Philipp Cimiano · Christian Chiarcos
John P. McCrae · Jorge Gracia

Linguistic Linked Data

Representation, Generation and Applications

Springer

Linguistic Linked Data

Philipp Cimiano • Christian Chiarcos •
John P. McCrae • Jorge Gracia

Linguistic Linked Data

Representation, Generation and Applications

Springer

Philipp Cimiano
Semantic Computing Group
Bielefeld University
Bielefeld
Germany

Christian Chiarcos
Angewandte Computerlinguistik
Goethe-University
Frankfurt am Main
Germany

John P. McCrae
Insight Centre for Data Analytics
National University of Ireland
Galway, Ireland

Jorge Gracia
Aragon Institute of Engineering
Research (I3A)
University of Zaragoza
Zaragoza, Spain

ISBN 978-3-030-30224-5 ISBN 978-3-030-30225-2 (eBook)
https://doi.org/10.1007/978-3-030-30225-2

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The four authors of this book have a long-standing collaboration that goes back to the year 2009 at least, when the project Monnet funded by the European Commission started. This was one of the first EC-funded projects concerned with investigating the relationship between language and ontologies and linked data in particular. Within this project, in which Jorge Gracia, John McCrae and Philipp Cimiano were direct collaborators, crucial foundations for the work described in this book were laid. On the one hand, the lemon model was developed as a direct result of the Monnet project. Further, seminal work on how to localize ontologies into multiple languages was carried out as part of the Monnet project. Within the LIDER project, also funded by the European Commission subsequently to Monnet, Jorge Gracia, John McCrae and Philipp Cimiano collaborated on developing guidelines for the modelling, generation and publication of linguistic linked data. Since January 2019, these activities are being continued in the context of the H2020 project 'Prêt-à-LLOD[1]: Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors', now involving all authors of this book and with a focus on the practical application of linguistic linked data technologies.

Independently of the line of work pursued within Monnet and LIDER, applications of linked data and semantic technologies to language resources and language technology have been developed in various other projects around this time. Most notably, this includes large-scale coordinated research actions, e.g. a project on 'Sustainability of Linguistic Data' funded by the German Research Foundation as a collaborative effort between three Collaborative Research Centres situated in Tübingen, Hamburg and Berlin/Potsdam, respectively. Out of this context, Christian Chiarcos began to apply semantic technologies, and in particular the ontology web language (OWL), to model linguistic annotations since 2005, and annotated corpora since 2009.

[1]The authors acknowledge funding by the European Commission under H2020 project Prêt-à-LLOD under grant agreement 825182.

These and related efforts by interested scholars and applicants of language resources and semantic technologies increasingly converged with the foundation of the Open Linguistics Working Group (OWLG), founded in October 2010, with Christian Chiarcos as one of its founding members, and the development of the Linguistic Linked Open Data (LLOD) cloud that grew out of this working group since early 2011. Around the same time, in July 2011, the Working Group on the Ontology-Lexicon Interface (Ontolex) was founded, with the Ontolex-lemon model as its output, and remarkable impact on the digital edition of lexical resources since then. With increasing interest in linked data beyond open resources, the term 'linguistic linked data' emerged as a generalization over 'linguistic linked open data'. Throughout this book, both terms are used interchangeably, and albeit the technology not being restricted to open resources, many prominent data sets are indeed available under open licenses.

The following four publications can be regarded as the seminal publications that defined the linguistic linked data paradigm:

- Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann. Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata. Springer, Heidelberg, 2012.
- Christian Chiarcos, Steven Moran, Pablo N. Mendes, Sebastian Nordhoff, and Richard Littauer. Building a Linked Open Data cloud of linguistic resources: Motivations and developments. In Iryna Gurevych and Jungi Kim (eds.), The People's Web Meets NLP. Collaboratively Constructed Language Resources. Springer, Heidelberg, 2013.
- Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum. Towards open data for linguistics: Lexical Linked Data. In Alessandro Oltramari, Piek Vossen, Lu Qin, and Eduard Hovy (eds.), New Trends of Research in Ontologies and Lexical Resources. Springer, Heidelberg, 2013.
- Jorge Gracia, Elena Montiel-Ponsoda, Philipp Cimiano, Asunción Gómez-Pérez, Paul Buitelaar, and John McCrae. Challenges for the multilingual Web of Data. Journal of Web Semantics, vol. 11, pp. 63–71. Elsevier B.V., 2012.

Since these seminal publications, a number of workshops have been organized on the topic including the well-known series of workshops on *Linked Data in Linguistics*, the series of workshops on the *Multilingual Semantic Web* as well as the *Summer Datathon on Linguistic Linked Open Data* series of summer schools, and the conference series on *Language, Data and Knowledge* (LDK). The authors of this book have all been key players in the organization of all these events.

The ideas developed in the above-mentioned initial collaborations roughly 10 years ago have been spreading at an initially unimagined way. What started as a rather naive and idealistic effort of improving the state of affairs concerning data reuse and interoperability, has turned into a standard approach for data sharing in computational linguistics, lexicography, typology, language research and digital humanities. The vocabularies that have emerged as part of the working groups related to the linguistic linked data paradigm, such as lemon, are widely used for the publication of language and linguistic resources.

The authors of this book deeply enjoy the uptake that their ideas have received. For all of us, it has been a big honour to be able to work together with so many people on the foundations of linguistic linked data.

This book is thus a result of the efforts of a whole community that has firmly pushed the ideas of sharing and reusing linguistic resources further and has worked out many details of the linked data approach in linguistics. This book would not have been possible without all these community efforts. The authors thus would like to dedicate this book to this passionate community that has vigorously believed in the ideas of open science and reused open standards and formats to improve the affairs of data sharing, publishing and reuse in linguistics by adopting the linked data principles put forth by Tim Berners-Lee. The linguistic linked data program is showing clear fruits by now, in that in using linked data principles to publish linguistic datasets and language resources, it is demonstrably easier to find and reuse datasets.

We hope you like this book!

Bielefeld, Germany — Philipp Cimiano
Frankfurt, Germany — Christian Chiarcos
Galway, Ireland — John P. McCrae
Zaragoza, Spain — Jorge Gracia
June 2019

Contents

Part I Preliminaries

1 Introduction 3
1.1 FAIR Principles 4
1.2 Linked Data as an Opportunity to Realize the FAIR Principles 4
References 8

2 Preliminaries 11
2.1 Introduction 11
2.2 Resource Description Framework 12
2.3 Serializing RDF 14
2.3.1 The N-Triples Language 14
2.3.2 Turtle 15
2.3.3 RDF/XML 16
2.3.4 RDFa 16
2.3.5 JSON-LD 16
2.4 RDF Semantics, RDFS and OWL 19
2.4.1 RDF Semantics 19
2.4.2 RDF Schema 21
2.4.3 Web Ontology Language (OWL) 22
2.5 The SPARQL Query Language 24
2.5.1 Publishing Data on the Web 26
2.6 Summary and Further Reading 27
References 27

3 Linguistic Linked Open Data Cloud 29
3.1 Background and Motivation 29
3.1.1 Linked Data 30
3.1.2 Linked Open Data 31
3.2 Linguistic Linked Open Data 32
3.2.1 The LLOD Cloud 33
3.2.2 Infrastructure and Metadata 36

3.3 LLOD Community 38
3.3.1 Summary and Further Reading 40
References 40

Part II Modelling

4 Modelling Lexical Resources as Linked Data 45
4.1 Introduction 45
4.2 The Core Model 46
4.3 Syntax and Semantics 48
4.4 Decomposition 53
4.5 Variation and Translation 54
4.6 Metadata 55
4.7 Applications 56
4.8 Summary and Further Reading 57
References 58

5 Representing Annotated Texts as RDF 61
5.1 Introduction 61
5.1.1 Tab-Separated Values: CoNLL TSV 61
5.1.2 Tree-Based Formats: TEI/XML 64
5.2 Annotating Web Resources 66
5.2.1 Web Annotation (Open Annotation) 67
5.2.2 Annotating Named Entities on the Web 69
5.3 Annotating Textual Objects 73
5.3.1 The NLP Interchange Format (NIF 2.0) 75
5.3.2 Provenance and Annotation Metadata in NIF 82
5.4 Summary and Further Reading 83
References 85

6 Modelling Linguistic Annotations 89
6.1 Introduction 89
6.2 Transforming Legacy Annotation Formats into RDF 91
6.2.1 CoNLL-RDF: Shallow Transformation of CoNLL into RDF 91
6.2.2 Querying and Manipulating CoNLL-RDF Annotations 94
6.3 Top-Down Modelling: Generic Data Structures 97
6.3.1 Linguistic Annotations in POWLA 98
6.3.2 Complementing NIF with POWLA 103
6.3.3 Transforming CoNLL-RDF to POWLA 111
6.4 Querying Annotated Corpora 113
6.5 Summary and Further Reading 118
References 120

7 Modelling Metadata of Language Resources 123
7.1 Introduction 123
7.2 Models for General Metadata 125
7.2.1 DC-Terms 126
7.2.2 DCAT 127
7.3 Modelling Metadata of LRs with Meta-Share.owl 130
7.4 Summary and Further Reading 134
References 135

8 Linguistic Categories 137
8.1 Introduction 137
8.2 The Case of Language Identifiers 140
8.2.1 ISO 639 Language Tags 140
8.2.2 IETF Language Tags 142
8.2.3 URI-Based Language Codes 144
8.3 General Repositories of Linguistic Reference Terminology 147
8.3.1 Data Category Modelling and Standardization in ISOcat 147
8.3.2 The General Ontology of Linguistic Description (GOLD) 148
8.3.3 Transition to the CLARIN Concept Registry and DatCatInfo 149
8.4 Application-Specific Terminology Repositories 150
8.4.1 LexInfo: Linguistic Categories for Lexical Resources 150
8.4.2 OLiA: Ontologies of Linguistic Annotation 151
8.4.3 Limits of Axiomatization 155
8.5 Summary and Further Reading 157
References 158

Part III Generation and Exploitation

9 Converting Language Resources into Linked Data 163
9.1 Introduction 163
9.2 General Methodology for Generating and Publishing LLD 164
9.3 Specification 164
9.4 Modelling 169
9.5 Generation 172
9.6 Linking 175
9.7 Publication 175
9.8 Exploitation 176
9.9 Guidelines for Particular Types of Language Resources 177
9.10 Inclusion into the LLOD Cloud 178
9.11 Summary and Further Reading 179
References 180

10 Link Representation and Discovery 181
10.1 Link Representation 181
10.1.1 Patterns for Creating Cross-Lingual Links at the Conceptual Level 184
10.1.2 Cross-Lingual Links at the Linguistic Level 185
10.2 Link Discovery 186
10.2.1 Problem Statement 187
10.2.2 Classification of Matching Techniques 188
10.2.3 Terminological Similarity 188
10.2.4 Structural Similarity 190
10.2.5 Cross-Lingual Linking 192
10.3 Linking Frameworks 193
10.4 Summary and Further Reading 194
References 194

11 Linked Data-Based NLP Workflows 197
11.1 Introduction 197
11.2 Implementing NLP Workflows Using NIF 199
11.2.1 Implementing a NIF-Compliant POS Tagging Service 200
11.2.2 Implementing a NIF-Based Dependency Parsing Web Service 201
11.2.3 Creating NLP Workflows with NIF-Based Services 204
11.3 Composing NLP Workflows with Teanga 204
11.3.1 Design and Implementation 205
11.3.2 Services in Teanga 205
11.3.3 Building Workflows 206
11.4 LAPPS Grid 208
11.5 Summary and Further Reading 209
References 210

Part IV Use Cases

12 Applying Linked Data Principles to Linking Multilingual Wordnets 215
12.1 Princeton WordNet 215
12.1.1 WordNet RDF 217
12.2 Global WordNet Interlingual Index 217
12.2.1 The Global WordNet Grid 218
12.2.2 Collaborative Interlingual Index 219
12.2.3 ILI Format 220
12.2.4 Linking WordNet with Wikipedia 223
12.3 Summary and Further Reading 224
References 224

13 Linguistic Linked Data in Digital Humanities 229
13.1 Introduction 229
13.2 Data Models and Vocabularies in DH 231
13.2.1 The Text Encoding Initiative 231
13.2.2 Simple Knowledge Organization System (SKOS) 234
13.2.3 CIDOC Vocabulary for Describing Object Metadata 236
13.2.4 The Canonical Text Service Protocol (CTS) 238
13.3 Case Studies and Applications of LLOD in Digital Humanities ... 241
13.3.1 Applying LLOD Methods in Prosopography 242
13.3.2 Using LLOD Techniques to Reference Geographical Information 244
13.3.3 Constructing a Database and Dictionary of Maya Hieroglyphic Writing 247
13.3.4 Facilitating the Study of Ancient Wisdom Literature 250
13.3.5 Encoding Chauliac's *Grande Chirurgie* with TEI and RDFa 253
13.4 Summary and Further Reading 256
References 258

14 Discovery of Language Resources 263
14.1 Introduction 263
14.2 Data Collection 265
14.2.1 META-SHARE 266
14.2.2 CLARIN 266
14.2.3 LRE Map 267
14.2.4 Linked Open Data Cloud/Datahub.io 267
14.2.5 Other Repositories 268
14.2.6 State of Play with Respect to Finding Language Resources on the Web 269
14.3 Modelling 269
14.4 Harmonization 270
14.4.1 Availability 271
14.4.2 Rights 271
14.4.3 Usage 272
14.4.4 Language 273
14.4.5 Type 273
14.4.6 Duplicate Detection 274
14.4.7 Data Completeness and Quality 275
14.5 Publishing Linghub with Yuzu 276
14.6 Summary 277
References 278

Part V Conclusions

15 Conclusion 283

A Selected Prefix Declarations 285

Acronyms and Abbreviations

This section lists the set of acronyms and abbreviations used in this book.

BioNLP	Biomedical Natural Language Processing
CC(-BY)(-NC)(-SA)(-ND)	Creative Commons (Attribution) (NonCommercial) (ShareAlike) (NonDerivative)
CCR	CLARIN Concept Registry
CIDOC	(French) *Comité International pour la Documentation*
CIDOC CRM	CIDOC Conceptual Reference Model
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
CoNLL	Conference on Computational Natural Language Learning
CSV	Comma-Separated Values
CTS	Canonical Text Service
cURL	Client for URLs/Curl URL Request Library
DC	Dublin Core
DCAT	Data Catalogue Vocabulary
DC-Terms	Dublin Core Terms
DH	Digital Humanities
DTD	Document Type Definition
DTS	Distributed Text Services
ELRA	European Language Resource Association
FAIR	Findable, Accessible, Interoperable and Re-usable
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
ID	Identifier
IOB(ES)	Inside-outside-beginning-end-start format
IRI	Internationalized Resource Identifier

ISO	International Standards Organization
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
KOS	Knowledge organization system
LAF	Linguistic Annotation Framework
LD	Linked data
LDC	Linguistic Data Consortium
lemon	Lexicon model for ontologies
LLD	Linguistic linked data
LLOD	Linguistic linked open data
LMF	Lexical Markup Framework
LOD	Linked Open Data
LOV	Linked Open Vocabularies
LR	Language resource
NER	Named entity recognition
NLP	Natural language processing
NIF	NLP Interchange Format
OLAC	Open Language Archives Community
OWL	Web Ontology Language
OWL (2) DL	Web Ontology Language (version 2), Description Logics
OWLG	Open Linguistics Working Group
POS	Part of speech
RDF	Resource Description Framework
RDFa	RDF in attributes
RDFS	RDF Schema
RST	Rhetorical Structure Theory
SGML	Standard Generalized Markup Language
SKOS	Simple Knowledge Organization System
SKOS-XL	Simple Knowledge Organization System eXtension for Labels
SPARQL	SPARQL Protocol and RDF Query Language
SRL	Semantic role labelling
SW	Semantic Web
TBX	TermBase eXchange format
TEI	Text Encoding Initiative
TSV	Tab-separated values
Turtle	Terse RDF Triple Language
UD	Universal Dependencies
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
W3C	World Wide Web Consortium
XML	eXtensible Markup Language
XSD	XML Schema Definition

Part I
Preliminaries

Chapter 1
Introduction

Digital language resources, comprising spoken and written material, are key to many fields, including linguistics research, lexicography, typology, the study of minority or extinct languages, but also to the development of machine-learned models for automated natural language processing (NLP).

Thus, many groups and institutions worldwide are active in the creation of language resources, comprising activities such as data collection, transcription of recordings, corpus creation, data annotation, quality control, etc.

The digital language resources that have been created so far and that will be created in the future represent an important cultural asset and treasure that not only allows us to develop NLP solutions or perform linguistic research today, but also to document the status of development of languages worldwide and preserve our way of thinking, our cultural identity, etc.

Language resources are thus an important cultural asset that need not only to be preserved, we need to also make sure that these resources can be reused as much as possible. In particular, a crucial issue is to maximize secondary reuse of language resources, that is ensuring that the data can be used by others for a different purpose than it was originally collected for. However, secondary reuse is in many cases hindered by a number of proprietary choices made by the data collector. Such choices include, for instance, the use of proprietary formats (either because no standard formats are available or because some formats require paying licenses for proprietary software, etc.). Other obstacles for secondary reuse are of a more conceptual nature including choices in data collection or annotation that limit the scope and applicability of the data in other contexts as well as mismatching conceptualizations of phenomena as reflected in annotation schemas. To maximize reuse, as a community we need guiding principles that can be followed when documenting, publishing and processing data.

P. Cimiano et al., *Linguistic Linked Data*,
https://doi.org/10.1007/978-3-030-30225-2_1

1.1 FAIR Principles

Secondary reuse of data is not only a concern within linguistics research. It is an issue that is relevant for any scientific discipline. In fact, the degree to which agreed-upon principles and standards for data management and reuse are available and followed on can be regarded as an indicator of maturity of a scientific discipline.

As a step towards increasing transparency and reproducibility in science, in 2016 a group of researchers around M.D. Wilkinson postulated the so-called FAIR Guiding Principles [1]. The acronym FAIR stands for Findable, Accessible, Interoperable and Re-usable:

- *Findability* implies that data and metadata are assigned globally unique and eternally persistent identifiers, and that the data is accompanied by rich metadata and that data is registered or indexed somewhere where it can be found.
- *Accessibility* implies that (1) data is retrievable by their identifier using an (2) open, free and universally implemented protocol, and (3) the protocol supports authentication and authorization if necessary.
- *Interoperability* implies that the data is described using a formal, accessible, shared and a standard data model to support sharing.
- Finally, *re-usability* implies accurate and relevant attributes, clear licensing and data usage terms and conditions, linking to provenance of data and the adherence to community standards.

The FAIR principles are clearly also relevant for linguistics research and there should be a broad interest in ensuring the FAIR principles for digital language resources to maximize their reuse. However, most of the solutions proposed so far fail on a number of FAIR principles.

1.2 Linked Data as an Opportunity to Realize the FAIR Principles

Language resources (dictionaries, terminologies, corpora, etc.) developed in the fields of corpus linguistics, computational linguistics and natural language processing (NLP) are often encoded in heterogeneous formats and developed in isolation from one another. This makes their discovery, reuse and integration for both the development of NLP tools and daily linguistic research a difficult and cumbersome task. In order to alleviate such an issue and to enhance interoperability of language resources on the Web, a community of language technology experts and practitioners has started adopting techniques coming from the field of linked data (LD). The LD paradigm emerged as a series of best practices and principles for exposing, sharing and connecting data on the Web [2].

The LD principles state that unique resource identifiers (URIs) should be used to name things in a way that allows people to look them up, to get useful information

for each of these resources and to discover related resources or entities. The four linked data principles are the following:

1. Use URIs as (unique) 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 Web standards such as the Resource Description Framework (RDF) and SPARQL.
4. Include links to other URIs, so that they can discover more things.

The first principle means that we assign a unique identifier (URI, [3]) to every element of a resource, i.e. each entry in a lexicon, each document in a corpus, every token in a corpus and to each data category that we use for annotation purposes. The benefit is that this makes elements, categories and annotations uniquely and globally identifiable in an unambiguous fashion. The second principle entails that any agent wishing to obtain information about the resource can contact the corresponding web server and retrieve this information using a well-established protocol (HTTP) that also supports different 'views' on the same resource. That is, computer agents might request a machine-readable format, while web browsers might request a human-readable and browsable view of this information as HTML. The third principle requires the use of standardized, and thus inter-operable data models for representing data (RDF, [4]) and querying linked data (SPARQL, [5]). The fourth principle fosters the creation of a network of language resources where objects of linguistic interest (words, senses, annotations) are connected to each other via links that express equivalence, relatedness, etc. and are linked to data categories defined in data category repositories such as ISOCat.

LD emerged in the context of the Semantic Web, an extension of the Web in which information is given *'well-defined' meaning*, *'better enabling computers and people to work in cooperation'* [6]. The LD principles have been applied to transform the current human-readable Web into a 'Web of Data' in which resources are linked across datasets and sites, and where facts and related knowledge are available for consumption by advanced, knowledge-based software agents as well as by humans through suitable interfaces.

The Semantic Web builds on so-called ontology languages, the Web Ontology Language (OWL)[1] in particular, to formally and axiomatically define the vocabulary used to describe data. The data model used to describe data is the Resource Description Framework (RDF),[2] which models data through the central notion of triples (s, p, o) consisting of a subject, a predicate and an object.

We mention below how the LD principles can support the realization of the FAIR principles for language data:

- Findability: First of all, by relying on URIs as globally unique identifiers, LD allows to unambiguously identify a particular data source as well as data element

[1] https://www.w3.org/OWL/.
[2] https://www.w3.org/RDF/.

contained in that resource. By relying on standard languages for description of content and metadata used for LD as well as by following the LD principles, language resources can be published in such a way that they can be indexed by semantic search engines and repositories that themselves can expose them in an appropriate fashion for their community members. LD provides mechanisms and vocabulary to describe information about a resource (metadata). This ensures that data can be searched and found more effectively.

- Accessibility: By following standard data models such as RDF and publishing data following the LD principles, homogeneity in data publication and thus data access can be achieved. By dereferencing URIs, people can get direct access to the content described in standard data formats and languages, being able to use standard tools for processing, querying and visualizing the data.
- Interoperability: LD fosters the reuse of existing ontologies and vocabularies and thus creates the basis for interoperability by encouraging the reuse of vocabulary elements existing already. As these vocabulary elements are formally described using ontology languages, this allows one to review and assess whether the meaning is appropriate when reusing the corresponding vocabulary elements, thus reducing ambiguity and making semantic choices transparent. Publishing and describing resources in a semantically non-ambiguous way creates the foundations for interoperability.
- Re-usability: By fostering reuse of semantically well-defined vocabularies and by adherence to standard data formats, LD has the potential to facilitate the reuse of data beyond its primary purpose. LD provides vocabularies for describing provenance information, terms of use and licensing conditions associated with data, a crucial aspect for data reuse.

As a consequence of the above advantages, imagine that for some linguistic study, all relevant datasets can be queried in the same manner for data describing a particular phenomenon under investigation. Such an integrated view over very different datasets is not possible given the current best practices in the management and sharing of language resources.

Given the advantages and the potential of LD for improving the usability and reusability of language and linguistic resources, since 10 years a research community has emerged that is studying how the LD principles can be applied to the modelling of linguistic data and language resources, taking into account the peculiarities of this domain of application. The community has been very active in developing vocabularies, best practices, tools, but also in understanding the benefits of the LD approach as well as systematizing the field.

As one aspect of this systematization effort, the community has developed early on the so-called *Linguistic Linked Open Data (LLOD) cloud*,[3] which is a depiction of the growing ecosystem of semantically connected linguistic datasets on the Web. The LLOD cloud is a community effort launched by The Open Knowledge Foun-

[3] http://linguistic-lod.org/llod-cloud.

dation's Working Group on Open Data in Linguistics (OWLG)[4] [7, 8] as a first step to bridge the gap between the advances in language technologies, and linguistics in general, and those taking place in the Semantic Web and artificial intelligence communities. Its main goal is to promote and track the use of LD in linguistics and facilitate the access to available language resources. Some recent advancements in LLD have also been driven by the activities developed within the framework of international projects such as LIDER,[5] FREME[6] and, more recently, Prêt-à-LLOD,[7] among others. Workshops, datathons and conferences such as the Multilingual Semantic Web Workshop,[8] the linked data in Linguistics Workshop (LDL),[9] the Workshop on Knowledge Extraction and Knowledge Integration (KEKI),[10] the Summer Datathon on Linguistic Linked Open Data,[11] the Conference on Language, Data and Knowledge (LDK),[12] the NLP&DBpedia Workshop Series,[13] among other initiatives, have encouraged interdisciplinary contributions and community gathering, and provide a perfect scenario to establish new collaborations along these lines of work.

As the interest of the Semantic Web and computational linguistics communities in LLD keeps increasing, and successive initiatives and workshops encourage and discuss their use and their potential benefits, the number of contributions that dwell on LLD grows rapidly. LD is increasingly being adopted by the computational linguistics and the digital humanities communities [7, 9–17], and an extensive number of efforts are now devoted towards the conversion of language resources to RDF.

This book describes how the LD principles can be applied to modelling language resources. The first part of this book until Chap. 3 provides foundations for understanding the remainder of the book. Chapter 2 in particular introduces the data models, ontology and query languages used as the basis of the Semantic Web and linked data. Chapter 3 provides a more detailed overview of the Linguistic Linked Data (LLD) Cloud as mentioned above.

The second part of the book focuses on modelling language resources using LD principles. Chapter 4 describes how to model lexical resources using Ontolex-lemon, the lexicon model for ontologies. Chapter 5 describes how to annotate and address elements of text represented in RDF. While Chap. 6 shows how to model

[4] http://linguistics.okfn.org/.
[5] http://lider-project.eu/.
[6] http://www.freme-project.eu/.
[7] http://www.pret-a-llod.eu/.
[8] http://msw4.insight-centre.org/.
[9] http://ldl2018.linguistic-lod.org/.
[10] http://keki2016.linguistic-lod.org/.
[11] http://datathon2017.retele.linkeddata.es/.
[12] http://ldk2017.org/.
[13] http://nlpdbpedia2015.wordpress.com/, http://nlpdbpedia2016.wordpress.com/.

annotations, Chap. 7 describes how to capture metadata of language resources. Chapter 8 shows how to represent linguistic categories and concludes Part II.

In the third part of the book, we describe how language resources can be transformed into LD in Chap. 9. Chapter 10 describes how links can be inferred and added to the data to increase connectivity and linking between different datasets. Chapter 11 discusses how to use LD resources for natural language processing.

The last part of the book, part IV, describes concrete applications of the technologies introduced in this book: that is representing and linking multilingual wordnets (Chap. 12), applications in digital humanities (Chap. 13) and discovery of language resources (Chap. 14).

References

1. M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, L.B. da Silva Santos, P.E. Bourne et al., The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3**, 160018 (2016)
2. C. Bizer, T. Heath, T. Berners-Lee, Linked data-the story so far. Int. J. Semant. Web Inf. Syst. **14**, 205 (2009)
3. T. Berners-Lee, R. Fielding, L. Masinter, Uniform Resource Identifier (URI): Generic Syntax (RFC 3986). Technical Report W3C (2005), http://www.ietf.org/rfc/rfc3986.txt
4. G. Klyne, J. Carroll, B. McBride, Resource Description Framework (RDF): Concepts and Abstract Syntax. Technical Report W3C Recommendation (2004), http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
5. S. Harris, A. Seaborne, SPARQL 1.1 query language. W3C recommendation, World Wide Web Consortium (2013)
6. T. Berners-Lee, J. Hendler, O. Lassila et al., The Semantic Web. Sci. Am. **284**(5), 28 (2001)
7. C. Chiarcos, S. Hellmann, S. Nordhoff, The Open Linguistics Working Group of the Open Knowledge Foundation, in *Linked Data in Linguistics* (Springer, Heidelberg, 2012), pp. 153–160
8. J. McCrae, C. Chiarcos, F. Bond, P. Cimiano, T. Declerck, The Open Linguistics Working Group: developing the Linguistic Linked Open Data cloud, in *Proceedings of the 10th Language Resources and Evaluation Conference (LREC)*, Portoroz, 2016, pp. 2435–2441
9. T. Declerck, P. Lendvai, K. Mörth, G. Budin, T. Váradi, Towards linked language data for digital humanities, in *Linked Data in Linguistics* (Springer, Berlin, 2012), pp. 109–116
10. C. Chiarcos, J. McCrae, P. Cimiano, Towards open data for linguistics: linguistic linked data, in *New Trends of Research in Ontologies and Lexical Resources* (Springer, Berlin, 2013), pp. 7–25
11. S. Hellmann, J. Lehmann, S. Auer, M. Brümmer, Integrating NLP using linked data, in *Proceedings of the International Semantic Web Conference (ISWC)* (Springer, Berlin, 2013), pp. 98–113
12. P. Cimiano, J.P. McCrae, T. Gornostay, B. Siemoneit, A. Lagzdins, Linked terminology: applying linked data principles to terminological resources, in *Proceedings of the 4th Biennial Conference on Electronic Lexicography (eLex)* (2015), pp. 1–11
13. J. McCrae, C. Fellbaum, P. Cimiano, Publishing and linking WordNet using lemon and RDF, in *Proceedings of the 3rd Workshop on Linked Data in Linguistics* (2014)
14. T. Flati, R. Navigli, Three birds (in the LLOD cloud) with one stone: BabelNet, Babelfy and the Wikipedia Bitaxonomy, in *Proceedings of SEMANTiCS* (2014)
15. I. El Maarouf, E. Alferov, D. Cooper, Z. Fang, H. Mousselly-Sergieh, H. Wang, The GuanXi network: a new multilingual LLOD for language learning applications, in *Proceedings of the*

2nd Workshop on Natural Language Processing and Linked Open Data (NLP&LOD2) (2015), p. 42
16. M. Villegas, M. Melero, N. Bel, J. Gracia, Leveraging RDF graphs for crossing multiple bilingual dictionaries, in *Proceedings of the 10th Language Resources and Evaluation Conference (LREC)*, Portoroz (2016)
17. E. González-Blanco, G. Del Río, C.I. Martínez Cantón, Linked open data to represent multilingual poetry collections. A proposal to solve interoperability issues between poetic repertoires, in *Proceedings of the 5th Workshop on Linked Data in Linguistics (LDL 2016): Managing, Building and Using Linked Language Resources*, Portoroz (May 2016)

Chapter 2
Preliminaries

Abstract This chapter introduces preliminaries that are essential to follow the content in the remainder of this book. First of all, we introduce the core data model of the Semantic Web and linked data, that is the Resource Description Framework, RDF. This format was designed in the 1990s and its core purpose is to represent data and knowledge in a Web-compatible fashion, taking into account that the Web can be regarded as a network of linked sites. RDF allows one to define networks of connected 'things' rather than a network of connected documents. We briefly introduce the semantics of RDF and also introduce the most popular serialization formats for RDF, that is N-Triples, Turtle, XML and JSON-LD. Glossing over many details, we briefly introduce the Web Ontology Language (OWL) as a vocabulary to describe ontological and terminological knowledge and SPARQL, the query language for RDF and linked data. Finally, we briefly discuss aspects of publishing linked data.

2.1 Introduction

Linked data is the term used to refer to *interlinked* collections of datasets published on the Web. To support publication of datasets and their linking, a number of standards have been developed, in particular by the World Wide Web Consortium (W3C) as part of the effort to provide standards and representation languages for a machine-readable Web in which information is given *'well-defined' meaning*, to *'better enable computers and people to work in cooperation'* (see [1]). Inspired by the fact that the Web is a network/linked graph of documents, the goal of the Semantic Web was not only to talk about 'documents' but also about the 'things' that exist in the world, elevating the latter to objects than one can actually talk about and describe in terms of their relations/connections to other objects that exist as well.

The basic data model behind the Semantic Web and linked data is the *Resource Description Framework* (RDF). As the name suggests, it is a data model that allows to describe resources, mainly via attributes and their relations to other resources. The data model relies on triples (s, p, o) connecting a so-called subject s to an

P. Cimiano et al., *Linguistic Linked Data*,
https://doi.org/10.1007/978-3-030-30225-2_2

object o via predicate (called property in RDF) p. An RDF document is a set of such triples. Alternatively, an RDF document can also be viewed as a directed, labelled graph where s and o correspond to vertices (nodes) and p is the label of an edge connecting node s to node o.

A number of tools and further models have been developed allowing to access, query and manipulate RDF data. For example, the *RDF Schema Language* (RDFS) allows to define further rules to infer additional triples from the data that are not explicitly mentioned in the data. The *Web Ontology Language* (OWL) further extends this reasoning capability allowing for a subset of First-Order Logic statements to be made, following the family of so-called *Description Logics* [2]. For example, one could define rules such as that "The *gender* property of any *Noun*, whose *language* value is *French*, has the value of either *masculine* or *feminine*."

The need to store and query RDF data is of course paramount to its usability. SPARQL, the *SPARQL Protocol and RDF Query Language*, was developed for this purpose in order to provide ways for querying RDF datasets, analogously to the use of SQL in traditional relational databases. Further, as most of the data on the Web is not in RDF, an important task consists in transforming it into RDF. In particular, we will look at the *JSON-LD* data model, which allows JSON documents to be interpreted as RDF documents.

2.2 Resource Description Framework

The Resource Description Framework (RDF) is a standard that was created for the representation of data on the Semantic Web. As mentioned above, an RDF document essentially consists of a set of triples $\langle s, p, o \rangle$, with s being the so-called subject and o being the object of the triple and p being the property relating the subject to the object. Subject and object are so-called *resources* that are typically represented using Uniform Resource Identifiers (URIs), resp. Internationalized Resource Identifiers (IRIs) as standardized by the World Wide Web Consortium.[1] By using URIs to identify resources, one can uniquely identify the entity denoted by these URIs. It is important to note that URIs as identifiers are global and thus shared across all Web documents existing worldwide. As an exception to using URIs at subject, predicate and object position (predicates are represented by URIs too!), we can use so-called blank nodes at subject and object position. While not totally accurate, for the sake of this book it is sufficient to understand blank nodes as existentially quantified variables, the scope of which is limited to a given RDF document. In contrast, URIs are logically speaking constants that are globally defined. We note that the use of blank nodes is often discouraged for a number of technical reasons (see [3], for instance).

[1] The original RDF specification required the use of URIs. RDF 1.1 requires IRIs, instead, that is the internationalized form of URIs with non-ASCII characters are supported.

A URI is a string of the following form [4]:

```
scheme:[//authority]path[?query][#fragment]
```

These are defined as follows

- **Scheme**: The scheme defines the protocol by which the resource may be located; it is usually one of the standard web protocols, e.g. `http`, `https` or `ftp`.
- **Authority:** This typically identifies the server where the resource is available normally of the form `user:password@host:port` where
  - **User and Password:** These are log-in details for the host. This is generally omitted as most URLs do not require a log-in to access.
  - **Host:** The name of the server that holds the resource either as an IP address or more frequently as a DNS name such as www.example.org.
  - **Port:** The port of the server to use. If omitted it is assumed that this will be the default port for the protocol, e.g. 80 for HTTP.
- **Path:** The scheme-specific locator for the resource. In the case of HTTP URLs, this is the path of the file on the server.
- **Query:** An optional extra path used for the dynamic generation of resources. URIs in RDF should generally not have a query string as resources represent fixed data.
- **Fragment** An identifier for locating the resource within a single file. The fragment is not normally passed to the server, but instead should be resolved by the client as a fragment normally refers to a resource that is a part of a larger document.

In RDF, predicates have a dual role. On the one hand they can be used to describe a subject (resource) by its relation to some other resource (object). The type of relation is then specified by the predicate. In this case the object is another resource denoted by a URI. Predicates can also be used to describe intrinsic properties of subject resources, thus playing the role of attributes. In this case the object can be a (typed) literal, which can be one of the following:

- A **plain literal** is just a (Unicode) string and should be used in limited contexts, i.e. for representing codes and identifiers.
- A **typed literal** has a type, typically from the XML Schema Types [5], although custom values may be defined (see [6]). This can be used for typical data values such as numeric, date and time values. Note that a plain literal is considered to have the XML Schema `string` type but is not equivalent to a typed literal with this type. This type must be a URI.
- A **Language-tagged literal** allows to add a specification of language by way of a so-called language tag, which is typically a two-letter code from ISO-639-1 [7, 8], but may be any IETF language tag (Sect. 8.2.2).

2.3 Serializing RDF

There are different formats for serializing RDF data so that it can be published on the Web. In this section we briefly present the most important RDF serializations, including N-Triples, Turtle, RDF/XML, RDFa and JSON-LD.

2.3.1 The N-Triples Language

An RDF graph consists of a set of triples that are contained in a single document. There are different possible serializations for RDF data. One of them is the N-Triples syntax, which lists all of the triples in their full form separated by the period symbol ".". URIs are typically given in their full form surrounded by angular brackets, e.g. `<http://www.example.org/resource#identifier>` and must be absolute (i.e. specify the scheme and path). Blank nodes start with `_:` and then a label that is an alphanumeric string. Literals are enclosed in double quotes, e.g. `"`, and may be followed by either a language tag with the `@` sign or a datatype with the `^^` symbol followed by a URI (in angular brackets).

An example of RDF data serialized in N-Triples format is given in Fig. 2.1. The example describes an English WordNet synset *06422547-n* that represents the concept `book`.

This refers to a particular document at a given URI:

```
http://wordnet-rdf.princeton.edu/rdf/id/06422547-n
```

Typing this URI into a browser will allow direct access to the data contained in Princeton WordNet. In the example, we see that the following facts are given

- There is a resource identified by `06422547-n` that is denoted by the URL above.
- It has the label "book" in English (`en`).

```
1  <http://wordnet-rdf.princeton.edu/rdf/id/06422547-n>
2    <http://www.w3.org/2000/01/rdf-schema#label>
3      "book"@en .
4  <http://wordnet-rdf.princeton.edu/rdf/id/06422547-n>
5    <http://wordnet-rdf.princeton.edu/ontology#partOfSpeech>
6      <http://wordnet-rdf.princeton.edu/ontology#noun> .
7  <http://wordnet-rdf.princeton.edu/rdf/id/06422547-n>
8    <http://wordnet-rdf.princeton.edu/ontology#hyponym>
9      <http://wordnet-rdf.princeton.edu/rdf/id/06423235-n> .
10 <http://wordnet-rdf.princeton.edu/rdf/id/06422547-n>
11   <http://wordnet-rdf.princeton.edu/ontology#hypernym>
12     <http://wordnet-rdf.princeton.edu/rdf/id/06423396-n> .
```

Fig. 2.1 A RDF document in N-triples format describing a synset from English WordNet

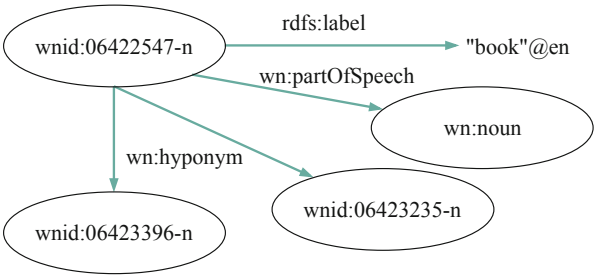

Fig. 2.2 A graphical depiction of the RDF data in Fig. 2.1 document

- Its part-of-speech value is noun.
- It is a hyponym of two resources identified as `06423235-n` and `06423396-n`. More information about this resource can be discovered by dereferencing the URIs given.

  This information is also depicted graphically in Fig. 2.2.

2.3.2 Turtle

While N-Triples is easy to parse, it can be excessively verbose. For this reason, a format called *Turtle* (*Terse RDF format*) was developed. Every document in N-Triples is also a Turtle document, but Turtle is more compact in that it allows for a number of abbreviations to avoid repetitive elements in the triple listing. Firstly, an abbreviation of URIs may be given to avoid repetition of long URIs, e.g.

```
1 @prefix wn: <http://wordnet-rdf.princeton.edu/ontology#> .
```

Then a URI can be given with a prefix followed by a colon and the suffix term without angular brackets, e.g. `wn:Synset`. In this case the URI is constructed by appending the target of the prefix to the value after the colon, so that the full URI, e.g. http://wordnet-rdf.princeton.edu/ontology#Synset is constructed. Secondly, triples may be separated by colon (`;`) or comma (`,`). In the case of separation with a colon, the subject is assumed to be fixed for the next triple and only the predicate and object need to be stated. In the case of a separation with a comma, both the subject and predict are fixed so that only the new object of the next triples needs to be stated. Thirdly, there are some further simplifications; for example, URIs may be given relatively and integers and decimals may be given as literals without quotes. Thus, the data in N-Triples format in Fig. 2.1 could be represented as the data Fig. 2.3 in Turtle.

```
@prefix wnid: <http://wordnet-rdf.princeton.edu/rdf/id/> .
@prefix wn: <http://wordnet-rdf.princeton.edu/ontology#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

wnid:06422547-n rdfs:label "book"@en ;
   wn:partOfSpeech wn:noun;
   wn:hyponym wnid:06423235-n , wnid:06423396-n .
```

Fig. 2.3 A simple RDF document in Turtle format for our WordNet synset example

2.3.3 RDF/XML

RDF is also frequently serialized in XML, and this format has been considered the default format for representing RDF. However, the RDF/XML serialization is generally very verbose and difficult for humans to understand. We will resort to the Turtle syntax for the remainder of this book as it is the most readable and concise syntax and thus suitable for a book format. For more details on the RDF/XML serialization, we refer the reader to [9].

2.3.4 RDFa

In addition, it is also possible to embed RDF markup within an HTML page, in an ePub document or in other types of XML documents. This is done with the *RDF in Attributes* (RDFa) specification [10]. In this case the URL for elements on a page may be specified with a special `about` attribute and the value of properties with the `property` attribute. In addition, links to URLs may be given with the `href` property as is usual in HTML. For example, we may give some metadata about WordNet as a resource, including Title, Author and Right as in Fig. 2.4. This generates the triples given in Turtle in Fig. 2.5.

2.3.5 JSON-LD

A recent development has been the recommendation of JSON-LD [11] as a new standard model for the representation of RDF data as JSON (JavaScript Object Notation). This model has a number of advantages:

1. JSON is a widely supported data model for which there exist a large number of libraries that support the interaction, including flexible object mapping libraries such as Jackson, Lift-JSON, etc.
2. JSON is easy for clients to access and has fewer security restrictions than other formats (e.g. XML).