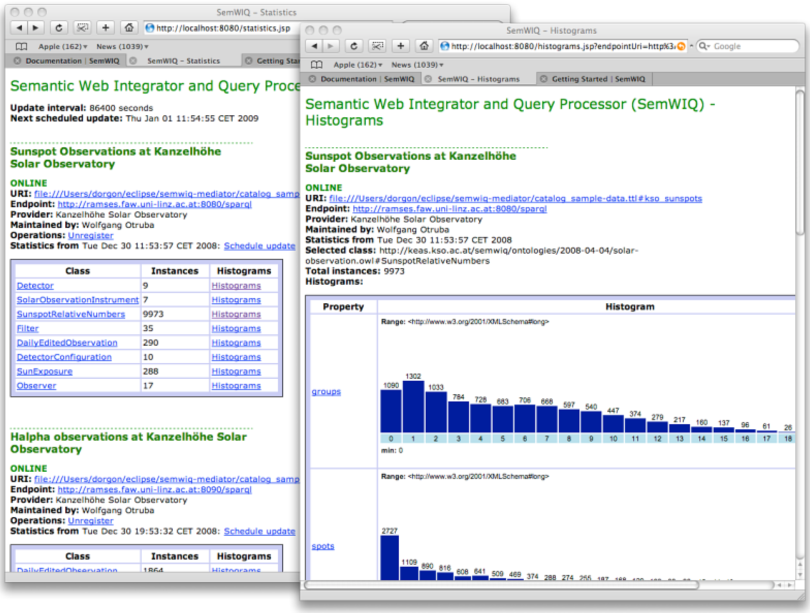
Listing 6.7: Usage of SCOVO in voiD.

Further, we have gathered that SCOVO is used in the RDFStats framework [110], see figure 6.7 (kudos to Andreas Langegger for the screenshot), that generates statistics for datasets behind SPARQL-endpoints and RDF documents.

6.2.6 Conclusion and Further Developments

We have proposed a vocabulary, SCOVO, and discussed good practice guidelines for publishing statistical data on the Web in this section. The framework aims at supporting people to publish their statistics on the Web of Data in an effective and efficient manner. The framework includes good practice rules stemming from our experience in publishing linked datasets. Finally, we have demonstrated the implementation of our framework and discussed practical issues with it.

However, there are limitations we are aware of. We advocate simplicity, hence there exist edge cases where it is hard to find an appropriate semantic modeling. Take for example our `scovo:Dataset` concept. In the current representation it is not explicitly



Semantic Web Integrator and Query Proce

Update interval: 86400 seconds
Next scheduled update: Thu Jan 01 11:54:55 CET 2009

Sunspot Observations at Kanzelhöhe Solar Observatory

ONLINE
URI: file:///Users/dorgon/eclipse/semwiq-mediator/catalog_samp
Endpoint: http://ramses.faw.uni-linz.ac.at:8080/sparql
Provider: Kanzelhöhe Solar Observatory
Maintained by: Wolfgang Otruba
Operations: Unregister
Statistics from Tue Dec 30 11:53:57 CET 2008: Schedule update

Class	Instances	Histograms
Detector	9	Histograms
SolarObservationInstrument	7	Histograms
SunspotRelativeNumbers	9973	Histograms
Filter	35	Histograms
DailyEditedObservation	290	Histograms
DetectorConfiguration	10	Histograms
SunExposure	288	Histograms
Observer	17	Histograms

Halpha observations at Kanzelhöhe Solar Observatory

ONLINE
URI: file:///Users/dorgon/eclipse/semwiq-mediator/catalog_samp
Endpoint: http://ramses.faw.uni-linz.ac.at:8090/sparql
Provider: Kanzelhöhe Solar Observatory
Maintained by: Wolfgang Otruba
Operations: Unregister
Statistics from Tue Dec 30 19:53:32 CET 2008: Schedule update

Class	Instances	Histograms
DailyEditedObservation	1864	Histograms

Semantic Web Integrator and Query Processor (SemWIQ) - Histograms

Sunspot Observations at Kanzelhöhe Solar Observatory

ONLINE
URI: file:///Users/dorgon/eclipse/semwiq-mediator/catalog_sample-data.ttl#kso_sunspots
Endpoint: http://ramses.faw.uni-linz.ac.at:8080/sparql
Provider: Kanzelhöhe Solar Observatory
Maintained by: Wolfgang Otruba
Statistics from Tue Dec 30 11:53:57 CET 2008
Selected class: http://keas.kso.ac.at/semwiq/ontologies/2008-04-04/solar-observation.owl#SunspotRelativeNumbers
Total instances: 9973
Histograms:

Property	Histogram
groups	**Range:** <http://www.w3.org/2001/XMLSchema#long>
spots	**Range:** <http://www.w3.org/2001/XMLSchema#long>

Figure 6.7: SCOVO driving the statistics in RDFStats.

defined how the overall range of the dataset is expressed. This may yield performance problems when one determines to figure out the overall range of a dataset. It is for sure possible to concatenate single dimensions used on the `scovo:Item`-level—for example concluding from the range of the four quarters `ex:Q12006` to `ex:Q42006` that the dataset actually is referring to the year 2006. Additionally, from the application of SCOVO in voiD we have learned that there is a demand for aggregates. The further developments that have taken place beyond SCOVO are described in the following section 6.3.

6.3 Beyond SCOVO

The Statistical Core Vocabulary (SCOVO) presented in the preceding section 6.2 has gained some interest and has also been used in in other reference implementations (VoID and RDFStats) as explained in section 6.2.5. The UK government and more specifically the Office for National Statistics (ONS)[14] as executive office of the UK Statistics Authority has expressed interest in SCOVO. As a result the author of this thesis has been invited to participate in the workshop *Publishing statistical datasets in SDMX and the semantic web* that has been organized by ONS in February 2010 in Sunningdale, United Kingdom. The aim of the workshop was to promote understanding of achievements and issues within each community (SDMX and Semantic Web), and to

[14] `http://www.ons.gov.uk`

build on this understanding to develop proposed guidelines for publishers of statistical datasets.

As a result, a working group has been started to combine the advantages of both the lightweight SCOVO approach and the massive SDMX standard. Intermediate results towards improved publication of statistical data on the Semantic Web have been presented by us in the LDOW'10 paper *Semantic Statistics: Bringing Together SDMX and SCOVO* [40]. The work has later on emerged into the *RDF Data Cube Vocabulary* which is on the W3C Recommendation Track to become a de-facto standard. In April 2012 a W3C Working Draft of the vocabulary[15] has been published.

The paper *Semantic Statistics: Bringing Together SDMX and SCOVO* [40] has been written together with Richard Cyganiak, Simon Field, Arofan Gregory, and Jeni Tennison. In this paper we have highlighted how to improve statistical data publishing on the Semantic Web. Many publishers of statistical data use the Statistical Data and Metadata Exchange (SDMX) standard [96] to represent statistics. The standard covers many aspects of statistical data publication and exchange. It includes for instance the representation of statistical data in flat files or XML, the definition of dimensions and attributes of observations as well as the discovery of statistical datasets through a central registry. Many organizations such as Eurostat, World Bank, World Health Organization (WHO), or International Monetary Fund (IMF) make use of this comprehensive standard, just to name a few. Also the Office for National Statistics (ONS) in the UK are expected to make their statistical data available using SDMX to allow an easy aggregation at international level. In addition, the UK Government has made a commitment in late 2009 to making public data available as Linked Data. This imposed a new challenge to statistics authorities as they are expected to publish their data both using SDMX and as Linked Data. On the way to unifying these two requirements also the above mentioned workshop *Publishing statistical datasets in SDMX and the semantic web* has been held. A very brief overview of SDMX follows and then we discuss how SDMX can be mapped onto Linked Data concepts. For data publishers it is of course important to build on existing standards and technology investments instead of replacing them.

6.3.1 SDMX

The Statistical Data and Metadata eXchange (SDMX) initiative was organized in 2001 by seven international organizations, the Bank for International Settlements (BIS), the

[15] `http://www.w3.org/TR/2012/WD-vocab-data-cube-20120405/`

European Central Bank (ECB), Eurostat, the International Monetary Fund (IMF), the Organization for Economic Co-operation and Development (OECD), the United Nations (UN) Statistics Division and the World Bank. These “Sponsoring Institutions” govern the SDMX activities. The goals for the initiative were to realize greater efficiencies in statistical practice, with a focus on employing current technology to enhance efficiency, improve quality, and address other challenges. These organizations all collect significant amounts of data, mostly from the national level, to support policy. They also disseminate data at the supra-national and international levels.

There have been several important results from this work: two versions of a set of technical specifications - ISO:TS 17369 (SDMX) - and the release of several recommendations for structuring and harmonizing cross-domain statistics, the SDMX Content-Oriented Guidelines. All of the products are available at `http://www.sdmx.org`. The standards are now being widely adopted around the world for the collection, exchange, processing, and dissemination of aggregate statistics by official statistical organizations. The UN Statistical Commission recommended SDMX as the preferred standard for statistics in 2007.

Several producers of important statistical datasets such as the Sponsoring Institutions, other statistical organizations, and central banks use SDMX. It is used for the support of data production and collection as well as for internal processing.

The scope of the SDMX standards is shown at a high level in figure 6.8. SDMX can be used at any level in the data provision or dissemination chain, be that at international level or within organizations, as it includes a view of the entire process of statistical production. This model is the result of implementation and analysis of statistical processes in many national, supra-national, and international organizations, and has been effectively used to support these functions in many implementations.

As shown in figure 6.8, Data Providers report Data and/or Metadata Sets which conform to certain business rules. The Data and/or Metadata Flow describes the provisioning characteristics over time and can be linked to one or more reporting categories. The Flow information needs to be linked to one Structure Definition. The Provision Agreement contains the details of the reporting of data or metadata by a specific Data Provider for a specific Data or Metadata Flow. Finally, also different Constraints can be applied to the Flow and/or Provision Agreement.

The SDMX Technical Specifications describe two major syntaxes: SDMX-EDI, which uses the flat-file UN/EDIFACT syntax; and SDMX-ML, which is broader in scope and offers XML formats for many types of statistical data and metadata. In both

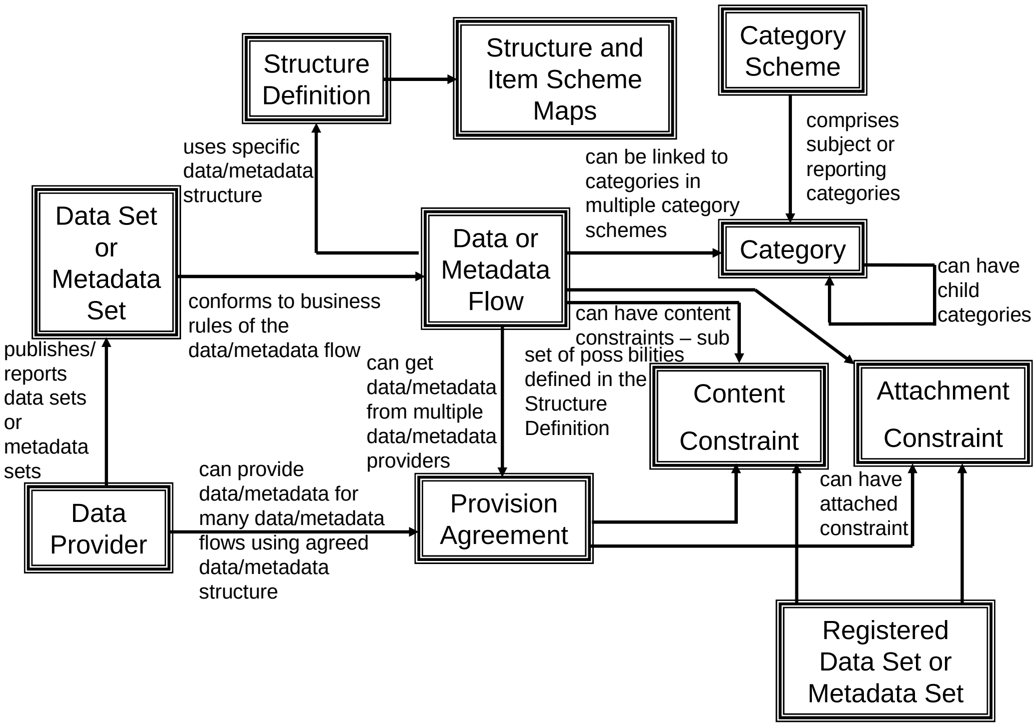

Structure Definition
Structure and Item Scheme Maps
Category Scheme
uses specific data/metadata structure
can be linked to categories in multiple category schemes
comprises subject or reporting categories
Data Set or Metadata Set
Data or Metadata Flow
Category
can have child categories
conforms to business rules of the data/metadata flow
can have content constraints – sub set of poss bilities defined in the Structure Definition
publishes/ reports data sets or metadata sets
can get data/metadata from multiple data/metadata providers
Content Constraint
Attachment Constraint
can provide data/metadata for many data/metadata flows using agreed data/metadata structure
Data Provider
Provision Agreement
can have attached constraint
Registered Data Set or Metadata Set

Figure 6.8: Scope of the SDMX standards. Source: [150, p. 52] © SDMX 2009

cases, users configure the formats to work with the statistical concepts of importance to their data and metadata, providing a flexible, generic basis from which to work. SDMX defines formats for exchanging statistical data and a service-based architecture for deploying and querying data with an optimized set of registry services. Tools are also becoming widely available for working with SDMX, as freeware, as open-source, and in statistical tools offered by commercial vendors.

6.3.2 SDMX and SCOVO

While SCOVO (cf. section 6.2) addresses the basic use case of publishing statistical data in linked data form, its minimalist design is limiting, and it does not support important scenarios that occur in statistical publishing and have led to the development of the SDMX information model, such as:

- definition and publication of the structure of a dataset independent from concrete data
- data flows which group together datasets that share the same structure, for example from different national data providers

- definition of "slices" through a dataset, such as an individual time series or cross-section, for individual annotation
- distinctions between dimensions, attributes and measures

There are also features of SDMX which are not directly addressed by SCOVO but can be expressed through other vocabularies or design patterns, such as:

- describing code lists, category schemes, and mappings between them using SKOS [151]
- describing metadata and access details about datasets using Dublin Core and voiD [2]
- describing organizations using FOAF [28]

Mapping SDMX to RDF can be realized partly by using SCOVO. Additional extensions are needed from other existing vocabularies and a new SDMX RDF vocabulary. In figure 6.9 an overview of the mapping from SDMX to RDF is given.

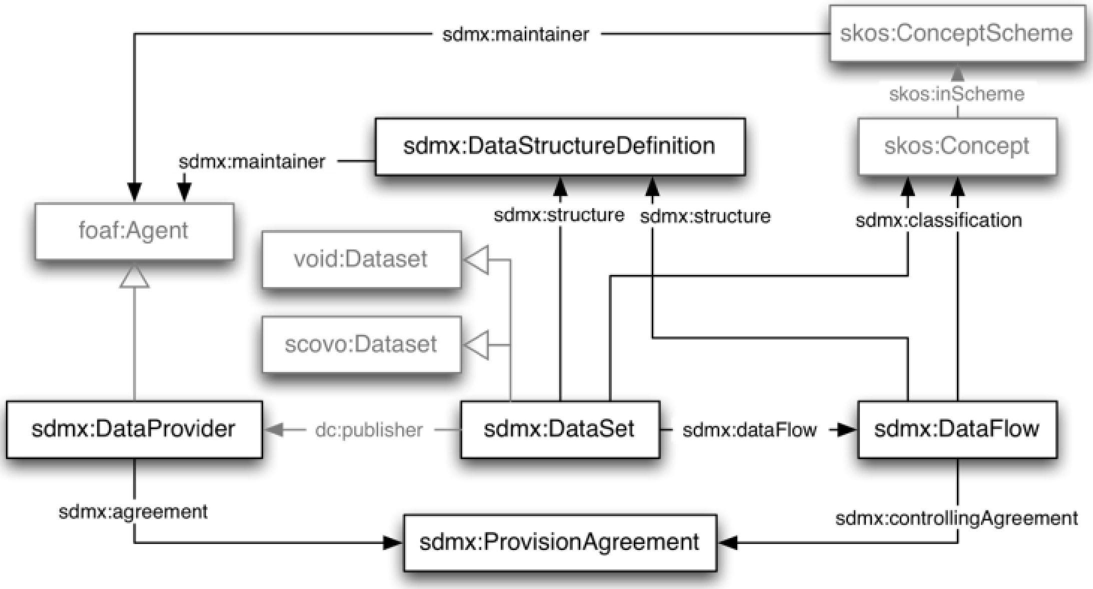


Figure 6.9: Overview of mapping SDMX to RDF

The `sdmx:Dataset` class is defined as a subclass of `scovo:Dataset` and `void:Dataset` which allows to describe the dataset and its access methods which also covers the basic capabilitites of the *registry* module in SDMX. Via the property `sdmx:structure` the connection to the Structure Definition can be realized. Organizations are represented as instance of `foaf:Agent` and Flows can also be represented. A detailed discussion of the Data Structure Definition, dataset representations, and content-oriented guidelines can be found in our paper [40] and are not reprinted here.

6.3.3 Conclusions and Further Work

There exist several approaches for realizing this mapping from SDMX to RDF. We have for instance demonstrated the mapping from the XML dialect of SDMX into RDF/XML via Extensible Stylesheet Language Transformation (XSLT). The generated Linked Data can be queried flexibly and different slices of the data can be retrieved. One of the potential issues with that approach is that a relatively large number of triples is created.

What we have learned is that both the statistical publishing and the Linked Data community share common aims in the direction of making data easily reusable. A considerable uptake in interest has also occurred and some further developments related to statistical Linked Data have been made. Most notably the work on the *RDF Data Cube Vocabulary* has to be mentioned, an effort mainly driven by the W3C Government Linked Data Working Group. The Data Cube Vocabulary is an RDF vocabulary for representing multi-dimensional "data cubes" in RDF. It is a further development of parts of SDMX-RDF presented above and represents a simplified version of the core of the SDMX Information Model. From `http://www.w3.org/TR/vocab-data-cube/` the latest draft of the Data Cube vocabulary can be retrieved.

6.4 RDB2RDF

As part of the activities of the W3C RDB2RDF Incubator Group we have prepared the report "A Survey of Current Approaches for Mapping of Relational Databases to RDF" [144] that has been written together by the author of this thesis with other members of the group.

In the survey we documented current approaches of mapping from relational databases (RDB) to RDF and categorized and compared the different approaches. For the comparison we defined a reference framework that shows the different possibilities and allows for an efficient categorization. The individual components of the reference framework are detailed below:

Creation of Mappings Mappings can either be created automatically, semi-automatically, or manually. Many automatic approaches consider a RDB record as RDF node, the corresponding RDB column name as RDF predicate and the RDB table cell as RDF object. A variation of the automatic generation of mappings between RDB and RDF is the use of an existing ontology to create simple mappings (cf. [90]).

Another approach is based on incorporating explicit domain semantics which can be expressed in a domain ontology. An existing ontology may be reused or a new ontology can be created based on some automated approach.

Mapping Representation and Accessibility of Mappings At the time of conducting the survey mappings between RDB and RDF were usually represented as XPath rules in a XSLT stylesheet, in XML-based declarative languages such as R2O [143] or in proprietary languages.

In addition, the newly created R2RML language [177] now also allows for expressing customized mappings that are represented as RDF graphs.

Mapping Implementation Two major approaches for implementing the mapping exist:

- Extract Transform Load (ETL): In the ETL approach the mappings to RDF are created (extracted and transformed) and then loaded into an RDF-capable store. The advantage of this approach is that it allows a stable performance and inference rules can be executed without compromising the query performance. The disadvantage is that changes in the underlying relational data need to be propagated and is therefore less-well suited for frequently changing data.

- Dynamic view: In the dynamic approach the mapping is done dynamically in response to query without pre-loading the RDF data. This way up-to-date data can be accessed at the potential cost of decreased query performance when inferences need to be made.

Query Implementation The implementation of the mapping also affects the query implementation to some extent. Basically, again two major possibilities exist:

- SPARQL → RDF: If RDF is readily available the query can be executed against the RDF store.

- SPARQL → SQL → RDB: The SPARQL query may also be translated to corresponding SQL queries and executed against the relational database (cf. e.g., [38]).

Application Domain Even though mapping tools are generally not domain-specific we included in our survey the application domain since domain semantics might be an important aspect if they are considered.

Data Integration Finally we included in our survey the different mapping approaches related to data integration, i.e. if additional datasets are considered and integrated in the mapping.

The full details of the survey are available in the report [144] and not reprinted here. The survey presented one of the two major report deliverables of the RDB2RDF Incubator Group. The second report is the Incubator Group report [119] which recommended that the W3C initiate a Working Group to standardize a language for mapping Relational Database schemes into RDF and OWL. The RDB2RDF Working Group has eventually been initiated and as one of the results R2RML (RDB to RDF Mapping Language) [177] is about to become a W3C Recommendation in 2012.

6.5 Summary

The creation of linkable data out of structured datasources was covered in this chapter which also addressed the two research questions *How can structured data be RDFized?* and *What are the potentials of public data and how can the data be represented?*. We presented our approach of generating linkable data out of structured data from EuroStat statistics in our riese use case. There we implemented a system that can transform relational data into RDF and we also introduced our approach of building Linked Data for both humans and machines. Motivated by our use case we also focused on statistical data as part of public data and introduced the Statistical Core Vocabulary (SCOVO) that can be used to represent statistical data as Linked Data. We also discussed the further developments that took SCOVO into account. Finally a brief overview of current approaches for mapping from relational databases (RDB) to RDF has been given. Regarding the general approach for RDFizing structured data we found two major mapping implementation strategies: Extract Transform Load (ETL) to transform data into RDF and load in a triplestore or a dynamic view that is created dynamically in response to a query.

Chapter 7

Unstructured Datasources

This section discusses briefly the creation of linkable data from unstructured datasources such as plain text on the example of Link2WoD. It has to be noted that the broad research area of information extraction and related subfields such as natural language processing or named-entity recognition are not directly in the main scope of this thesis but research results are applied. Therefore this chapter is included for completeness as the term extraction module is an essential part in Link2WoD. Special thanks go to Helmut Mülner who contributed to the development of the term extraction module in Link2WoD.

In the following section 7.1 a short introduction to the broad related research areas is given. Our applied approach in Link2WoD is further on discussed in section 7.2.

7.1 General Approaches

In order to make unstructured data such as natural language text linkable and become part of Linked Data it needs to be processed accordingly and the inherent structured information needs to be made explicit. In the research area of information extraction several approaches have already been proposed and developed. Most activities focus on the processing of human language text through natural language processing (NLP). However, also the extraction of structured data from multimedia content can be considered as information extraction task.

Especially on the Web (of Documents) large amounts of unstructured data exist and the challenge is to make use of this tremendously large information base. Approaches such as Linked Data already aim at providing structured data which makes the (automated) use of this data considerably easier. In order to create structured data (and potentially later on even Linked Data) several approaches exist. Due to the complexity many of them focus on specific application areas. In contrast to information retrieval which concerns how to identify relevant documents from a document collection, information extraction already produces structured data that is ready for post-processing.

Generally speaking, an information extraction task takes an information artifact such as text as input. Optional inputs are training corpora and/or target data structures. If a data structure is not supplied this is also considered an open information extraction task (cf. [53]). In cases where a data structure is supplied this is also referred to as closed information extraction. The output is a populated data structure and optionally also an information extraction pattern. Different systems and algorithms exist. A rather recent overview of the entire field of information extraction is provided by Sarawagi in [146].

The major types of structured data that is typically extracted in an information extraction task are:

- Entities: Being in the scope of named-entity recognition the task is to identify elements in a text that can be categorized (e.g., as person, organization, etc.).
- Relationships: Relationship extraction tries to identify relations between entities that in the most cases can be represented as a triple `<entity1> <relation> <entity2>`. This may include relations that are explicitly expressed such as `<Wolfgang> <locatedIn> <Graz>` extracted from "Wolfgang lives in the city of Graz" or also more complex relations that are spread over a document and which may also include the analysis of sentiments or opinions.
- Structural information: Some algorithms also target the extraction of more structural information from text such as the extraction of ontologies (e.g., [116]) or the extraction of tables (e.g., [92]).

Approaches include manually defined rules or learning-based systems that require training data. Different classifiers can be used such us generative naïve Bayes classifiers or discriminative maximum entropy models. Also many different sequence models have been proposed including Conditional Markov Models (CMMs), Hidden Markov Models (HMMs), and Conditional Random Fields (CRFs). CRFs [109] are considered as state-of-the-art methods. Also several other as well as hybrid approaches exist.

Different approaches are mainly compared by their accuracy and computational complexity. A survey of Web information extraction systems has been conducted by

Chang et al. [33] which compares some of the major approaches.

As the research area of information extraction is not in the main scope of this thesis the interested reader is referred to the literature referenced above for further details. In the following section we will briefly discuss how we have dealt with unstructured data in the context of our Link2WoD demonstrator.

7.2 Application in Link2WoD

We have partly reported about the creation of linkable data in the context of Link2WoD in our paper "Towards a Commercial Adoption of Linked Open Data for Online Content Providers" [74] which has been written together with Alexander Stocker, Harald Mayer, Helmut Mülner, and Ilir Ademi. It has been presented by the author of this thesis at the 6th International Conference on Semantic Systems (I-SEMANTICS 2010) in Graz. The author of this thesis was the research and development lead of Link2WoD and contributed most parts of the work. Helmut Mülner contributed to the development of the term extraction module in Link2WoD.

In the Link2WoD demonstrator three different modules are responsible for processing input content that takes the form of unstructured data, i.e. plain text. The three modules are:

- Term extraction & classification
- Linked Data consumption & interlinking
- Linked Data provision

The *term extraction & classification* module does named-entity recognition for a set of pre-defined categories and based on the input text provides as output identified entities. The *Linked Data consumption & interlinking* module (cf. section 10.3) takes this output to find related information from Linked Data sources and creates links to these external Linked Data sources. Finally, the *Linked Data provision* module (cf. section 13.4) makes the combined information available as Linked Data.

The *term extraction & classification* module that has been developed for Link2WoD is targeted at German-language content as the cooperating content provider produces the majority of its content in German. Even though there exist several term extraction solutions for the English language (which can also be plugged in and used in the prototype) there is still a lack of well-performing, generally available tools for German.

Through its modular structure the prototype can support term extraction solutions for any language that are available from third-party providers.

The method developed in Link2WoD for term extraction follows a simple approach: All common terms that occur in general dictionary are removed from the text. As a result, only "interesting' terms such as names, toponyms, and domain-specific terms are left over. This approach requires a rather complete dictionary containing all morphologically possible word formations. We were able to acquire a list of words from a German dictionary and created a complete dictionary that includes an entire set of all possible word-forms. The dictionary lookup was implemented as minimum deterministic finite automaton which enables a good computational performance. With the dictionary the following queries can for instance be executed quite efficiently:

- Exact query of a word
- Query of all words with a given prefix
- Query terms that contain placeholders for one or more characters

One of the challenges in the German language is the frequent occurrence of compounds (i.e. lexemes that consist of more than one stem). Through compounding almost arbitrarily long words can be created, especially in agglutinative languages such as German. Extreme examples of compounding in German are the "Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz" (literally "beef labeling supervision duties delegation law") which is a law of the German state of Mecklenburg-Vorpommern of 1999, dealing with the supervision of the labeling of beef. But also the famous example of the "Donaudampfschiffahrtsgesellschaftskapitän" (literally "Danube steamboating association captain") demonstrates the exaggerated possibilities that have to be dealt with. In order to address such issues two approaches for the decomposition of compounds have been developed:

1. A fast but inaccurate method which tries to identify the longest possible parts of a word in a list of words, without using grammar rules.
2. A slower but more accurate method which only considers grammatically valid word-forms in the decomposition.

However, this approach for entity extraction does not work for proper names that are homonymous to common words (e.g., the person "Claudia Schmied" where the German term "Schmied" also refers to a "blacksmith") or compounds including such words (e.g., the city of "Salz-Burg" where the German term "Burg" refers to a "castle"). We have addressed this issue by integrating whitelists of terms such as toponyms, first names, or company names. With a rule-based approach we were also able to increase the accuracy. Specific rules have for instance been applied to identify person names or organizations. For person names we have implemented the approach described by Volk and Clematide [165] which relies on the observation that there is a rather stable set of personal first names. They also found that a person's last name is usually introduced in a text with either his/her first name, with a title, or a word describing his/her profession or function. We also integrated the approach described in [165] of recognizing company names based on keywords that indicate the occurrence of a company name. Additional accuracy improvements have been accomplished by integrating specific corpora of toponyms.

As a short sum-up the *term extraction & classification* module combines the following approaches:

- **Blacklist:** All terms and possible word formations that occur in a general dictionary are removed from the text. As a result this step delivers "interesting" words such as names, toponyms, and domain-specific terms. All possible morphological formations have to be considered and compounds are also a special challenge in the German language.
- **Whitelist:** Further a whitelist with relevant terms is applied.
- **Rules:** With a rule-based approach also person and company names are identified (e.g. first name followed by noun → person name).

The module can be used as a standalone command-line application or integrated as a Web service via the Simple Object Access Protocol (SOAP) [169] in other applications. In Link2WoD we integrated the *term extraction & classification* module via the Web service approach. The Web service takes a plain text document as input and produces the following output:

- Identified entity
- Categorization (location, person, organization, etc.)
- Occurrence
- Confidence value

The output of this module sets the base for further use of the analyzed text in the *Linked Data consumption & interlinking* module (cf. section 10.3) which takes the identified entities to find related information from Linked Data sources and creates links to them.

7.3 Summary

In this chapter the research question of *How can linkable data be extracted from unstructured data?* has been briefly addressed. We highlighted the general approaches to make unstructured data such as natural language text linkable and become part of Linked Data. We also reported about the application in our *Link2WoD* use case. As part of Link2WoD the *term extraction & classification* module does named-entity recognition for a set of pre-defined categories and based on the input text provides as output identified entities. It is targeted at German-language content and is based on a combination of blacklists, whitelists and rules.

Part III

Interrelating Linked Data

Chapter 8

General Considerations on the Interrelation

Interrelating Linked Data by setting links between different data sources is one of the Linked Data principles. It allows to discover more information about a resource. Linked Data browsers, crawlers, and applications can automatically follow these links and retrieve additional information from various sources. As discussed in chapter 3 we identified in our initial study in 2008 that there is a need to improve and increase the interlinks between different datasets.

A brief overview of general considerations on the interrelation of Linked Data has been given by the author of this thesis in a section about "Linking data" in the article "Linked Data and multimedia: the state of affairs" [149]. The article has been written together with Bernhard Schandl, Bernhard Haslhofer, Tobias Bürger, and Andreas Langegger. It has been published in the international journal "Multimedia Tools and Applications". Our further relevant already published work is referenced in the corresponding sections.

Two basic distinctions of interrelation can be made between *coreference resolution* and *reference enhancement*. *Coreference resolution* aims at identifying coreferent instances in different sources that describe the same real-world entity. It is also known in the database research community as "record linkage" and several other names also corefer to this challenge. Even though minor distinctions may be made, it can be said on an abstract level that several terms such as "coreference resolution", "record linkage", "entity disambiguation", "duplicate detection", or "record matching" are coreferent instances of this same real-world entity. The problem has for a long time been recognized in the database research community (cf. [183]) and the initial idea of record linkage even goes back to a 1946 article called "Record Linkage" [49]. In the Linked Data environment coreference resolution is an important approach and in addition also *reference enhancement* is of interest. With reference enhancement we refer to not only finding coreferent instances of the same entity in different datasources but instead extend the search for additional references that are related to a resource in question and provide further information.

An example for coreference resolution is given in Figure 8.1. In general it can be assumed that there exist two different datasets ($D1$ and $D2$) that have partially overlapping information. Two data instances (r_{D1} and s_{D2}) from each dataset refer to the same real-world entity. When this relationship is discovered it can be expressed with the property `owl:sameAs`. As a more concrete example for instance we have DBpedia as dataset $D1$ and GeoNames as dataset $D2$. It is identified that the DBpedia entry about Graz (`dbpedia:graz`, cf. r_{D1}) is coreferencing (via `owl:sameAs`) to the GeoNames description of Graz (`geonames:2778067`, cf. s_{D2}), i.e. both resources are referring to the same real-world entity, namely the city of Graz.

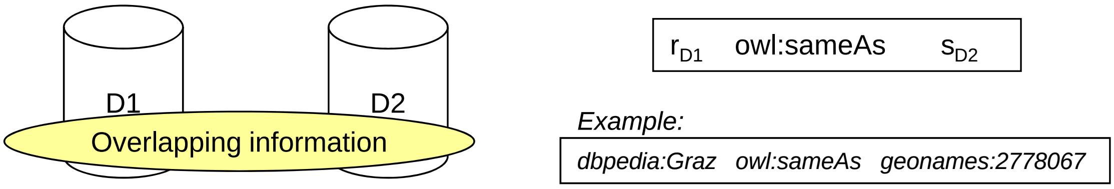

D1
D2
Overlapping information
r_{D1} owl:sameAs s_{D2}
Example:
dbpedia:Graz owl:sameAs geonames:2778067

Figure 8.1: Example of Linked Data coreference resolution.

In Figure 8.2 an example of reference enhancement is depicted. Again we have two different datasets ($D1$ and $D2$). However, this time the interrelation between the different datasets is not based on co-occurrence but instead a link is created that introduces additional information and semantics also by using a broad variety of different properties, such as `rdfs:seeAlso` for instance. The property used for the interrelation now also describes in more detail the kind of relationship between the two instances r_{D1} and s_{D2}. Any URI can be used as interlinking property but it is desirable to use terms from a well-known vocabulary. As a concrete example it would for instance also be discovered that Graz (e.g. represented by `dbpedia:graz`, with DBpedia being dataset $D1$) is of `rdf:type` place (e.g. `http://schema.org/Place`, with schema.org being dataset $D2$). In the following sections several exemplary approaches are discussed.

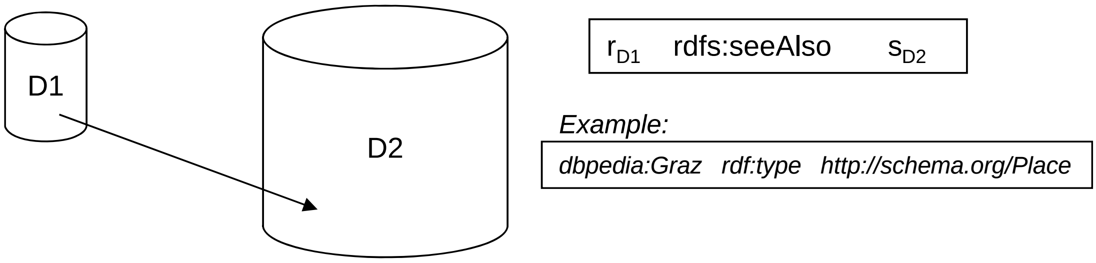

D1
D2
r_{D1} rdfs:seeAlso s_{D2}
Example:
dbpedia:Graz rdf:type http://schema.org/Place

Figure 8.2: Example of Linked Data reference enhancement.

Links between different Linked Data sources on the Web of Data are important to allow the discovery of more information for both humans and machines. While on the Web of Documents links between different documents are usually untyped, it is desired to set typed links in the Web of Data. Using different properties enables a more fine-grained automated analysis of data across different datasets. The property that is used to link data depends on the application domain and the intended semantics. Additional complexity in interrelating Linked Data comes from the use of different ontologies where also a lot of research has been done in the area of ontology matching and this topic is not in the scope of this thesis. Regarding the general approach for Linked Data link generation it can be stated that methods for database integration offer a certain guidance but due to the nature of Linked Data cannot be directly applied to it without complications. Instead, refined techniques are required that take the specifics of Linked Data into account. Data is represented in RDF and structured according to some RDF schema or ontology which can also be exploited. Generally speaking the interrelation of Linked Data can be done via

- user based approaches (cf. chapter 9) or
- automated approaches (cf. chapter 10).

User based approaches rely on a user manually creating the links. This can be done without any tool support and we have also investigated and developed some tools to aid in the manual creation of links between different datasets. These manually created links can subsequently also be used as input for automated methods. To create a substantial number of links a considerable manual effort is required. With a large number of users a manual approach can also lead to many created (potentially high-quality) links but users either need to be incentivized to create the links or the link creation should happen as a by-product of some other routine task. Human users are especially good at creating reference enhancements that interrelate different entities. With automated approaches it is possible to dramatically increase the speed and amount of link creation in comparison to manual approaches. Automated approaches are best suited for identifying coreferent instances across different sources and different strategies exist. Methods for record linkage from the database domain can partially be applied and refined for Linked Data suitability.

As many interlinks between different datasets are created automatically based on some pre-defined algorithm, there exist several challenges. One of the issues is data quality where an automated approach needs to identify the correct instances to be interrelated. Given as an example an instance with the name "Obama" it is not entirely defined by this single piece of information which real-world entity it refers to. It can be the president of the United States but also a small city with this name in Japan. Additional information and metrics are needed to improve the automated interlinking. There exists also the challenge of creating interlinks with different, strongly-typed and meaningful properties. However, automated approaches mainly target the coreference resolution which basically can only result in an interrelation via `owl:sameAs`. To identify semantically better defined relationships between different datasets also specific techniques need to be applied. As our initial study in 2008 [85] also revealed there is a relatively low variety in the properties used for interlinking.

Summary General considerations on the interrelation were presented in this chapter. Interrelating Linked Data by setting links between different data sources is one of the Linked Data principles. It allows to discover more information about a resource. We distinguish between *coreference resolution* where different datasets contain information about the same entity and *reference enhancement* where links to additional information are provided. Regarding the research question *Why is there a low variety of different properties used for the interlinking?* we found that reference enhancement is needed to increase the variety of properties. As this is based on relationship discovery this is a more complex task than coreference resolution. We also introduced the two approaches for interrelating which can be done user based or automated.

The following chapter 9 will introduce some of our user based approaches for interrelating Linked Data where we also considered strategies for increasing the variety of different properties for interlinking. In chapter 10 we briefly discuss general approaches for automated interlinking and go into detail for our examples in riese and *Link2WoD*.

Chapter 9

User Based Approaches

One approach of creating interrelations between different datasets is to exploit the capabilities of a human user and let the user manually create links between different datasets. Humans are especially good at creating associations and in some examples we wanted to make use of this. Manual link creation is possible without tool support but a very tedious task. We have developed some tools that make it easier fur human users to create links. With a small number of users a manual approach is only sensible for small datasets because of the required effort. However, if a large deployment base can be reached and sufficient users are attracted it is also possible to use the so called "wisdom of the crowd" to even interlink bigger datasets. Several factors influence the potential success of such an approach. The usability of the tool plays an important role and creating links should be made easy for the user. Ideally, the interlinking is a side-product of a process or interaction that the user already does to achieve some other goal apart from plain link creation. In addition there need to be appealing incentives for the user to invest some time for creating links with other datasets.

Even though human users can contribute to the creation of high-quality interlinks some attention has to be paid towards trust and privacy issues as we have already highlighted in section 5.3. If a system for manual link generation is open to the public there is the risk of somebody introducing intentionally or also unintentionally wrong links. During the development of our user contributed interlinking (UCI) approach we have identified that an editorial process similar to the one at Wikipedia can contribute to the identification of incorrect user contributed information. In addition users should be authenticated and in the UCI approach we have considered to use the distributed OpenID [141] system. In the example use case of the *Intelligent Media Annotation & Search* (IMAS) system from SALERO (cf. section 9.2) we have addressed this issue by only giving accounts to trusted users and tracking user contributions.

With riese ("RDFizing and Interlinking the EuroStat Data Set Effort") [73] we have contributed to the LOD cloud by adding the Eurostat data. We also introduced a new way of enriching datasets called "User Contributed Interlinking" (UCI), which is a generally applicable Wiki-style approach enabling users to add semantic (that is: typed) links between data items on a URI-basis. This approach is described in detail in section 9.1. We also investigated more specific applications for multimedia data where we investigated user based interlinking of multimedia data and details are described in section 9.2 for two demonstration use cases.

9.1 User Contributed Interlinking (UCI)

With *User Contributed Interlinking* (UCI) we proposed a new way of creating semantic links and introduced it for riese ("RDFizing and Interlinking the EuroStat Data Set Effort"). We reported about UCI in our paper "Building Linked Data For Both Humans and Machines" [73] that has been written together with Yves Raimond and Michael Hausenblas. The paper has been presented by the author of this thesis at the Linked Data on the Web (LDOW) 2008 workshop in Beijing, China. In our paper "Interlinking of Resources with Semantics" [82] we briefly discussed a generalized UCI approach in a demonstrator called "**i**nterlinking of **r**esources with **s**emantics" (IRS). The paper has been written together with Michael Hausenblas and presented as a poster at the 5th European Semantic Web Conference (ESWC) 2008. The remainder of this section contains an extracted and adapted reprint of our paper "Scripting User Contributed Interlinking" [83] where more details of UCI and IRS have been discussed. The paper has been written together with Michael Hausenblas and Yves Raimond. It has been presented at the 4th Workshop on Scripting for the Semantic Web (SFSW08).

As already discussed, RDF links can either be set manually or generated by automated linking algorithms for large datasets. For the latter case Raimond et.al. [140] have shown that simple interlinking algorithms produce rather poor results. Naive approaches trying to perform a simple literal lookup are likely to fail; for instance, when trying to interlink data from the geographical domain with GeoNames it is possible to do a simple literal lookup using the search facility provided by GeoNames. However, when querying for the city Vienna almost 20 results will be returned as there exist that many cities named Vienna around the world. Advanced approaches such as described in [140] are needed to disambiguate similar matches and finally create appropriate interlinks. Still, there is no guarantee that the automatically generated interlinks are truly relevant. Moreover the automated process is also restricted to predefined datasets implying that only a subset of the data available on the Semantic Web is considered when looking for potential interlinks.

Many interlinking algorithms found in current LOD datasets are largely based on very simple templates. This means that a huge number of interlinks can be generated, however, the quality of these links in terms of their respective 'semantic strength' is somewhat limited. It is well known that humans are good at associations, so we basically propose with UCI to let humans do the hard part of the interlinking.

9.1.1 UCI in riese

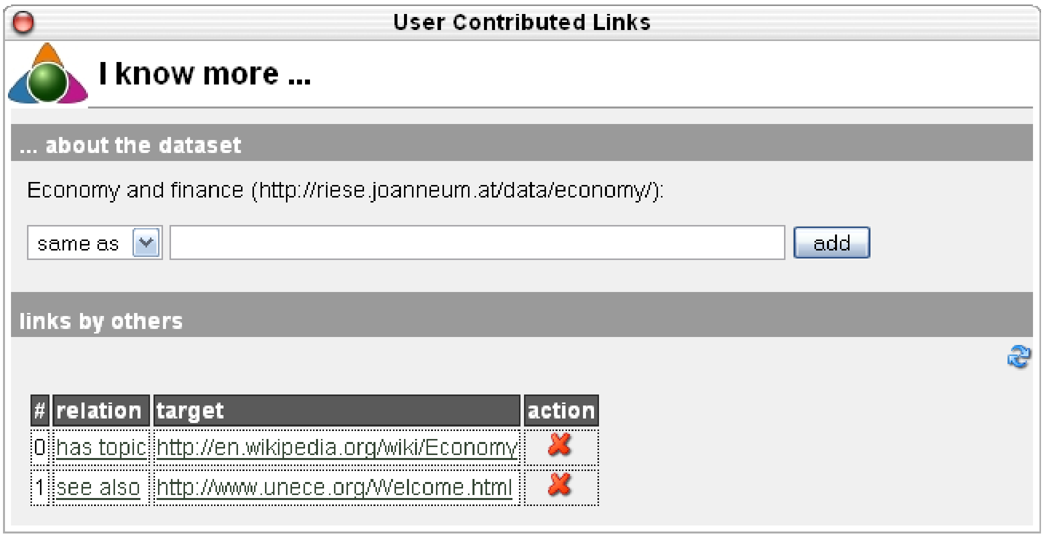
The User Contributed Interlinking (UCI) module enables the user to add (and remove for that matter) additional links to a certain data page. As the user must specify the type (cf. the drop-down box in Figure 9.1) it is ensured that only valid triples are introduced to the system—the subject of the RDF statement is always the page where the 'Related' box is on; the predicate is determined through the type selection. The object (named target in our context) is supplied by the user. A REST-based interface for adding UCI-triples automatically is available as well. Regarding the acceptance of the UCI, i.e. enabling users to contribute semantically typed links, we refer to the success story of Wikipedia [145] and strive for considerable community involvement. In riese, we therefore tried to implement many of the success factors of Wikipedia, such as openness or ease of editing.

The UCI part of riese, the UCI-interface, can be understood as an agent in the sense of [10]. The UCI-interface allows to list, add, and remove user-contributed semantic links from each of the statistical data items.

Operation	**Query String**
list semantic links of the data item `sURI`	`?src=sURI`
add a semantic link to the data item `sURI`	`?src=sURI&property=pURI&target=tURI`
remove a semantic link from the data item `sURI`	`?src=sURI&property=pURI&target=tURI&remove`

Table 9.1: Supported operations of the UCI-interface.

The operations supported by the current version of the UCI-interface are listed in



User Contributed Links

I know more ...

... about the dataset

Economy and finance (http://riese.joanneum.at/data/economy/):

same as ▾ add

links by others

#	relation	target	action
0	has topic	http://en.wikipedia.org/wiki/Economy	✗
1	see also	http://www.unece.org/Welcome.html	✗

Figure 9.1: The UCI module—users can provide own links.

Table 9.1. Note that the base service URI `http://riese.joanneum.at/interlinking/uci-interface.php` is assumed. With an additional `format` parameter the output format can be controlled. The default format is XHTML, an RDF/XML representation can be obtained using `format=RDF`.

To avoid concurrent editing a simple lock mechanism has been implemented. In case two users simultaneously want to add a semantic link to a data item, an according “please-hold-the-line” message is displayed.

It has to be noted that the UCI data is kept in a separate document—that is, a separate RDF/XML document, `uci-store.rdf`, per data item—in order to allow updates independently from statistical-data updates.

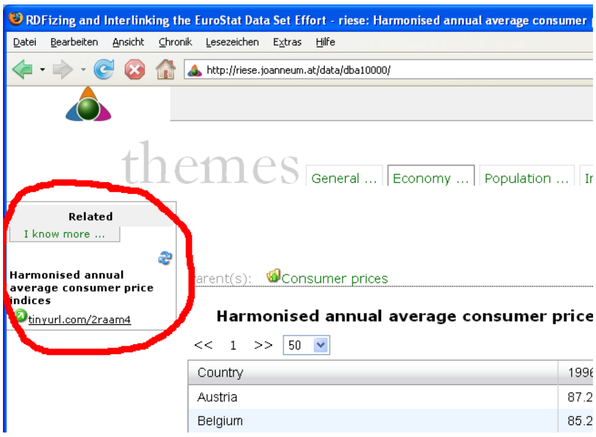
```
@PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@PREFIX foaf: <http://xmlns.com/foaf/0.1/> .
@PREFIX riesed: <http://riese.joanneum.at/data/>.

<riesed:economy>
 rdfs:seeAlso <http://www.unece.org/Welcome.html> ;
 foaf:topic <http://dbpedia.org/resource/Economy> .
```

Listing 9.1: An example result from a UCI query.

To obtain, for example, an RDF representation of the UCI data for the data item `http://riese.joanneum.at/data/economy`, one would use the query string `?src=http:`

//riese.joanneum.at/data/economy/&format=RDF. The result (rendered in RDF/N3 for better readability, here) would then be as shown in listing 9.1.

UCI User Interface On the client side we have implemented an user interface that controls the UCI-interface using AJAX (the UCI-UI). The Yahoo! User Interface Library (YUI) has been utilized for panels, events, etc. but also for the asynchronous communication. The UCI data is merged into the UCI user interface at rendering time. Figure 9.2 shows the UCI user interface "launch pad": For each data item a user may choose to add semantic links using the "I know more" button, effectively launching the UCI-UI.



RDFizing and Interlinking the EuroStat Data Set Effort - riese: Harmonised annual average consumer

http://riese.joanneum.at/data/dba10000/

Harmonised annual average consumer price indices

tinyurl.com/2raam4

arent(s): Consumer prices

Harmonised annual average consumer price

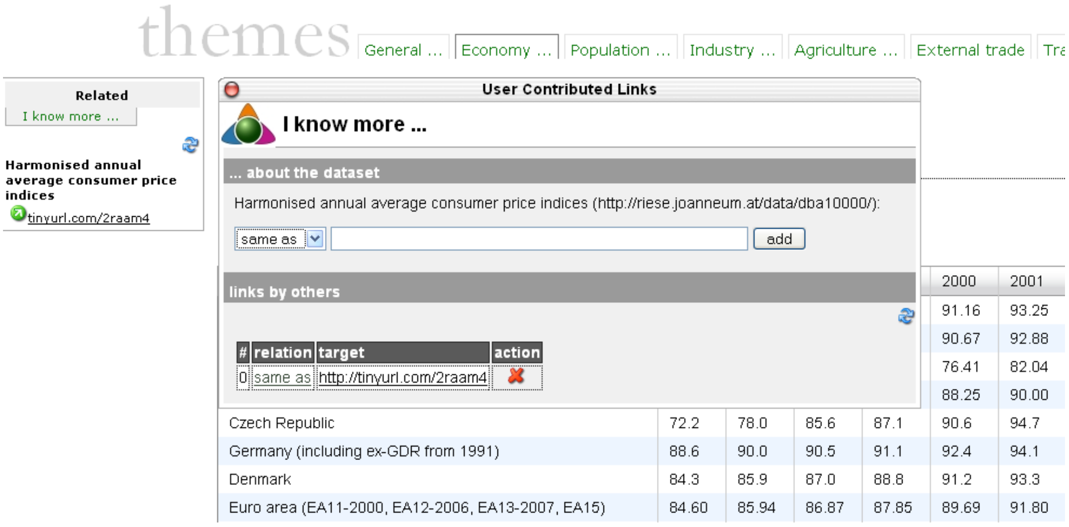
Country	1996
Austria	87.2
Belgium	85.2

Figure 9.2: UCI in riese - Launch pad.

In figure 9.3 the main UCI panel is depicted. Users can view, add, and remove semantic links with it.

Note how the subject of the RDF statement is implicitly set to the data item from which it has been fired. Currently three semantic link types (properties) are supported (`owl:sameAs`, `rds:seeAlso`, and `foaf:topic`). We decided to control this part of the RDF statement as well strictly to (i) make it easier to use for the average user from the street, and (ii) to avoid issues when following-your-nose. Finally, the object of the RDF statement is the open part of the UCI data. With open we mean that is is up to the user to determine what URI to paste in. However, people are encouraged to use URIs pointing to RDF (or GRDDL-able) resources.

Handling Data Updates One issue that we came across was the handling of data updates when users added new links to a dataset. Even though newly added links can be



Harmonised annual average consumer price indices
tinyurl.com/2raam4

Harmonised annual average consumer price indices (http://riese.joanneum.at/data/dba10000/):

#	relation	target	action
0	same as	http://tinyurl.com/2raam4	✗

					2000	2001
					91.16	93.25
					90.67	92.88
					76.41	82.04
					88.25	90.00
Czech Republic	72.2	78.0	85.6	87.1	90.6	94.7
Germany (including ex-GDR from 1991)	88.6	90.0	90.5	91.1	92.4	94.1
Denmark	84.3	85.9	87.0	88.8	91.2	93.3
Euro area (EA11-2000, EA12-2006, EA13-2007, EA15)	84.60	85.94	86.87	87.85	89.69	91.80

Figure 9.3: UCI in riese - Main panel.

retrieved when a resource description is accessed we added a notification functionality that allows data consumers to be notified about new data. When new data is available, one way to signal this is to subscribe to a news feed. We chose Atom [128] as the news feed format, as a corresponding RDF vocabulary (AtomOwl [11]) exists.

On the riese updates page (`http://riese.joanneum.at/updates/`) the data news feed is made available in AtomOwl. The AtomOwl feed in turn is serialized as XHTML+RDFa; see listing 9.2 for an excerpt of the updates page.

Using AtomOwl over XHTML+RDFa allows both humans and machines to consume the data updates properly. A human user directly accessing the page is able to view the updates, a Semantic Web agent capable of understanding XHTML+RDFa can process the feed entries for its purposes. A real-world example of how to use the AtomOwl-feed is provided in the following. In this experiment we have programmed SPARQLBot to access and query the AtomOwl embedded in the riese updates page. SPARQLBot offers a Web-based interface to define commands, which in turn maps to a SPARQL query (shown in listing 9.3).

Eventually, the same procedure can be applied to other scenarios, for example, when attempting to consume news feeds in an online news-reader.

9.1.2 Towards Generalizing User Contributed Interlinking

With the User Contributed Interlinking (UCI) we have proposed a novel approach for creating high-quality interlinks by relying on the users. The UCI approach is motivated

```
<body about=”http://riese.joanneum.at/updates” instanceof=”awol:
   Feed”>
 <div rel=”awol:title” instanceof=”awol:Content”>
  <span property=”awol:body”>updates</span>
 </div>
 <div id=”main−updates”>
  <ul rel=”awol:entry” instanceof=”awol:Entry”>
   <li rel=”awol:title” instanceof=”awol:Content”>
    <span property=”awol:body”>Compensation of employees − NACE J
       −K − Current prices − Millions of euro − SA</span>:
    <span rel=”awol:link” instanceof=”awol:Link”>
     <a rel=”awol:to” href=”http://riese.joanneum.at/data/na075”>
     http://riese.joanneum.at/data/na075</a>
     </span>
   </li>
  </ul>
```

Listing 9.2: An AtomOwl data update example in XHTML+RDFa.

by the observation that generic, template-based algorithms (such as described in [140]) are limited regarding the quality of the typed links.

For large datasets such as riese where the entire European statistics are brought to the Semantic Web it might appear impractical at first sight to manually generate interlinks to other datasets. It is obvious that it is not feasible to have one person dedicated to manually looking for adequate related sources. However, by applying the Wiki-principle we want to initiate a crowdsourcing process that encourages users to contribute to linked datasets with similar enthusiasm as they already show in the case of Wikipedia. It has to be noted that the proposed UCI-feature is the first of its kind. The current implementation as it can be found in riese is meant to bootstrap the community-involvement in the area of linked datasets. It should be adapted to other datasets as well. Based on the experiences gained with the first release of UCI the system and the related processes will be refined. User acceptance is the critical success factor of UCI and therefore we aimed at implementing as many of the best practices of Wikipedia as possible.

Sanger [145] was actively involved in the beginning of Wikipedia and has identified several factors that led to the great success of the platform such as openness and ease of editing. By inviting everybody to contribute we clearly highlight the openness of UCI. In addition we are working on enhancing the user experience by constantly improving the user interface design and keeping the user requirements at an absolute minimum as

```
PREFIX aowl: <http://bblfish.net/work/atom-owl/2006-06-06/#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?headline ?feed WHERE {
?feed rdf:type aowl:Feed ;
         aowl:entry ?entry .
?entry  aowl:title ?eTitle .
?eTitle aowl:body ?headline .
}
LIMIT 10
```

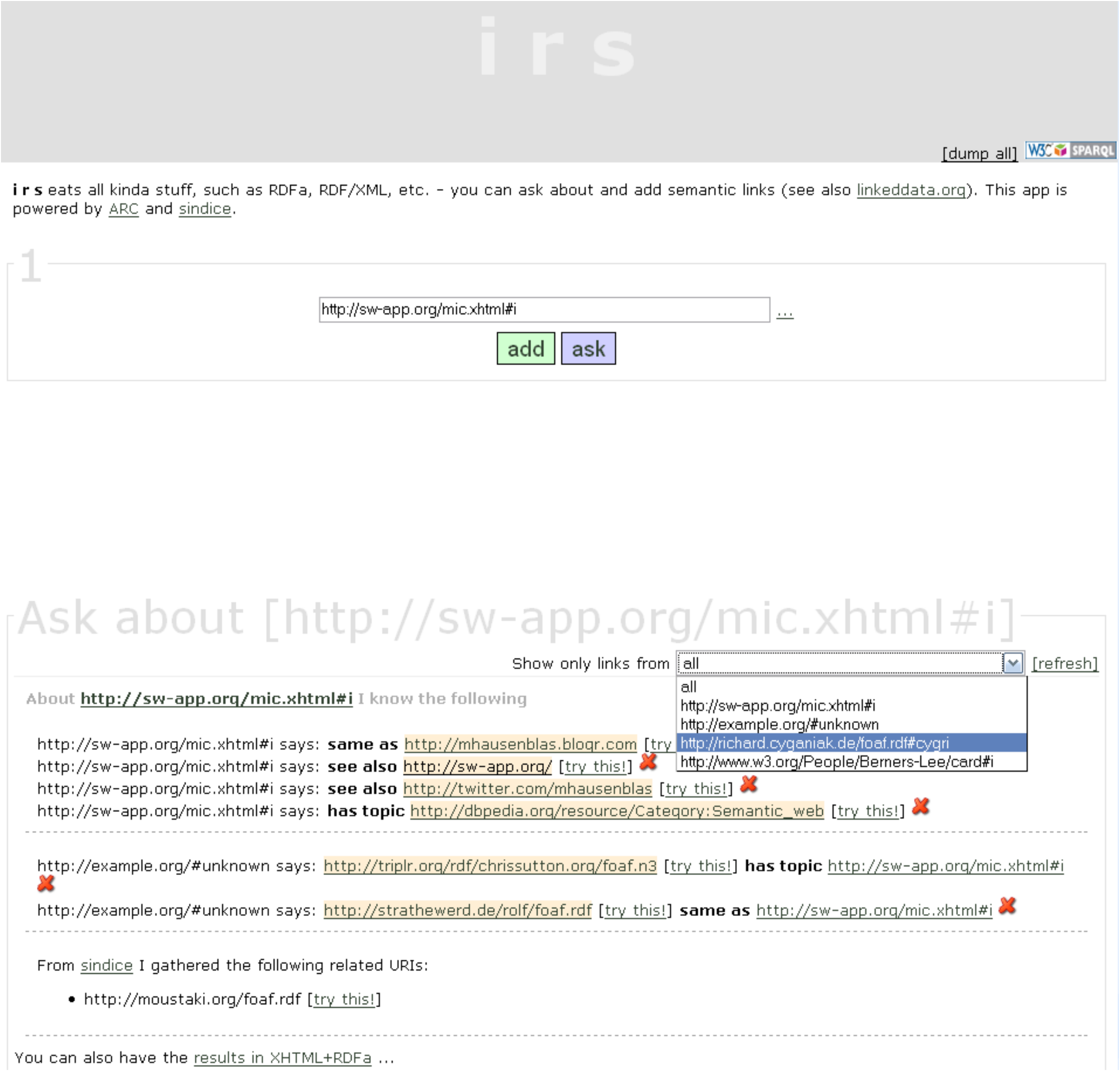
Listing 9.3: A SPARQL query for data updates on riese.

for instance no registration is required for using the UCI.

One of the disadvantages of common Wikis as identified in [132] is the limitation that "Wiki content is generally not available in a machine-processable format". With UCI we directly address this issue as the target outcome RDF is machine-processable per se. There are nevertheless still challenges left, such as reaching a critical mass of contributors by providing appropriate incentives or addressing data provenance issues. However, we would like to see a conversion of the strong community-engagement from Web 2.0 to the Semantic Web and contribute to the initiation of this transformation by providing useful tools such as the UCI.

As a next step, we have prototypically implemented a generalized UCI in a demonstrator called IRS (which stands for **i**nterlinking of **r**esources with **s**emantics). The IRS demonstrator was available online for testing purposes and as a proof-of-concept. A screen shot of IRS is shown in figure 9.4; it enables users to create semantic links (currently `owl:sameAs`, `rds:seeAlso`, and `foaf:topic`), to ask about existing links and to preview the (RDF) content. Further, a simple version of provenance tracking is offered: By placing the statements into a named graph (default is `http://example.org/#unknown`), one can track down who stated what. A simple off-the-shelf SPARQL-endpoint is also available in IRS.

While the Semantic Web itself may be regarded as a (backbone) infrastructure, developers of Semantic Web applications have to be aware of issues arising with it. We have presented a Wiki-style approach for user contributed (semantic) interlinking (UCI) in general, along with a discussion of tangible results. With UCI we have showcased an approach potentially increasing the end-user involvement in the Semantic Web. The acceptance of such features by the community is crucial and tools should be improved in order to provide an enjoyable user experience. For multimedia content we have devel-


i r s
[dump all]
W3C SPARQL
1
add
ask
all
http://sw-app.org/mic.xhtml#i
http://example.org/#unknown
http://richard.cyganiak.de/foaf.rdf#cygri
http://www.w3.org/People/Berners-Lee/card#i
[refresh]

i r s eats all kinda stuff, such as RDFa, RDF/XML, etc. - you can ask about and add semantic links (see also linkeddata.org). This app is powered by ARC and sindice.

http://sw-app.org/mic.xhtml#i ...

Ask about [http://sw-app.org/mic.xhtml#i]

Show only links from all

About **http://sw-app.org/mic.xhtml#i** I know the following

http://sw-app.org/mic.xhtml#i says: **same as** http://mhausenblas.blogr.com [try
http://sw-app.org/mic.xhtml#i says: **see also** http://sw-app.org/ [try this!]
http://sw-app.org/mic.xhtml#i says: **see also** http://twitter.com/mhausenblas [try this!]
http://sw-app.org/mic.xhtml#i says: **has topic** http://dbpedia.org/resource/Category:Semantic_web [try this!]

http://example.org/#unknown says: http://triplr.org/rdf/chrissutton.org/foaf.n3 [try this!] **has topic** http://sw-app.org/mic.xhtml#i

http://example.org/#unknown says: http://strathewerd.de/rolf/foaf.rdf [try this!] **same as** http://sw-app.org/mic.xhtml#i

From sindice I gathered the following related URIs:

- http://moustaki.org/foaf.rdf [try this!]

You can also have the results in XHTML+RDFa ...

Figure 9.4: A demonstrator for a generalized UCI: IRS.

oped three demonstration applications that should provide an improved user experience and these approaches are discussed in the following section.

9.2 User Based Multimedia Interlinking

User contributed interlinking has already been motivated in the previous section and a general approach has been presented. To maximize user participation we have created some tools that improve the user experience and make link creation easier. For an e-learning platform we investigated approaches for using tags to create Linked Data and with the CaMiCatzee demonstrator we showcased how semantic links could enter multimedia platforms on the Web. With the SALERO Intelligent Media Annotation & Search system we finally also introduced this user based link creation approach into an application for multimedia asset management.

9.2.1 Interlinking E-Learning Platforms

As an example for interlinking social e-learning platforms we looked into the ELGG platform and developed a module called SID (Semantically Interlinked Data). This module enables the creation of Linked Data from tagged and published user generated content on the platform. The related research has been published in the paper "Weaving Social E-Learning Platforms Into the Web of Linked Data" [156] which has been written together with Selver Softic and Behnam Taraghi.

The motivation for creating the SID module for an e-learning platform is two-fold: By creating links to Linked Data sources the generated content of individuals gets enhanced and with more semantics the content also becomes more valuable for all users. In many cases user generated content such as text, images, or video is accompanied with associated metadata that can also be in the form of tags. Tagging of content is a relatively common phenomenon observed with user generated content. These tags can help an individual user in retrieving and categorizing content. However, when tags created by different users are considered some potential challenges exist. Terms used for tags can be ambiguous as for instance the term "apple" can refer to the fruit or the company with this name. Another challenge is introduced by the individual diversity of tag choice when different terms are used by different users to refer to the same real-world entity and issues related to coreference resolution arise. The SID approach addresses both issues by enhancing descriptions and tags with semantics as a possible solution. Through the linking of tags to datasets like DBpedia the content descriptions are given a well-defined meaning. If the meaning of a user chosen tag is ambiguous, the user is provided with a list of different meanings (supplied from Linked Data sources) to choose from.

Social e-Learning platforms like ELGG[1] combine the advantages of social networks, Web 2.0 principles and educational techniques. They represent very interactive platforms but are limited to the area of their community. The exchange of information between two installations of the same platform (e.g. between two collaborating universities) cannot be done easily as there is no standardized exchange format used and it would require the implementation of specific Web services. The use of RDF and Linked

[1] http://www.elgg.org/

Data in particular could solve this problem and enable the exchange and integration of information across platform boundaries. As a first step with SID a module was implemented that focuses on tag-based semantically interlinked content.

The ELGG platform is one of many open source e-Learning environments used by numerous educational institutions and organizations around the world. Based on Web 2.0 technologies and its modular structure users can establish digital identities, connect with other users, collaborate with them and discover new resources through their connections. The common characteristic of Web 2.0 applications is the central role of the users. They are the actual content creators who share information with other users in form of blogs or some other type of content. Tagging plays an important role in this area. Similar data are interlinked to each other within the system through tags. The search functionality of the platform bases on this principle, so that searching for items results in retrieval of resources that are tagged with those items as keywords. The fact that tags do not have a defined meaning and are considered merely as literal strings is a disadvantage for the entire application. Semantically interlinking tags with resources which describe tags such as DBpedia could lead to more appropriate search results.

Our proposed interlinking model for tag based semantic enhancement of user contributed content in social e-Learning platforms and its interlinking into the Linking Open Data cloud follows the recommendations for the usage of tripartite models [125] containing information about users, tags, and related resources [105]. As simplicity is one of our model design goals the MOAT and Tag ontologies appear to be appropriate. The MOAT ontology offers to define meanings and to differ between different meanings of a tag. Additionally the context and scope of a tag can be explicitly described. The interlinking model for social e-Learning platforms makes use of the three wide spread ontologies MOAT, SIOC, and FOAF and simple relations between them (cf. example for ELGG Community in figure 9.5).

The proposed interlinking model was applied in TUGLL (TU Graz Learn Land), an ELGG platform running at Graz University of Technology. It contains several thousands of users and tagged blog entries. For the purpose of RDFizing and interlinking we developed the SID (Semantically Interlinked Data) module containing classes for exporting FOAF profiles, SIOC posts and MOAT tag meanings. The resulting RDF can be retrieved directly and also via a SPARQL endpoint that is provided. The semantic tagging mechanism provides an auto-complete function for disambiguation as depicted in figure 9.6 that the user can optionally turn on. The feature makes use of the web

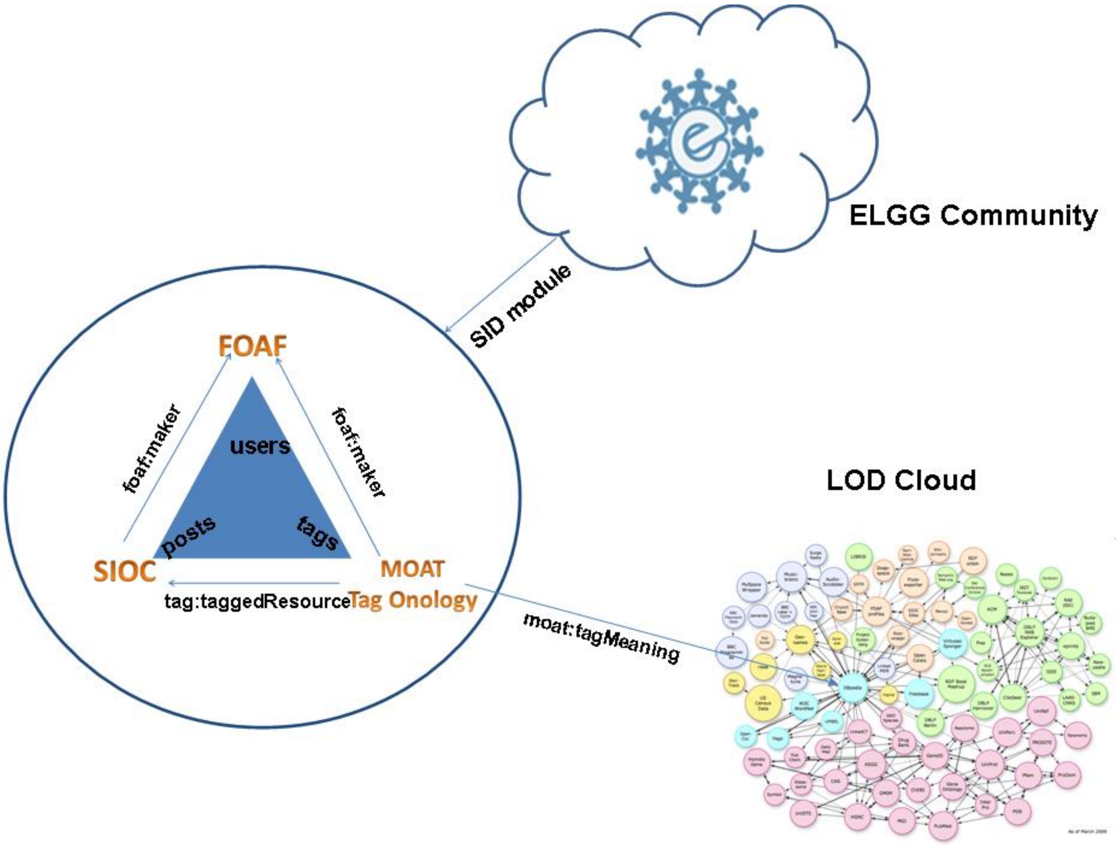
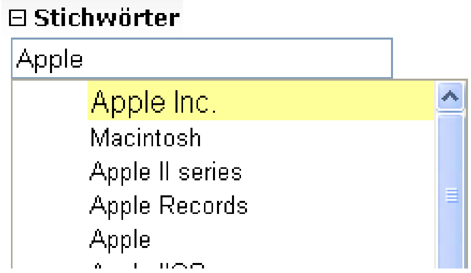


Figure 9.5: Model for tag-based interlinking of social content in ELGG.

service based DBpedia lookup API[2]. The auto-complete function leads to a meaningful tagging in the entire system as keywords are set more precisely and also the context is well-defined. One of the important advantages gained through this RDFizing and interlinking of the content is the interoperability. It can provide other systems with the public data (triples) as an open service.



⊟ Stichwörter

Apple

Apple Inc.
Macintosh
Apple II series
Apple Records
Apple

Figure 9.6: Auto completed semantic tagging.

As preliminary result a simple user scenario represented through a recommender

[2] http://lookup.dbpedia.org/

widget for a single resource (post, bookmark, slide, file, RSS ,video etc.) was implemented containing related resources tagged with same tag meaning or alternatively users having resources tagged with the same meaning. In Listing 9.4 an exemplary query for the latter case used by widget to gather information is shown. All further relevant information can also be easily retrieved in this simple way.

```
SELECT DISTINCT ?pers WHERE {

  ?reftag a tags:RestrictedTagging;
          tags:taggedResource
            <http://tugll.tugraz.at/medien07/weblog/778.html>;
          moat:tagMeaning ?meaning_ref.

  ?tag a  tags:RestrictedTagging;
          foaf:maker ?pers;
          tags:taggedResource ?res;
          moat:tagMeaning ?meaning.

FILTER(?meaning = ?meaning_ref && ?res!= <http://tugll.tugraz.at/
   medien07/weblog/778.html>)
}
```

Listing 9.4: Selecting related persons who tagged their resources with the same meaning.

The presented approach for RDFizing and interlinking a platform like ELGG shows several benefits: The problem of sharing information inside the community and to the outer world is diminished. The knowledge base of the community gets enhanced by reliable sources. Data present in RDF can be used for internal analysis or for integration of Semantic Web modules into the learning environment. Even though there are still open issues to be solved, with the proposed approach a large community finds its entrance to the Web of Data.

9.2.2 Catch Me If You Can (CaMiCatzee) Demonstrator

With *Catch Me If You Can* (CaMiCatzee) we have developed a multimedia data interlinking concept demonstrator. The goal is to show how user contributed content in this example as images from the flickr platform can be interlinked by users with other data. We have presented this approach in the paper “Interlinking Multimedia Data” [81] at the Linking Open Data Triplification Challenge at the International Conference on Semantic Systems (I-Semantics08). The paper has been written together with Michael Hausenblas.

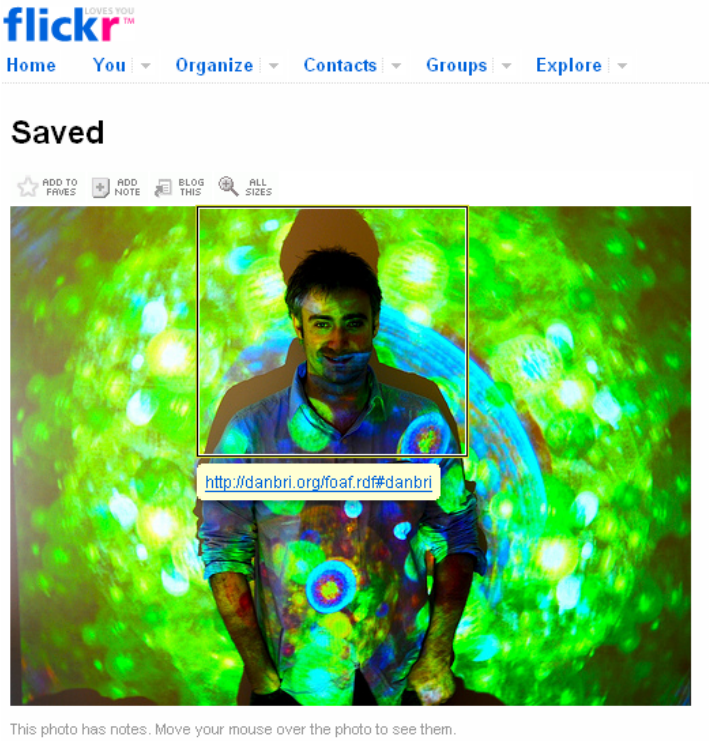
Our development of the CaMiCatzee demonstrator was inspired by the popularity of social media sites (such as flickr) and our User Contributed Interlinking (UCI) approach. Although social media sites provided at the time of creating the demonstrator already features for tagging and commenting, the outcome was mainly shallow metadata. In addition, Linked Data sets primarily addressed textual resources and the interlinking was based on rather simple string matching algorithms. Multimedia resources had been neglected so far. When referring to multimedia resources interlinking, we do not talk about global metadata such as the creator or a title; we rather focus on a fine-grained interlinking, for example, objects in a picture. Clearly, the advantage is having high-quality semantic links from a multimedia asset to other data, hence allowing to connect to the linked datasets.

In flickr it is possible to annotate parts of a picture using so called “notes”. As the primary domain, we chose people depictions. Typically, flickr notes contain a string stating, e.g., “person X is depicted in this picture”. However, there is no straightforward way to relate this information with other data, such as FOAF data, locations, and contextual information (conference, holiday, etc.). This is where we step in: we apply the UCI principle by harnessing the fine-grained annotation capabilities of flickr in order to let people semantically annotate pictures. An exemplary annotation is shown in figure 9.7 where a note has been added to the image on flickr containing the person’s URI. This can be done for all images where adding notes is enabled.

In the initial release of the multimedia interlinking demonstrator “Catch Me If You Can” (CaMiCatzee) the query for depictions can be performed using one of the following:

- the URI of a person’s FOAF document,
- a person’s URI, or
- simply a person’s name.

In the latter two cases a matching FOAF URI will be retrieved from sindice, a semantic indexer. Subsequently flickr is queried for matching annotations (on the person URI extracted from the FOAF document) yielding all pictures containing the desired person. Additionally, in the “full report”, the flickr tags of a picture are evaluated and used as a base for introducing `rdfs:seeAlso` links; this overview is offered in XHTML+RDFa, allowing consumption by both humans and machines.



Saved

http://danbri.org/foaf.rdf#danbri

This photo has notes. Move your mouse over the photo to see them.

Figure 9.7: Example of adding a note with the person's URI in a flickr image.

The system's architecture is depicted in figure 9.8 which shows external sources and services in the blue box, the CaMiCatzee server in the green box, and the CaMiCatzee client in the black box. Among the external sources is flickr where also the annotation is done and the user generated content is stored. Sindice is used to retrieve URIs for identifying persons if a person URI is not directly supplied in the query. On the CaMiCatzee server several modules are responsible for interacting with the external services and rendering the view for the CaMiCatzee client with XHTML+RDFa as well as for providing the SPARQL endpoint.

With this demonstrator we have shown the basic concept of user based interlinking of images. Our prototype had one major shortcoming in terms of usability. Even though we showed that identifying people on images via semantic links is feasible, it is not very user-friendly having to insert URIs in image notes. We have addressed this issue in the user interface design for the SALERO *Intelligent Media Annotation & Search* (IMAS) system which is presented in the following section.

As a side note, it is interesting to see a similar approach being quite popular some time after we have proposed this demonstrator. The social network Facebook for instance offers a photo tagging mechanism and in 2011 additional features such as tagging photos with Facebook Pages [56] have been launched. In the Facebook eco-system it is

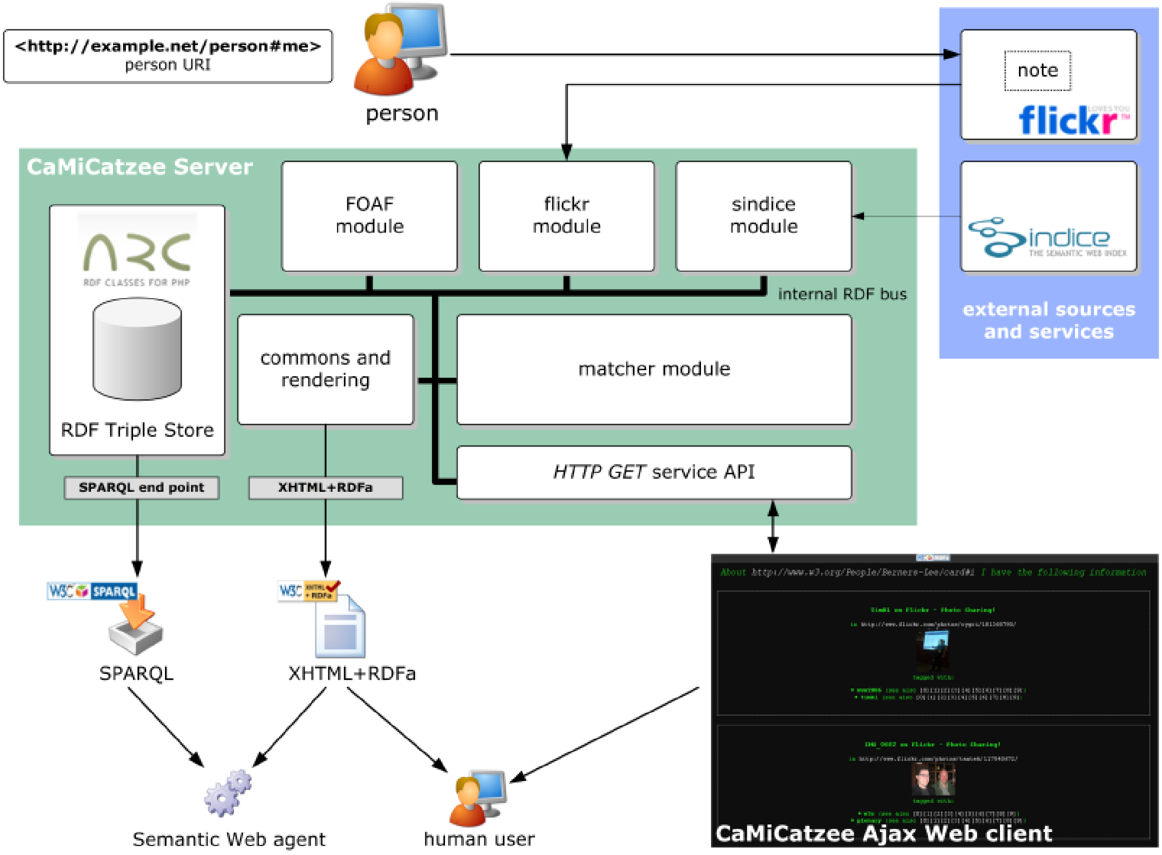


Figure 9.8: The CaMiCatzee system architecture.

therefore possible to easily create rich annotations of photos that contain links to URIs of people or other real-world entities (on Facebook). The user simply clicks on a region in the photo and can start typing the name of the person or other entity in a textbox. As the user types, suggestions of Facebook persons or Facebook Pages are offered in a dropdown box for the user to choose from. In some sense this feature on Facebook combines our approaches from the SID module shown in the preceding section and the CaMiCatzee demonstrator presented in this section, however not in an open system such as Linked Open Data but instead locked into the Facebook world.

9.2.3 SALERO Intelligent Media Annotation & Search[3]

A further exemplary implementation of the User Contributed Interlinking approach has been implemented in the *Intelligent Media Annotation & Search* (IMAS) system which has been created in the course of the SALERO (Semantic AudiovisuaL Entertainment Reusable Objects) project. The IMAS system provides a fast and easy to use approach to create semantic annotations and relationships of media resources. It is targeted at the media industry to provide efficient ways for organizing and retrieving media assets. The IMAS system has been described in the paper "SALERO Intelligent Media Annotation & Search" [181] which has been written together with Wolfgang Weiss, Tobias Bürger, Robert Villa, and Punitha Swamy. It has been presented at the International Conference on Semantic Systems (I-SEMANTICS) 2008. In the paper "Statement-based Semantic Annotation of Media Resources" [180] further details of the system along with the results of a usability test have been discussed. The paper has been written together with Wolfgang Weiss, Tobias Bürger, Robert Villa, and Punitha P. It has been presented at the 4th International Conference on Semantic and Digital Media Technologies (SAMT 2009) and has been published in Springer's Lecture Notes in Computer Science. This section contains parts of an adapted reprint of the original paper with kind permission from Springer Science and Business Media.

The management of media resources in media production is a continuous challenge due to growing amounts of content. We present a statement-based semantic annotation approach which allows fast and easy creation of semantic annotations of media resources. The approach is implemented in the *Intelligent Media Annotation & Search*[4] (IMAS) system. An integral part of the work being done in SALERO is the management of media objects with semantic technologies which is addressed by the IMAS system by enabling their semantic annotation and retrieval. The use of semantic technologies reduces the problem of ambiguity in search by using existing, well-defined vocabularies, it allows us to do query expansions and to deal with multilinguality. Being a further development of the User Contributed Interlinking (UCI) approach we also focused on usability issues to optimize the ease-of-use of the system.

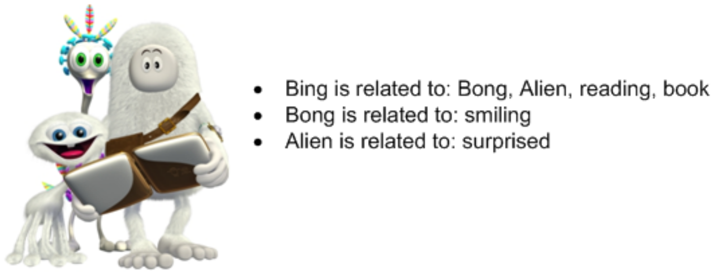
During prototypical development iterations of our system we have experienced, that most paradigms applied in semantic annotation tools are not suitable for inexperienced

[3] This section contains parts of an adapted reprint of "W. Weiss, T. Bürger, R. Villa, P. Punitha, and W. Halb. *Statement-Based Semantic Annotation of Media Resources.* In T.-S. Chua, Y. Kompatsiaris, B. Mérialdo, W. Haas, G. Thallinger, and W. Bailer, editors, *Semantic Multimedia*, volume 5887 of *Lecture Notes in Computer Science*, pages 52–64." © Springer-Verlag Berlin / Heidelberg 2009, with kind permission from Springer Science and Business Media under license number 2915330201249.

[4] http://salero.joanneum.at/imas/

users who are typically used to keyword-based tagging and suffer from information overload when confronted with complex annotation tasks and user interfaces. Our aim was thus to develop an approach which is faster and easier to use for our targeted user group, while making a compromise in complexity of full semantic annotations. Adhering to Linked Data principles it is possible to describe the content of each media resource and to relate media resources to other media resources as well as other entities from Linked Open Data sets.

The IMAS end user application is an integrated Web-based application which can be used to annotate and search for media objects. It allows to annotate arbitrary resources which are stored in preconfigurable media repositories. Media resources are annotated by creating statements and by relating them to other media resources. Annotation statements contain semantic elements which are defined in an annotation ontology. An exemplary image with statements is illustrated in figure 9.9.



- Bing is related to: Bong, Alien, reading, book
- Bong is related to: smiling
- Alien is related to: surprised

Figure 9.9: IMAS example image with statements.

For the creation of such statements and thus for the user based creation of links between the media resource and semantic descriptions is shown in figure 9.10. The three different possibilities in the user interface are:

1. combining concepts via drag-and-drop,
2. selecting concepts consecutively, and
3. using the text box as a command line interface with auto-completion.

Input option three is optimally suited for frequent users and input options one and two are ideal for users who rarely create annotations.

An additional possibility to annotate the content of media resources in the IMAS system is to relate them to each other. Hereby, we can describe that one media resource is, for instance, a part, a revision, or a transformation of another media resource. This allows us to use statements of the related media resources, to keep track of revisions

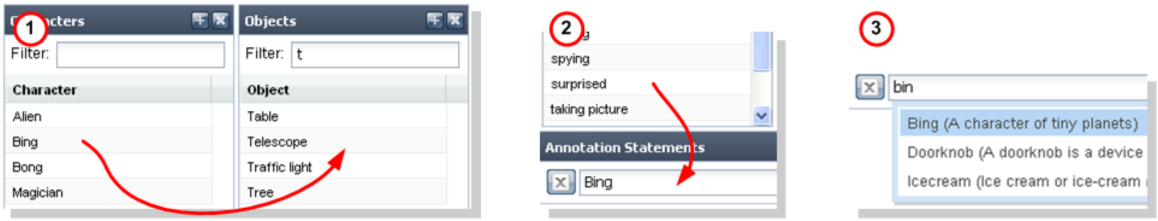
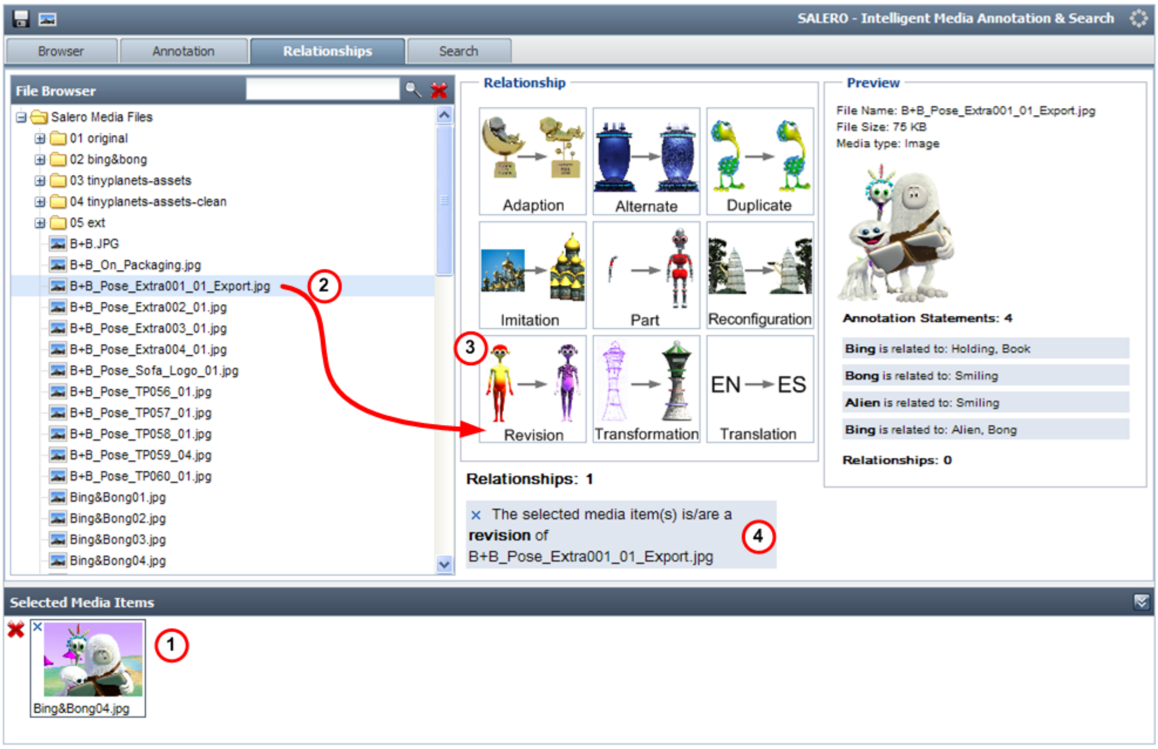


Figure 9.10: IMAS creation of statements.

of the media resources or to suggest alternatives in the search result. To create relationships (see also figure 9.11) of selected media resources (1), the user drags a source media resource from the file browser (2) and drops it on the desired relationship (e.g. the revision) of the relationship panel (3). This action finally creates the following annotation (4): `< SourceResource is_a_revision_of TargetResource >`.



The selected media item(s) is/are a **revision** of B+B_Pose_Extra001_01_Export.jpg

Bing is related to: Holding, Book
Bong is related to: Smiling
Alien is related to: Smiling
Bing is related to: Alien, Bong

Figure 9.11: IMAS creation of relationships.

As the IMAS system in a production environment can only be used by authenticated users only trusted parties can introduce new annotations and links into the system. This greatly reduces the risk of having intentionally wrong or misleading information in the system. In addition, user interactions are also logged which allows to easily identify users who abused the system.

A usability study of the IMAS tool was also conducted which included the completion of annotation tasks by test users in IMAS and two other tools (Google Picasa and PhotoStuff). Google Picasa allows free-text annotations of images whereas PhotoStuff supports complex semantic annotations. In the study we measured the task completion times for annotating media objects and the free-text approach of Picasa was the fastest, followed by the IMAS system. Creating annotations in PhotoStuff required the most time. Test users liked the auto-completion feature in the IMAS system and preferred this light-weight semantic tagging approach. The full details of the usability study and further information about IMAS are described in our paper [180].

With the IMAS system we have demonstrated a domain-specific implementation of a user based interlinking approach which provides an easy-to-use interface that is also suited for non-technical experts and hides the underlying complexity.

9.3 Summary

User based approaches for interrelating Linked Data were discussed in this chapter. In this way also the two related research questions *How can the manual creation of links be supported?* and *How can user based approaches be realized?* are addressed. Manual link creation is possible without tool support but a very tedious task. We have developed some tools that make it easier for human users to create links. With riese we have contributed to the LOD cloud by adding the EuroStat data. We also introduced a new way of enriching datasets called “User Contributed Interlinking” (UCI), which is a Wiki-style approach enabling users to add typed links between data items on a URI-basis. We have also prototypically implemented a generalized UCI in a demonstrator called IRS (which stands for **i**nterlinking of **r**esources with **s**emantics). To maximize user participation we have created some tools that improve the user experience and make link creation easier. For an e-learning platform we investigated approaches for using tags to create Linked Data and with the CaMiCatzee demonstrator we showcased how semantic links could enter multimedia platforms on the Web. With the SALERO Intelligent Media Annotation & Search (IMAS) system we finally also introduced this user based link creation approach into an application for multimedia asset management. In this way we demonstrated a domain-specific implementation of a user based interlinking approach which provides an easy-to-use interface that is also suited for non-technical experts and hides the underlying complexity. We found that usability is an important success factor and creating links should be made easy for the user combined with integration in routine processes and the provision of appealing incentives.

Chapter 10

Automated Approaches

With automated approaches it is possible to interrelate different Linked Data sets with relatively little manual effort. As we have already discussed, automation is best suitable for identifying coreferent instances of the same real-world entity across different datasets. This is also related to the problem of coreference resolution in the database community which needs to be adapted for the Linked Data world. It is quite common that information about the same real-world entity is contained in different data pools, i.e. these potentially partial information snippets reside in different "data silos". When it can be discovered, that two instances in different datasources are coreferent a relationship between these instances can be established. Setting links is also one of the Linked Data principles [14] to allow the discovery of more information with the "follow-your-nose" approach. If for instance a dataset D_1 contains information about a resource r_1 it might be the case that a different dataset D_2 contains information about a resource r_2 which both refer to the same entity and therefore more valuable information can be gained by combining these two datasets.

This chapter discusses some general strategies stemming from the database area and gives a brief overview of relevant metrics and techniques. Specifics of Linked Data are discussed and finally our approaches that we applied in riese and *Link2WoD* are presented.

10.1 Introduction

For the automated identification of coreferent instances across different datasets different terms exist such as "record linkage", "duplicate detection", or "coreference resolution". A general survey about duplicate detection in databases is presented by Elmagarmid et al. in [50] and Winkler [183] also provided an overview of record linkage with a focus on statistical data processing. The origins of the term record linkage are in the public health area where files of individual patients had to be found in different systems based on the name, date of birth and other information of the patient. Many sophisticated approaches for coreference resolution have been developed in the past decades and applied in various areas.

10.1.1 Data Preprocessing

In typical applications of coreference resolution a data preprocessing step is conducted where data is being normalized or standardized. In this step the aim is to improve the data quality so as to make it better comparable and usable. When entries are stored in a uniform manner and key identifiers follow a standardized format the identification of coreferent instances will perform considerably better. As an example in Table 10.1 three coreferent entries in different datasets are presented:

Dataset	**Name**	**DOB**	**Country**
D1	W. Halb	18OCT83	AT
D2	Halb, Wolfgang	1983-10-18	Austria
D3	Wolfgang Halb	10/18/83	AUT

Table 10.1: Example for data to be preprocessed.

As this example shows, the different formatting of the individual entries and fields makes it difficult to immediately identify that all three entries in the three different datasets actually contain the same information which is just represented differently. One important part of data preprocessing is data standardization through which it can be achieved that information of a certain type is represented in a specific format. Information usually can be stored in many different ways and a uniform representation is usually required for successful coreference resolution. When information is not standardized this can lead to wrong results because the common identifying information cannot be compared. Even though international standards for representing very commonly used information such as dates or countries exist, not every application also applies these standards for various reasons. For more specialized types of information it is also quite likely that no uniformly agreed standard exists at all which can lead to even more different representations of the same information entity. Additional problems are introduced by variant spellings (which can for instance quite often be observed in addresses) or misspellings (e.g. “Likned Data” instead of “Linked Data”).

Where internationally agreed standards for information representation exist, they should also be used to format the corresponding information accordingly. Date information for instance should be standardized according to ISO 8601 [95] which follows the pattern YYYY-MM-DD where years (Y) should be represented as four-digit number, separated by a hyphen (-) from the two-digit month (M) number and finally the two-digit day (D) number. The standardized representation in the example would consequently be `1983-10-18`. By adhering to this standard dates can be represented unambiguously. A similar example are countries which can also be represented according to ISO 3166-1 [97] in alpha-2 notation (i.e. two-letter country codes that are with some exceptions also used for for the Internet’s country code top-level domains). The standardization of names typically consists of the identification of name components such as first name and last name.

10.1.2 Matching Techniques

After the data preprocessing has been completed the individual information objects are represented in a way that they can be compared. In the following it needs to be identified which information objects should be compared as it would for instance not make sense to compare dates with names. In databases this requires either a manual mapping of fields or some automated approach to identify corresponding fields in a database. In the case of Linked Data this identification is eased when common vocabularies are used for properties such as for instance `foaf:givenName` to describe the given name of a person or `wgs84_pos#lat` to describe the latitude coordinates of a geo-location. Nevertheless, even with standardized data the coreference resolution is not trivial. Ambiguous information resources that carry the same label or share other common characteristics might refer to different real-world entities (e.g. a resource with the name “Obama” might refer to the president of the United States but also a small city with this name in Japan). Another challenge are typographical variations of string data that can be due to misspellings or different naming conventions. Different string similarity techniques exist to deal with this issue.

String similarity techniques String similarity techniques that are character-based work well for typographical errors. The *edit distance* measure is one of the basic techniques and also often referred to as the *Levenshtein distance* [115]. Given two strings s_1

and s_2 the edit distance is defined as the minimum number of edit operations of single characters needed to transform string s_1 into string s_2. Three different types of edit operations exist:

- inserting a character into the string,
- deleting a character from the string, or
- replacing a character.

Typically each operation has the same cost but variations exist where different weights are given to different operations. The lower the edit distance, the more similar the two strings are.

Another string similarity measure that is often used for coreference resolution is the *Jaro-Winkler distance* [182] which is best suited for short strings. It is a variant of the Jaro distance [100] which was mainly used for comparing first and last names. The Jaro distance d_j of two given strings s_1 and s_2 is defined as:

$$d_j = \frac{1}{3}\left(\frac{c}{|s_1|} + \frac{c}{|s_2|} + \frac{c - t/2}{c}\right)$$

where c is the number of common characters and t is the number of transpositions. Common are all characters $s_1[j]$ and $s_2[j]$ for which $s_1[i] = s_2[j]$ and $|i - j| \leq \frac{1}{2}\min(|s_1|, |s_2|)$. The number of transpositions t is the number of matching characters with different sequence order when comparing each character of s_1 with its matching characters in s_2. The Jaro-Winkler distance d_w is further on based on the Jaro distance d_j and defined as:

$$d_w = d_j + l \cdot p \cdot (1 - d_j)$$

where l is the number of common prefix characters up to a maximum of 4 characters and p is a constant scaling factor that as a default has the value 0.1 according to [182]. Variations exist and as long as p does not exceed the value of 1, the Jaro-Winkler distance is normalized where a score of 0 means that the strings do not match and at a score of 1 both strings match exactly.

There also exist further string similarity metrics that can be used to compare two strings. However, the Levenshtein, Jaro, and Jaro-Winkler distance metrics presented above are widely used and can also be applied in combination with additional approaches.

Set similarity techniques Techniques that are based on sets or tokens work well where naming conventions lead to different arrangements of words (e.g., "Wolfgang Halb" vs. "Halb Wolfgang"). In such cases character-based string similarity metrics as presented above will less strongly identify the similarity. General approaches taking set similarity into account are for instance cosine similarity, TF-IDF (term frequency - inverse document frequency), or the Jaccard measure. These approaches consider strings to be multisets of words or tokens. The Jaccard similarity coefficient is defined as the size of the intersection divided by the size of the union of two word sets S and T:

$$\frac{|S \cap T|}{|S \cup T|}$$

Similarly, the TF-IDF approach also considers common terms in the word sets but weights are given to terms, where rare terms from the set have higher weights. Again, different variations of the approaches exist that try to capture specific aspects such as for instance spelling errors, etc. However, as Bilenko et al. [19] have shown there is no single metric that is best suitable for all kinds of data sets. Metrics that perform very well for some data sets can have poor results on others.

Duplicate record detection approaches While the string-based approaches presented above target the comparison of individual strings (which relates to individual fields in databases) there also exist approaches that take an entire record into account to create better coreference resolution results. Converting this database view to Linked Data means that string-based similarity approaches target individual triples sharing a common property whereas record based approaches investigate all information about a given resource.

Machine learning approaches have for instance been successfully used so far which rely on training data. It is also possible to apply certain domain knowledge in identification approaches and combine with distance metrics.

10.1.3 Linked Data Specifics

As has been pointed out before, the research related to coreference resolution has mainly focused on the database community and some specifics of Linked Data exist which have to be taken into account when automatically interrelating Linked Data. Some tools and approaches for the link creation have been proposed and are briefly discussed below.

One of the major differences to traditional databases is that in Linked Data different datasets reside in different sources and potentially apply different vocabularies. This makes the integration process different. Among the differences are also the following aspects:

Schema Typically, relations/tables in a database have a defined structure where the attributes are pre-defined. This relates to both source and target schema for coreference resolution in databases. Linked Data does not follow such a strict structure and allows a resource to have an arbitrary amount of properties. This implies that Linked Data resource descriptions need not necessarily follow a common structure even if they are contained in the same data set.

Ontologies and vocabularies Linked Data also encourages the use of ontologies and vocabularies which carry additional information that can be exploited.

Accessibility Due to the nature of Linked Data we have to deal with a distributed system and apart from only considering coreference resolution algorithms also the availability of network resources has to be taken into account. Further information about a resource might be linked to an additional dataset which needs to be resolved. In a distributed system also performance aspects play an important role and the behavior of remote systems can most likely not be controlled.

As Raimond et al. [140] have discussed, very naive approaches that only look for literal matches in different data sets are likely to fail because too many ambiguous resources will be retrieved. Additional features such as restrictions to certain categories can improve the result. For the use case of music datasets they have also proposed a graph based matching algorithm which produced promising results. Instead of only comparing single similarities of resources themselves also the similarity of their neighbors was taken into account.

More generalized approaches for interrelating Linked Data have also been developed in parallel to our efforts. At the time of starting with our riese implementations no generalized Linked Data link generation software was available. In the following some selected approaches are presented before our approaches that we have applied in riese and *Link2WoD* are discussed in the subsequent sections.

Silk The Silk framework [167] has been released in 2009 and since then been further developed. The interlinking tool is parameterized by the Silk Link Specification

Language (Silk LSL) where the user specifies the type of resources to link and the comparison techniques to use. Datasets need to be accessible via a SPARQL endpoint. Link conditions can be expressed that take various similarity metrics (along with definable thresholds) into account. The type of property to create the link can be specified and the output is a linkset containing the links between the two given input datasets. The original Silk framework uses similar approaches as we have applied in the riese interlinking already in 2008 where we link two datasets based on instance similarities. In the course of the LOD2 project Silk has been further developed and now offers also a user interface in the workbench for easy creation of linkage specifications and a MapReduce version based on Hadoop for deployment on computing clusters.

Consistent Reference Service (CRS) The Consistent Reference Service (CRS) [98, 99] is utilized in the Resilience Knowledge Base (RKB) Explorer which is a Semantic Web application aimed at presenting a unified view of different data sources. CRS uses the concept of bundles to represent URI equivalence lists. These bundles are created based on an ad-hoc Java program and examples are available for some domains. A new program needs to be written for additional datasets and should contain mechanisms for resource selection and it can apply any (similarity) metric that is supported through Java code. This makes the approach flexible but the development of new matching programs is only possible for skilled Java programmers.

ODDLinker The Open Data Dataset Linker (ODDLinker) [77, 78] has been created in the course of the Linked Movie Data Base (LinkedMDB) project. It works with relational databases and exposes the result as RDF, thus is suited for relational data exposed as RDF but not for Linked Data natively. The introduced LinQL language is an extension of SQL and allows a declarative specification of linkage requirements.

KnoFuss The KnoFuss architecture [127] aims at supporting data integration and implements a component-based approach that allows a flexible selection of different methods. It also supports the reuse of existing ontology alignments. The comparison process is driven by a dedicated ontology for each dataset and only works with local copies of the data sets. The output is a merged dataset.

Concept Aggregation Framework for Structuring Information Aspects of Linked Open Data (CAF-SIAL) The CAF-SIAL framework introduced by Latif (cf. [112]) structures and presents information from Linked Data sources. It includes a component to aggregate relevant concepts from DBpedia and has been applied for identifying persons in two use cases for interlinking metadata of a digital journal and in an Expertise Mining System.

As this short overview has shown there is a certain interest in creating interlinking tools for Linked Data. As one of the earliest approaches we have also applied an interlinking technique in riese which is discussed in the following section.

10.2 Interlinking in riese

The *RDFizing and Interlinking the EuroStat Data Set Effort* (riese) has already been presented in section 6.1 where we discussed general aspects and the generation of linkable data. Further on in section 9.1.1 we discussed UCI, our user based approach for interrelating riese data sets. The current section presents our automated approach that has been applied in riese. This has partly also been reported in our paper "Building Linked Data For Both Humans and Machines" [73] that has been written together with Yves Raimond and Michael Hausenblas.

Leaving the mapping of the Eurostat data into RDF apart, it is equally important to apply the follow-your-nose principle [158], hence creating interlinks to other datasets. We have implemented a mapping module in the riese software stack that can be configured with a mapping configuration. For creating interlinks in riese we have basically used the following approach:

1. Restrict the source dataset to possible candidates for interlinking to the target dataset;
2. For each qualifying item in the source dataset look up a potentially matching resource in the target dataset;
3. Restrict the results by appropriate classifications or identifiers;
4. Create the interlink.

This general approach is also reflected in the mapping configuration which supports the following specifications:

- Source: Specification of the source dataset and identification of the resources to be interlinked

- Target: Specification of the target dataset(s) and identification of potential resources to be interlinked

- LinkingSpec: Specification of how links can be created

The linking specification (LinkingSpec) can further on use the following features:

- Similarity: Using one of the supported similarity metrics (literal) values can be compared. It requires an identification of the RDF triple object to be compared. Optionally, a threshold value for the similarity metric can be given. Supported distance metrics are the Levenshtein, Jaro, and Jaro-Winkler distances.

- Match: Exact matches can also be requested. This can be used for matching values from both datasets. We also support graph path expressions to a limited extent which enables flexible specifications.

- Restrict: Further conditional restrictions can also be applied to ease the creation of links where only minor variations of a rule are required.

The mapping configurations can be loaded in the riese software and are applied to corresponding datasets. A dataset registry keeps information about configured datasets. Local data from riese is accessed locally within SWI Prolog directly on the mapping information which can reduce access times and increase performance. External datasets are supported either via a (remote) SPARQL endpoint or via a dump of the dataset that can also be kept locally. When remotely accessing a SPARQL endpoint the time requirements for the mapping process are usually higher as the performance also depends on the remote system that is not in our control. It has the advantage though to always access up-to-date data. We also experienced that external services are not always available and service level agreements that guarantee a certain availability generally do not exist. Reasons for the unavailability of external resources include network issues but also maintenance downtimes at external third parties. Some datasets also impose restrictions on the number of queries allowed in a given amount of time or restrict the size of returned result sets. To overcome this limitation a local copy of the remote data can also be held to increase the speed of the mapping process and to be immune against network problems and third party service disruptions. The disadvantage of using local copies however is the issue of maintaining an up-to-date copy and being informed about updates.

For the actual mapping we had differing experiences and found cases where it is relatively straight forward to create high-quality interrelations whereas there also exist more diffcult ones. One example of interlinking that can be done with almost perfect precision and recall is the interlinking between country descriptions in riese and GeoNames. Countries can be identified based on a standardized representation. For the identification of countries the ISO-3166 standard [97] exists. GeoNames makes use of the ISO-3166 alpha2 country codes and Eurostat also supports these standardized codes in the Nomenclature of Territorial Units for Statistics (NUTS) [55] which is a geocode standard for statistical purposes.

The NUTS specification has been introduced by Eurostat and is used by many official institutions in the European Union. Strictly speaking NUTS is a three-level hierarchical classification system (levels 1, 2, and 3) but there also exists a NUTS level 0 specification that every NUTS code begins with and which corresponds to country codes in the ISO-3166 alpha2 standard. The subdivision of the country is then referred to with one number. A second or third subdivision level is referred to with another number each. Below these three NUTS levels also two levels of Local Administrative Units (LAU) have been defined. The correspondence between the NUTS levels and the national administrative units are also defined in the specification. In Germany for instance a NUTS level 1 region relates to states ("Länder") whereas NUTS level 1 regions are groups of states ("Gruppen von Bundesländern") in Austria.

Coming back to the mapping of countries between Eurostat or riese and GeoNames we can state that the use of the ISO-3166 alpha2 country codes (e.g., "AT") instead of the label (e.g., "Austria") assures that exactly the same resource is addressed in both datasets. Note that using ISO-3166 codes for identifying country descriptions in different datasets was already used by Voss [168] and others. An example configuration for such an exact matching is shown in listing 10.1.

In the practical implementation this means that first of all the source dataset riese is restricted to only geographical features of NUTS level 0. According to the nomenclature used it is possible to exactly identify country descriptions in the source dataset. Then the GeoNames search Webservice (i.e. the target dataset) is queried using the standardized codes. The result from the target dataset is then further restricted to return only countries, i.e. entries having a specified GeoNames feature code (e.g., "A.PCLI"). Finally all the matches are being interlinked by inserting a new triple into the source dataset which relates the resources using `owl:sameAs`. In this case it is possible to create exact matching high-quality interlinks.

However, the case of countries is an example where it is in particular possible to create exact matches because common identifiers in the two datasets allow for deterministic identification of the interrelations. Another example from the geographical domain

```
<Source dataset="riese" var="a">
  ?a rdf:type dimension:geo[NUTS0]
</Source>
<Target dataset="geonames" var="b">
  ?b gn:featureCode gn:A.PCLI
</Target>
<LinkingSpec property="owl:sameAs">
  <Match>
    <Feature1 res="?a/dimension:geo[NUTS0]" />
    <Feature1 res="?b/gn:countryCode" />
  </Match>
</LinkingSpec>
```

Listing 10.1: Mapping configuration snippet for creating exact country links based on ISO-3166 alpha2

already shows that the creation of interlinks is not that straightforward. As the example in figure 10.1 shows, the toponym "Vienna" can have different meanings. By only looking at the label of the geo-location it is not possible to identify the corresponding entry for instance in GeoNames.

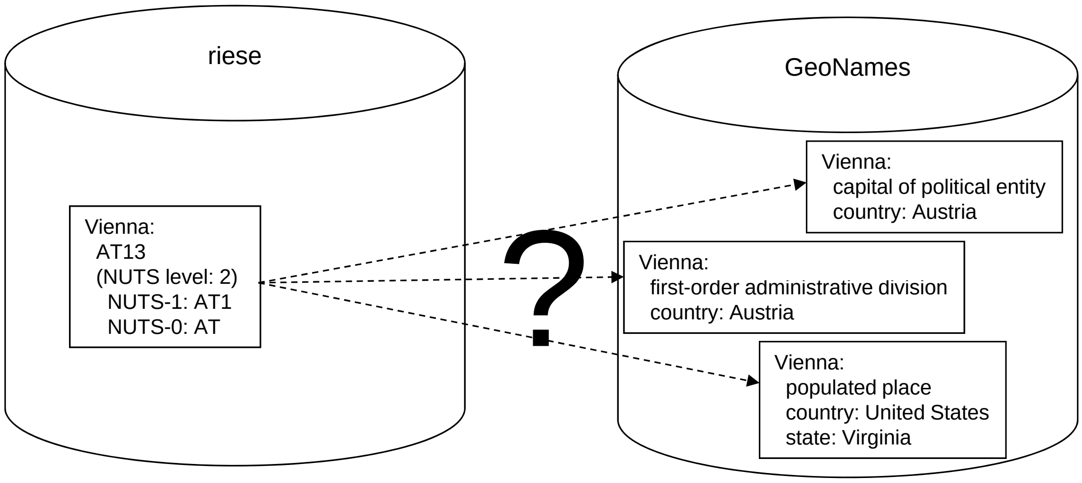


Figure 10.1: Matching Vienna from riese to GeoNames.

In GeoNames different entries labeled with "Vienna" exist. Doing a simple literal match or comparison does not work in such a case. Around the world several geo-locations carry this name (in the figure just a small subset is visualized for demonstration purposes). Even in Austria different entries exist which is due to the ambiguity of the toponym "Vienna" for Austria alone. There exists the city named "Vienna" which is also a state ("Bundesland") with the same name. However, especially in statistical applications it is paramount to reduce ambiguities and use exact identifications. An example of how this issue can be addressed and how a exact interrelation can be created is shown in listing 10.2.

```
<Source dataset=”riese” var=”a”>
  ?a rdf:type dimension:geo
</Source>
<Target dataset=”geonames” var=”b”>
  ?b gn:featureClass gn:A
</Target>
<LinkingSpec property=”owl:sameAs”>
  <Similarity metric=”jaroWinkler”>
    <Feature1 res=”?a/rdfs:label” />
    <Feature2 res=”?b/gn:name” />
    <Threshold accept=”0.97” />
  </Similarity>
  <Match>
    <Feature1 res=”?a/dim:geo/parentFeature[NUTS0]” />
    <Feature1 res=”?b/gn:countryCode” />
  </Match>
  <Restrict>
    <Condition res=”?a/dim:geo/parentFeature[NUTS0]” val=”AT”>
      <Feature1>
        ?a rdf:type dimension:geo[NUTS2]
      </Feature1>
      <Feature2>
        ?b gn:featureCode gn:A.ADM1
      </Feature2>
    </Condition>
  </Restrict>
</LinkingSpec>
```

Listing 10.2: Mapping configuration snippet for creating further geo links in riese

The mapping configuration in this example can be used to resolve and interrelate several geographical features from riese to GeoNames - for simplification only the most relevant aspects are shown. At first, the source and target resources are identified. Then the linking specification defines how related resources can be found which are equivalent and will be linked using the property `owl:sameAs`. Note, that also other properties and specifications can be used to interrelate different datasets. We use a similarity metric, in this case the Jaro-Winkler distance to consider string similarities with typographical variations, and define a threshold. Even though we can expect that entries in both datasets use the same naming, this example also allows the identification of related resources where slight naming variations exist. Further on a `Match` condition is specified where we take additional information about the resource into account. A matching country reference is required which already greatly reduces potential ambiguity. Considering the example shown in figure 10.1 this would result in discarding potential matches from GeoNames that are not in Austria. Finally the `Restrict` specifications show that we can introduce additional conditions on how to disambiguate potential matches from the target dataset. For instance when looking at riese entries from Austria we specify that NUTS level 2 features relate to resources in geonames that are of `gn:featureCode` with the type `gn:A.ADM1` (first-order administrative division) which represents a primary administrative division of a country, such as a state. This exact mapping is for instance also defined in the NUTS correspondence specifications. Additional restrict conditions can also be added such as for instance shown in listing 10.3 for Germany where Eurostat NUTS level 1 entries relates to states which are first-order administrative divisions in GeoNames.

```
    <Condition res="?a/dim:geo/parentFeature[NUTS0]" val="DE">
      <Feature1>
        ?a rdf:type dimension:geo[NUTS1]
      </Feature1>
      <Feature2>
        ?b gn:featureCode gn:A.ADM1
      </Feature2>
    </Condition>
```

Listing 10.3: Additional configuration snippet for restrict conditions

Using these mapping specifications we were able to create exact interrelations between geographical dimensions in riese and geo-related Linked Data sets such as GeoNames. For many categories there exist verbal definitions of mappings from NUTS regions to geographical regions and we configured our mapping accordingly. However, as regions for statistical reporting in NUTS do not always have a corresponding general geographical match we could not create exact interrelations for all NUTS regions but also introduced approximate matches which have been semantically described. For the matching of geographical references we also considered additional features that are available (at least with approximate figures) in both the target and the source dataset such as population counts or area sizes.

We also created further interlinks to datasets such as DBpedia, CIA Factbook and Wikicompany. The case of geographical interlinks has been described above. For additional categories we primarily considered riese dimensions for the interlinking but also looked at entire riese Datasets. For most of the non-geographic interlinks we considered labels of the resource along with category information. For instance for the riese dimension `dimension:aircraft` we required potential DBpedia matches to be of `rdf:type dbpedia-owl:Aircraft`. If we only looked up the label of e.g., *A300* this could result in DBpedia in several resources such as the A300 road in Great Britain or the desired Airbus aircraft A300. With the additional category information we were able to create more appropriate results. For some cases it is possible to find exact `owl:sameAs` matches whereas in some cases exact coreference resolution is not possible but instead reference enhancements can be discovered.

By introducing these interlinks users of riese not only benefit from a larger interlinked dataspace but also by being able to produce even more flexible and powerful queries. In the following section 10.3 we present our further work towards automated interlinking that we have applied in Link2WoD.

10.3 Interlinking in Link2WoD

In the context of the Link2WoD demonstrator we have implemented and integrated a *Linked Data consumption & interlinking* module which takes extracted entities from an unstructured source as an input. These entities are supplied from the *term extraction & classification* module (cf. section 7.2). We have already partly reported about the interlinking in Link2WoD already in the paper “Towards a Commercial Adoption of Linked Open Data for Online Content Providers” [74] which has been written together with Alexander Stocker, Harald Mayer, Helmut Mülner, and Ilir Ademi. It has been presented by the author of this thesis at the 6th International Conference on Semantic Systems (I-SEMANTICS 2010) in Graz. The author of this thesis was the research and development lead of Link2WoD and contributed most parts of the work.

The *Linked Data consumption & interlinking* module retrieves additional information about the identified terms and creates interlinks to Linked Data sources. Disambiguation between different concepts that share the same label is one of the biggest challenges. The term “Berlin” can for instance refer to the German capital or any of the other cities called “Berlin” (there exist approximately 100 places with that name globally). Another example is the term “Obama” which may refer to the American president, his wife, or a city in Japan.

The general workflow of the module starts with a search for potentially relevant information in Linked Data sources. In the first phase this is done via a query to DBpedia and GeoNames, two datasets that already contain a lot of general knowledge and geographic information. Even though the use of further datasets is easily possible, this first step aims at identifying relevant concepts for the found terms where these two Linked Data hubs already provide sufficient information for general news articles. The query is based on the method developed for riese (cf. section 10.2) which uses the extracted terms and takes the original document's language into account as well as textual similarity measures (e.g., Levenshtein distance [115] or Jaro-Winkler distance [182]). However, due to the nature of the input data it is not possible to use certain features of our riese approach: The input consists of terms that do not carry much structured additional information except for potentially a categorization as person, toponym, organization. Further information is not available and therefore specific conditions in the `Match` and `Restrict` section of the Linking specification cannot be evaluated. This approach works nevertheless where unambiguous matches in the target datasets can be found.

In some cases it is not possible to find only one distinct concept but rather many potential matches are found and thus disambiguation needs to be done. As mentioned above, in the current use case there is almost no structured additional information about the term available directly as it is simply part of a news article. This implies that approaches for interrelating Linked Data that have been discussed before such as our interlinking approach in riese, SILK [167] or the Consistent Reference Service (CRS) [98, 99] cannot be applied in this context. These approaches rely on further information about a resource that can be compared in two different datasets, i.e. there needs to be some overlap between the source and target datasets which is not the case in our scenario where the source dataset contains no more structured information about a concept than its label.

The information about the category of the article (e.g. politics, local news, etc.) can aid in narrowing potential concepts though. Additional information that is available and can be exploited is the co-occurrence of terms in one article. By looking at all potential concepts for all extracted terms contained in one document it is possible to construct what we call a *Linked Data Entity Space*. Relations between ambiguous concepts are modeled which allows to identify concepts that are most closely related, i.e. that have the shortest conceptual distance. Figure 10.2 exemplifies this concept for the case of the identified terms "Obama" and "United States" where for the term "Obama" more potential matches can be found:

To construct the *Linked Data Entity Space* we take the following approach:

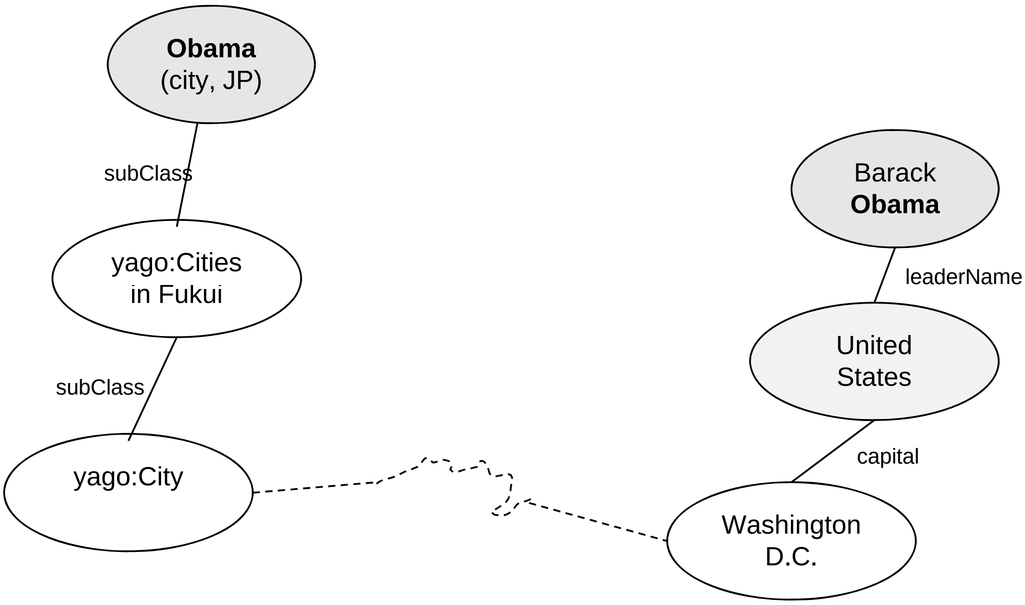


Figure 10.2: Linked Data Entity Space exemplification.

1. Try to resolve all identified terms from the term extraction in DBpedia and store references to resources that can be retrieved with a given confidence.
2. Store the terms that are ambiguous and cannot directly be resolved with the required confidence.
3. Identify the smallest graph of potential matches and already identified resources.
4. Select the concepts that are most closely related.

For the selection of potential matches we restrict the potential matches to a certain amount of n potential matches where we have used $n = 5$ for our Link2WoD demonstrator which achieved good results in terms of accuracy of the results and computational performance. The top selected matches are based on the *entity rank* feature that is available in Openlink Virtuoso which can also be used in SPARQL queries. The term *entity rank* refers to a quantity describing the relevance of a URI in an RDF graph. It is defined by the count of references to a given subject in Virtuoso, weighed by the rank of the referrers and the outbound link count of referrers (cf. [133]). The ranking is precomputed for a given RDF graph and can be accessed in a SPARQL query via `<LONG::IRI_RANK> (?s)`. Note that is not standard SPARQL though but an extremely helpful proprietary feature of Virtuoso.

The identification of the smallest graph for a given set of ambiguous terms was also realized via dedicated SPARQL queries to retrieve the distance (i.e. number of edges) between already identified resources and potential matches. To compute this figure we utilize again the capabilities of Virtuoso which allows an arbitrary subquery to be made transitive (cf. [51]). This way we can simply identify the distance between two given resources which is efficiently implemented in Virtuoso. A maximum number of transitive steps can be defined to reduce compute time and we found that a maximum number of 3 steps (i.e. looking up to a maximum distance in the graph of 3) delivered good accuracy and acceptable computational performance. In general, OWL allows a property to be defined as transitive. For a situation where we have the triples `{ S1 P O1 }` and `{ O1 P O2 }` and `P` being transitive, the fact `{ S1 P O2 }` can be implied. In Virtuoso it is even possible to consider arbitrary properties in the query as transitive. The use of further complex conditions to define relatedness is also supported. These features considerably ease the effort required to run advanced analyses.

For toponyms geographic distance metrics are also taken into account. Toponym disambiguation based on geographic distances is extremely powerful for certain categories such as local news. If different ambiguous geo-locations for a given term are discovered we use again one of the features natively available in Virtuoso to compute the geographical distance between the potential geographical matches under consideration. The built-in function `st_distance` (cf. [134]) can retrieve the distance between two given resources that carry geographical information. As an example consider we have a given an identified toponym of "Munich" (as `dbpedia:Munich`) and discovered ambiguous matches for "Berlin" (e.g., `dbpedia:Berlin` and `dbpedia:Berlin,_Connecticut`). The following example query in listing 10.4 can directly be used at a Virtuoso instance that hosts the DBpedia dataset (e.g., at `http://dbpedia.org/sparql`):

```
SELECT (bif:st_distance(?loc1,?loc2)), (bif:st_distance(?loc1,?
   loc3))
WHERE
  {
    <http://dbpedia.org/resource/Munich>
      geo:geometry ?loc1 .
    <http://dbpedia.org/resource/Berlin>
      geo:geometry ?loc2  .
    <http://dbpedia.org/resource/Berlin,_Connecticut>
      geo:geometry ?loc3 .
  }
```

Listing 10.4: Query example for retrieving geographic distances between resources.

This query will return that the distance from `dbpedia:Munich` to `dbpedia:Berlin` is only 502 kilometer whereas from `dbpedia:Munich` to `dbpedia:Berlin,_Connecticut` is 6341 kilometers. By using the results from these analyses we are able to select the geographical resources that are geographically closely related.

To ensure stable performance we did not rely for these queries on remote SPARQL endpoints of DBpedia which are out of our control and can behave differently based on demand by other third parties or network traffic. We installed a local instance of Virtuoso containing a copy of the DBpedia data and synchronize the data continuously. This allows us to achieve a stable performance in a system that is our under control.

Further on, when the concept has been identified it is possible to retrieve further information from different data sources. The information gained from Linked Data sources can also be extremely useful for queries to other repositories containing for instance user contributed multimedia content. It is possible to find more appropriate matches as the query can be enhanced with more metadata for finding relevant images and videos. Especially for concepts related to geographical locations it is possible to supply coordinates that have been retrieved from DBpedia or GeoNames in the query when searching for images on Flickr or videos on Youtube.

An evaluation of our approach is discussed in section 13.5 along with a discussion of how this data can be consumed in the media industry in general and for the Link2WoD use case specifically.

10.4 Summary

This chapter focused on automated approaches and therefore addressed the research question *How can links be created automatically?* Some general strategies stemming from the database area were introduced and a brief overview of relevant metrics and techniques was given. Specifics of Linked Data were discussed and our approaches that we applied in riese and *Link2WoD* have been presented. In riese we used a mapping approach that is based on a manual specification. It is configured by defining potential source and target resources along with a linking specification containing similarity metrics and conditional restrictions. Using these specifications we were able to create exact interrelations between geographical dimensions in riese and geo-related Linked Data sets such as GeoNames. We also created further interlinks to datasets such as DBpedia, CIA Factbook and Wikicompany. As part of the Link2WoD demonstrator we have implemented and integrated a *Linked Data consumption & interlinking* module which takes extracted entities from an unstructured source as an input. For the disambiguation of entities we introduced what we call a *Linked Data Entity Space*. Relations between ambiguous concepts are modeled which allows to identify concepts that are most closely related, i.e. that have the shortest conceptual distance. For toponyms geographic distance metrics were also taken into account. We discovered that toponym disambiguation based on geographic distances is extremely powerful for certain categories such as local news.

Part IV

Consuming Linked Data

Chapter 11

General Considerations on Consumption

After having discussed the creation of linkable data and interrelation of Linked Data in the previous sections we now consider the consumption of Linked Data in this final part of the conceptual framework for Linked Data generation and utilization. In the past years significant amounts of Linked Data have been made available and there exist different possibilities for exploiting this large pool of Linked Data. A basic distinction of two major Linked Data consumption schemes can be made between the main purpose of a Linked Data application which may primarily target

- machine interpretable consumption of the data and/or
- human users targeted visualizations of Linked Data sources.

It is quite apparent that a plain representation of Linked Data in RDF/XML or some other equivalent format is not well suited for human users. Instead, Linked Data applications can make use of this machine interpretable information base and use this data for solving tasks that in the end support human tasks. Linked Data on a plainly technical level can therefore be considered as a backbone and end-user applications take care of human-friendly visualizations. As an analogy, also relational databases are driving many websites on the traditional Web of Documents but in the end appealing visualizations are created by web applications - human users do not have to query the relational database directly or are confronted with plain database dumps. Instead, traditional Web applications use all the data in the background and only display what is relevant for a human. However, for programmatic access by machines the data contained in a relational database can be offered via web services or specialized APIs. Similarly, in the Web of Data, i.e. the Linked Data environment, the plain data should be used in the background and also for machine access the pure RDF data along with access mechanisms such as via SPARQL endpoints is very valuable. Developers will want to have access to the underlying data and Linked Data offers all the mechanisms to support the development of effective Linked Data applications that can exploit the combined power of interlinked data sets. For the average mainstream end-user, these underlying technical details are usually not of interest but a convenient and user-friendly visualization is desired. In the end, it might even be the case that a Linked Data application behaves like a traditional Web application except for the Linked Data application to use more sophisticated data sources and provide more appropriate information.

A rough differentiation of Linked Data applications can also be made in the following categories:

- Plain RDF Linked Data providers,
- Linked Data browsers,
- Linked Data search engines, and
- specific end-user applications.

It might also be the case that a Linked Data application combines features from some of the categories. Our riese demonstration application for instance is targeted at both machine and human consumption. Especially with the XHTML+RDFa approach employed in riese it is possible to serve both user groups with the same source. Machine users can extract the related RDF from the page and human users are provided with a convenient HTML visualization inside their web browser. From the four categories of Linked Data applications listed above the first three mainly target developers (for inclusion and reuse of resources) and users with advanced technical skills. Specific end-user applications usually can be designed in such a user-friendly way that they are also suitable for inexperienced human users. The different categories of Linked Data applications are explained below:

Plain RDF Linked Data providers At the core of Linked Data is the data itself. To be considered a Linked Data set a data provider should follow the Linked Data principles [14]. Third parties can exploit this data in their general-purpose or domain-specific applications. Access to the data can be given on the base of individual URIs, via a SPARQL endpoint or as a complete dump of the dataset. APIs are also offered in some cases that allow easy integration of the data in other web application projects. Generic developer tools for interacting with Linked Data are also available for the major programming languages.

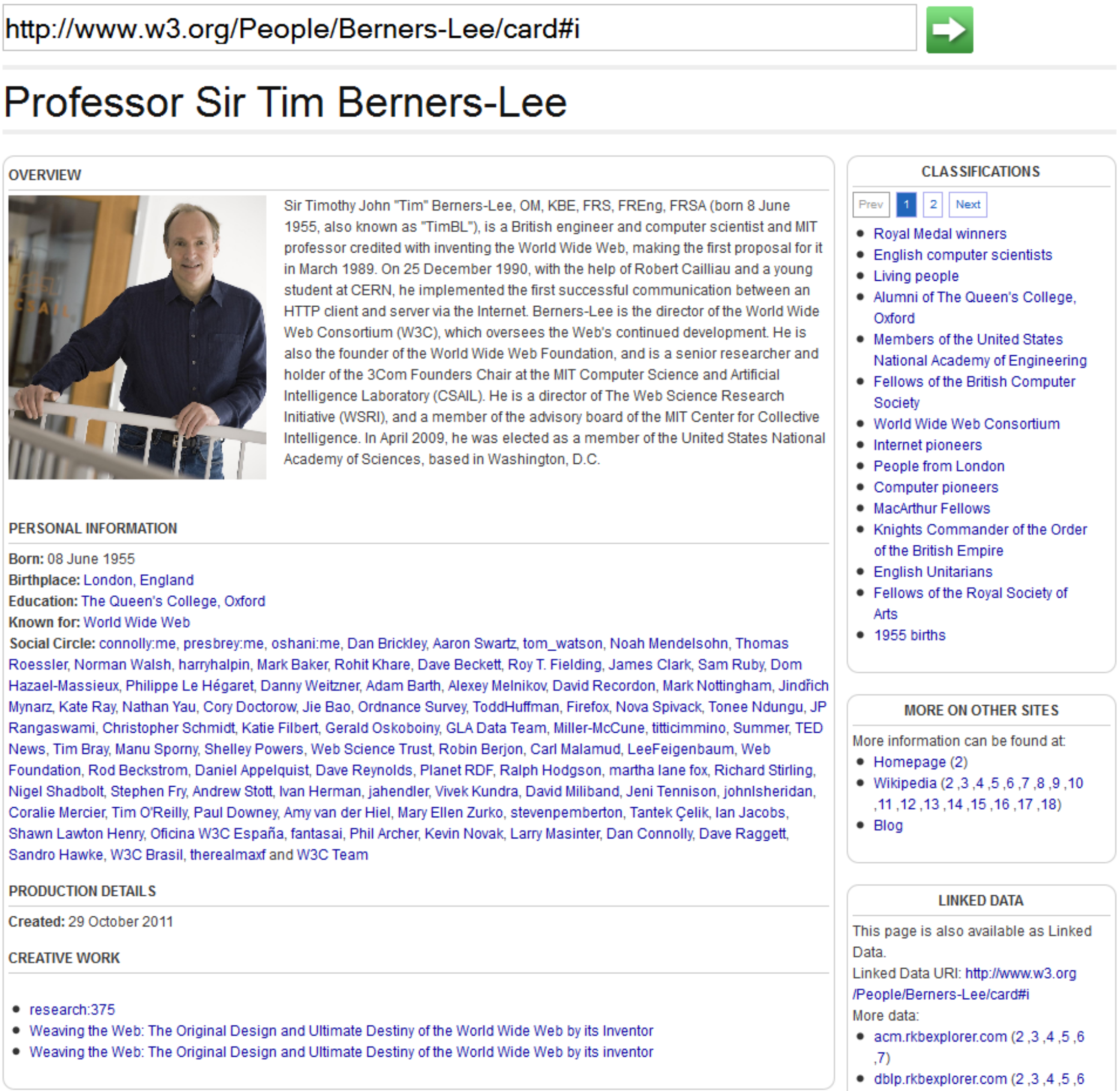
Linked Data browsers Similar to traditional Web browsers these Linked Data browsers allow to view Linked Data that is provided by plain Linked Data providers. Instead of rendering HTML such browsers visualize RDF data in a human-readable way and users can navigate between different Linked Data sources. One of the first browsers of this kind is the Tabulator [17] which originated from Tim Berners-Lee's idea in 2005. Tabulator is available as extension for the Firefox browser and allows the viewing of resources. It also includes an experimental feature to edit resources and automatically send SPARQL updates. Further browsers are for instance Marbles [12], LinkSailor[1], or Sig.ma [162] which render Linked Data visualizations on the server side and can thus be used in (almost) every browser without installing additional software.

Both LinkSailor and Sig.ma can also be considered Linked Data aggregation or mash-up engines. An exemplary visualization of a resource presented through LinkSailor is shown in figure 11.1. LinkSailor displays information about a Linked Data resource and also integrates information from other sources it links to. It is provided by Talis and uses the Talis platform. The aggregated view has a clean user interface but lacks provenance information, i.e. it is not directly shown which resource contains the original data.

Sig.ma was created as a demonstration of live, on the fly Web of Data mashup. The Sig.ma mashup engine combines information by following links from a given resource and also includes other related information it retrieves on the Web of Data. Information is also served via the Sindice search engine and Sig.ma evaluates data that is available as RDF but also data that is embedded in Web pages, e.g. as RDFa or using some other microformats. In the result view the different data sources are listed and the mashup view provides provenance information about where the information displayed originates from.

Linked Data search engines Different search engines for Linked Data have been developed which crawl Linked Data and harvest data by applying the follow-your-noise principle, i.e. following RDF links. Similar to traditional search engines most Linked Data search engines allow a keyword-based query but the Linked Data versions allow

[1] http://linksailor.com



http://www.w3.org/People/Berners-Lee/card#i

Professor Sir Tim Berners-Lee

OVERVIEW

Sir Timothy John "Tim" Berners-Lee, OM, KBE, FRS, FREng, FRSA (born 8 June 1955, also known as "TimBL"), is a British engineer and computer scientist and MIT professor credited with inventing the World Wide Web, making the first proposal for it in March 1989. On 25 December 1990, with the help of Robert Cailliau and a young student at CERN, he implemented the first successful communication between an HTTP client and server via the Internet. Berners-Lee is the director of the World Wide Web Consortium (W3C), which oversees the Web's continued development. He is also the founder of the World Wide Web Foundation, and is a senior researcher and holder of the 3Com Founders Chair at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). He is a director of The Web Science Research Initiative (WSRI), and a member of the advisory board of the MIT Center for Collective Intelligence. In April 2009, he was elected as a member of the United States National Academy of Sciences, based in Washington, D.C.

PERSONAL INFORMATION

Born: 08 June 1955
Birthplace: London, England
Education: The Queen's College, Oxford
Known for: World Wide Web
Social Circle: connolly:me, presbrey:me, oshani:me, Dan Brickley, Aaron Swartz, tom_watson, Noah Mendelsohn, Thomas Roessler, Norman Walsh, harryhalpin, Mark Baker, Rohit Khare, Dave Beckett, Roy T. Fielding, James Clark, Sam Ruby, Dom Hazael-Massieux, Philippe Le Hégaret, Danny Weitzner, Adam Barth, Alexey Melnikov, David Recordon, Mark Nottingham, Jindřich Mynarz, Kate Ray, Nathan Yau, Cory Doctorow, Jie Bao, Ordnance Survey, ToddHuffman, Firefox, Nova Spivack, Tonee Ndungu, JP Rangaswami, Christopher Schmidt, Katie Filbert, Gerald Oskoboiny, GLA Data Team, Miller-McCune, titticimmino, Summer, TED News, Tim Bray, Manu Sporny, Shelley Powers, Web Science Trust, Robin Berjon, Carl Malamud, LeeFeigenbaum, Web Foundation, Rod Beckstrom, Daniel Appelquist, Dave Reynolds, Planet RDF, Ralph Hodgson, martha lane fox, Richard Stirling, Nigel Shadbolt, Stephen Fry, Andrew Stott, Ivan Herman, jahendler, Vivek Kundra, David Miliband, Jeni Tennison, johnlsheridan, Coralie Mercier, Tim O'Reilly, Paul Downey, Amy van der Hiel, Mary Ellen Zurko, stevenpemberton, Tantek Çelik, Ian Jacobs, Shawn Lawton Henry, Oficina W3C España, fantasai, Phil Archer, Kevin Novak, Larry Masinter, Dan Connolly, Dave Raggett, Sandro Hawke, W3C Brasil, therealmaxf and W3C Team

PRODUCTION DETAILS

Created: 29 October 2011

CREATIVE WORK

- research:375
- Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor
- Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its inventor

CLASSIFICATIONS

- Royal Medal winners
- English computer scientists
- Living people
- Alumni of The Queen's College, Oxford
- Members of the United States National Academy of Engineering
- Fellows of the British Computer Society
- World Wide Web Consortium
- Internet pioneers
- People from London
- Computer pioneers
- MacArthur Fellows
- Knights Commander of the Order of the British Empire
- English Unitarians
- Fellows of the Royal Society of Arts
- 1955 births

MORE ON OTHER SITES

More information can be found at:

- Homepage (2)
- Wikipedia (2 ,3 ,4 ,5 ,6 ,7 ,8 ,9 ,10 ,11 ,12 ,13 ,14 ,15 ,16 ,17 ,18)
- Blog

LINKED DATA

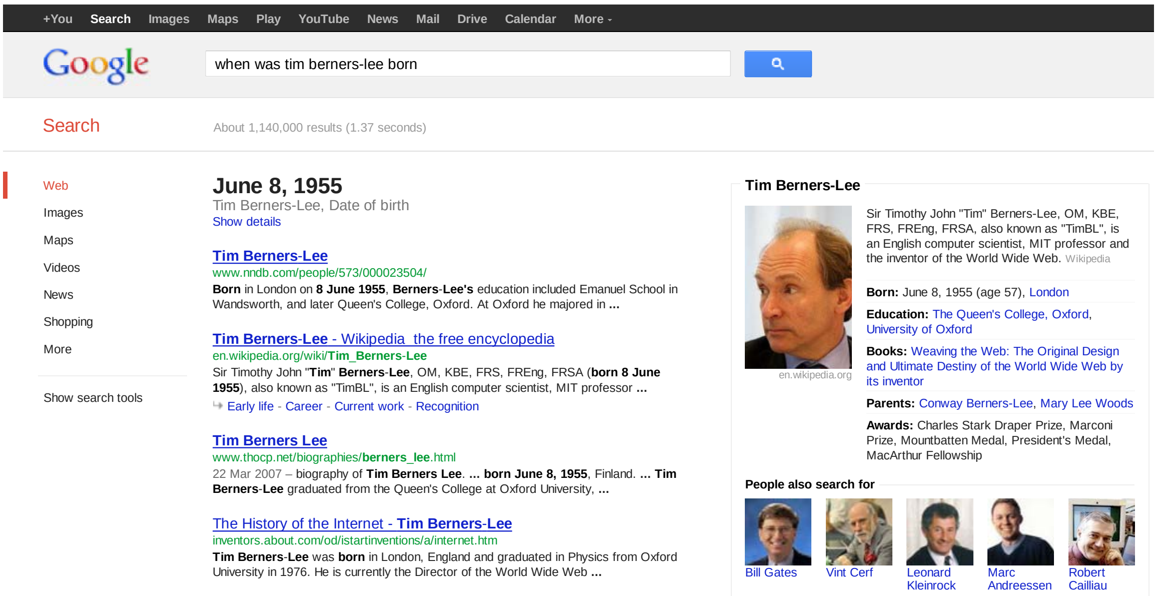
This page is also available as Linked Data.
Linked Data URI: http://www.w3.org /People/Berners-Lee/card#i
More data:

- acm.rkbexplorer.com (2 ,3 ,4 ,5 ,6 ,7)
- dblp.rkbexplorer.com (2 ,3 ,4 ,5 ,6

Figure 11.1: LinkSailor example screenshot displaying information about Tim Berners-Lee.

more interaction with the retrieved data. Search results from different sources can be combined and filtered individually. Sig.ma presented above can also be considered a Linked Data search engine. Further engines include Falcons [34] or the Semantic Web Search Engine (SWSE) [88] which provide user interfaces that are also targeted at human consumption. In addition, search engines or Semantic Web indexes like Sindice [163] also provide APIs that can be used by applications to access resources that reference a given keyword or URI.

The benefits of structured data and semantically enhanced information are also exploited at "traditional" search engines such as Google. In 2012 for instance, Google introduced the Knowledge Graph [154] which makes use of structured data such as from Freebase or Wikipedia and also takes user search queries into account. An example search result is shown in figure 11.2 where a box on the right side displays extracted structured information. This is combined with earlier announced features such as Rich Snippets [65] or Google Squared [121]. This also follows the support of structured information by big search engine players where for instance in 2011 Google, Bing, and Yahoo! have jointly launched schema.org [70] to create and support a common vocabulary for structured data markup on web pages. With all these steps the Semantic Web and Linked Data vision also find an entrance to mainstream web usage.



when was tim berners-lee born

About 1,140,000 results (1.37 seconds)

June 8, 1955
Tim Berners-Lee, Date of birth
Show details

Tim Berners-Lee
www.nndb.com/people/573/000023504/
Born in London on **8 June 1955**, **Berners-Lee's** education included Emanuel School in Wandsworth, and later Queen's College, Oxford. At Oxford he majored in **...**

Tim Berners-Lee - Wikipedia the free encyclopedia
en.wikipedia.org/wiki/**Tim_Berners-Lee**
Sir Timothy John "**Tim**" **Berners-Lee**, OM, KBE, FRS, FREng, FRSA (**born 8 June 1955**), also known as "TimBL", is an English computer scientist, MIT professor **...**
Early life - Career - Current work - Recognition

Tim Berners Lee
www.thocp.net/biographies/**berners_lee**.html
22 Mar 2007 – biography of **Tim Berners Lee**. **...** **born June 8, 1955**, Finland. **...** **Tim Berners-Lee** graduated from the Queen's College at Oxford University, **...**

The History of the Internet - **Tim Berners-Lee**
inventors.about.com/od/istartinventions/a/internet.htm
Tim Berners-Lee was **born** in London, England and graduated in Physics from Oxford University in 1976. He is currently the Director of the World Wide Web **...**

Tim Berners-Lee

Sir Timothy John "Tim" Berners-Lee, OM, KBE, FRS, FREng, FRSA, also known as "TimBL", is an English computer scientist, MIT professor and the inventor of the World Wide Web. Wikipedia

en.wikipedia.org

Born: June 8, 1955 (age 57), London

Education: The Queen's College, Oxford, University of Oxford

Books: Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its inventor

Parents: Conway Berners-Lee, Mary Lee Woods

Awards: Charles Stark Draper Prize, Marconi Prize, Mountbatten Medal, President's Medal, MacArthur Fellowship

People also search for

Bill Gates
Vint Cerf
Leonard Kleinrock
Marc Andreessen
Robert Cailliau

Figure 11.2: Google Knowledge Graph example result for Tim Berners-Lee.

Specific end-user applications Apart from general-purpose approaches for the consumption of Linked Data such as via Linked Data browsers or search engines there also exist specific applications for certain domains. Recently, several applications in the domain of government data have emerged in the course of open government data initiatives around the globe but there also exist applications for various other domains that exploit Linked Data and provide a user-friendly interface. One of the government data example applications is the US Global Foreign Aid mashup [44] shown in figure 11.3. It is part of the TWC LOGD Portal that supports the deployment of Linked Open Government Data (LOGD) and provides different mashups. The mashup shows foreign aid data from the United States Agency for International Development (USAID), the

Department of Agriculture and the Department of State, mashed up with information from the New York Times API and CIA World Factbook.

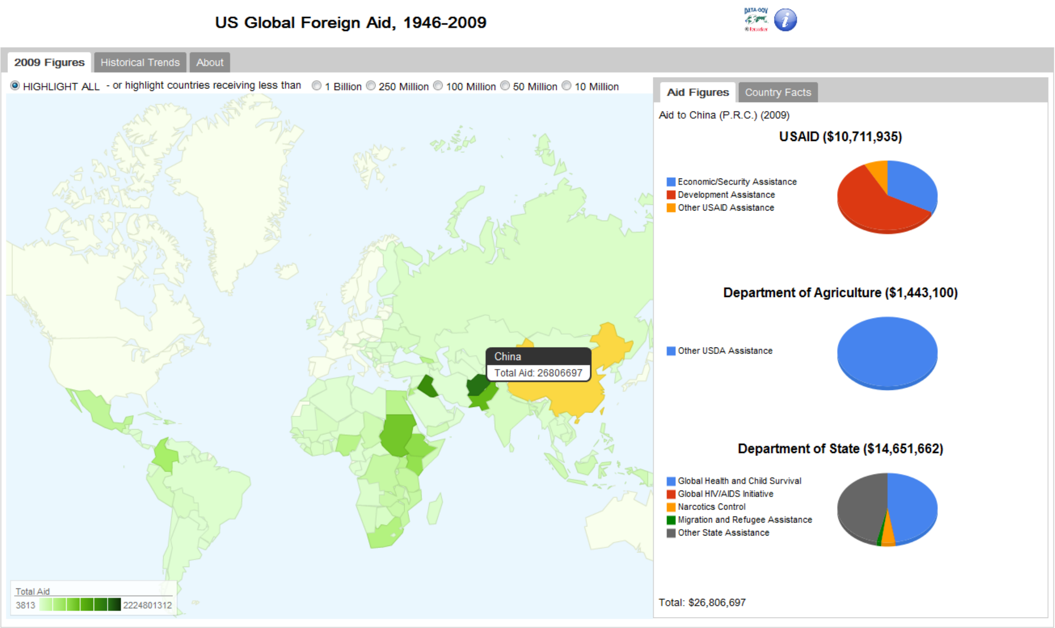

US Global Foreign Aid, 1946-2009
2009 Figures
Historical Trends
About
● HIGHLIGHT ALL - or highlight countries receiving less than ○ 1 Billion ○ 250 Million ○ 100 Million ○ 50 Million ○ 10 Million
China
Total Aid: 26806697
Total Aid
3813
2224801312
Aid Figures
Country Facts
Aid to China (P.R.C.) (2009)
USAID ($10,711,935)
Economic/Security Assistance
Development Assistance
Other USAID Assistance
Department of Agriculture ($1,443,100)
Other USDA Assistance
Department of State ($14,651,662)
Global Health and Child Survival
Global HIV/AIDS Initiative
Narcotics Control
Migration and Refugee Assistance
Other State Assistance
Total: $26,806,697

Figure 11.3: US Global Foreign Aid mashup - example screenshot.

Another prominent use case are the websites of the British Broadcasting Corporation (BBC) for BBC Music, BBC Programmes and the BBC Wildlife Finder (cf. [108, 139]) which are driven by and also provide Linked Data. The human end-user benefits from a website that contains interrelated information from different sources within the BBC and also from external datasets. Historically, the individual parts of the BBC were not well connected and information has been duplicated at various places. With the newly employed approach the websites have been based on a Linked Data framework where every item that the BBC has an interest in was assigned an individual URI. Internally to the BBC these identifiers are used to manage information for the website but also externally to link to datasets such as DBpedia, Musicbrainz, or GeoNames. Users can consume a HTML view but for developers also different representations of the underlying information are available for instance as RDF/XML.

In this chapter we have discussed general aspects of Linked Data consumption and presented some examples. In the following chapters we focus on the solutions that have been developed in the course of this thesis.

Chapter 12

Government Data

Government data is the domain of one of our two major use cases. With riese we have made EuroStat statistics available as Linked Data and the considerations of creating linkable data and interrelating it have been discussed in previous sections already. In general, Linked Data has seen quite widespread adoption in the area of government data which is also driven by Open Government Data (OGD) initiatives around the world. In Austria the Open Government Data Austria initiative aims at supporting and fostering such developments. In the United Kingdom the first large-scale application of Linked Data for OGD was deployed. Sir Tim Berners-Lee among others assisted the UK government in the realization of OGD and along these efforts also the 5-star concept [14] of Linked Open Data has been articulated (cf. also section 5) to encourage (governmental) data providers.

Linked Open Data and its surrounding technologies are well suited for the publication and consumption of government data as they can cater to the needs of all involved stakeholders. In the following section 12.1 we discuss some general aspects regarding the consumption of governmental Linked Open Data and in section 12.2 finally present how our riese demonstrator can be used for Linked Data consumption.

12.1 Consuming Governmental Linked Data

The consumption of governmental Linked Data follows in principle the same approaches as for the consumption of general-purpose Linked Data. Minor differences exist in the nature of the provided data as government data in many cases contains statistical data and in the detailed definition of target audiences. The main audience groups of governmental Linked Data and general consumption characteristics are provided below.

General public The general public is likely interested in accessing the data in a user-friendly way inside their web browser. A HTML rendering of the data will be most suitable. Information should be easily retrievable and also combine relevant information sources. With Linked Data the underlying information can be organized and powerful queries are also supported. Detailed data can be made available for interested users as well.

Developers Developers and tech-savvy users can get direct access to the data via the multiple access methods supported by Linked Data. New applications can be built that exploit the provided data and combine it with other sources. Several contests for creating applications based on open government data have already been held to stimulate developers.

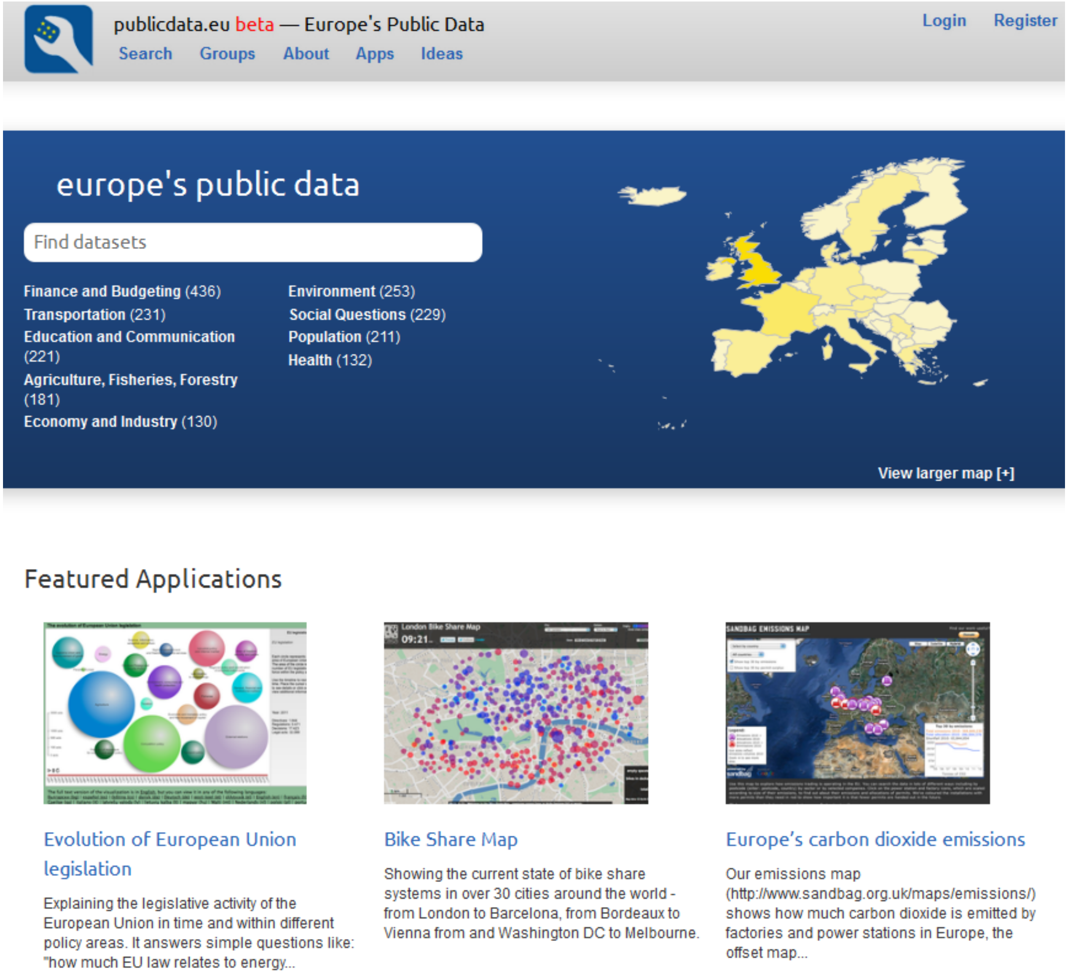
Public administration users Linked Data can also bring benefits for the public administration in creating data that is better reusable and increasing the interoperability of data between different governmental organizations.

Sir Tim Berners-Lee [15] has suggested to use Linked Data as the "interconnection bus" for open government data and named three typical reasons for putting government data online which are all best served by using Linked Data techniques [15]:

1. Increasing citizen awareness of government functions to enable greater accountability;
2. Contributing valuable information about the world; and
3. Enabling the government, the country, and the world to function more efficiently.

He further emphasized the importance of using standards such as RDF or SPARQL and suggests the use of persistent URIs following the Linked Data principles. Regarding the overall approach he admits that a top-down approach where national strategies define open government data rules can be quite time-consuming and he therefore suggests a bottom-up procedure where suitable datasets should just be published as a start. Prominent examples such as `data.gov.uk` in the United Kingdom or `data.gov` in the United States already follow a national strategy and publish significant volumes of Linked Data.

With the first beta release of `publicdata.eu` [117] in 2011 a portal was launched that indexes data from different open government data initiatives around Europe. A screenshot of the portal is shown in figure 12.1. National level data is for instance included from the United Kingdom (`data.gov.uk`), France (`data-publica.com`), Sweden (`opengov.se`) and on regional level for instance from Vienna (`data.wien.gv.at`), Paris (`opendata.paris.fr`), or the Italian Piedmont region (`dati.piemonte.it`). Additional data is continuously being integrated. Most notably, the portal supports Linked Data and provides data as RDF as well as a SPARQL endpoint for querying the data catalog. It is developed by the Open Knowledge Foundation and is based on its open-source data portal platform CKAN (the Comprehensive Knowledge Archive Network). CKAN also catalogs almost all Linked Data sets. Several applications have already been built based on publicly available data and many of these applications are featured on the portal.



europe's public data

Finance and Budgeting (436)
Transportation (231)
Education and Communication (221)
Agriculture, Fisheries, Forestry (181)
Economy and Industry (130)
Environment (253)
Social Questions (229)
Population (211)
Health (132)

Featured Applications

Evolution of European Union legislation

Explaining the legislative activity of the European Union in time and within different policy areas. It answers simple questions like: "how much EU law relates to energy...

Bike Share Map

Showing the current state of bike share systems in over 30 cities around the world - from London to Barcelona, from Bordeaux to Vienna from and Washington DC to Melbourne.

Europe's carbon dioxide emissions

Our emissions map (http://www.sandbag.org.uk/maps/emissions/) shows how much carbon dioxide is emitted by factories and power stations in Europe, the offset map...

Figure 12.1: Publicdata.eu portal entry page - screenshot.

Driven by several European initiatives such as the ones stated above, the European Commission has communicated in late 2011 a strategy [54] that supports open government data and the Commission has also announced a portal that will make data resources of the Commission and other European institutions and agencies easily accessible and usable. It is part of the Digital Agenda and the EU 2020 strategy to put Europe's economies onto a high and sustainable growth path. The Commission has announced measures to adapt the legal framework for data reuse, financially support open data and corresponding portals as well as fostering the information exchange between European Member States.

In the United States the `data.gov` effort has already followed the priority Open Government Initiative and makes data available that is generated and held by the Federal Government. It follows a standards based approach and contributes more than 6 billion triples [161] as Linked Data. Via access to the query points several applications have also been developed. As a joint effort from the United States and India the Open Government Platform (OGPL) [63] has been released as limited alpha version in May 2012 which should contribute to the global open government movement. The support of Linked Data technologies has been announced and this contributes to further uptake in the consumption of governmental Linked Data.

At the World Wide Web Consortium (W3C) several standardization efforts have been started. The Government Linked Data (GLD) Working Group [172] started in 2011 and aims at providing standards and other information for publishing governmental data effectively as Linked Data. Among the group's activities the *RDF Data Cube Vocabulary* has to be noted which takes up some of our work on modeling statistical data (cf. chapter 6). The group will also target best practice advice for governmental Linked Data publishing along with recommendations for further standard vocabularies useful in the government data domain. In parallel the eGovernment Interest Group has been chartered at W3C which has as mission "to build and strengthen the community of people who use or promote the use of W3C technologies to improve Government" [171]. It follows a broader scope than the GLD Working Group and also addresses issues outside of Linked Data but both groups coordinate their efforts within W3C's eGovernment activity.

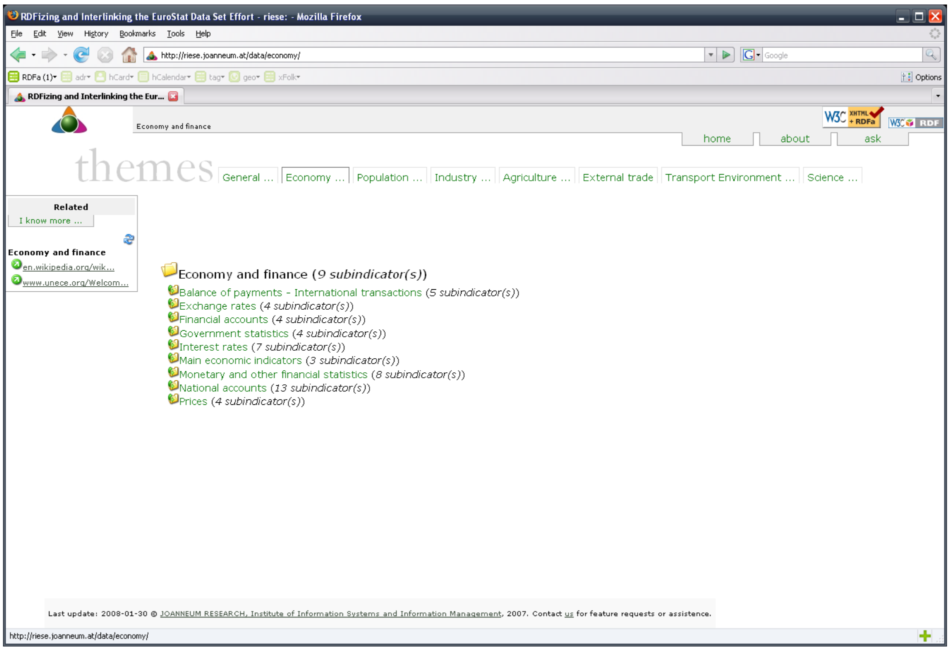
As this short overview of governmental Linked Data consumption shows, several activities have been started and quite widespread adoption of Linked Data principles for open government data can be witnessed. As developments are proceeding at a rapid pace both on the technology and policy level it will be exciting to follow the upcoming trends. In the following section 12.2 we discuss the usage of governmental Linked Data in our riese project which was one of the first governmental Linked Data sets.

12.2 Using riese

With the *RDFizing and Interlinking the Eurostat data Set Effort* (riese) we have contributed one of the first governmental Linked Data sets already in 2008. It aims at being useful for both humans and machines. While at the time of creating riese most other LOD datasets were targeted mainly at machine consumption we tried to satisfy both humans and machines. In chapter 6 we have already discussed how linkable data can be created from structured sources such as it is the case with the Eurostat statistical data. We have also presented the Statistical Core Vocabulary (SCOVO) which is a further development of the riese core schema for generally expressing statistical information and also discussed further related developments. In section 9.1 we have discussed *User Contributed Interlinking* (UCI) - a new approach that we introduced in riese which lets users interrelate Linked Data. Automated interlinking approaches in the context of riese have finally been discussed in section 10.2.

Here we finally present how Linked Data in riese can be consumed by both humans and machines. We have also reported partly about this in our paper "Building Linked Data For Both Humans and Machines" [73]. Both human and machine users would presumably start at the top-level page in order to get an overview of the available data. In Fig. 12.2 the hierarchical rendering of a selected Eurostat theme (the 'Economy' theme) is depicted.

A machine accessing the same page would have another view, namely focusing on the embedded RDF, exemplarily shown in listing 12.1. The data view is contained in a combined XHTML+RDFa representation where both the underlying data and the visual representation reside. XHTML+RDFa is a W3C Recommendation [175] and allows to embed rich metadata in the form of RDF into a web document. Through standalone RDF distilling applications the data can be extracted but also several RDF programming tools allow the automated use of RDF contained in XHTML+RDFa documents. Also search engines such as Google consider embedded RDFa data and use this information for extracting semantically rich data which can for instance be displayed in Google's Rich Snippets where the displayed search results already contain some additional relevant information. It is notable that although both humans and machines access the same resource, different parts are relevant. This is made possible through the deployment in XHTML+RDFa. The browser will render a nice user interface, the


RDFizing and Interlinking the EuroStat Data Set Effort - riese: - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://riese.joanneum.at/data/economy/
Google
RDFa (1) adr hCard hCalendar tag geo xFolk
Options
RDFizing and Interlinking the Eur...
W3C XHTML + RDFa
W3C RDF
home
about
ask
themes
General ...
Economy ...
Population ...
Industry ...
Agriculture ...
External trade
Transport Environment ...
Science ...
http://riese.joanneum.at/data/economy/

Economy and finance

Related

I know more ...

Economy and finance

en.wikipedia.org/wik...

www.unece.org/Welcom...

Economy and finance (*9 subindicator(s)*)

- Balance of payments - International transactions (*5 subindicator(s)*)
- Exchange rates (*4 subindicator(s)*)
- Financial accounts (*4 subindicator(s)*)
- Government statistics (*4 subindicator(s)*)
- Interest rates (*7 subindicator(s)*)
- Main economic indicators (*3 subindicator(s)*)
- Monetary and other financial statistics (*8 subindicator(s)*)
- National accounts (*13 subindicator(s)*)
- Prices (*4 subindicator(s)*)

Last update: 2008-01-30 © JOANNEUM RESEARCH, Institute of Information Systems and Information Management, 2007. Contact us for feature requests or assistence.

Figure 12.2: The Eurostat theme ’Economy’ viewed by a human user.

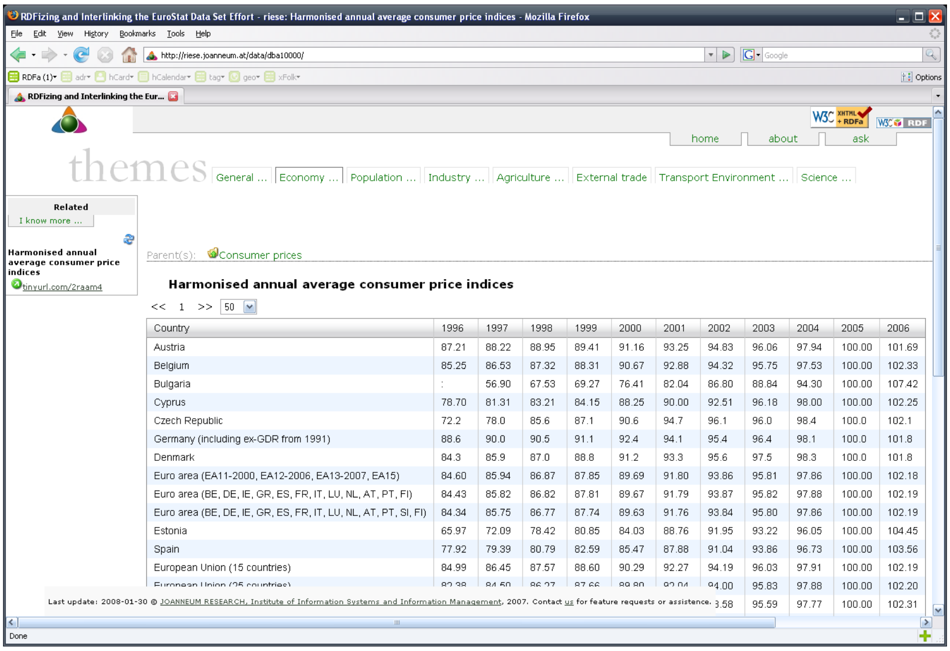
machine gets what it deserves: triples.

```
<body
 about=”http://riese.joanneum.at/data/economy”
 instanceof=”riese:Dataset”>
 ...
<div id=”main-ind”>
  ...
 <a href=”http://riese.joanneum.at/data/bop”
    rel=”skos:narrower”>
 Balance of payments - International transactions
 </a>
</div>
```

Listing 12.1: The Eurostat theme ’Economy’ viewed by a machine.

With the development of the HTML5 standard [174] a similar mechanism to RDFa has been introduced with the Microdata [173] concept. Microdata also allows to embed data in web documents but with less complexity than RDFa. Microdata is based on a JSON data model whereas RDFa uses the RDF data model. In a recent comparison [160] between the two approaches it has been highlighted that RDFa is better suited for multi-lingual content, explicit datatype definitions, or the use of different vocabularies in the same document. Basically the combination of both approaches is possible and the W3C evaluates different strategies for combining the benefits of both techniques.

In riese the data is structured according to the riese core schema (cf. section 6.1.4) and the user can select a certain dataset she is interested in to further explore this data. A screenshot of the visualization inside a browser of a single table is depicted in figure 12.3. Again, the underlying RDF data is also contained in the same document. Machines can access this data and also discover more relevant with the follow-your-noise approach of resolving links to other datasets internally to riese and also externally in the Linked Open Data environment.


RDFizing and Interlinking the EuroStat Data Set Effort - riese: Harmonised annual average consumer price indices - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://riese.joanneum.at/data/dba10000/
Google
RDFa (1) adr hCard hCalendar tag geo xFolk
Options
RDFizing and Interlinking the Eur...
W3C XHTML + RDFa
W3C RDF
home
about
ask
themes
General ...
Economy ...
Population ...
Industry ...
Agriculture ...
External trade
Transport Environment ...
Science ...
Related
I know more ...
Harmonised annual average consumer price indices
tinyurl.com/2raam4
Parent(s):
Consumer prices
<< 1 >> 50
Done

Harmonised annual average consumer price indices

Country	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Austria	87.21	88.22	88.95	89.41	91.16	93.25	94.83	96.06	97.94	100.00	101.69
Belgium	85.25	86.53	87.32	88.31	90.67	92.88	94.32	95.75	97.53	100.00	102.33
Bulgaria	:	56.90	67.53	69.27	76.41	82.04	86.80	88.84	94.30	100.00	107.42
Cyprus	78.70	81.31	83.21	84.15	88.25	90.00	92.51	96.18	98.00	100.00	102.25
Czech Republic	72.2	78.0	85.6	87.1	90.6	94.7	96.1	96.0	98.4	100.0	102.1
Germany (including ex-GDR from 1991)	88.6	90.0	90.5	91.1	92.4	94.1	95.4	96.4	98.1	100.0	101.8
Denmark	84.3	85.9	87.0	88.8	91.2	93.3	95.6	97.5	98.3	100.0	101.8
Euro area (EA11-2000, EA12-2006, EA13-2007, EA15)	84.60	85.94	86.87	87.85	89.69	91.80	93.86	95.81	97.86	100.00	102.18
Euro area (BE, DE, IE, GR, ES, FR, IT, LU, NL, AT, PT, FI)	84.43	85.82	86.82	87.81	89.67	91.79	93.87	95.82	97.88	100.00	102.19
Euro area (BE, DE, IE, GR, ES, FR, IT, LU, NL, AT, PT, SI, FI)	84.34	85.75	86.77	87.74	89.63	91.76	93.84	95.80	97.86	100.00	102.19
Estonia	65.97	72.09	78.42	80.85	84.03	88.76	91.95	93.22	96.05	100.00	104.45
Spain	77.92	79.39	80.79	82.59	85.47	87.88	91.04	93.86	96.73	100.00	103.56
European Union (15 countries)	84.99	86.45	87.57	88.60	90.29	92.27	94.19	96.03	97.91	100.00	102.19
European Union (25 countries)	82.38	84.50	86.27	87.66	89.80	92.04	94.00	95.83	97.88	100.00	102.20
							3.58	95.59	97.77	100.00	102.31

Last update: 2008-01-30 © JOANNEUM RESEARCH, Institute of Information Systems and Information Management, 2007. Contact us for feature requests or assistence.

Figure 12.3: A single data table in XHTML+RDFa.

In addition to directly accessing individual resources also the possibility of retrieving a complete dump of the data is offered. This is especially useful for machine access through semantic indexers as they are interested in the entire dataset and by providing all data in a single file the data can be obtained much faster. It eliminates the need of resolving each URI. Moreover a query interface is offered via a SPARQL endpoint. Tech-savvy users can use this interface to make sophisticated queries over the entire dataset and developers can benefit from this query interface when they are looking for data to integrate or link to in their own applications.

In riese it is not only possible to passively consume the information: With the User Contributed Interlinking (UCI) module presented already in section 9.1.1 users can actively contribute information in the form of links to other resources.

With riese presented in this section we have shown that Linked Data applications are well suited for consuming governmental data such as statistics. Users can access the data in a user-friendly way through their web browser and especially through the use of XHTML+RDFa it is possible to serve both humans and machines from the same source. Third party application developers can tap into this information source, reuse it in their own application or mashup and enrich content via the Linked Open Data ecosystem.

12.3 Summary

One of the application areas for Linked Data is the domain of government data discussed in this chapter. Our riese use case also belongs to this domain. Driven by Open Government Data (OGD) initiatives around the world, Linked Data has seen quite widespread adoption in this area. In the United Kingdom the first large-scale application of Linked Data for OGD was deployed. In late 2011 the European Commission also communicated a strategy to support open government data. Related efforts from governments around the world and standardization efforts at the World Wide Web Consortium also reflect the large interest in the area. In riese we have presented an approach for consuming government data that is suitable for both humans and machines. The data view is contained in a combined XHTML+RDFa representation where both the underlying data and the visual representation reside. In addition a complete dump of the dataset and query access via SPARQL is offered.

Chapter 13

Media Industry[1]

With the Link2Wod demonstrator we have targeted the media industry and the intention of the tool is to support editors at online content providers in easily integrating Linked Data sources into their original content. We have reported about Link2WoD, its usage and how it can support online content providers in the paper "Towards a Commercial Adoption of Linked Open Data for Online Content Providers" [74] which has been written together with Alexander Stocker, Harald Mayer, Helmut Mülner, and Ilir Ademi. It has been presented by the author of this thesis at the 6th International Conference on Semantic Systems (I-SEMANTICS 2010) in Graz. The author of this thesis was the research and development lead of Link2WoD and contributed most parts of the work. Thanks go to Alexander Stocker for contributing to the business aspects, Helmut Mülner who contributed to the development of the term extraction module in Link2WoD and to Ilir Ademi who contributed to the development of the Link2WoD standalone web technology preview demonstrator. Sections 13.1, 13.2, 13.4, 13.5, and 13.6 in this chapter are partly reprinted from the paper by permission of the Association for Computing Machinery, Inc.

In the paper "Harnessing Semantic Web technologies for solving the Dilemma of Content Providers" [178] written together with Claudia Wagner, Peter Scheir, and Alexander Stocker we have addressed how Semantic Web technologies can help content providers to open up their content for integration and reuse by third parties without

[1] Sections 13.1, 13.2, 13.4, 13.5, and 13.6 in this chapter are partly reprinted from the paper "W. Halb, A. Stocker, H. Mayer, H. Mülner, and I. Ademi. *Towards a commercial adoption of linked open data for online content providers.* In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 16:1–16:8. ACM, New York, NY, USA, 2010. ISBN 978-1-4503-0014-8. `http://doi.acm.org/10.1145/1839707.1839727`." ©2010 Association for Computing Machinery, Inc. Reprinted by permission under license number 2933590117669.

losing revenue. The paper has been presented by the author of this thesis at the 6th International Conference on Semantic Systems (I-SEMANTICS 2010) in Graz. The research lead for this paper was Claudia Wagner and the author of this thesis contributed to Linked Data aspects and experiences from the development of Link2WoD. Section 13.3 is based on this work.

13.1 Introduction and Motivation

This section provides a short introduction into the online content industry and motivates the needs of online editors for a new technology supporting their online editorial process.

The current business model of commercial content providers, e.g., online newspapers, is comparatively transparent: Content providers aim to produce useful contents, publish these contents on their portals and indirectly monetize them by serving advertisements. Hence, revenues of online providers largely depend on the number of users consuming online content (their reach) and indirectly also on their session length. The two most common advertising pricing models are the cost per impression (CPI) and the cost per click (CPC) model. Advertisements are either served directly by the content provider or by third party services. For a more detailed discussion on business models, cf. e.g., [118] or [135].

Valuable online content, which is suitable for monetization through advertisement, is in practice created by a specialist, the online editor. An online editor usually investigates upcoming topics, aggregates (online) content from various sources, including user generated content taken from blogs, or professional content from press agencies, and merges all these small junks together shaping a fascinating story, capable of drawing the attention of the user. Needless to say, online editors require a "good nose" for how to create such content, which is preferably consumed by people on the Web. Unlike any other role in the online content industry, professional online editors depend and rely more on Web technology. Research has shown that people on the web are rather scanning content than reading everything in detail which is contradictory to readers of classical newspapers [89]. Though, very little is known about the needs of people regarding nature, structure and presentation of online content, and we may just imagine, what differs good online content from bad. In practice, the amount of revenue generated from advertisements may serve as indicator, determining quality and appropriateness of online content.

Human resources are always scarce, which implies that they have to be utilized efficiently and effectively. From talks with commercial online content providers we learned that professional online editors spend most of their time investigating interesting and suitable material, which may enrich their editorial content to become more fascinating and valuable to their audience and/or enables them to produce editorial content quicker, enriched with multimedia objects and hyperlinks to advanced material. The latter is, where third party content will play an important role. Accurately considering these aspects may increase the attention of existing content consumers, keeping them on site for a longer period as well as attracting new consumers, positively affecting the amount of revenues gained from ads embedded in the content. Anyway, to prove these hypotheses is not the goal of our current paper.

Our research problem may be outlined as follows: Professional online editors are facing at least two business challenges:

- They need to be very efficient in their core business, i.e. developing their editorial content and getting it published instantly on the Web.
- They have to produce up to date (multimedia) content, particularly capable of drawing the attention of humans.

We investigated current Web technology, foremost Linked Data, to support online editors in creating professional online content. The next section takes a detailed look into our approach discussing the potentials of Linked Data for business and transforming them to our specific situation.

13.2 Towards a Technology-oriented Solution

In this section we will motivate Linked Data as a proper technology-oriented solution for business, capable of dealing with both introduced challenges. As a subtopic of the Semantic Web, Linked Data, based on four simple rules, has gained much attention in research. In a nutshell, the vision of the Linked Data community is to first facilitate the generation of semantically enriched Data (Linked Data) and as a result of this data-supply others will come and build intelligent applications on top of it. Such a strategy is supposed to be a very pragmatic solution for the well-known chicken-egg problem of the Semantic Web.

We find that as a current Semantic Web technology, Linked Data is capable of generating benefits in at least three different business scenarios:

1. Enterprises may adopt Linked Data to interlink their own content, increasing its accessibility for humans and machines.

2. Enterprises may adopt Linked Data to integrate third party content into their own portals, as the Web may "know" more than they do.

3. Enterprises may adopt Linked Data to prepare their own content for third party adoption, enhancing its reusability and visibility.

These three general benefits will especially affect commercial content providers, as they are dealing with huge amounts of data usually stored in data silos. A data store is called a "data silo" if it is so tightly dependent on a specific environment that it is impossible to reciprocally use and share information with other information management systems within or across organizational boundaries (cf. [93]). The so-called "silo effect" is also currently popular in the business and organizational communities to describe a lack of communication and common goals between departments in an organization. Experts agree that these silos (whether on an organizational or information technology level) greatly reduce efficiency and should be reduced, whereas cooperation should instead be fostered wherever possible (cf. [37]).

As a result of this siloing, content providers usually find themselves again in at least one of the following scenarios:

1. They may operate more than one content portal, facing the need to better integrate and interlink their data to achieve better accessibility, i.e. to allow enhanced search and retrieval across portals.

2. They may want to integrate third party data to enrich their own editorial content with open structured data, choose open - instead of licensed - content to reduce costs, and develop valuable intelligent applications based on open data.

3. They may provide their own content to be used by third parties, thereby achieving a significant increase in visibility and reach, raising the general reusability of their own content, and leading to third party adoption coming along with search-engine related benefits.

Figure 13.1 depicts how online content providers may use Linked (Open) Data and benefit from it.

From this it follows that adopting Linked Data may definitively result in a paradigm shift for the professional editing business: Online editors may benefit much from Linked

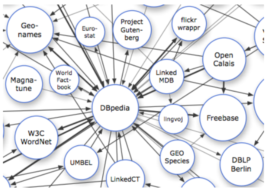
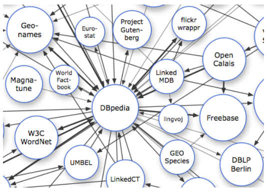
	Data provider	Data consumer	Benefit for Enterprise
1	Enterprise (Portal)	 Geo-names Euro-stat Project Guten-berg flickr wrappr Open Calais Magna-tune World Fact-book Linked MDB DBpedia lingvoj Freebase W3C WordNet UMBEL GEO Species DBLP Berlin LinkedCT	Increased visibility of own content Increased reusability of own content due to consistent data structure Search Engine Optimization
2	 Geo-names Euro-stat Project Guten-berg flickr wrappr Open Calais Magna-tune World Fact-book Linked MDB DBpedia lingvoj Freebase W3C WordNet UMBEL GEO Species DBLP Berlin LinkedCT	Enterprise (Portal)	Enriching own content with open structured data Utilizing free data instead of fee-based data Develop intelligent application based on open data
3	Enterprise (Portal 1)	Enterprise (Portal X)	Increase accessability of own content Allow structured exchange of data between isolated portals Allow search and retrieval across different domains

Figure 13.1: Three scenarios for the use of Linked Data.

Data as this new technology will enable them to perform better and to create more valuable content. Specifically, Linked Data delivers the following benefits to online editors:

- With Linked Data it is easy to integrate own or third party data to generate appropriate and up-to-date content on the one hand, and
- there is already a plethora of different Linked Data sources available, that provides information ranging from geographical to statistical data which is ready for integration.

Whenever an enterprise is dealing with Linked data, the question arises, whether Linked Data has more value to the human end user, than "ordinary" data. The Linked Data Value Chain provides a lightweight model, to conceptualize the business perspective of Linked Data and to make the value adding process to the data transparent.

The Linked Data Value Chain [113] comprises three different concepts, Participating Entities, Linked Data Roles, and Types of Data. Participating entities are persons, enterprises, associations, and research institutes owning at least one of the following roles:

- Raw Data Providers provide any kind of data.

- Linked Data Providers provide any kind of data in a Linked Data format. They consume Raw Data provided by Raw Data Providers and transform it into Linked Data.

- Linked Data Application Providers provide Linked Data Applications. They consume Linked Data provided by Linked Data Providers, process it within their applications and transform it into Human-Readable-Data.

- End users are humans who (like to) consume Human-Readable-Data, which is a human-readable presentation of Linked Data provided by Linked Data Application Providers.

The Linked Data Value Chain distinguishes between three types of Data, Raw Data, Linked Data and Human Readable Data. According to this concept, the value of the data increases with every data transformation step and human readable data has the highest value for the human enduser.

- Raw Data is any kind of structured or unstructured data that has not yet been converted into Linked Data

- Linked Data is any data in a Resource Description Framework (RDF) format interlinked with other RDF data.

- Human-Readable-Data is any kind of data which is prepared for the consumption of humans.

With respect to the introduced Linked Data Value Chain, online content providers currently and almost exclusively act as Raw Data Providers. Mentionable exceptions are the BBC [108] and the New York Times [111]. Linked Data technology will enable enterprises to act as Linked Data Providers, providing Linked Data for third parties, and to act as Linked Data Application providers, consuming data from third parties and providing more valuable Human-Readable-Data on top of it for the human enduser.

Based on the illustrated initial situation of commercial content providers, the current business challenges of online editors and the concept of the Linked Data Value Chain, we developed a first prototype, serving professional content providers.

13.3 Opening Content for Integration and Reuse

As has been outlined above, the business model of most online content providers is based on selling advertisements and therefore they aim at attracting a large number of users. By making content available for reuse or integration by third parties the concern arises that others could potentially generate business value from the original content. In [178] we have shown ways of how content providers can open up their content without losing revenues.

We have argued that content providers could reach more users by opening up their content and making it available for reuse by others. However, we also identified three major reasons for content providers to restrict the access to their resources:

1. Advertisements: In most cases advertisements are integrated in the original web page and if only the content (without advertisements) was reused this would result in a loss of revenues for the original content provider.

2. Licenses: Content sometimes is sub-licensed from other third parties which limits the legal scope of distributing the content.

3. Measuring attractiveness: Performance indicators such as page impressions or unique users are a crucial measure in the online advertising business and are based on site level but are not measured for individual bits of content.

As a potential solution approach we suggest to include descriptions about how content can be used and reused. The Open Digital Rights Language (ODRL) [94] is a vocabulary which allows to express terms and conditions for certain resources. In our paper [178] we show several examples of how content can be described to require content reusers to also include advertisements for the benefit of the original content producer or how certain usage restrictions can be expressed. This way the first two concerns regarding advertisements and licenses could be addressed. For measuring attractiveness of reused content it would be possible to restrict the reuse to reuse by reference, i.e. not allowing to copy the content but only linking to the original content which would drive the users again to the original source where advertisements can be displayed and site performance indicators can be tracked as usual. The other option would be to express via policies that counting mechanisms (such as a “counting pixel” or tracking JavaScripts) need to be included when copying content.

These proposed approaches are possibilities to reduce the concerns of content providers related to opening their content as Linked Data for instance. The Link2WoD demonstrator allows to include such usage policies in the created output.

13.4 Prototype

Motivated by the findings stated in section 13.2 a prototype has been developed that is intended to support editors at online content providers but can also be used as a general-purpose tool in other domains. It enriches editorial content with further information from Linked Data sources and also generates a Linked Data version of the editorial content. A distinguishing feature of the tool is that it supports multilingual content which poses additional challenges in the context of the Web of Data where English is still the predominant language. Currently the prototype is being tested at a cooperating content provider.

13.4.1 Use Case

The primary use case of the developed prototype is to support the editorial process at online content providers. In order to meet the needs of the industry we collaborated closely with one of the major content providers in Austria. The tool can be integrated in a provider's content management system and while an editor creates an article, the text is analyzed for interesting terms and the tool automatically suggests further information from the Web of Data. This includes data from various Linked Data sources such as textual content, links, images, or video. The editor instantaneously receives additional information about the article she is writing and can then decide which external content should be included in the article. This manual decision step has been introduced following feedback from editors and content providers as they prefer to have more control over the published content. It further allows to improve content suggestions based on previous selections and preferences. As an added benefit the final article is enriched with more exciting content and provides the reader with further information without having to leave the provider's pages. This increases the attractiveness of the content provider with further potential positive effects on visit durations, returning users, as well as page and ad impressions leading to higher ad revenues for the content provider.

```
<body about=”http://test.joanneum.at/news/xyz”>
<h1 property=”dc:title”>Von Graz aus ...</h1>
...
<div resource= ”http://dbpedia.org/resource/Larnaca” rel=”dcterms:relation”>
<span property=”rdfs:label” xml:lang=”de”>Larnaka</span>
<span property=”dbpprop:abstract” xml:lang=”de”>Larnaka, auch Larnaca, griechisch Larnaka ...</span>
...
</div>
<div resource=”http://sws.geonames.org/146400/” rel=”dcterms:relation”>
...
</div>
```

Listing 13.1: RDFa snippets of an exemplary Linked Data output

13.4.2 Modules

The tool consists of three different modules that work together as a unified solution that can be integrated into an existing content management system as a service or can be used as a stand-alone web application. The term extraction and classification module (cf. chapter 7) identifies interesting and relevant terms which act as an input for the Linked Data consumption & interlinking module (cf. section 10.3) where additional information from Linked Data sources is collected. The Linked Data Provision module makes the discovered information available as Linked Data to the public.

13.4.3 Linked Data provision

The final article contains related information from Linked Data sources as well as image and video content from Flickr and Youtube. It also includes an RDF representation that is embedded in the webpage via RDFa [175], a practice for building linked data for both humans and machines as described in [73]. Some exemplary snippets of this output are shown in Listing 13.1 where a machine processable description of the article, its relations with further information sources and additional information is provided along with the human-targeted output as shown in figure 13.2. The recently added support of RDFa content by important global search engine providers such as Yahoo! and Google [36] underlines the industry's appreciation of this approach. For further processing by applications an RDF/XML version is also provided.

	English language support	German language support	Catego-rizations	Integration of external content
Our prototype	via extensions	yes	yes	yes (Linked Data, multimedia content)
Calais	yes	-	yes	-
Zemanta	yes	limited	-	yes (select sources)
Ontos API	yes	yes	yes	-

Table 13.1: Comparison of different systems for semantic enrichment

13.4.4 In Use

The prototype can be used in a standalone web application or it can be integrated in a content management system. In figure 13.2 a screenshot of the demonstrator web application is shown. In the text editor area the original text is inserted, the panel on the right side then automatically suggests further information from Linked Data sources that the editor can simply select with a single click. The information will be integrated in the final online news article and RDF describing the news article as well as links to other resources is also supplied.

13.5 Evaluation

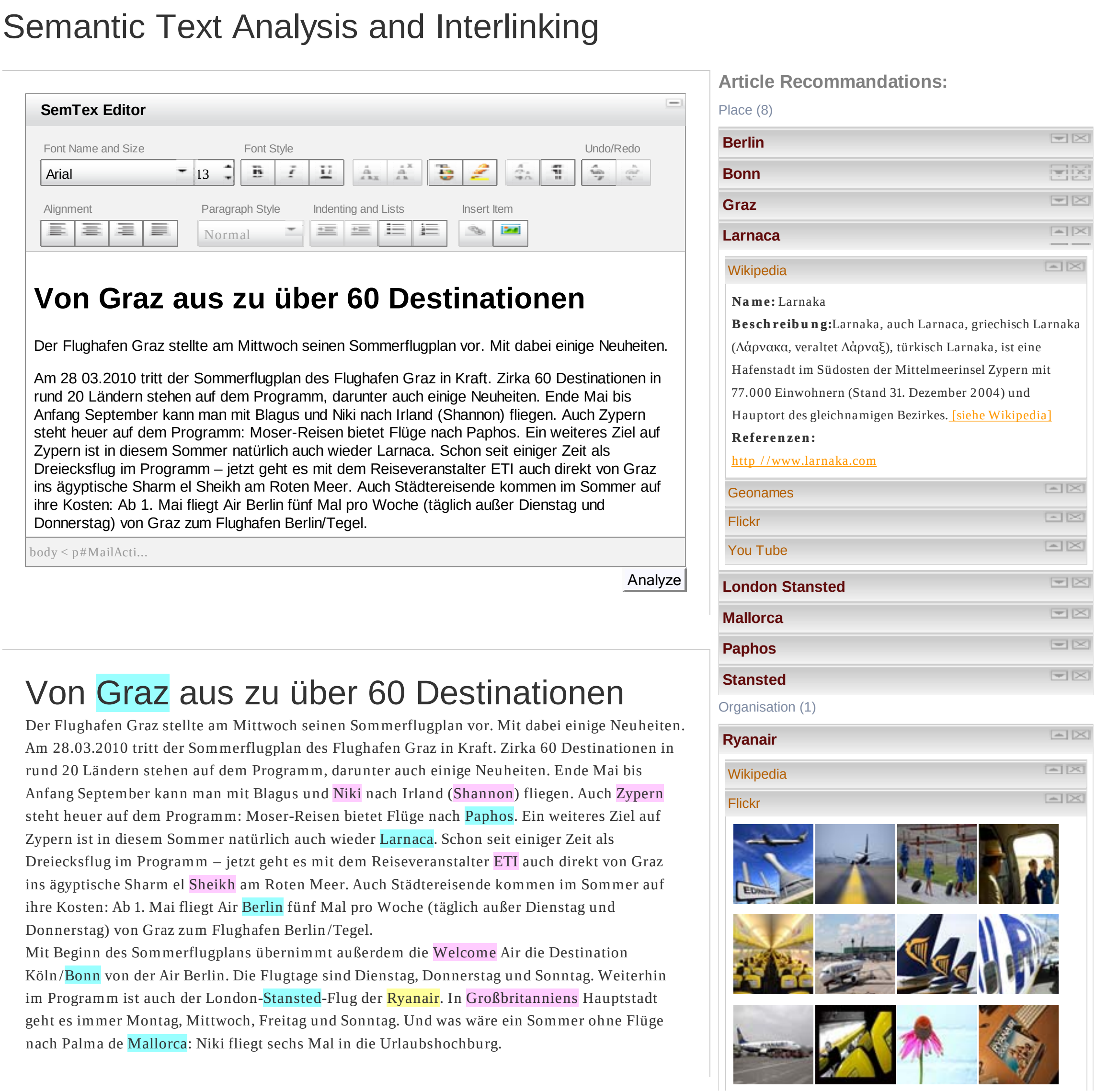
Although our intention was to investigate paradigms of the commercial adoption of Linked Data, and even though our prototype is still under development, we have conducted a preliminary evaluation and included a benchmark against other systems for semantic content enrichment. These early tests already indicate that the achieved performance of our our prototype is superior compared to similar systems that aim at enhancing (editorial) content. The most distinguishing features of our solution are the support of the German language, recommendation of multimedia content and the relationship of the recommended objects to concepts extracted from the text.

One goal of our evaluation was to benchmark our prototype against three popular applications for semantic enrichment as listed in table 13.1 where demonstrators are publicly available on the Web: Calais[2], Ontos API[3], and Zemanta[4]. However, as can be

[2] http://viewer.opencalais.com/

[3] http://try.ontos.com/

[4] http://www.zemanta.com/demo/



Semantic Text Analysis and Interlinking

Von Graz aus zu über 60 Destinationen

Der Flughafen Graz stellte am Mittwoch seinen Sommerflugplan vor. Mit dabei einige Neuheiten.

Am 28 03.2010 tritt der Sommerflugplan des Flughafen Graz in Kraft. Zirka 60 Destinationen in rund 20 Ländern stehen auf dem Programm, darunter auch einige Neuheiten. Ende Mai bis Anfang September kann man mit Blagus und Niki nach Irland (Shannon) fliegen. Auch Zypern steht heuer auf dem Programm: Moser-Reisen bietet Flüge nach Paphos. Ein weiteres Ziel auf Zypern ist in diesem Sommer natürlich auch wieder Larnaca. Schon seit einiger Zeit als Dreiecksflug im Programm – jetzt geht es mit dem Reiseveranstalter ETI auch direkt von Graz ins ägyptische Sharm el Sheikh am Roten Meer. Auch Städtereisende kommen im Sommer auf ihre Kosten: Ab 1. Mai fliegt Air Berlin fünf Mal pro Woche (täglich außer Dienstag und Donnerstag) von Graz zum Flughafen Berlin/Tegel.

Name: Larnaka

Beschreibung: Larnaka, auch Larnaca, griechisch Larnaka (Λάρνακα, veraltet Λάρναξ), türkisch Larnaka, ist eine Hafenstadt im Südosten der Mittelmeerinsel Zypern mit 77.000 Einwohnern (Stand 31. Dezember 2004) und Hauptort des gleichnamigen Bezirkes. [siehe Wikipedia]

Referenzen:

http //www.larnaka.com

Von Graz aus zu über 60 Destinationen

Der Flughafen Graz stellte am Mittwoch seinen Sommerflugplan vor. Mit dabei einige Neuheiten. Am 28.03.2010 tritt der Sommerflugplan des Flughafen Graz in Kraft. Zirka 60 Destinationen in rund 20 Ländern stehen auf dem Programm, darunter auch einige Neuheiten. Ende Mai bis Anfang September kann man mit Blagus und Niki nach Irland (Shannon) fliegen. Auch Zypern steht heuer auf dem Programm: Moser-Reisen bietet Flüge nach Paphos. Ein weiteres Ziel auf Zypern ist in diesem Sommer natürlich auch wieder Larnaca. Schon seit einiger Zeit als Dreiecksflug im Programm – jetzt geht es mit dem Reiseveranstalter ETI auch direkt von Graz ins ägyptische Sharm el Sheikh am Roten Meer. Auch Städtereisende kommen im Sommer auf ihre Kosten: Ab 1. Mai fliegt Air Berlin fünf Mal pro Woche (täglich außer Dienstag und Donnerstag) von Graz zum Flughafen Berlin/Tegel.

Mit Beginn des Sommerflugplans übernimmt außerdem die Welcome Air die Destination Köln/Bonn von der Air Berlin. Die Flugtage sind Dienstag, Donnerstag und Sonntag. Weiterhin im Programm ist auch der London-Stansted-Flug der Ryanair. In Großbritanniens Hauptstadt geht es immer Montag, Mittwoch, Freitag und Sonntag. Und was wäre ein Sommer ohne Flüge nach Palma de Mallorca: Niki fliegt sechs Mal in die Urlaubshochburg.

Figure 13.2: Screenshot of the prototype

seen from the comparison table, some tools have restrictions which resulted in a limited evaluation: We found out that Calais currently does not support German-language content at all. Using a translation service as an intermediary step and processing the translated output - an English version of the German-language article - might at the first glance solve the language problem. As this would indirectly result in an evaluation of the translation service within our evaluation, we decided not to include Calais at all as its functionality does not match with our requirements. We furthermore learned that Ontos API is currently in fact capable of extracting terms, but does not provide any content enrichment which resulted in a drop-out of Ontos API, too. As only Zemanta provides all required functionality for a valid comparison, we are limited to benchmark our prototype with Zemanta.

We developed a six-step procedure for our preliminary evaluation:

1. In the first step we selected a set of German-language articles from a local commercial news provider. Our preliminary evaluation contained only ten articles, including articles in the domains of politics, business, sports, and culture. For a more comprehensive future evaluation we have access to more than 3,000 articles from the newspaper.

2. We manually extracted persons, places, and organizations from each article. Furthermore, we extracted terms which are important for a particular article as they describe the content.

3. Thereafter, we processed the content of each article with our prototype as well as with Zemanta. This resulted in an automatic extraction of terms which we evaluate in the next step.

4. We did a quantitative comparison between manually and automatically extracted terms for both applications, our prototype and Zemanta.

5. We manually analyzed quantity and quality of the content recommended from both applications. Thereby, we especially wanted to explore whether and how recommended objects relate to the terms (concepts) extracted from the articles.

6. We compared the recommendation-ability of our prototype to the results of Zemanta. As the ability to suggest content for enrichment differs in various aspects, this evaluation is mainly done qualitatively.

In the following, we present the lessons learned from our preliminary evaluation. Figures 13.3 and 13.4 present the quantitative results of the term extraction evaluation

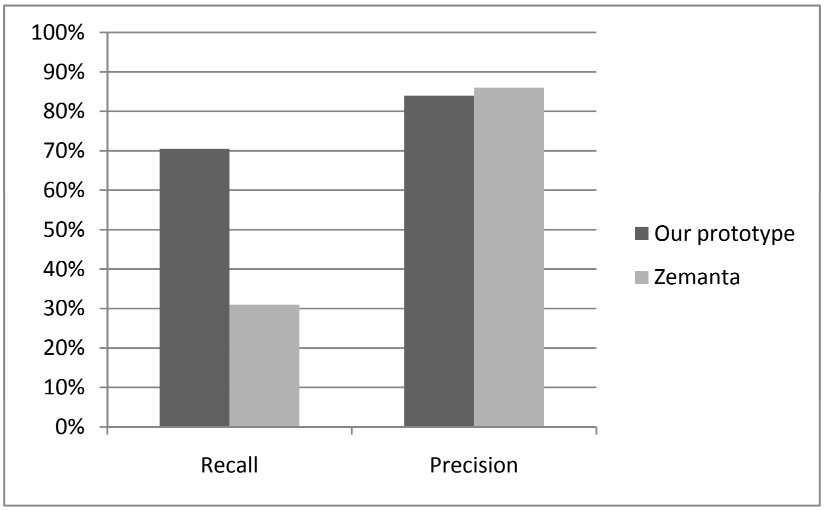


Figure 13.3: Recall and precision for all terms

(step 1 - 4): From our investigation we learned that our prototype generated satisfactorily results. The recall of all relevant terms (cf. figure 13.3) is considerably higher in our prototype and even though Zemanta has identified less relevant terms the precision values are comparable. This shows that our prototype is capable of retrieving more correct relevant terms which allows online editors to be more flexible when enriching their content with content from third party sources.

Figure 13.4 shows the recall values for places, people, and organizations. Our prototype outperforms Zemanta when extracting people and places: In this context it is important to mention that Zemanta can extract internationally known persons very well, but struggles when trying to extract local politicians or businessmen. When it comes to extracting organizations, our prototype is on a comparable level with Zemanta.

Evaluating the ability of our prototype to suggest text, pictures and videos for content enrichment (step 5 - 6) is more challenging, especially as it differs from those of Zemanta in various aspects: While our prototype links suitable data for any of the selected terms using Wikipedia, GeoNames, Flickr, and Youtube, Zemanta only recommends content for the entire article. Therefore, data recommended by our prototype is of finer granularity and of higher semantics. This aspect makes it almost impossible to compare the quality of content linked by our prototype against the content Zemanta recommends. From the requirements of online editors, we learned that content should preferably be linked to the terms (concepts) in the text and not to the entire article. Evaluating the quantity of content linked, we found out that our prototype suggests a plethora of media and text objects compared to Zemanta. However, both quantity

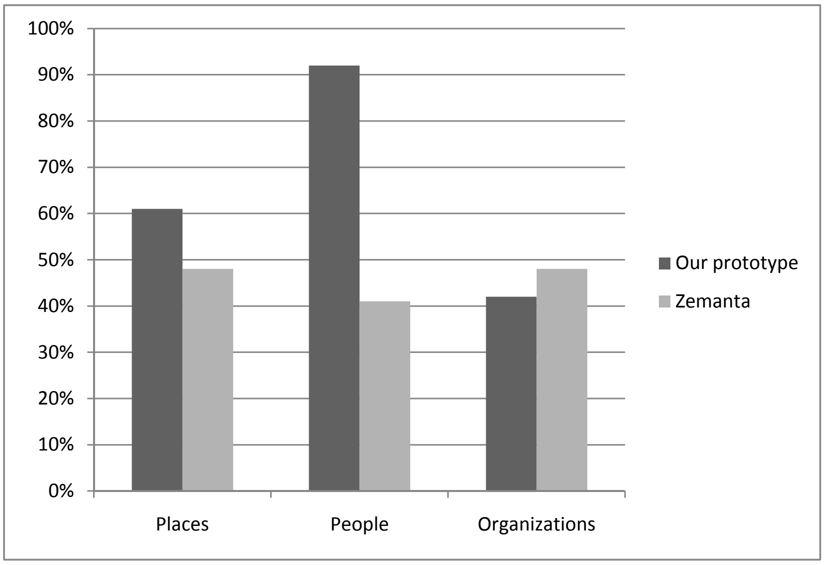


Figure 13.4: Recall of entities by category

and quality of content can be further improved: We noticed, that mainly pictures from Flickr and videos from Youtube did not always match with extracted concepts.

13.6 Challenges

Over the past months Linked Data technology has matured, various applications have been created, and new data is constantly being added. However, during the development of the prototype we have identified three important challenges that still need attention and should be addressed to make Linked Data applications competitive and allow their use in professional production environments. It has to be noted that most of the issues do not only apply to Linked Data but the Web in general.

Distributed infrastructure: One of the advantages of Linked Data is that it can be accessed via HTTP URIs and queries on remote SPARQL [137] endpoints that are made available by the data providers themselves or Linked Data consolidators (such as the Virtuoso instance hosting Linked Open Data[5]). Even though it is appealing to use data from the Web without having to care about own infrastructure the downside is that one becomes dependent on the data providers' infrastructure. Response times for queries depend on server load and it could happen that a resource is unavailable for any (technical) reason that cannot be influenced by the data consumer. One of the solutions

[5] http://lod.openlinksw.com/

to this issue is hosting a copy of relevant Linked Data on infrastructure that is under one's own control - which in turn creates the need for getting informed about dataset changes where different approaches exist and a commonly agreed strategy still needs to evolve[6]. From a business perspective one could also identify the need for Linked Data providers/consolidators that offer Service Level Agreements that contractually guarantee the availability of relevant Linked Data for professional users.

Data quality: For professional users the quality of data can be an important issue that should be taken into account by data publishers if they want their data to be reused.

Legal issues: Dataset publishers should be explicit about licensing terms of their data to protect their rights or make the data legally safely usable by others (cf. [41]).

Even though further challenges in the context of Linked Data (such as provenance, trust, archiving, etc.) would be worthwhile dealing with it seems that solving the issues presented above could foster an even greater acceptance of solutions based on Linked Data for professional use.

13.7 Conclusions

From discussions with online editors we learned much about the nature of their business and their need for quick but professional production of valuable and up-to-date online content. We drove our motivation to investigate the adoption of Linked Data for commercial content providers from the requirements of online editors. The business opportunities underlying Linked Data may fulfill their requirements and potential concerns about content reuse by third parties have been addressed by proposing the use of policies. Our Link2WoD prototype is aimed at supporting online editors during their editorial process and beyond. Results of a preliminary evaluation have been presented and showed that our prototype is already able to outperform other solutions. When implementing our first prototype, we also learned much about challenges impairing a successful adoption of Linked Data for professional production environments. We aim at improving Link2WoD based on our learnings and integrating the next version of our prototype into the content management system of a major Austrian content provider enabling us to perform further tests of the system.

[6] see `http://esw.w3.org/topic/DatasetDynamic` for a list of proposals to describe Linked Data updates and changes

Chapter 14

Process Modeling and Execution

To demonstrate the flexibility of Linked Data and its potential usage and consumption in various domains we also briefly report in this chapter about our related experiences in the course of the SERSCIS ("Semantically Enhanced, Resilient and Secure Critical Infrastructure Services") project[1]. SERSCIS aims to develop adaptive service-oriented technologies for creating, monitoring and managing secure, resilient and highly available information systems underpinning critical infrastructures.

Adaptivity of business processes is an important function that enables businesses to remain productive and limit disruptions to vital processes even under changing and challenging circumstances. Especially in the area of critical infrastructures it is essential to keep business processes alive and adapt to service disruptions and other factors that may occur at any time. One of the solutions developed in SERSCIS is to make the execution of processes that are described with the widely used Web Services Business Process Execution Language (WS-BPEL) [130] adaptive. WS-BPEL is as a short form also often referred to only as Business Process Execution Language (BPEL).

We have reported about our research results in the following papers: The paper "Agile Service Oriented Architecture with Adaptive Processes Using Semantically Annotated Workflow Templates" [75] has been written together with Herwig Zeiner, Bernhard Jandl, Harald Lernbeiß, and Christian Derler. The author of this thesis presented the paper at the 2010 IEEE International Conference on Web Services and contributed towards the semantic and Linked Data aspects. Further details have also been presented in the paper "Making Business Processes Adaptive Through Semantically Enhanced Workflow Descriptions" [184] written together with Herwig Zeiner, Harald Lernbeiß,

[1] http://www.serscis.eu/

Bernhard Jandl, and Christian Derler. Details on the underlying semantic modeling framework and associated software tools have been given in the paper "Dynamic Service Orchestration Using Human and Machine Interpretable System Knowledge with Associated Graphical Software Tools" [29] written together with Neil Briscombe, Herwig Zeiner, Stuart Bertram, Mike Kirton, and Christian Derler.

In the following section 14.1 we briefly discuss some of the basic related principles of Web service business process modeling and execution. Linked Data and semantics related aspects of our SERSCIS approach are discussed in section 14.2. Further details are given in the respective papers cited above and additional information is also available from the SERSCIS homepage.

14.1 WS-BPEL Basics

With WS-BPEL it is possible to model and execute business processes with Web services. Basically, Web services aim at achieving interoperability between applications by using Web standards. In order to integrate (potentially heterogeneous) systems it is necessary to integrate their complex interactions by using a standard process integration model. These models should consider that business interactions involve stateful sequences of message exchanges between different parties that can also influence each other. BPEL offers a language to define such a "service orchestration". Two different types of process definitions are supported:

- **Abstract Processes** serve as a behavioral interface and are not intended to be executed. They can be used by executable processes for instance to hide internal process details from a business partner or to define process templates that capture the essential process logic without execution details.

- **Executable Processes** contain all necessary specifications and can be executed by a BPEL engine such as Apache ODE (Orchestration Director Engine) [6].

BPEL has gained the status of a standard workflow language in many enterprise settings and therefore many different tools and commercial as well as open source execution engines exist. Based on interactions between the process and its partners BPEL defines a model and a grammar for describing the behavior of a business process. Interactions between partners occur exclusively through Web Service interfaces and a `partnerLink` encapsulates the structure of the relationship at the interface level. A BPEL process defines the orchestration of multiple service interactions between partners to achieve a business goal.

Even though BPEL contains a well-defined mechanism for handling failure situations, there is no built-in support for dynamic adaptation of a workflow. When a BPEL encoded business process is deployed in an execution environment the workflow cannot be adapted easily and dynamically at runtime. Service endpoints or references to other services cannot be changed without editing and redeploying the deployment description. Such a restart step requires a downtime of the system which is not acceptable in most environments. A dynamic replacement of services in the process is required and not generally available. Attempts for making BPEL more adaptive and flexible have already been presented (cf. e.g., [103, 126, 35]) but have restrictions in either flexibility of service descriptions and selections or support of stateful process dynamics. Entirely semantic dynamic service selection based on OWL-S [120], SA-WSDL [148], or WSML [42] allows greater flexibility at the cost of precision and recall in the service selection (cf. [106]). However, in several situations it is necessary to use the appropriate services only. This applies particularly to a critical infrastructure setting such as in SERSCIS. Our solution to obtain the maximum flexibility is to make use of semantically annotated workflow templates that have a reduced expressivity compared to pure semantic web service descriptions and is described in the following section.

14.2 The SERSCIS Approach[2]

Usually BPEL engines only allow limited flexibility and adaptability to changing circumstances. An executable BPEL process is linked to specific services and provides only limited possibility of changing services at runtime. In our approach, we make BPEL processes adaptive by introducing the ability to execute BPEL processes that select the concrete services only when they are needed. One of the prerequisites for dynamically exchanging and selecting services at runtime is a way of specifying suitable services in an workflow template.

The architectural approach taken by SERSCIS is based on service-oriented architecture as the basic concept of software infrastructure. Failure or underperformance of

[2] This section is partly reprinted from the paper "W. Halb, H. Zeiner, B. Jandl, H. Lernbeiß, and C. Derler. *Agile Service Oriented Architecture with Adaptive Processes Using Semantically Annotated Workflow Templates.* In *Proceedings of the 2010 IEEE International Conference on Web Services*, ICWS '10, pages 632–633. IEEE Computer Society, Washington, DC, USA, 2010. ISBN 978-0-7695-4128-0. doi:10.1109/ICWS.2010.42." © 2010 IEEE. Reprinted in accordance with item 2 of "Retained Rights/Terms and Conditions" in "IEEE Copyright and Consent Form" rev. 062802.

any of the interlinked ICT systems owing to faults, mismanagement or (cyber-)attack compromise the ability of any or all the interconnected businesses to plan their use of resources, to maintain high levels of efficiency, and to continue providing information needed by others. The SERSCIS approach is to develop service-oriented technologies for creating, monitoring and managing ICT systems, allowing dynamic adaptation to manage changing situations, and to counter the risk amplification effect of interconnectedness.

The role of the workflow execution environment of the SERSCIS system is to execute business processes that have been defined in form of semantically annotated workflow templates. It is built around the Apache ServiceMix enterprise service bus (ESB) and uses Apache CXF as well as Apache ODE. Apache CXF is a highly configurable and extendable service framework. It is based on the concept of interceptor chains, where each interceptor can read and manipulate an incoming or outgoing message at different stages. This concept is perfect for inserting standard and custom built interceptors into the chain in order to realize adaptive processes as follows: Security interceptors extract relevant information from the headers of incoming messages. There they validate the messages and check with a policy decision point whether the operation associated with the message is allowed to pass on to the requester by considering the state of the resource that is affected. A binding interceptor selects an appropriate service for invocations that refer to unbound partner links and a fault interceptor catches and logs messages representing failures on the system level. Apache ODE (Orchestration Director Engine) is a BPEL execution engine that is easily integrated into any Java Business Integration (JBI) environment since the JBI distribution is an integral part of the ODE project.

14.2.1 Semantically Annotated Workflow Template

One of the prerequisites for dynamically exchanging and selecting services at runtime and helps us specifying suitable services in an abstract workflow is what we call workflow template. The BPEL file plus the WSDL files it depends on constitute an abstract workflow. Neither a binding to a specific protocol nor to concrete services is required in there. A deployment descriptor represents the “grounding” of the workflow. It defines how to route the in- and outbound messages of the workflow. Specifically the ODE/JBI deployment descriptor requires a `<provide>` section for each partner link exposed to the outside and an `<invoke>` section for each `partnerLink` used to call an external service.

In a normal ODE/ServiceMix installation these sections simply map partner links to service end points. In the SERSCIS environment the `invoke` partner links are bound to an end point of a component. This component will execute the mapping dynamically based on semantic annotations containing information according to Linked Data principles [14] that specify how to map a partner link. These annotations are implemented as URIs pointing to Linked Data resources on an enterprise-level where semantic information about how to resolve the link are accessible. The mapping specifications are following the four Linked Data rules as (i) URIs are used for naming them, (ii) they can be accessed (at least on the enterprise-level) via HTTP, (iii) information is represented as RDF, and (iv) they link to other URIs - in our case to the respective resources and ontologies describing system interdependencies and security aspects. The annotation must be included with the deployed workflow in order to be discoverable during the run-time service binding process. Receiving partner links must be annotated with information on how to expose them.

14.2.2 Stateful Dynamic Service Binding

Having partner links bound to concrete services is required for an executable workflow instance. Being able to do this dynamically is a main source for system resilience. Binding services adds additional state information to a workflow instance (such as: which concrete service is bound to which partner link). The process of (late) binding a partner link to a concrete service involves several steps, which take the history (and thus the binding data base) into account. These steps are shown on a high level in figure 14.1. The "state interface" mentioned in this figure is an interface that can optionally be implemented by a workflow in order to allow the execution environment to request state information (e.g. if a service has already been initialized, scheduled, etc.) from the workflow instance itself. This feature is used to allow a workflow to override the default behavior of the execution environment for service binding, which is to lookup workflow metadata for a hint on how to proceed or as a last resort to let a binding unmodified once established.

The first step in service binding is to find a set of suitable services. A service is suitable, if the semantic class of services matches, it implements the required interface, is usable under the current agreements and the non-functional characteristics comply with the applicable workflow management policies. A starting set of services is built by querying the resource registry triple store, which contains semantically annotated entries of all available services. This ensures that the results belong to applicable service

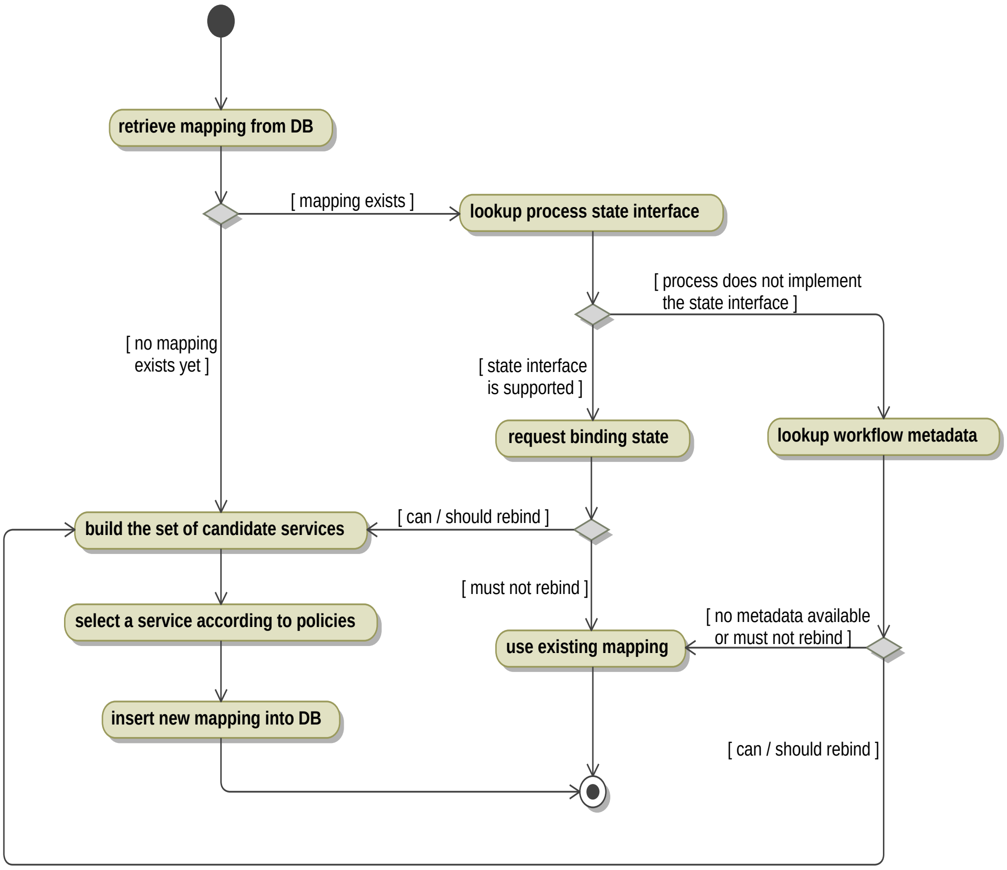

retrieve mapping from DB
[mapping exists]
lookup process state interface
[process does not implement the state interface]
[no mapping exists yet]
[state interface is supported]
request binding state
lookup workflow metadata
build the set of candidate services
[can / should rebind]
select a service according to policies
[must not rebind]
[no metadata available or must not rebind]
use existing mapping
insert new mapping into DB
[can / should rebind]

Figure 14.1: High level view of dynamic service binding process

classes. From the intermediary result all the services that cannot be used due to missing interfaces, missing usage rights or violated policies are eliminated. Service selection can then be optimized locally or globally, where each strategy has its drawbacks, and the best solution is expected to be found in the middle, where global optimization is done stepwise according to hints the designer of the workflow template has inserted in the form of annotations, leaving the rest of service bindings to just local optimization.

14.3 Conclusion

We have presented an approach for adaptive BPEL processes that is realized with semantically annotated workflow templates that contain dynamic mapping specifications as Linked Data. It can reliably be used in critical infrastructure settings. With this final example of Linked Data consumption we have shown that Linked Data concepts can be flexibly applied in various settings ranging from the open internet to closed enterprise settings.

Part V

Conclusions and Outlook

Chapter 15

Concluding Remarks

In this thesis the research work of the author in the field of Linked Open Data has been presented. Throughout the thesis the question *How can Linked Data generation and utilization be optimized?* has been addressed and several solutions for creating, interrelating and consuming Linked Data on the Web have been presented.

Part I of the thesis presented an introduction to the field and discussed the foundations. In chapter 1 the research was motivated and the underlying research questions have been presented. It also contained a listing of the authors scientific contributions and the published work. The organization of the thesis along the conceptual framework for Linked Data generation and utilization has also been presented.

In chapter 2 the basic principles of the Semantic Web and Linked Open Data (LOD) have been presented. Foundational elements of the Semantic Web have been discussed along the Semantic Web Stack and an introduction to the basics of LOD has been given. The Linked Data principles were presented and we also discussed the evolution of LOD with illustrations of the LOD cloud.

The research question *What is the size of the Semantic Web?* has been addressed in chapter 3. As a starting point for the research presented in this thesis we conducted a study in 2008 to assess the state of the Semantic Web. We have identified two different types of datasets, namely single-point-of-access datasets (such as DBpedia), and distributed datasets (e.g. the FOAF-o-sphere). At least for the single-point-of-access datasets it seems that automatic interlinking yields a high number of semantic links, however of rather shallow quality. Our finding was that not only the number of triples is relevant, but also how the datasets both internally and externally are interlinked. The outcomes of this study have been used to align our further research activities.

The two major guiding use cases of this thesis have been presented in chapter 4. The first use case riese is in the application area of government data and has been started by us in 2008 to make EuroStat statistics available as Linked Data. The second use case *Link2WoD* targets the media industry and has been developed with the intention to support editors at online content providers but can also be used as a general-purpose tool for enriching unstructured data with Linked Data.

Part II of this thesis focused on creating linkable data. The research question of *Which data is linkable?* has been covered in chapter 5 where general considerations on creating linkable data have been presented. We found that in principle every sort of source data can become Linked Data and the biggest benefits of Linked Data can be gained with data that is potentially reused. We also discussed general approaches of generating Linked Data. Regarding the overall strategy for generating linkable and Linked Data we stated the two possibilities of entirely manually creating the RDF or using some tool to support in the creation. For tool-supported creation we made a distinction between two different major kinds of source data, namely structured and unstructured data that have been addressed in subsequent chapters. Further considerations that might need to be taken into account such as the issue of trusting interlinked multimedia data have also been discussed and the abstract "Provenance-Trust-Privacy" (PTP) model has been presented.

The creation of linkable data out of structured datasources was covered in chapter 6 which also addresses the two research questions *How can structured data be RDFized?* and *What are the potentials of public data and how can the data be represented?*. We presented our approach of generating linkable data out of structured data from EuroStat statistics in our riese use case. There we implemented a system that can transform relational data into RDF and we also introduced our approach of building Linked Data for both humans and machines. Through the use of XHTML+RDFa we showed that the same document resource can deliver a good human user interface experience and machine access to the underlying data. Motivated by our use case we focused on statistical data as part of public data and introduced the Statistical Core Vocabulary (SCOVO) that can be used to represent statistical data as Linked Data. We also discussed the further developments that took SCOVO into account. Finally a brief overview of current approaches for mapping from relational databases (RDB) to RDF has been given. Regarding the general approach for RDFizing structured data we found two major mapping implementation strategies: Extract Transform Load (ETL) to transform data into RDF and load in a triplestore or a dynamic view that is created dynamically in response to a query.

In chapter 7 the research question of *How can linkable data be extracted from unstructured data?* has been briefly addressed. We highlighted the general approaches to make unstructured data such as natural language text linkable and become part of Linked Data. We also reported about the application in our *Link2WoD* use case. As part of Link2WoD the *term extraction & classification* module does named-entity recognition for a set of pre-defined categories and based on the input text provides as output identified entities. It is targeted at German-language content and is based on a combination of blacklists, whitelists and rules.

Part III is focused on interrelating Linked Data. General considerations on the interrelation are presented in chapter 8. Interrelating Linked Data by setting links between different data sources is one of the Linked Data principles. It allows to discover more information about a resource. We distinguish between *coreference resolution* where different datasets contain information about the same entity and *reference enhancement* where links to additional information are provided. Regarding the research question *Why is there a low variety of different properties used for the interlinking?* we found that reference enhancement is needed to increase the variety of properties. As this is based on relationship discovery this is a more complex task than coreference resolution. We also introduced the two approaches for interrelating which can be done user based or automated.

User based approaches for interrelating Linked Data are discussed in chapter 9. In this way also the two related research questions *How can the manual creation of links be supported?* and *How can user based approaches be realized?* are addressed. Manual link creation is possible without tool support but a very tedious task. We have developed some tools that make it easier for human users to create links. With riese we have contributed to the LOD cloud by adding the EuroStat data. We also introduced a new way of enriching datasets called "User Contributed Interlinking" (UCI), which is a Wiki-style approach enabling users to add typed links between data items on a URI-basis. We have also prototypically implemented a generalized UCI in a demonstrator called IRS (which stands for **i**nterlinking of **r**esources with **s**emantics). To maximize user participation we have created some tools that improve the user experience and make link creation easier. For an e-learning platform we investigated approaches for using tags to create Linked Data and with the CaMiCatzee demonstrator we showcased how semantic links could enter multimedia platforms on the Web. With the SALERO Intelligent Media Annotation & Search (IMAS) system we finally also introduced this user based link creation approach into an application for multimedia asset management. In this way we demonstrated a domain-specific implementation of a user based interlinking approach which provides an easy-to-use interface that is also suited for non-technical experts and hides the underlying complexity. We found that usability is an important success factor and creating links should be made easy for the user combined with integration in routine processes and the provision of appealing incentives.

Chapter 10 focused on automated approaches and therefore addressed the research question *How can links be created automatically?* Some general strategies stemming from the database area were introduced and a brief overview of relevant metrics and techniques was given. Specifics of Linked Data were discussed and our approaches that we applied in riese and *Link2WoD* have been presented. In riese we used a mapping approach that is based on a manual specification. It is configured by defining potential source and target resources along with a linking specification containing similarity metrics and conditional restrictions. Using these specifications we were able to create exact interrelations between geographical dimensions in riese and geo-related Linked Data sets such as GeoNames. We also created further interlinks to datasets such as DBpedia, CIA Factbook and Wikicompany. As part of the Link2WoD demonstrator we have implemented and integrated a *Linked Data consumption & interlinking* module which takes extracted entities from an unstructured source as an input. For the disambiguation of entities we introduced what we call a *Linked Data Entity Space*. Relations between ambiguous concepts are modeled which allows to identify concepts that are most closely related, i.e. that have the shortest conceptual distance. For toponyms geographic distance metrics were also taken into account. We discovered that toponym disambiguation based on geographic distances is extremely powerful for certain categories such as local news.

Part IV focused on consuming Linked Data. In chapter 11 we discussed general considerations on the consumption and addressed the research question *How can Linked Data be consumed by humans and machines?* We made a basic distinction between two main purposes of Linked Data applications that may target machine interpretable consumption of the data and/or human users targeted visualizations of Linked Data sources. As we learned from our riese use case the XHTML+RDFa approach is especially useful for providing Linked Data to both humans and machines from the same resource. In addition we also explained the different categories of Linked Data applications: Plain RDF Linked Data providers, Linked Data browsers, Linked Data search engines, and specific end-user applications. In the remaining chapters of that part we addressed the research question *Which application areas are well suited for Linked Data consumption?*

One of the application areas for Linked Data is the domain of government data discussed in chapter 12. Our riese use case also belongs to this domain. Driven by Open Government Data (OGD) initiatives around the world, Linked Data has seen quite widespread adoption in this area. In the United Kingdom the first large-scale application of Linked Data for OGD was deployed. In late 2011 the European Commission also communicated a strategy to support open government data. Related efforts from governments around the world and standardization efforts at the World Wide Web Consortium also reflect the large interest in the area. In riese we have presented an approach for consuming government data that is suitable for both humans and machines. The data view is contained in a combined XHTML+RDFa representation where both the underlying data and the visual representation reside. In addition a complete dump of the dataset and query access via SPARQL is offered.

Chapter 13 discusses the consumption of Linked Data in the media industry that we have targeted with the Link2WoD demonstrator. From discussions with online editors we learned much about the nature of their business and their need for quick but professional production of valuable and up-to-date online content. We drove our motivation to investigate the adoption of Linked Data for commercial content providers from the requirements of online editors. The business opportunities underlying Linked Data may fulfill their requirements and potential concerns about content reuse by third parties have been addressed by proposing the use of policies. Our Link2WoD prototype is aimed at supporting online editors during their editorial process and beyond. Results of a preliminary evaluation have been presented and showed that our prototype is already able to outperform other solutions. When implementing our first prototype, we also learned much about challenges impairing a successful adoption of Linked Data for professional production environments.

In chapter 14 we demonstrated the flexibility of Linked Data and its potential usage and consumption in various domains such as for instance for process modeling and execution. We have presented an approach for adaptive processes encoded in the Business Process Execution Language (BPEL). The approach was realized with semantically annotated workflow templates that contain dynamic mapping specifications as Linked Data. With this final example of Linked Data consumption we have shown that Linked Data concepts can be flexibly applied in various settings ranging from the open Internet to closed enterprise settings.

A final outlook including general trends and ideas for future work is given in the following chapter 16 which concludes this thesis.

Chapter 16

Outlook

This final chapter concludes the thesis and provides an outlook. First some general trends related to the area of Semantic Web and Linked Data are discussed. Then ideas for future work are explored.

16.1 General Trends

Over the recent years a gradual increase of Semantic Web applications could be witnessed. New solutions keep appearing that make use of information on the Web and aid human users in their daily tasks. Tim Berners-Lee's vision of the Semantic Web from 1998 is about to become reality and several parts of the vision have by now already been realized. Information on the Web is continuously becoming more machine processable and intelligent software can combine this information and interact with it to assist human users in their information needs and various kinds of tasks.

The Linking Open Data (LOD) community project that started in 2007 also attracted considerable scientific attention and several research projects have addressed related aspects to help in the realization of Linked Data and the Semantic Web vision. As the examples presented throughout the thesis have shown, a wide range of different solutions has been implemented. With LOD starting as a small community project primarily among academia soon a greater audience could be reached and also commercial users started to develop Linked Data applications. In several domains the adoption of Linked Data has been started and is actively being pursued. Especially in the area of public and government data several efforts are currently ongoing. Also in the area of libraries and the media industry growing interest in Linked Data can be observed. This trend further expands rapidly into other application areas.

Many notable services and applications have recently been released that can be considered Semantic Web applications in a broader sense, even though these solutions might not follow the entirety of the original Semantic Web vision or completely adopt the Semantic Web Stack. Released in 2012 the Google Knowledge Graph can almost be considered a true Semantic Web application as it displays structured information on the search result page by taking semantics into account. The underlying information is presumably even partly Linked Data but concrete technical details have not yet been reported. What is interesting in that regard is that Google as a search engine can also make use of a tremendously large amount of user search behavior to fine-tune the results. Simply by using the search engine users are indirectly providing valuable feedback that can be exploited. Another interesting application is the intelligent virtual assistant Siri introduced on Apple's iPhone. In order to accomplish tasks for the iPhone user and retrieve desired information, Siri also partly makes use of Semantic Web technologies and greatly benefits from being able to access structured and well-defined data sources.

As a general trend it can be observed that Semantic Web and Linked Data technologies are making their way into mainstream applications and in order to achieve user acceptance the underlying technical complexity needs to be largely hidden from the users.

16.2 Ideas for Future Work

Even though a widespread adoption of Linked Data applications is already commencing there are still several open issues that need to be addressed and therefore carry the potential for future research work. Some of the issues are listed below.

User Interfaces and Interaction Early attempts for realizing Linked Data applications that combine and integrate information from different sources have shown some limitations regarding user interface design and user interaction patterns. Navigating large amounts of data is still a research challenge, even though some attractive end user interfaces for specific environments have been presented. General Linked Data interaction paradigms need to evolve and take the specifics of Linked Data into account.

Distributed Infrastructure Linked Data relies on the principle that data resides in individual datasets that are connected through links. Current access mechanisms satisfy many needs but the distribution also brings potential drawbacks. Data consumers need to rely on third party data providers to also actually deliver the data when it is needed. Currently no guarantees are made that data remains accessible - which is also the case for the traditional Web of Documents. However, in the Linked Data environment the unavailability of resources may have a significant impact if applications need to combine data from different sources. There are several potential issues like server downtimes, slow response times when demand is high, truncated result sets from queries or other restrictions imposed by the dataset provider. As we have also briefly outlined in section 13.6 already, there might be a need for Service Level Agreements that contractually guarantee the availability of relevant Linked Data for professional users or some other solutions that can increase the availability of Linked Data. Further challenges include the synchronization of data updates across various sources and the implications this has. In addition the issue of trust remains a challenge in distributed systems that still needs to receive more attention in the Linked Data related research.

Integration and Tool Support It is apparent that efficient mechanisms are needed to create and maintain Linked Data. In several cases a Linked Data version of some original data is only generated with some timely distance. Consequently, a Linked Data view might not reflect the up-to-date data and performant techniques are needed to enable realtime access. Also the creation of Linked Data needs to be eased and ideally data producers should not have to care about technical details but get support from tools that can automatically expose the data.

Data Quality and Maintenance Everybody can contribute to Linked Data which is also one of its success factors. However, as increasingly applications exploit this data for advanced uses the question of data quality and reliability needs to be addressed. Data quality not only concerns the intrinsic quality of one dataset but also the quality of links to other datasets. This further raises issues related to the maintenance of links and the handling of data updates.

Legal issues For applications that consume Linked Data it is also important to access specifications about the associated conditions and licenses when reusing data. Initiatives such as Creative Commons provide a framework of licenses for creative works but not all aspects, especially copyright in relation to data, are treated uniformly across all jurisdictions. Public domain or “no rights reserved” licenses such as CC0 might not suit all dataset providers. In order to provide legal solutions for open data also the

Open Data Commons project has been started by the Open Knowledge Foundation and different licenses ranging from public domain (Public Domain Dedication and License - PDDL) to attribution (Attribution License - ODC-By) and share-alike (Open Database License - ODC-ODbL) are available. Still not all dataset publishers are explicit about licensing terms even though it would help them either in protecting their rights or making the data legally safely reusable by others. Moreover, future (user interface) research might be necessary for cases where attribution is required and data from many different sources is combined.

There are also additional open research challenges in addition to those presented above that will be influenced by emerging use cases and usage patterns. We are already seeing increasing application of Linked Data and when the remaining challenges are addressed this will enable a further evolutionary move towards exploiting the full potential of the Web.

Part VI

Appendix

Appendix A

Author's Contributions

A.1 Publications

W. Halb, Y. Raimond, and M. Hausenblas. *Building Linked Data For Both Humans and Machines.* In C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, editors, *Proceedings of the WWW08 Linked Data on the Web Workshop (LDOW 2008)*, volume 369 of *CEUR Workshop Proceedings.* Beijing, China, 2008. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-369/paper06.pdf. ISSN 1613-0073.

M. Hausenblas, W. Halb, and Y. Raimond. *Scripting User Contributed Interlinking.* In *Proceedings of the 4th Workshop on Scripting for the Semantic Web (SFSW08)*, volume 368 of *CEUR Workshop Proceedings.* Tenerife, Spain, 2008. http://CEUR-WS.org/Vol-368/paper6.pdf.

M. Hausenblas and W. Halb. *Interlinking of Resources with Semantics (Poster).* In *5th European Semantic Web Conference (ESWC2008).* 2008.

M. Hausenblas, W. Halb, Y. Raimond, and T. Heath. *What is the Size of the Semantic Web?* In *Proceedings of the 4th International Conference on Semantic Systems (I-SEMANTICS 2008)*, pages 9–16. Graz, Austria, September 2008.

M. Hausenblas and W. Halb. *Interlinking Multimedia Data.* In *Linking Open Data Triplification Challenge at the International Conference on Semantic Systems (I-Semantics08).* 2008.

W. Halb and M. Hausenblas. *select * where { :I :trust :you }: How to Trust Interlinked Multimedia Data.* In S. Auer, S. Dietzold, S. Lohmann, and J. Ziegler, editors, Proceedings of the International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW'08), volume 417 of *CEUR-WS*, pages 59–65.

Koblenz, Germany, December 2008. http://ceur-ws.org/Vol-417/paper6.pdf.

S. S. Sahoo, W. Halb, S. Hellmann, K. Idehen, T. Thibodeau, S. Auer, J. Sequeda, and A. Ezzat. *A Survey of Current Approaches for Mapping of Relational Databases to RDF*, 2009. W3C RDB2RDF Incubator Group January 08 2009.

M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayers. *SCOVO: Using Statistics on the Web of Data.* In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, editors, *The Semantic Web: Research and Applications*, volume 5554 of *Lecture Notes in Computer Science*, pages 708–722. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-02120-6. doi:10.1007/978-3-642-02121-3_52.

S. Softic, B. Taraghi, and W. Halb. *Weaving Social E-Learning Platforms Into the Web of Linked Data.* In *Proceedings of I-KNOW '09 and I-SEMANTICS '09*, pages 559–567. Graz, Austria, 2009.

W. Weiss, T. Bürger, R. Villa, P. Swamy, and W. Halb. *SALERO Intelligent Media Annotation & Search.* In *Proceedings of I-KNOW '09 and I-SEMANTICS '09*, pages 622–629. Graz, Austria, 2009.

W. Weiss, T. Bürger, R. Villa, P. Punitha, and W. Halb. *Statement-Based Semantic Annotation of Media Resources.* In T.-S. Chua, Y. Kompatsiaris, B. Mérialdo, W. Haas, G. Thallinger, and W. Bailer, editors, *Semantic Multimedia*, volume 5887 of *Lecture Notes in Computer Science*, pages 52–64. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-10542-5. doi:10.1007/978-3-642-10543-2_7.

R. Cyganiak, S. Field, A. Gregory, W. Halb, and J. Tennison. *Semantic Statistics: Bringing Together SDMX and SCOVO.* In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, editors, *Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010*, volume 628 of *CEUR Workshop Proceedings.* CEUR-WS.org, 2010. http://ceur-ws.org/Vol-628/ldow2010_paper03.pdf.

W. Halb, H. Zeiner, B. Jandl, H. Lernbeiß, and C. Derler. *Agile Service Oriented Architecture with Adaptive Processes Using Semantically Annotated Workflow Templates.* In *Proceedings of the 2010 IEEE International Conference on Web Services*, ICWS '10, pages 632–633. IEEE Computer Society, Washington, DC, USA, 2010. ISBN 978-0-7695-4128-0. doi:10.1109/ICWS.2010.42.

W. Halb, A. Stocker, H. Mayer, H. Mülner, and I. Ademi. *Towards a commercial adoption of linked open data for online content providers.* In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 16:1–16:8.

ACM, New York, NY, USA, 2010. ISBN 978-1-4503-0014-8. http://doi.acm.org/10.1145/1839707.1839727.

H. Zeiner, W. Halb, H. Lernbeiß, B. Jandl, and C. Derler. *Making business processes adaptive through semantically enhanced workflow descriptions.* In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 27:1–27:3. ACM, New York, NY, USA, 2010. ISBN 978-1-4503-0014-8. http://doi.acm.org/10.1145/1839707.1839741.

C. Wagner, P. Scheir, A. Stocker, and W. Halb. *Harnessing semantic web technologies for solving the dilemma of content providers.* In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 20:1–20:5. ACM, New York, NY, USA, 2010. ISBN 978-1-4503-0014-8. http://doi.acm.org/10.1145/1839707.1839733.

N. Briscombe, H. Zeiner, W. Halb, S. Bertram, M. Kirton, and C. Derler. *Dynamic Service Orchestration Using Human and Machine Interpretable System Knowledge with Associated Graphical Software Tools.* In *Proceedings of the Workshop on Open Knowledge Models (OKM-2010) at the 17th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*. Lisbon, Portugal, 2010.

B. Schandl, B. Haslhofer, T. Bürger, A. Langegger, and W. Halb. *Linked Data and multimedia: the state of affairs. Multimedia Tools and Applications*, 59:523–556, 2012. ISSN 1380-7501. doi:10.1007/s11042-011-0762-9.

A.2 Activities

The author of this thesis has been active in several activities related to the topics covered in this thesis.

A.2.1 Conference, Workshop and Event (Co-)Organization

- Member of the Organizing Committee of "Web of Data Practitioners Days" (WOD-PD 2008), 22-23 October 2008, Vienna, Austria

- Member of the Organizing Committee of "Linked Data Camp Vienna", 30 November & 1 December 2009, Vienna, Austria

- Assistant to the Organizing Committee of the "4th International Conference on Semantic and Digital Media Technologies" (SAMT 2009), 2-4 December 2009, Graz, Austria

- Member of the Organizing Committee of "Open Government Meetup Graz", 31 August 2010, Graz, Austria

- Member of the Organizing Committee of "Linked Open Geodata Meetup", 6 September 2011, Graz, Austria

A.2.2 Tutorials and Talks

- Invited Talk "PrestoPrime and Linked Data" at "7th International Workshop on Content-Based Multimedia Indexing" (CBMI 2009), 5 June 2009, Chania, Greece

- Tutorial "Web of Data in the Context of Multimedia" together with Bernhard Haslhofer, Bernhard Schandl, Andreas Langegger, and Tobias Bürger at "4th International Conference on Semantic and Digital Media Technologies" (SAMT 2009), 2 December 2009, Graz, Austria

- Lightning Talk "Statistical Linked Data" at "Open Government Data Austria Meetup", 8 April 2010, Vienna, Austria

- Tutorial "Workshop Pt. 2: Linked Data Demonstrators - Frameworks, Mashups, Mobile Applications" together with Tobias Bürger at "6th International Conference on Semantic Systems" (I-SEMANTICS 2010) Linked Data Camp, 2 September 2010, Graz, Austria

- Invited Expert at LOD2 project plenary meeting, 21 March 2012, Vienna, Austria

A.2.3 Reviewing, Program Committee and Moderation

- Member of the Program Committee for "1st Workshop on Trust and Privacy on the Social and Semantic Web" (SPOT2009) co-located with the "6th European Semantic Web Conference" (ESWC2009), 1 June 2009, Heraklion, Greece

- Co-moderator of OGD2011 pre-conference workshop "OGD and Businesses", 14 February 2011, Vienna, Austria

- Session chair of "Technology & Infrastructure" session at "Open Government Data Conference 2011" (OGD 2011), 16 June 2011, Vienna, Austria

- Reviewer for IEEE Intelligent Systems Magazine (ISSI-2011-09-0121), Special Issue Mar/Apr 2012 "Linked Open Government Data"
- Member of the Program Committee at iPraxis track co-located with "8th International Conference on Semantic Systems" (I-SEMANTICS 2012), 5-7 September 2012, Graz, Austria

A.2.4 W3C Activities

- Member of the W3C RDB2RDF Incubator Group and contribution to survey report of the group
- Member of the W3C RDB2RDF Working Group
- Member of the W3C Library Linked Data Incubator Group

Bibliography

[1] B. Adida, M. Birbeck, S. McCarron, and S. Pemberton. *RDFa in XHTML: Syntax and Processing.* W3C Working Draft 18 October 2007, W3C Semantic Web Deployment Working Group, 2007. (Cited on page 61.)

[2] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. *Describing Linked Datasets - On the Design and Usage of voiD, the 'Vocabulary of Interlinked Datasets'.* In *WWW 2009 Workshop: Linked Data on the Web (LDOW2009).* Madrid, Spain, 2009. (Cited on pages 37, 57, 82 and 87.)

[3] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. *Describing Linked Datasets with the VoID Vocabulary*, March 2011. `http://www.w3.org/TR/void/`. W3C Interest Group Note. (Cited on pages 37 and 56.)

[4] J. F. Allen and G. Fergusson. *Actions and Events in interval temporal logic.* Technical Report TR521, University of Rochester, Computer Science Department, 1994. `http://citeseer.ist.psu.edu/allen94actions.html`. Available at `http://citeseer.ist.psu.edu/allen94actions.html`. Last accessed February 2008. (Cited on page 74.)

[5] J. Antoch. *Environment for statistical computing.* *Computer Science Review*, 2(2):113–122, 2008. (Cited on page 70.)

[6] Apache Software Foundation. *Apache ODE*, 2011. `http://ode.apache.org/`. (Cited on page 176.)

[7] P. Assini. *NESSTAR: A Semantic Web Application for Statistical Data and Metadata.* In *International Workshop Real World RDF and Semantic Web Applications, 11th International World Wide Web Conference (WWW2002).* 2002. (Cited on pages 60 and 70.)

[8] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. *DBpedia: A Nucleus for a Web of Open Data.* In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*, pages 722–735. 2007. (Cited on pages 26, 39, 46 and 60.)

[9] D. Ayers. *Evolving the Link. IEEE Internet Computing*, 11(3):94–96, 2007. (Cited on page 30.)

[10] D. Ayers. *Graph Farming. IEEE Internet Computing*, 12(1):80–83, 2008. (Cited on page 105.)

[11] D. Ayers and H. Story. *AtomOwl Vocabulary Specification* . Namespace Document, Atom Owl Working Group, 2006. (Cited on page 108.)

[12] C. Becker and C. Bizer. *Exploring the Geospatial Semantic Web with DBpedia Mobile. Web Semantics: Science, Services and Agents on the World Wide Web*, 7(4):278 – 286, 2009. ISSN 1570-8268. doi:10.1016/j.websem.2009.09.004. (Cited on page 147.)

[13] T. Berners-Lee. *Semantic Web Road map*, October 1998. `http://www.w3.org/DesignIssues/Semantic.html`. (Cited on pages 4, 15 and 21.)

[14] T. Berners-Lee. *Linked Data*, July 2006. `http://www.w3.org/DesignIssues/LinkedData.html`. Updated 2009-06-18. (Cited on pages 4, 22, 23, 45, 47, 123, 146, 151 and 179.)

[15] T. Berners-Lee. *Putting Government Data online*, 2009. `http://www.w3.org/DesignIssues/GovData.html`. (Cited on page 152.)

[16] T. Berners-Lee. *Tim Berners-Lee on the next Web*, February 2009. `http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html`. Talk at TED2009. (Cited on page 6.)

[17] T. Berners-lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets. *Tabulator: Exploring and analyzing linked data on the semantic web.* In *In Proceedings of the 3rd International Semantic Web User Interaction Workshop*. 2006. (Cited on page 147.)

[18] T. Berners-Lee, J. Hendler, and O. Lassila. *The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American*, 284(5):34–43, May 2001. doi:10.1038/scientificamerican0501-34. (Cited on page 4.)

[19] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg. *Adaptive Name Matching in Information Integration. IEEE Intelligent Systems*, 18(5):16–23, September 2003. ISSN 1541-1672. doi:10.1109/MIS.2003.1234765. (Cited on page 127.)

[20] C. Bizer. *Quality-Driven Information Filtering in the Context of Web-Based Information Systems.* Ph.D. thesis, Freie Universität Berlin, 2007. (Cited on page 56.)

[21] C. Bizer and R. Cyganiak. *Quality-driven information filtering using the WIQA policy framework. Web Semant.*, 7(1):1–10, January 2009. ISSN 1570-8268. doi:10.1016/j.websem.2008.02.005. `http://dx.doi.org/10.1016/j.websem.2008.02.005`. (Cited on page 57.)

[22] C. Bizer, R. Cyganiak, and T. Heath. *How to Publish Linked Data on the Web.* `http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial/`, 2007. (Cited on page 50.)

[23] C. Bizer, T. Heath, D. Ayers, and Y. Raimond. *Interlinking Open Data on the Web (Poster).* In *4th European Semantic Web Conference (ESWC2007)*, pages 802–815. 2007. (Cited on pages 50 and 70.)

[24] C. Bizer, T. Heath, and T. Berners-Lee. *Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009. doi:10.4018/jswis.2009081901. (Cited on page 37.)

[25] C. Bizer, A. Jentzsch, and R. Cyganiak. *State of the LOD Cloud*, 2011. `http://www4.wiwiss.fu-berlin.de/lodcloud/state/`. Version 0.3, 09/19/2011. (Cited on page 24.)

[26] D. Booth. *URI Declaration Versus Use.* `http://dbooth.org/2007/uri-decl/`, 2008. (Cited on page 76.)

[27] D. Brickley and R. Guha. *RDF Vocabulary Description Language 1.0: RDF Schema.* W3C Recommendation, RDF Core Working Group, 2004. (Cited on pages 20 and 64.)

[28] D. Brickley and L. Miller. *FOAF Vocabulary Specification.* `http://xmlns.com/foaf/0.1/`, 2004. (Cited on pages 49 and 87.)

[29] N. Briscombe, H. Zeiner, W. Halb, S. Bertram, M. Kirton, and C. Derler. *Dynamic Service Orchestration Using Human and Machine Interpretable System Knowledge with Associated Graphical Software Tools.* In *Proceedings of the Workshop on Open Knowledge Models (OKM-2010) at the 17th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*. Lisbon, Portugal, 2010. (Cited on pages 11, 13 and 176.)

[30] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. *Graph structure in the Web. Computer Networks: The International Journal of Computer and Telecommunications Networking*, 33(1-6):309–320, 2000. (Cited on page 30.)

[31] T. Bürger and M. Hausenblas. *Why Real-World Multimedia Assets Fail to Enter the Semantic Web.* In *International Workshop on Semantic Authoring, Annotation and Knowledge Markup (SAAKM07)*. Whistler, Canada, 2007. (Cited on page 53.)

[32] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler. *Named Graphs, Provenance and Trust.* In *In WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 613–622. ACM Press, 2005. (Cited on page 56.)

[33] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan. *A Survey of Web Information Extraction Systems. IEEE Trans. on Knowl. and Data Eng.*, 18(10):1411–1428, October 2006. ISSN 1041-4347. doi:10.1109/TKDE.2006.152. (Cited on page 93.)

[34] G. Cheng and Y. Qu. *Searching Linked Objects with Falcons: Approach, Implementation and Evaluation. International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):49–70, 2009. doi:10.4018/jswis.2009081903. ISSN: 1552-6283, EISSN: 1552-6291. (Cited on page 148.)

[35] K. Christos, C. Vassilakis, E. Rouvas, and P. Georgiadis. *QoS-Driven Adaptation of BPEL Scenario Execution.* In *Proc. 2009 IEEE Intl. Conference on Web Services*, pages 271–278. IEEE Computer Society, Washington, DC, USA, 2009. ISBN 978-0-7695-3709-2. doi:http://dx.doi.org/10.1109/ICWS.2009.80. (Cited on page 177.)

[36] D. Connolly. *Search Engines take on Structured Data*, May 2009. `http://www.w3.org/QA/2009/05/structured_data_and_search_eng.html`. W3C Blog. (Cited on page 167.)

[37] M. Côté. *A matter of trust and respect. CA magazine*, March 2002. (Cited on page 162.)

[38] R. Cyganiak. *A relational algebra for SPARQL.* Technical report, Digital Media Systems Laboratory. HP Laboratories Bristol, 2005. HPL-2005-170. (Cited on page 90.)

[39] R. Cyganiak, R. Delbru, and G. Tummarello. *Semantic Web Crawling: A Sitemap Extension.* http://sw.deri.org/2007/07/sitemapextension, 2007. (Cited on page 79.)

[40] R. Cyganiak, S. Field, A. Gregory, W. Halb, and J. Tennison. *Semantic Statistics: Bringing Together SDMX and SCOVO.* In C. Bizer, T. Heath, T. Berners-Lee, and M. Hausenblas, editors, *Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010*, volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010. http://ceur-ws.org/Vol-628/ldow2010_paper03.pdf. (Cited on pages 11, 12, 84 and 88.)

[41] I. Davis. *Linked Data and the Public Domain*, July 2009. http://blogs.talis.com/nodalities/2009/07/linked-data-public-domain.php. Nodalities blog. (Cited on page 173.)

[42] J. de Bruijn, D. Fensel, U. Keller, M. Kifer, H. Lausen, R. Krummenacher, A. Polleres, and L. Predoiu. *Web Service Modeling Language (WSML)*, 2005. http://www.w3.org/Submission/WSML/. W3C Member Submission 3 June 2005. (Cited on page 177.)

[43] L. Ding and T. Finin. *Characterizing the Semantic Web on the Web.* In *5th International Semantic Web Conference, ISWC 2006*, pages 242–257. 2006. (Cited on page 30.)

[44] L. Ding, T. Lebo, J. S. Erickson, D. DiFranzo, G. T. Williams, X. Li, J. Michaelis, A. Graves, J. G. Zheng, Z. Shangguan, J. Flores, D. L. McGuinness, and J. A. Hendler. *TWC LOGD: A portal for linked open government data ecosystems. Web Semantics: Science, Services and Agents on the World Wide Web*, 9(3):325 – 333, 2011. ISSN 1570-8268. doi:10.1016/j.websem.2011.06.002. (Cited on page 149.)

[45] L. Ding, L. Zhou, T. Finin, and A. Joshi. *How the Semantic Web is Being Used:An Analysis of FOAF Documents.* In *38th International Conference on System Sciences*. 2005. (Cited on page 30.)

[46] Dublin Core Metadata Initiative. *DCMI Metadata Terms.* http://dublincore.org/documents/dcmi-terms/, 2008. (Cited on page 64.)

[47] Dublin Core Metadata Initiative. *Dublin Core Metadata Element Set, Version 1.1.* http://dublincore.org/documents/dces/, 2008. (Cited on pages 49 and 64.)

[48] E. Dumbill. *Description of a Project (DOAP) vocabulary.* http://usefulinc.com/ns/doap, 2005. (Cited on page 64.)

[49] H. L. Dunn. *Record Linkage. American Journal of Public Health and the Nations Health*, 36(12):1412–1416, December 1946. doi:10.2105/AJPH.36.12.1412. (Cited on page 100.)

[50] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. *Duplicate Record Detection: A Survey. IEEE Trans. on Knowl. and Data Eng.*, 19(1):1–16, January 2007. ISSN 1041-4347. doi:10.1109/TKDE.2007.9. (Cited on page 124.)

[51] O. Erling and I. Mikhailov. *SPARQL and Scalable Inference on Demand*, 2008. http://virtuoso.openlinksw.com/whitepapers/SPARQL%20and%20Scalable%20Inference%20on%20Demand.pdf. OpenLink Software Virtuoso White Paper. (Cited on page 139.)

[52] K. S. Esmaili and H. Abolhassani. *A Categorization Scheme for Semantic Web Search Engines.* In *4th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-06)*. Sharjah, UAE, 2006. (Cited on page 30.)

[53] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. *Open information extraction from the web. Commun. ACM*, 51(12):68–74, December 2008. ISSN 0001-0782. doi:10.1145/1409360.1409378. http://doi.acm.org/10.1145/1409360.1409378. (Cited on page 92.)

[54] European Commission. *Open data. An engine for innovation, growth and transparent governance*, December 2011. http://ec.europa.eu/information_society/policy/psi/docs/pdfs/opendata2012/open_data_communication/de.pdf. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. COM(2011) 882 final. (Cited on page 154.)

[55] Eurostat. *Regions in the European Union. Nomenclature of territorial units for statistics NUTS 2006 /EU-27*, volume Methodologies and working papers of *Gen-*

eral and Regional Statistics. Office for Official Publications of the European Communities, 2007. ISBN 978-92-79-04756-5, ISSN 1977-0375, Cat. No. KS-RA-07-020-EN-N. (Cited on page 132.)

[56] Facebook. *Feature Launch: Photo tagging for Pages*, 2011. `http://www.facebook.com/note.php?note_id=10150168953654822`. (Cited on page 117.)

[57] L. Feigenbaum. *Modeling Statistics in RDF - A Survey and Discussion*, March 2008. `http://www.thefigtrees.net/lee/blog/2008/03/modeling_statistics_in_rdf_a_s.html`. (Cited on page 79.)

[58] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. *RFC 2616: Hypertext Transfer Protocol – HTTP/1.1.* Technical report, IETF, 1999. `http://tools.ietf.org/html/rfc2616`. (Cited on page 48.)

[59] R. T. Fielding. *Architectural styles and the design of network-based software architectures.* Ph.D. thesis, 2000. (Cited on page 23.)

[60] T. Finin, L. Ding, R. Pan, A. Joshi, P. Kolari, A. Java, and Y. Peng. *Swoogle: Searching for knowledge on the Semantic Web.* In *AAAI 05 (intelligent systems demo)*. 2005. (Cited on page 30.)

[61] J. Gantz and D. Reinsel. *Extracting Value from Chaos.* IDC, June 2011. `http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf`. (Cited on page 3.)

[62] P. Gearon, A. Passant, and A. Polleres. *SPARQL 1.1 Update*, 2012. `http://www.w3.org/TR/sparql11-update/`. W3C Working Draft 05 January 2012. (Cited on page 20.)

[63] General Services Administration. *Data.gov Releases Open Source Software*, May 2012. `http://www.data.gov/opengovplatform`. (Cited on page 154.)

[64] Geonames. *Geonames Ontology.* `http://www.geonames.org/ontology/`, 2007. (Cited on page 64.)

[65] K. Goel, R. V. Guha, and O. Hansson. *Introducing Rich Snippets.* Google Webmaster Central Blog, May 2009. `http://googlewebmastercentral.blogspot.co.at/2009/05/introducing-rich-snippets.html`. (Cited on page 149.)

[66] J. Golbeck. *Combining Provenance with Trust in Social Networks for Semantic Web Content Filtering.* In *Provenance and Annotation of Data, International Provenance and Annotation Workshop, IPAW 2006, Chicago, IL, USA, May 3-5, 2006, Revised Selected Papers*, volume 4145 of *Lecture Notes in Computer Science*, pages 101–108. Springer, 2006. (Cited on page 56.)

[67] C. Golbreich and E. K. Wallace. *OWL 2 Web Ontology Language. New Features and Rationale*, 2009. `http://www.w3.org/TR/owl2-new-features/`. W3C Recommendation 27 October 2009. (Cited on page 20.)

[68] A. Grossenbacher. *Semantic Web: Basics, RDF, DC and the description of a statistical site.* `http://tinyurl.com/2d5gta`, 2007. (Cited on pages 61 and 70.)

[69] T. R. Gruber. *A translation approach to portable ontology specifications. KNOWLEDGE ACQUISITION*, 5:199–220, 1993. (Cited on page 20.)

[70] R. Guha. *Introducing schema.org: Search engines come together for a richer web.* Google Official Blog, June 2011. `http://googleblog.blogspot.co.at/2011/06/introducing-schemaorg-search-engines.html`. (Cited on page 149.)

[71] A. Gulli and A. Signorini. *The Indexable Web is More than 11.5 Billion Pages.* In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 902–903. 2005. (Cited on page 30.)

[72] W. Halb and M. Hausenblas. *select * where { :I :trust :you }: How to Trust Interlinked Multimedia Data.* In S. Auer, S. Dietzold, S. Lohmann, and J. Ziegler, editors, *Proceedings of the International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW'08)*, volume 417 of *CEUR-WS*, pages 59–65. Koblenz, Germany, December 2008. `http://ceur-ws.org/Vol-417/paper6.pdf`. (Cited on pages 10, 12, 50 and 57.)

[73] W. Halb, Y. Raimond, and M. Hausenblas. *Building Linked Data For Both Humans and Machines.* In C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee, editors, *Proceedings of the WWW08 Linked Data on the Web Workshop (LDOW 2008)*, volume 369 of *CEUR Workshop Proceedings.* Beijing, China, 2008. `http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-369/paper06.pdf`. ISSN 1613-0073. (Cited on pages 9, 12, 13, 59, 69, 70, 79, 104, 130, 155 and 167.)

[74] W. Halb, A. Stocker, H. Mayer, H. Mülner, and I. Ademi. *Towards a commercial adoption of linked open data for online content providers.* In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 16:1–16:8. ACM, New York, NY, USA, 2010. ISBN 978-1-4503-0014-8. http://doi.acm.org/10.1145/1839707.1839727. (Cited on pages 11, 13, 93, 136 and 159.)

[75] W. Halb, H. Zeiner, B. Jandl, H. Lernbeiß, and C. Derler. *Agile Service Oriented Architecture with Adaptive Processes Using Semantically Annotated Workflow Templates.* In *Proceedings of the 2010 IEEE International Conference on Web Services*, ICWS '10, pages 632–633. IEEE Computer Society, Washington, DC, USA, 2010. ISBN 978-0-7695-4128-0. doi:10.1109/ICWS.2010.42. (Cited on pages 11, 13 and 175.)

[76] A. Harth, A. Polleres, and S. Decker. *Towards a social provenance model for the Web.* In *Workshop on Principles of Provenance (PrOPr), Edinburgh, Scotland.* 2007. (Cited on page 56.)

[77] O. Hassanzadeh and M. Consens. *Linked Movie Data Base.* In *Linking Open Data Triplification Challenge at the International Conference on Semantic Systems (I-Semantics08)*. 2008. (Cited on page 129.)

[78] O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller, and M. Wang. *A framework for semantic link discovery over relational data.* In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1027–1036. ACM, New York, NY, USA, 2009. ISBN 978-1-60558-512-3. doi:10.1145/1645953.1646084. (Cited on page 129.)

[79] M. Hausenblas, W. Bailer, T. Bürger, and R. Troncy. *Deploying Multimedia Metadata on the Semantic Web (Poster).* In *2nd International Conference on Semantics And digital Media Technologies (SAMT 07)*. 2007. (Cited on page 57.)

[80] M. Hausenblas, W. Bailer, and H. Mayer. *Deploying Multimedia Metadata in Cultural Heritage on the Semantic Web.* In *First International Workshop on Cultural Heritage on the Semantic Web, collocated with the 6th International Semantic Web Conference (ISWC07)*. Busan, South Korea, 2007. (Cited on page 61.)

[81] M. Hausenblas and W. Halb. *Interlinking Multimedia Data.* In *Linking Open Data Triplification Challenge at the International Conference on Semantic Systems (I-Semantics08)*. 2008. (Cited on pages 10, 13, 50 and 115.)

[82] M. Hausenblas and W. Halb. *Interlinking of Resources with Semantics (Poster).* In *5th European Semantic Web Conference (ESWC2008).* 2008. (Cited on pages 10, 13 and 104.)

[83] M. Hausenblas, W. Halb, and Y. Raimond. *Scripting User Contributed Interlinking.* In *Proceedings of the 4th Workshop on Scripting for the Semantic Web (SFSW08)*, volume 368 of *CEUR Workshop Proceedings.* Tenerife, Spain, 2008. http://CEUR-WS.org/Vol-368/paper6.pdf. (Cited on pages 10, 13, 69 and 104.)

[84] M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayers. *SCOVO: Using Statistics on the Web of Data.* In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, editors, *The Semantic Web: Research and Applications*, volume 5554 of *Lecture Notes in Computer Science*, pages 708–722. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-02120-6. doi:10.1007/978-3-642-02121-3_52. (Cited on pages 10, 12 and 40.)

[85] M. Hausenblas, W. Halb, Y. Raimond, and T. Heath. *What is the Size of the Semantic Web?* In *Proceedings of the 4th International Conference on Semantic Systems (I-SEMANTICS 2008)*, pages 9–16. Graz, Austria, September 2008. (Cited on pages 5, 7, 10, 12, 29 and 102.)

[86] M. Hausenblas, W. Slany, and D. Ayers. *A Performance and Scalability Metric for Virtual RDF Graphs.* In *3rd Workshop on Scripting for the Semantic Web (SFSW07).* Innsbruck, Austria, 2007. (Cited on pages 32, 66, 71 and 79.)

[87] T. Heath and E. Motta. *Revyu.com: a Reviewing and Rating Site for the Web of Data.* In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*, pages 895–902. 2007. (Cited on page 60.)

[88] A. Hogan, A. Harth, J. Umbrich, S. Kinsella, A. Polleres, and S. Decker. *Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine. Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):365 – 401, 2011. ISSN 1570-8268. doi:10.1016/j.websem.2011.06.004. (Cited on page 148.)

[89] K. Holmqvist, J. Holsanova, M. Barthelson, and D. Lundqvist. *Reading or Scanning? A Study of Newspaper and Net Paper Reading.* In J. Hyönä, R. Radach,

and H. Deubel, editors, *The Mind's Eye*, pages 657 – 670. North-Holland, Amsterdam, 2003. ISBN 978-0-444-51020-4. doi:10.1016/B978-044451020-4/50035-9. (Cited on page 160.)

[90] W. Hu and Y. Qu. *Discovering simple mappings between relational database schemas and ontologies.* In *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ISWC'07/ASWC'07, pages 225–238. Springer-Verlag, Berlin, Heidelberg, 2007. ISBN 3-540-76297-3, 978-3-540-76297-3. (Cited on page 89.)

[91] A. Hunt and D. Thomas. *The pragmatic programmer : from journeyman to master.* Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 2000. (Cited on page 63.)

[92] M. Hurst. *The Interpretation of Tables in Texts.* Ph.D. thesis, University of Edinburgh, 2000. (Cited on page 92.)

[93] J. Hurwitz, C. Baroud, R. Bloor, and M. Kaufman. *Service Oriented Architecture for Dummies*, chapter 13: Where's the data?, pages 153–166. Wiley Publishing Inc., Hoboken, NJ, USA, 2007. (Cited on page 162.)

[94] R. Iannella, S. Guth, D. Paehler, and A. Kasten. *ODRL Version 2.0 Core Model*, April 2012. `http://www.w3.org/community/odrl/two/model/`. W3C ODRL Community Group Final Specification. (Cited on page 165.)

[95] ISO. *ISO 8601:2004(E)8. Data elements and interchange formats — Information interchange — Representation of dates and times.* Standard No. ISO 8601:2004(E), 2004. (Cited on page 125.)

[96] ISO. *Statistical data and metadata exchange (SDMX).* Standard No. ISO/TS 17369:2005, 2005. (Cited on pages 70 and 84.)

[97] ISO. *ISO 3166-1:2006. Codes for the representation of names of countries and their subdivisions – Part 1: Country codes.* Standard No. ISO 3166-1:2006, 2006. (Cited on pages 125 and 132.)

[98] A. Jaffri, H. Glaser, and I. Millard. *URI identity management for semantic web data integration and linkage.* In *Proceedings of the 2007 OTM Confederated international conference on On the move to meaningful internet systems - Volume Part II*, OTM'07, pages 1125–1134. Springer-Verlag, Berlin, Heidelberg, 2007. ISBN 3-540-76889-0, 978-3-540-76889-0. (Cited on pages 129 and 137.)

[99] A. Jaffri, H. Glaser, and I. Millard. *Managing URI Synonymity to Enable Consistent Reference on the Semantic Web.* In P. Bouquet, H. Halpin, H. Stoermer, and G. Tummarello, editors, *Proceedings of the 1st International Workshop on Identity and Reference on the Semantic Web (IRSW2008)*, volume 422 of *CEUR Workshop Proceedings*. CEUR-WS.org, Tenerife, Spain, 2008. http://ceur-ws.org/Vol-422/irsw2008-submission-6.pdf. (Cited on pages 129 and 137.)

[100] M. A. Jaro. *UNIMATCH: A Record Linkage System: User's Manual.* Technical report, U.S. Bureau of the Census, Washington, D.C., 1976. (Cited on page 126.)

[101] A. Jentzsch, R. Cyganiak, and C. Bizer. *Next version of the LOD cloud diagram. Please provide input, so that your dataset is included*, September 2010. http://lists.w3.org/Archives/Public/semantic-web/2010Sep/0017.html. Announcement on semantic-web@w3.org mailing list. (Cited on pages 26 and 37.)

[102] T. Käfer, J. Umbrich, A. Hogan, and A. Polleres. *Towards a Dynamic Linked Data Observatory.* In *Linked Data on the Web Workshop (LDOW2012), in conjunction with the 21st International World Wide Web Conference (WWW 2012)*. 2012. (Cited on page 37.)

[103] D. Karastoyanova, A. Houspanossian, M. Cilia, F. Leymann, and A. Buchmann. *Extending BPEL for run time adaptability.* In *Proc. EDOC Enterprise Computing Conference 2005*, pages 15 – 26. Sept. 2005. doi:10.1109/EDOC.2005.14. (Cited on page 177.)

[104] M. Kifer and H. Boley. *RIF Overview*, 2010. http://www.w3.org/TR/rif-overview/. W3C Working Group Note 22 June 2010. (Cited on page 21.)

[105] H. L. Kim, A. Passant, J. G. Breslin, S. Scerri, and S. Decker. *Review and Alignment of Tag Ontologies for Semantically-Linked Data in Collaborative Tagging Spaces.* In *Proceedings of the 2008 IEEE International Conference on Semantic Computing*, ICSC '08, pages 315–322. IEEE Computer Society, Washington, DC, USA, 2008. ISBN 978-0-7695-3279-0. doi:10.1109/ICSC.2008.79. http://dx.doi.org/10.1109/ICSC.2008.79. (Cited on page 113.)

[106] M. Klusch, A. Leger, D. Martin, M. Paolucci, A. Bernstein, and U. Küster. *3rd International Semantic Service Selection Contest - Performance Evaluation of Semantic Service Matchmakers.* http://www-ags.dfki.uni-sb.de/~klusch/s3/s3-2009-summary.pdf, 2009. (Cited on page 177.)

[107] G. Klyne, J. J. Carroll, and B. McBride. *RDF/XML Syntax Specification (Revised).* W3C Recommendation, RDF Core Working Group, 2004. (Cited on page 56.)

[108] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee. *Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Connections.* In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, editors, *The Semantic Web: Research and Applications*, volume 5554 of *Lecture Notes in Computer Science*, pages 723–737. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-02120-6. doi:10.1007/978-3-642-02121-3_53. (Cited on pages 150 and 164.)

[109] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.* In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001. ISBN 1-55860-778-1. (Cited on page 92.)

[110] A. Langegger and W. Wöss. *RDFStats - An Extensible RDF Statistics Generator and Library.* In *Database and Expert Systems Application, 2009. DEXA '09. 20th International Workshop on*, pages 79–83. IEEE, 2009. doi:10.1109/DEXA.2009.25. (Cited on page 82.)

[111] R. Larson and E. Sandhaus. *NYT to Release Thesaurus and Enter Linked Data Cloud*, June 2009. `http://open.blogs.nytimes.com/2009/06/26/nyt-to-release-thesaurus-and-enter-linked-data-cloud/`. (Cited on page 164.)

[112] A. Latif. *Discovery, Triplification and Consumption of Pertinent Resources from Linked Open Data.* Ph.D. thesis, Graz University of Technology, Graz, Austria, 2011. (Cited on page 129.)

[113] A. Latif, A. U. Saeed, P. Hoefler, A. Stocker, and C. Wagner. *The Linked Data Value Chain: A Lightweight Model for Business Engineers.* In A. Paschke, H. Weigand, W. Behrendt, K. Tochtermann, and T. Pellegrini, editors, *Proceedings of I-Semantics 2009. 5th International Conference on Semantic Systems*, pages 568–577. Journal of Universal Computer Science, 2009. (Cited on page 163.)

[114] G. Lestina, W. LaPlant, D. Gillman, and M. Appel. *Technical Development of the Proposed Statistical Metadata Standard*. Report, Bureau of the Census, 1996. (Cited on page 70.)

[115] V. I. Levenshtein. *Binary codes capable of correcting deletions, insertions and reversals*. *Soviet Physics Doklady*, 10(8):707–710, 1966. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965. (Cited on pages 125 and 137.)

[116] Y. Li and K. Bontcheva. *Hierarchical, perceptron-like learning for ontology-based information extraction*. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 777–786. ACM, New York, NY, USA, 2007. ISBN 978-1-59593-654-7. doi:10.1145/1242572.1242677. (Cited on page 92.)

[117] F. Lindenberg. *publicdata.eu: Data, Apps and 800,000 Triples*, June 2011. `http://lod2.okfn.org/2011/06/16/publicdata-eu-data-apps-and-800000-triples/`. LOD2 Open Knowledge Foundation Blog. (Cited on page 153.)

[118] K. Lyons, C. Playford, P. R. Messinger, R. H. Niu, and E. Stroulia. *Business Models in Emerging Online Services*. In M. L. Nelson, M. J. Shaw, and T. J. Strader, editors, *Value Creation in E-Business Management*, volume 36 of *Lecture Notes in Business Information Processing*, pages 44–55. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-03131-1. doi:10.1007/978-3-642-03132-8_4. (Cited on page 160.)

[119] A. Malhotra. *W3C RDB2RDF Incubator Group Report*, 2009. `http://www.w3.org/2005/Incubator/rdb2rdf/XGR-rdb2rdf/`. (Cited on page 90.)

[120] D. Martin, M. Burstein, J. Hobbs, O. Lassila, D. McDermott, S. McIlraith, S. Narayanan, M. Paolucci, B. Parsia, T. Payne, E. Sirin, N. Srinivasan, and K. Sycara. *OWL-S: Semantic Markup for Web Services*, 2004. `http://www.w3.org/Submission/OWL-S/`. W3C Member Submission 22 November 2004. (Cited on page 177.)

[121] M. Mayer and J. Menzel. *More Search Options and other updates from our Searchology event*. Google Official Blog, May 2009. `http://googleblog.blogspot.co.at/2009/05/more-search-options-and-other-updates.html`. (Cited on page 149.)

[122] S. McClean, W. Grossmann, and K. Froeschl. *Towards Metadata-Guided Distributed Statistical Data Processing*. In *Proc. of New Techniques and Technologies for Statistics (NTTS)*. 1998. (Cited on page 70.)

[123] D. L. McGuinness and F. van Harmelen. *OWL Web Ontology Language Overview.* W3C Recommendation, OWL Working Group, 2004. (Cited on pages 20 and 64.)

[124] Microsoft. *Naming a File.* `http://msdn2.microsoft.com/en-us/library/aa365247.aspx`, 2008. (Cited on page 68.)

[125] P. Mika. *Ontologies are us: A unified model of social networks and semantics. Web Semant.*, 5(1):5–15, March 2007. ISSN 1570-8268. doi:10.1016/j.websem.2006.11.002. `http://dx.doi.org/10.1016/j.websem.2006.11.002`. (Cited on page 113.)

[126] O. Moser, F. Rosenberg, and S. Dustdar. *VieDAME - flexible and robust BPEL processes through monitoring and adaptation.* In *ICSE Companion '08: Companion of the 30th international conference on Software engineering*, pages 917–918. ACM, New York, NY, USA, 2008. ISBN 978-1-60558-079-1. doi:http://doi.acm.org/10.1145/1370175.1370186. (Cited on page 177.)

[127] A. Nikolov, V. S. Uren, E. Motta, and A. N. D. Roeck. *Handling Instance Coreferencing in the KnoFuss Architecture.* In P. Bouquet, H. Halpin, H. Stoermer, and G. Tummarello, editors, *Proceedings of the 1st International Workshop on Identity and Reference on the Semantic Web (IRSW2008)*, volume 422 of *CEUR Workshop Proceedings*. CEUR-WS.org, Tenerife, Spain, 2008. `http://ceur-ws.org/Vol-422/irsw2008-submission-1.pdf`. (Cited on page 129.)

[128] M. Nottingham and R. Sayre. *The Atom Syndication Format.* RFC 4287, Network Working Group, 2005. (Cited on page 108.)

[129] N. Noy and A. Rector. *Defining N-ary Relations on the Semantic Web.* W3C Working Group Note, W3C Semantic Web Best Practices and Deployment Working Group, 2006. (Cited on page 71.)

[130] OASIS Web Services Business Process Execution Language (WSBPEL) Technical Committee. *Web Services Business Process Execution Language Version 2.0*, 2007. `http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html`. OASIS Standard. (Cited on page 175.)

[131] OECD. *Management of Statistical Metadata at the OECD.* Report, Organisation for Economic Co-operation and Development (OECD), 2006. (Cited on page 70.)

[132] D. E. O'Leary. *Wikis: 'From Each According to His Knowledge'. Computer*, 41(2):34–41, 2008. (Cited on page 110.)

[133] OpenLink Software. *Faceted Views over Large-Scale Linked Data*, 2009. http://www.openlinksw.com/dataspace/dav/wiki/Main/VirtuosoFacetsViewsLinkedData. Virtuoso Open-Source Wiki. (Cited on page 138.)

[134] OpenLink Software. *OpenLink Virtuoso Universal Server: Documentation - 16.15. RDF and Geometry*, 2009. http://docs.openlinksw.com/virtuoso/rdfsparqlgeospat.html. (Cited on page 139.)

[135] A. Osterwalder and Y. Pigneur. *An e-business model ontology for modeling e-business.* In *Proceedings of 15th Bled Electronic Commerce Conference. e-Reality: Constructing the e-Economy.* Bled, Slovenia, 2002. (Cited on page 160.)

[136] H. Papageorgiou, F. Pentaris, E. Theodorou, M. Vardaki, and M. Petrakos. *Modeling statistical metadata. Scientific and Statistical Database Management, 2001. SSDBM 2001. Proceedings. Thirteenth International Conference on*, pages 25–35, 2001. doi:10.1109/SSDM.2001.938535. (Cited on page 70.)

[137] E. Prud'hommeaux and A. Seaborne. *SPARQL Query Language for RDF.* Technical report, W3C RDF Data Access Working Group, 2008. W3C Recommendation 15 January 2008. (Cited on pages 20, 79 and 172.)

[138] Y. Raimond and S. Abdallah. *The Event Ontology.* http://motools.sourceforge.net/event/event.html, 2007. (Cited on page 64.)

[139] Y. Raimond, T. Scott, S. Oliver, P. Sinclair, and M. Smethurst. *Use of Semantic Web technologies on the BBC Web Sites.* In D. Wood, editor, *Linking Enterprise Data*, pages 263–283. Springer US, 2010. ISBN 978-1-4419-7665-9. doi:10.1007/978-1-4419-7665-9_13. (Cited on page 150.)

[140] Y. Raimond, C. Sutton, and M. Sandler. *Automatic Interlinking of Music Datasets on the Semantic Web.* In *WWW 2008 Workshop: Linked Data on the Web (LDOW2008).* Beijing, China, 2008. (Cited on pages 104, 109 and 128.)

[141] D. Recordon and D. Reed. *OpenID 2.0: a platform for user-centric identity management.* In *Proceedings of the second ACM workshop on Digital identity management*, DIM '06, pages 11–16. ACM, New York, NY, USA, 2006. ISBN 1-59593-547-9. doi:10.1145/1179529.1179532. http://doi.acm.org/10.1145/1179529.1179532. (Cited on page 103.)

[142] L. Richardson and S. Ruby. *RESTful Web Services. Web services for the real world.* O'Reilly Media, Sebastopol, CA, USA, First edition, 2007. ISBN 9780596529260. (Cited on page 23.)

[143] J. B. Rodriguez and A. Gómez-Pérez. *Upgrading relational legacy data to the semantic web.* In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 1069–1070. ACM, New York, NY, USA, 2006. ISBN 1-59593-323-9. doi:10.1145/1135777.1136019. (Cited on page 89.)

[144] S. S. Sahoo, W. Halb, S. Hellmann, K. Idehen, T. Thibodeau, S. Auer, J. Sequeda, and A. Ezzat. *A Survey of Current Approaches for Mapping of Relational Databases to RDF*, 2009. W3C RDB2RDF Incubator Group January 08 2009. (Cited on pages 10, 12, 88 and 90.)

[145] L. Sanger. *The Early History of Nupedia and Wikipedia: A Memoir.* In C. DiBona, M. Stone, and D. Cooper, editors, *Open Sources 2.0: The Continuing Evolution.* O'Reilly, 2005. (Cited on pages 105 and 109.)

[146] S. Sarawagi. *Information Extraction. Found. Trends databases*, 1(3):261–377, March 2008. ISSN 1931-7883. doi:10.1561/1900000003. (Cited on page 92.)

[147] L. Sauermann and R. Cyganiak. *Cool URIs for the Semantic Web.* W3C Interest Group Note 31 March 2008, W3C Semantic Web Education and Outreach Interest Group, 2008. `http://www.w3.org/TR/cooluris/`. (Cited on pages 17, 23, 77 and 79.)

[148] SAWSDL Working Group. *Semantic Annotations for WSDL and XML Schema*, 2007. `http://www.w3.org/TR/sawsdl/`. W3C Recommendation 28 August 2007. (Cited on page 177.)

[149] B. Schandl, B. Haslhofer, T. Bürger, A. Langegger, and W. Halb. *Linked Data and multimedia: the state of affairs. Multimedia Tools and Applications*, 59:523–556, 2012. ISSN 1380-7501. doi:10.1007/s11042-011-0762-9. (Cited on pages 11 and 99.)

[150] SDMX. *SDMX User Guide*, January 2009. `http://sdmx.org/wp-content/uploads/2009/02/sdmx-userguide-version2009-1-71.pdf`. (Cited on page 86.)

[151] Semantic Web Deployment Working Group. *SKOS Simple Knowledge Organization System Reference.* W3c recommendation, Semantic Web Deployment Work-

ing Group, 2008. http://www.w3.org/TR/skos-reference/. (Cited on pages 49, 64, 73 and 87.)

[152] N. Shadbolt, T. Berners-Lee, and W. Hall. *The Semantic Web Revisited. IEEE Intelligent Systems*, 21(3):96–101, May 2006. doi:10.1109/MIS.2006.62. (Cited on pages 4 and 22.)

[153] Y. Shafranovich. *RFC 4180: Common Format and MIME Type for Comma-Separated Values (CSV) Files.* Technical report, IETF, 2005. http://tools.ietf.org/html/rfc4180. (Cited on page 46.)

[154] A. Singhal. *Introducing the Knowledge Graph: things, not strings.* Google Official Blog, May 2012. http://googleblog.blogspot.co.at/2012/05/introducing-knowledge-graph-things-not.html. (Cited on pages 22 and 149.)

[155] S. Softic and M. Hausenblas. *Towards Opinion Mining Through Tracing Discussions on the Web.* In *Social Data on the Web (SDoW 2008) Workshop at the 7th International Semantic Web Conference*. Karlsruhe, Germany, 2008. (Cited on page 69.)

[156] S. Softic, B. Taraghi, and W. Halb. *Weaving Social E-Learning Platforms Into the Web of Linked Data.* In *Proceedings of I-KNOW '09 and I-SEMANTICS '09*, pages 559–567. Graz, Austria, 2009. (Cited on pages 10, 13 and 112.)

[157] H. Stuckenschmidt and F. van Harmelen. *Information Sharing on the Semantic Web.* Springer, 2005. (Cited on pages 61 and 70.)

[158] E. Summers. *Following your nose to the Web of Data. Information Standards Quarterly*, 20(1), 2008. (Cited on pages 63 and 130.)

[159] J. Tauberer. *The 2000 U.S. Census: 1 Billion RDF Triples.* http://www.rdfabout.com/demo/census/, 2007. (Cited on pages 60, 65, 70 and 80.)

[160] J. Tennison. *Microdata and RDFa Living Together in Harmony*, August 2011. http://www.jenitennison.com/blog/node/165. (Cited on page 156.)

[161] G. Thomas and J. Hendler. *The Web is Evolving*, November 2010. http://www.data.gov/communities/node/116/blogs/142/vid/3583. Data.gov Semantic Web Blog. (Cited on page 154.)

[162] G. Tummarello, R. Cyganiak, M. Catasta, S. Danielczyk, R. Delbru, and S. Decker. *Sig.ma: Live views on the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web*, 8(4):355 – 364, 2010. ISSN 1570-8268. doi:10.1016/j.websem.2010.08.003. `http://www.sciencedirect.com/science/article/pii/S1570826810000624`. (Cited on page 147.)

[163] G. Tummarello, R. Delbru, and E. Oren. *Sindice.com: Weaving the Open Linked Data. Proceedings of the 6th International Semantic Web Conference 2007 (ISWC2007)*, 4825:552–565, 2007. (Cited on pages 30, 79 and 148.)

[164] UN. *Guidelines for Statistical Metadata on the Internet.* Report, United Nations Economic Commission for Europe (UNECE), 2000. (Cited on page 70.)

[165] M. Volk and S. Clematide. *Learn - Filter - Apply - Forget. Mixed Approaches to Named Entity Recognition.* In *Proceedings of the 6th International Workshop on Applications of Natural Language to Information Systems*, NLDB'01, pages 153–163. GI, 2001. ISBN 3-88579-332-6. (Cited on page 95.)

[166] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer. *Semantic Wikipedia.* In *15th International Conference on World Wide Web, WWW 2006*, pages 585–594. 2006. (Cited on page 31.)

[167] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. *Discovering and Maintaining Links on the Web of Data.* In *Proceedings of the 8th International Semantic Web Conference*, ISWC '09, pages 650–665. Springer-Verlag, Berlin, Heidelberg, 2009. ISBN 978-3-642-04929-3. doi:10.1007/978-3-642-04930-9_41. `http://dx.doi.org/10.1007/978-3-642-04930-9_41`. (Cited on pages 128 and 137.)

[168] J. Voss. *Encoding changing country codes in RDF with ISO 3166 and SKOS.* In *International Conference on Metadata and Semantics Research (MTSR07).* 2007. (Cited on page 132.)

[169] W3C. *SOAP Specifications*, 2007. `http://www.w3.org/TR/soap/`. (Cited on page 95.)

[170] W3C. *Extensible Markup Language (XML) 1.0 (Fifth Edition)*, 2008. `http://www.w3.org/TR/2008/REC-xml-20081126/`. W3C Recommendation 26 November 2008. (Cited on page 17.)

[171] W3C. *eGovernment Interest Group Charter*, 2011. `http://www.w3.org/egov/IG/charter-2011`. (Cited on page 154.)

[172] W3C. *Government Linked Data Working Group Charter*, 2011. http://www.w3.org/2011/gld/charter. (Cited on page 154.)

[173] W3C. *HTML Microdata*, 2012. http://www.w3.org/TR/microdata/. W3C Working Draft 29 March 2012. (Cited on page 156.)

[174] W3C. *HTML5 - A vocabulary and associated APIs for HTML and XHTML*, 2012. http://www.w3.org/TR/html5/. W3C Working Draft 29 March 2012. (Cited on page 156.)

[175] W3C. *XHTML+RDFa 1.1 - Support for RDFa via XHTML Modularization*, 2012. http://www.w3.org/TR/xhtml-rdfa/. W3C Recommendation 07 June 2012. (Cited on pages 155 and 167.)

[176] W3C OWL Working Group. *OWL 2 Web Ontology Language. Document Overview*, 2009. http://www.w3.org/TR/owl-overview/. W3C Recommendation 27 October 2009. (Cited on page 20.)

[177] W3C RDB2RDF Working Group. *R2RML: RDB to RDF Mapping Language*, 2012. http://www.w3.org/TR/2012/WD-r2rml-20120529/. W3C Working Draft 29 May 2012. (Cited on pages 89 and 90.)

[178] C. Wagner, P. Scheir, A. Stocker, and W. Halb. *Harnessing semantic web technologies for solving the dilemma of content providers.* In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 20:1–20:5. ACM, New York, NY, USA, 2010. ISBN 978-1-4503-0014-8. http://doi.acm.org/10.1145/1839707.1839733. (Cited on pages 11, 13, 159 and 165.)

[179] T. D. Wang. *Gauging Ontologies and Schemas by Numbers.* In *4th International Workshop on Evaluation of Ontologies for the Web (EON2006).* 2006. (Cited on page 30.)

[180] W. Weiss, T. Bürger, R. Villa, P. Punitha, and W. Halb. *Statement-Based Semantic Annotation of Media Resources.* In T.-S. Chua, Y. Kompatsiaris, B. Mérialdo, W. Haas, G. Thallinger, and W. Bailer, editors, *Semantic Multimedia*, volume 5887 of *Lecture Notes in Computer Science*, pages 52–64. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-10542-5. doi:10.1007/978-3-642-10543-2_7. (Cited on pages 10, 13, 119 and 122.)

[181] W. Weiss, T. Bürger, R. Villa, P. Swamy, and W. Halb. *SALERO Intelligent Media Annotation & Search.* In *Proceedings of I-KNOW '09 and I-SEMANTICS '09*, pages 622–629. Graz, Austria, 2009. (Cited on pages 10, 13 and 119.)

[182] W. E. Winkler. *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.* In *Proceedings of the Section on Survey Research*, pages 354–359. 1990. (Cited on pages 126 and 137.)

[183] W. E. Winkler. *Overview of Record Linkage and Current Research Directions.* Research Report Series Statistics #2006-2, Statistical Research Division, U.S. Census Bureau, 2006. http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf. (Cited on pages 100 and 124.)

[184] H. Zeiner, W. Halb, H. Lernbeiß, B. Jandl, and C. Derler. *Making business processes adaptive through semantically enhanced workflow descriptions.* In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, pages 27:1–27:3. ACM, New York, NY, USA, 2010. ISBN 978-1-4503-0014-8. http://doi.acm.org/10.1145/1839707.1839741. (Cited on pages 11, 13 and 175.)