Open data — linked data — linked open data — linguistic linked open data (LLOD): a general introduction

Christian Chiarcos, Antonio Pareja-Lora

Angaben zur Veröffentlichung / Publication details:

Chiarcos, Christian, and Antonio Pareja-Lora. 2019. "Open data — linked data — linked open data — linguistic linked open data (LLOD): a general introduction." In *Development of linguistic linked open data resources for collaborative data-intensive research in the language sciences*, edited by Antonio Pareja-Lora, Maria Blume, Barbara C. Lust, and Christian Chiarcos, 1–17. Cambridge, MA: The MIT Press.
https://doi.org/10.7551/mitpress/10990.003.0003.

1 Open Data—Linked Data—Linked Open Data—Linguistic Linked Open Data (LLOD): A General Introduction

Christian Chiarcos and Antonio Pareja-Lora

Background: Scientific Principles and Openness

In recent decades, the scientific community has become increasingly aware of the importance of openness—for software (open source), publications (open access), structured data (open knowledge), and data collections in general (Open Data). Here, we focus on the latter aspect. Indeed, publishing data collections under open resources has become routine in modern-day research. In this initial chapter, we elaborate on motivations and conventions for publishing Open Data in linguistics and related areas.

The Open Data movement in linguistics—as well as in all areas of study in science, computation, and humanities—draws on three main motivations: (1) responsibility, (2) reproducibility, and (3) reusability.

1. The scientific process—the generation of novel insights, the establishment and revision of paradigms of thought and scientific methodologies, and their documentation, dissemination, and critical reflection—is driven by societal, economic, and ecological need to understand and to develop our past, present, and future. In this sense, scientific research comes with both a privilege and a responsibility: Any projects are supported by public funding, and in return their results should (and in fact are often required to) become available to the public. In the last few decades, this has contributed to the rise of open access in scientific publications, and, along with it, to open source licensing of scientific code and data.
2. Another motivation for the increasing importance of Open Data in research is inherent to the scientific method: Scientific hypotheses must be testable, scientific theories should be verifiable, and published results should be replicable. For data-driven disciplines such as empirical branches of linguistics, verification presupposes the availability of empirical data, while replicability requires access to the original data that the research builds on. Although various distribution and publication models are suitable for this purpose—and have in fact been implemented by agencies such as the Linguistic Data Consortium (LDC) or the European Language Research Association (ELRA); by community portals

such as Perseus,[1] the Cuneiform Digital Library Initiative,[2] and The Language Archive;[3] or within distributed community efforts such as the Universal Dependencies,[4] and UniMorph[5]—publication under an open source license posits the lowest possible barrier for reusability, accessibility, and dissemination of research data.

3. A third practical motivation for publishing (and using) scientific data is the immense effort put into creating such resources and the potential gains of sharing and reusing existing data. In several areas of linguistics, this pertains to primary data, such as recordings, transcripts, and written text; as an extreme example, data collections for languages at the fringe of extinction and/or spoken in remote areas of the world are irreplaceable.

Regardless of the initial motivation, reusability (whether for replication studies, new applications, or novel experiments) is the ultimate goal of publishing Open Data. But secondary reuse of data is not only a concern within linguistics research. It is also an issue relevant to any scientific discipline. In fact, the degree to which an area of research develops and follows agreed-upon principles and standards for the management of data, with respect to its goal of fostering reproducibility, can be regarded as an indicator of its maturity as a scientific discipline.

For linguistics, progress in this direction involves challenges at numerous levels, ranging from political, ethical, and legal issues—for example, community conventions for handling national and international copyright, and privacy issues (for experimental data or field recordings)—to community-wide rules of best practice for documentation, maintenance, and distribution; and beyond those, to the technical question of how to represent, access, and integrate existing data collections.

As a technology, Linked Data allows us to integrate heterogeneous data collections hosted by different data providers, and thus naturally complements the call to Open Data in both science and society. Linked *Open* Data (LOD) describes their conjoint application to a dataset. In application to linguistically relevant datasets, *Linguistic Linked Open Data* (LLOD) describes conventions and a community that has emerged since 2010 whose most prominent outcome is the *Linguistic Linked Open Data cloud* diagram. In this volume, we describe the application of Linked (Open) Data to linguistic data, in particular from the angle of language acquisition.

Open Data in Science

The Open Data movement represents a global change of mind for our understanding of economy, society, and science. In the twenty-first century, a novel paradigm that facilitates both transparency and openness has been emerging. In politics, this has been manifested, for example, in an increased number of Freedom of Information Acts or in the use of Right to Information Laws, among nearly 70 countries in 2006 (Banisar 2006) and more than 100 countries in 2018 (Banisar 2018).

Likewise, the scientific *communis opinio* is increasingly shifting from closed (private) data to Open Data. For its successful implementation, open science does, however, require community standards on how to perform, document, license, and access data publications.

To improve transparency and reproducibility of scientific research, a group of researchers collaborating with M. D. Wilkinson formulated the FAIR Guiding Principles in 2016 (Wilkinson et al. 2016).

F *Findability* implies (1) that data and metadata are assigned globally unique and eternally persistent identifiers, (2) that the data are accompanied by rich metadata, and that (3) the data are registered or indexed in a site where they can be found.

A *Accessibility* implies (1) that data are retrievable by their identifier using an (2) open, free, and universally implemented protocol, and (3) that the protocol supports authentication and authorization if necessary.

I *Interoperability* implies that the data are described using a formal, accessible, shared, and broadly applicable language for knowledge representation.

R *Reusability* implies addition of accurate and relevant attributes, clear licensing and data usage terms and conditions, a linking to provenance of data, and adherence to community standards.

Linked Data represents a technical framework that allows users to tackle these challenges both in general and for the specific needs of linguistics and language technology.

Linked Data

Much of today's data are available in scattered repositories and in diverse formats. In fact, many potentially valuable datasets are being created or shared in data formats intended for human consumption rather than for automated processing. As an example, electronic edition via PDF (Portable Document Format) is still considered state of the art in various disciplines in the humanities; and regularly, spreadsheet or office software is used to create and to fill forms and tables of those PDF documents, without any formal data structures.

Likewise, a popular piece of software in linguistics is optimized for human consumption rather than for machine readability: The Field Linguist's Toolbox[6] provides word- and morpheme-level glossing functionalities. Its underlying format, however, is a plain text format, and the alignment between different layers of morpheme annotation is done by means of whitespaces. However, its current font has an impact on the width of the text displayed, and whitespace alignment between, say, morpheme segmentation and morpheme glossing, or between morpheme segmentation and word segmentation, can only be replicated if the exact widths of each character and each whitespace in the underlying font are known. Unfortunately, many fonts use variable character width, so that, in general, Toolbox segmentation cannot be reliably interpreted or converted into other formats.

These difficulties correspond to problems and needs associated with the Web of Documents in general. First, it is not machine-readable because the data are unstructured. Second, the data are disconnected. Only documents are linked and the meanings of the links are not clear. Third, only a text search is currently feasible.

A proposed solution to these problems is to complement the Web of Documents with the Web of Data, guided by Linked Data principles. The term "Linked Data" was originally published in 2006 as a Design Issue by Tim Berners-Lee (2006) and provides a set of four rules of best practice to be followed for the publication of data on the web. In a slightly reformulated form, these rules are reproduced below.

1. Uniform Resource Identifiers: Use URIs for identifying data and relations.
2. Resolvable via HTTP(S): Use HTTP(S) URIs so that people can look up those names.
3. Standardized formats: For any URI in a dataset, provide useful information using RDF-based standards.
4. Links: Include links to other URIs, so that users can discover more things.

A Uniform Resource Identifier (URI; Berners-Lee et al. 2005) is a compact sequence of characters that identifies an abstract or physical resource. An absolute URI begins with a protocol or a scheme name (e.g., https) followed by an authority (e.g., en.wikipedia.org) and a path (e.g., /wiki/Linguistic_Linked_Open_Data), followed by an optional query (headed by ?) and a fragment (headed by #, e.g., #Linguistic_Linked_Open _Data):

```
https://en.wikipedia.org/wiki/Linguistic _ Linked _ Open _ Data
#Linguistic _ Linked _ Open _ Data
```

This example illustrates that the typical form of a URI in a Linked Data context is a Uniform Resource Locator (URL; Berners-Lee et al. 1994). URLs define a subset of URIs that not only identify a resource, but also provide a means of locating it by describing its primary access mechanism (in this case, the HTTPS protocol). The URI standard is complemented by Internationalized Resource Identifiers (IRIs; Duerst and Suignard 2005), which extend the scope of permissible characters to Unicode: Non-ASCII characters are mapped to ASCII escape sequences by means of the URI percent encoding, as for example the symbol $\bar{g}$ (Unicode character U+1E21, UTF-8 E1B8A1) as %E1%B8%A1.

The third rule prescribes the use of certain standards. In its original formulation, the standards RDF (data model) and SPARQL (query language) were named. Subsequently, however, additional standards have been developed. Therefore, we interpret this rule nowadays in a way that every data format for which a W3C-standardized interpretation as RDF data exists should be a viable option. This includes native RDF serializations such as Turtle,[7] JSON-LD,[8] or RD F/XML;[9] languages that permit the embedding of RDF content;[10] mapping languages to produce RDF data from other formats;[11] languages that are defined on the basis of RDF;[12] and RDF-based query languages.[13] As data from various sources (CSV files, XML, relational databases, RDF-native data) can be seamlessly con-

verted between different RDF serializations, RDF-based representation formalisms enable data, information consumers, and processors alike to access, interpret, and transform information in a flexible, serialization-independent manner.

The RDF data model formalizes labeled directed multi-graphs, that is, nodes (RDF resources) and relations (RDF properties) that hold between them. Both nodes and relations are identified by means of URIs, and a triple of source node ("subject"), relation ("property") and target node ("object") constitutes a statement:

```
<https://en.wikipedia.org/wiki/Linguistic _ Linked _ Open _ Data>
<http://xmlns.com/foaf/spec/primaryTopic>
<http://dbpedia.org/resource/Linguistic _ Linked _ Open _ Data>
. # . marks end of statement, comments after #
```

This example is written in Turtle notation, with whitespace-separated full URIs and . to mark the end of the statement. In addition, Turtle provides a number of practical shorthands, for example the introduction of prefixes. The following Turtle fragment is thus equivalent:

```
PREFIX wpedia: <https://en.wikipedia.org/wiki/>
PREFIX foaf: <http://xmlns.com/foaf/spec/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
wpedia:Linguistic _ Linked _ Open _ Data
foaf:primaryTopic dbpedia:Linguistic _ Linked _ Open _ Data .
```

RDF triples can also take another form, where a source node ("subject") is assigned a literal value rather than a target node:

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
wpedia:Linguistic _ Linked _ Open _ Data
rdfs:label "Linguistic Linked Open Data"@en.
```

Several statements can also be conjoined by means of a semicolon ; (same subject, different property, different object) or a comma (same subject, same property, different object):

```
wpedia:Linguistic _ Linked _ Open _ Data
foaf:primaryTopic dbpedia:Linguistic _ Linked _ Open _ Data ;
rdfs:label "Linguistic Linked Open Data"@en.
```

The fourth rule requires some actual linking, that is, the creation of cross-references between different, distributed datasets, thus enabling a Web of Data to arise along and beside the Web of Documents. This is illustrated in the example above, where a Wikipedia URL and a DBpedia URI are being connected with the RDF property foaf:primaryTopic. The key difference between RDF links and HTML hyperlinks is that the former are semantically typed. Thus, a machine-readable, semantically defined graph representation is created for them, which is not only useful for resource integration on the Web of Data, but also a very generic data structure that finds immediate application in linguistics.

Actually the linking mechanism provides interesting possibilities for scientific datasets, including permitting immediate access to remote datasets and terminology bases. In this way, it becomes possible to share identifiers and to identify concepts and entities corresponding with each other, and thus to harmonize distributed datasets not only on the level of format and means of access, but also on a conceptual level, by means of the use of (or reference to) existing vocabularies. Domain terminology provided in an ontology, for example, can be linked to generic knowledge bases such as the DBpedia,[14] and subsequently enriched with DBpedia information. For instance, assume that we have both a definition of "(technological) singularity" in an English thesaurus and its linking with the English DBpedia:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX my: <http://please.de/fine/by/yourself#>
my:singularity owl:sameAs dbpedia:Technological _ singularity.
```

As the English DBpedia provides a German label, we can immediately return the German labels to our thesaurus concepts and thus apply them to the analysis of another language. This is implemented in the following SPARQL query:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?mySingularity ?germanLabel
WHERE {  # for all owl:sameAs links
?mySingularity owl:sameAs ?dbpediaResource.
# find the rdfs labels of the objects
?dbpediaResource rdfs:label ?germanLabel.
FILTER(lang(?germanLabel,'de'))
# and limit the result to German language
}
```

Likewise, large-size databases—of, say, genes, proteins, geographical names, or even movie titles—can be linked over different languages and integrated with each other, so that information from various sources complements each other. There are several reasons for publishing Linked Data: First, it allows ease of discovery through linking. Second, it is easy to consume by both humans and machines. Third, it reduces redundant research and supports collaboration. Fourth, it adds value, visibility, and impact.

Of course, Linked Data is not constrained to *Open* Data, but, obviously, publishing data under open licenses facilitates their accessibility for subsequent adaptation and enrichment. Yet, it is important to remember that not all Linked Data are open and that licensed data can still profit from using standards (enriched with links to Linked Data and/or accessed by standard tools).

Linked Open Data

The definition of Linked Open Data (LOD) is Linked Data that are openly licensed. In 2010, Tim Berners-Lee (Berners-Lee 2006) extended his original Linked Data description with a second component on Open Data. Linked Open Data (LOD) is Linked Data that are released under an open license, such as defined by the Open Definition,[15] where "open means **anyone** can **freely access, use, modify, and share** for **any purpose** (subject, at most, to requirements that preserve provenance and openness)."

For promotional reasons, the degree of LOD compliance is expressed by a star scheme, whereby a data publisher receives 1 to 5 stars (*), according to the following requirements:

```
   *    data available as Open Data on the web (e.g., as a scan)
  **    if * using machine-readable, structured format (e.g., DOCX)
 ***    if ** using non-proprietary format (e.g., HTML)
****    if *** using open, RDF-based standards
*****   if **** plus linking with other people's data
```

In addition, data publishers are encouraged to publish data along with their metadata and to register these metadata in major catalogs such as http://datahub.io/, or, for linguistic data, in http://linghub.org. From these repositories, the LOD (resp., LLOD) diagrams are being generated.

Linked Open Data has become a trend in scientific research and infrastructures during the 2010s, with prominent resources such as DBpedia (Lehmann et al. 2009), developed within an open-source project with the same name that aimed at extracted structured data from Wikipedia and related resources. DBpedia version 2016–10 includes extractions in 134 languages with a total of over 13 billion RDF statements (triples). With more and more datasets being linked with DBpedia and other LOD datasets, a Linked Open Data cloud has emerged, and as a visualization of the growing Web of *Open* Data, this process has been documented with a series of LOD cloud diagrams.[16] As of October 2018, the diagram contained 1,229 datasets with 16,125 links (figure 1.1). Primary applications of RDF technology and LOD resources are concerned with resource integration and also with resource reuse. Hence, major components of the LOD cloud diagram are term bases such as statistical government data, or biomedical databases, and indeed the key advantage of RDF technology and LOD resources is their high level of reusability and accessibility. SPARQL 1.1 supports the concept of federation: By means of the SERVICE keyword, it is possible to consult external SPARQL endpoints (RDF databases with web interfaces) as part of a query against a local triple (or quad) store.

In fact, resources can be freely shared and cloned, and redundant copies can contribute to the sustainability of LOD datasets independently from the institution that originally provided those data or their technical infrastructures.

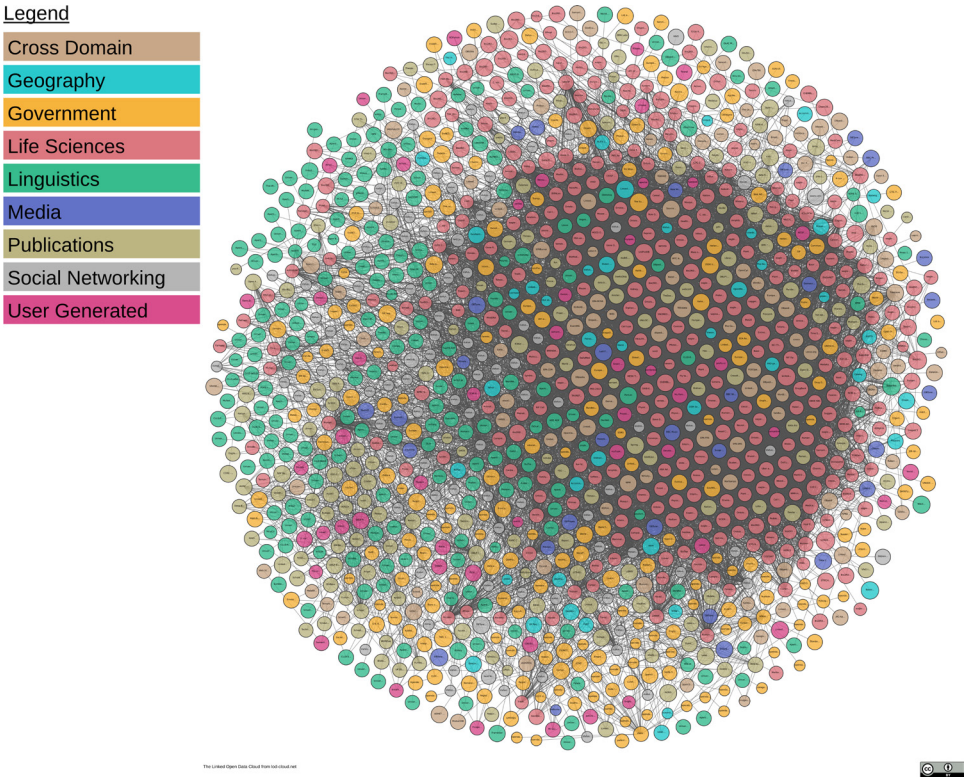

Figure 1.1
LOD diagram, version of Oct. 31, 2018, https://lod-cloud.net.

Indeed, such redundant copies are normally created as LOD are processed: While the SPARQL 1.1 protocol allows users to access remote SPARQL endpoints by means of the SERVICE keyword, and remote RDF dumps by means of LOAD, both come with a certain degree of overhead, and thus lead to runtime reductions for applications that consume Linked Data. Real-world applications of LOD thus normally work on local copies, instead, so that redundant and distributed copies are created as a side effect of LOD-based applications.

For scientific applications, another factor of LOD is important—that is, that different applications can refer to the same term in the same database. Thus results, data, and annotations can all be traced over different datasets while information about them can be put in relation with each other. Of course, the same applies, even to a larger extent, to vocabularies used for different resources. With increased reusability and reuse of scientific datasets, the datasets serve as models for the vocabulary of subsequent resources, and indeed research in community-based vocabulary development has intensified in recent years.

We also must admit that LOD comes with a number of technical challenges. LOD and RDF technology both provide a high-level view as well as a generic technology for processing and integrating different data sources, and of course "genericity" does come with a price. The potential of RDF and RDF-based technology in comparison to classical relational databases can thus be compared to the gains and challenges of high-level programming languages (such as Java or Python) in comparison to low-level programming languages (such as machine code or assembler). However, for many problems, processing RDF (cf. Python) will be considerably slower than using an implementation-specific SQL dialect (cf. assembler), though it excels in portability, reusability, and development effort. In particular, RDF is superior at dealing with sparse and heterogeneous data, but for densely populated databases, RDF technology is slow in comparison with classical relational database technology. Unlike SQL, RDF technology allows users to reach out beyond a data silo and to seamlessly link data with external resources.

One specific challenge in this context is that links between resources and resources themselves were created for different purposes, according to different methodologies and are maintained by different providers. This can lead to inconsistencies in the interpretation and in the quality of statements (triples) they provide. An increasingly important aspect is thus the tracing of provenance and related metadata, so that scientific and industry applications alike can (and should) inspect the composition of data aggregated from LOD and must not blindly rely on their correctness.

In summary, Linked Open Data is enabling a change of data and information readers and processors in that it enables us to abstract from resource-specific formats and representations and technologies, and then to integrate information over distributed datasets. Linked *Open* Data represents the core of the emerging Web of Data and thus enables a global change of data and information management and processing. LOD comes with rich technological support, in terms of portable means of access and representation (W3C-standardized data models, formats, protocols, and query languages), in terms of technical support with off-the-shelf databases, and in terms of the existence of a considerable developer and user community. At the same time, many scientific challenges in relation to LOD core techniques seem to have been solved, so that the focus in LOD research has moved from foundations and basic standards to applications. A recent development in this regard is the publication of domain-specific sub-clouds, which since August 2018 have been available as LOD addenda diagrams. Linguistic Linked Open Data represents one such area of application.

Linked Open Data in Linguistics

As is true of any field of scientific research, the FAIR principles are relevant for linguistics, language studies, and natural language processing—that is, for the digital language resources they produce and build on—and indeed Bird and Simons (2003) formulated comparable

requirements and best practice recommendations for language resources 15 years ago, which we have reorganized and slightly reworded below according to the FAIR principles.

As far as technical and legal aspects are concerned, RDF and (Linguistic) Linked (Open) Data provide an ideal framework to implement these requirements. In the enumeration below, this is illustrated with a ± ranking ranging from − to +++.[17]

F findability

existence at a data provider++ : Register language resources at a major resource portal. In a Linguistic Linked Open Data context, this would be LingHub (http://linghub.org/) or one of the resource portals it builds on.[18]

relevance/discovery+ : Provide metadata according to community-approved conventions and vocabularies.

persistence+ : Provide persistent identifiers to language resources (e.g., a persistent URL) and unique identifiers for components of a language resource.

long-term preservation+ : Provide long-term preservation by hosting at an institution committed to that purpose.

A accessibility

open format+++ : Provide data in an open format supported by multiple tools.

complete access+ : Provide direct access to the full data and documentation.

unimpeded access+++ : Provide documentation about the methods of access.

universal access+++ : Provide universal access to every interested user.

I interoperability

terminology++ : Map linguistic terms and markup elements to a common ontology.

format documentation+++ : Provide data in a self-describing format (including XML, RDF, JSON).

machine-readable format+++ : Use open standards such as those provided by the W3C (Unicode, XML, etc.).

human-readable format+ : Provide human-readable versions of the material.

R reusability

rich content++ :[19] Provide rich and linguistically relevant content.

accountability+ : Fully document both the resource and its source data.

provenance+ : Provide provenance and attribution metadata.

immutability+ : Provide immutable, fixed versions of a resource, with appropriate versioning.

legal documentation+++ : Document intellectual property rights of all components of the language resource.

research license+++ : Ensure that the resource may be used for research purposes.

complete preservation+++ : Make sure that all aspects of the language resource and its documentation remain accessible in the future (i.e., independent from any particular software).

Current accessibility challenges arise in the different formats and schemes of documents, their distribution, and the dispersed nature of metadata collections. There have long been efforts to recognize and address these problems, but these activities were never coordinated. In particular, RDF was used, but resources were rarely linked to other resources in the Web of Data. So a community needed to be built. Since 2010, the increasing popularity of applying RDF to language resources and the potential for creating links between different datasets led (1) to the formation of the Open Linguistics Working Group of Open Knowledge International[20] and, subsequently (2) to the emergence of a Linguistic Linked Open Data (LLOD) cloud, as well as (3) to the development of community conventions for the publication of linguistically relevant datasets on the Web of Data.

Open Knowledge International is a nonprofit organization, founded in 2004, that promotes open knowledge in all its forms (e.g., publication of government data in the UK and USA); it provides infrastructural support for several working groups. The Open Linguistics Working Group of the Open Knowledge Foundation (OWLG) was organized in October 2010 in Berlin, Germany, and assembled a network of individuals interested in linguistic resources and/or their publication under open licenses. The OWLG is multidisciplinary and has infrastructure in the forms of a mailing list and a website.[21] Its most important activities are the organization of community events such as workshops, datathons/summer schools and conferences, and the ongoing development of the Linguistic Linked Open Data (sub-) cloud, currently maintained under http://linguistic-lod.org/.

The Linguistic Linked Open Data (LLOD, figure 1.2) cloud is a collection of linguistic resources that have been published under open licenses as Linked Data. It is decentralized in its development and maintenance and was developed as a community effort in the context of the Open Linguistics Working Group of the Open Knowledge Foundation. Initially, the OWLG maintained a list of open or representative resources; in January 2011, this group marked possible synergies between these resources in the first draft of a LLOD cloud diagram. At this time, it was merely a vision, and the draft included non-open resources as placeholders for other resources to come, though none have been realized. In the closing chapter of their contributed volume on Linked Data in Linguistics, Chiarcos, Nordhoff, and Hellmann (2012) provided a hypothetical linking for selected datasets from NLP, Semantic Web, and linguistic typology described in the book. In September 2012, the

LLOD cloud diagram materialized as a result of the first datathon on Multilingual Linked Open Data for Enterprises (MLODE-2012). Since 2012, more data and more rigid quality constraints have been added, collaborations with national and international research projects have been established, and related W3C community groups have emerged.

With the increasing popularity of LLOD, in August 2014 "linguistics" was recognized as a top-level category of the colored LOD cloud diagram, with LLOD resources formerly having been classified into other categories. In August 2018, a copy of the LLOD cloud diagram was incorporated into the LOD cloud diagram as a domain-specific addendum. Within the LOD cloud, Linguistic Linked Open Data is growing at a relatively high rate. While the annual growth of the LOD cloud (in terms of new resources added) over the last two years has been at 10.2% on average for the LOD cloud diagram, the LLOD cloud diagram itself has been growing at 19.3% per year (cf. figure 1.3).

Aside from its maintaining the LLOD cloud diagram, the OWLG aims to promote open linguistic resources by raising awareness and collecting metadata, and aims to facilitate a wide range of community activities by hosting workshops, using its extensive mailing list, and

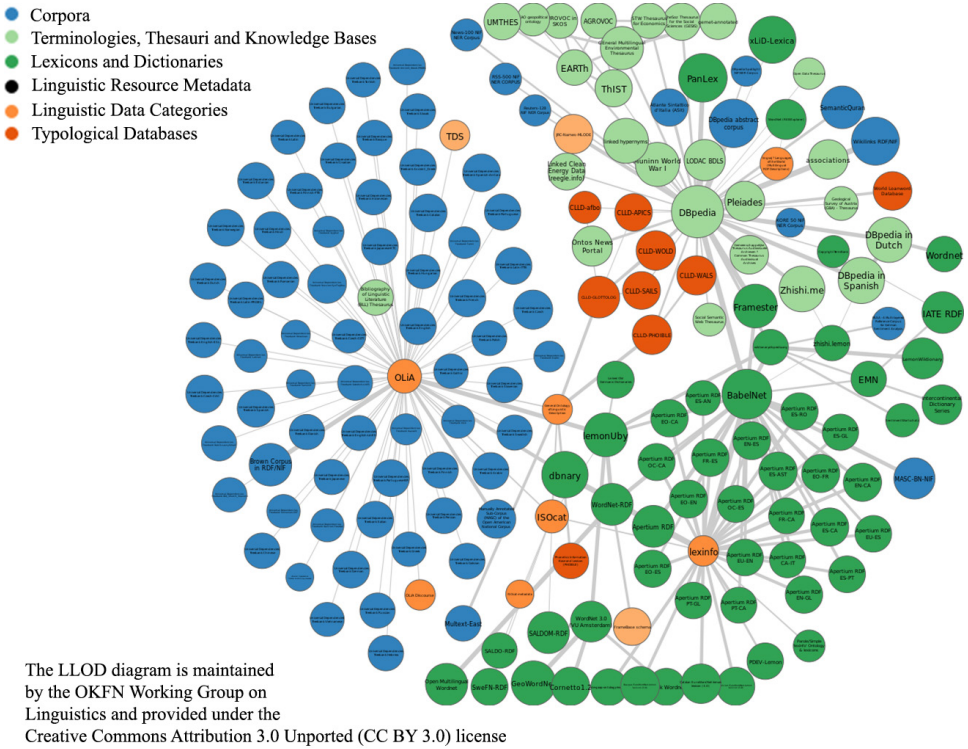

Figure 1.2
Linguistic Linked Open Data (LLOD) cloud diagram, version of August 2017, http://linguistic-lod.org.

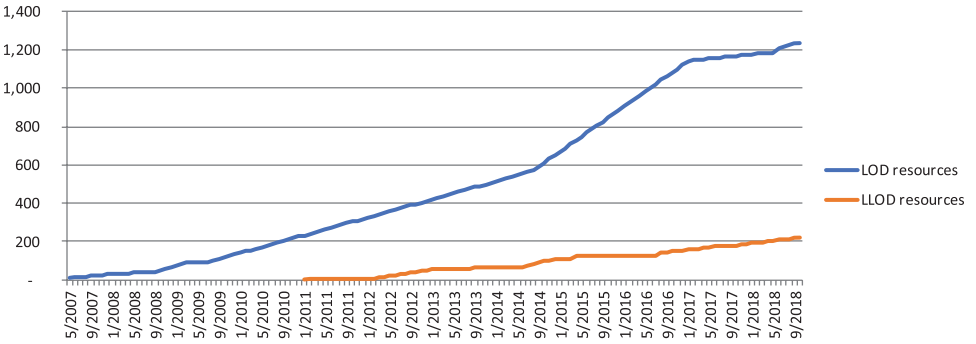

Figure 1.3
Number of resources in the LOD and LLOD cloud diagrams, corresponding, respectively, to the periods 2007–2018 and 2011–2018.

creating various publications. In doing so, they facilitate exchange between and among more specialized community groups, such as the W3C community groups (for instance, the Ontology-Lexica Community Group (OntoLex),[22] the Linked Data for Technology Working Group [LD4LT],[23] or the Best Practices for Multilingual Linked Open Data Community Group [BPM- LOD]).[24]

At the time of writing, the most vibrant of these W3C community groups is the Onto-Lex group, which is developing specifications for lexical data in a LOD context; this need correlates with the high popularity among LLOD resources of the OntoLex vocabulary (Cimiano, McCrae, and Buitelaar 2016). Whereas specifications for lexical resources are relatively mature, as are term bases for either language varieties (de Melo 2015; Nordhoff and Hammarström 2011) or linguistic terminology (Aguado de Cea, Álvarez de Mon, Gómez-Pérez, and Pareja-Lora 2004; Chiarcos 2008; Chiarcos and Sukhareva 2015), the process of developing widely applied data models for other types of language resources, such as corpora and data collections in general, is still ongoing. To a certain extent, this volume aims to contribute to this discussion and its future development.

Chances, Challenges, and Prospects

The individual contributions herein document progress made in the field of Linguistic Linked Open Data since 2012 (Chiarcos et al. 2012). One important difference in comparison to developments in that year—a time when the community was largely building on small-scale experiments and imagining a bright vision of the future—is that providers of existing infrastructures and of existing platforms are increasingly getting involved in both the process and the discussion; this is reflected by the contributors to this volume.

The general situation is that a remarkable amount of Linguistic Linked Open Data is already available, an amount that continues to steadily grow. In a longer perspective, we

can expect additional data providers to offer an L(O)D view on their data, and to support RDF serializations such as JSON-LD as interchange formats. However, LOD's further growth and popularity depend crucially on the development of applications that are capable of either consuming these data in a linguist-friendly fashion or of enriching local data with wide-ranging web resources.

At the time of writing, working with RDF normally requires a certain level of technical expertise—at minimum, basic knowledge of SPARQL and of at least one RDF format. The authors' personal experience in teaching university courses shows that linguists *can* be successfully trained to acquire both. However, this is not normally done and is unlikely to ever be part of the linguistics core curriculum. This may change once designated textbooks on Linked Open Data for NLP and for linguistics become available, but for the time being a priority for this effort and the wider community remains to provide concrete applications tailored to the needs of linguists, lexicographers, researchers in NLP, and knowledge engineers.

Promising approaches in this direction do exist: Existing tools can be complemented with an RDF layer to facilitate their interoperability. This is the scope of several chapters in this volume. Likewise, LLOD-native applications are possible—for instance, to use RDFa (RDF in attributes; Herman et al. 2015) to complement an XML workflow with SPARQL-based semantic search by means of web services (Tittel et al. 2018); to provide aggregation, enrichment, and search routines for language resource metadata (Chiarcos et al. 2016; McCrae and Cimiano 2015); to use RDF as a formalism for annotation integration and data management (Burchardt et al. 2008; Pareja-Lora 2012; Chiarcos et al. 2017); or to use RDF and SPARQL for manipulating and evaluating linguistic annotations (Chiarcos, Khait et al. 2018; Chiarcos, Kosmehl et al. 2018). While these applications demonstrate the potential of LOD technology in linguistics, they come with a considerable entry barrier, and they address the advanced user of RDF technology rather than a typical linguist. Even though concrete applications do exist, the path remains long to reaching the level of user-friendliness that occasional users of this technology might expect.

A notable exception in this regard is LexO (Bellandi, Giovannetti, and Piccini 2018), a graphical tool for collaboratively editing lexical and ontological resources that natively build on the OntoLex vocabulary and RDF; LexO was designed to conduct lexicographical work in a philological context (for instance, creating the *Dictionnaire des Termes Médico-botaniques de l'Ancien Occitan*). Other projects whose objective is to provide LLOD-based tools for specific areas of application have been recently approved, so progress in this direction is happily to be expected within the next years.[25]

Acknowledgments

This chapter originates from a joint presentation given by Antonio Pareja-Lora, Martin Brümmer, and Christian Chiarcos at the 2015 LSA workshop titled "Development of Linguistic Linked Open Data (LLOD) Resources for Collaborative Data-Intensive Research in the Language Sciences." On the one hand, the work of the first author has been partially

supported by the German Federal Ministry for Science and Education (BMBF) in the context of the Research Group *Linked Open Dictionaries* (LiODi, 2015–2020). On the other hand, the work of the second author has been partially supported by the projects RedR+Human (Dynamically Reconfigurable Educational Repositories in the Humanities, ref. TIN2014-52010-R) and CetrO+Spec (Creation, Exploration and Transformation of Educational Object Repositories in Specialized Domains, ref. TIN2017-88092-R), both financed by the Spanish Ministry of Economy and Competitiveness.

Notes

1. Greek and Latin literature, http://www.perseus.tufts.edu.
2. Ancient Mesopotamian philology, https://cdli.ucla.edu.
3. Data archive about languages worldwide, https://tla.mpi.nl/.
4. Cross-linguistically comparable syntax annotations, https://universaldependencies.org/.
5. Cross-linguistically comparable morpheme inventories, http://unimorph.github.io/.
6. https://software.sil.org/toolbox/.
7. https://www.w3.org/TR/turtle/.
8. https://www.w3.org/TR/json-ld/.
9. https://www.w3.org/TR/rdf-syntax-grammar/.
10. This includes HTML+RDFa (https://www.w3.org/TR/html-rdfa/), XHTML+RDFa (https://www.w3.org/TR/xhtml-rdfa/), or XML+RDFa (https://www.w3.org/TR/rdfa-core/).
11. Using standards such as CSV2RDF (https://www.w3.org/TR/csv2rdf/), the RDB to RDF Mapping language R2RML (https://www.w3.org/TR/r2rml/), or the Direct Mapping of Relational Data to RDF (https://www.w3.org/TR/rdb-direct-mapping/).
12. Including the Web Ontology Language OWL (https://www.w3.org/TR/2012/REC-owl2-mapping-to-rdf-20121211/) or the Simple Knowledge Organization System SKOS (https://www.w3.org/2009/08/skos-reference/skos.html).
13. For example, SPARQL (https://www.w3.org/TR/sparql11-query/) or SHACL (https://www.w3.org/TR/shacl/).
14. https://wiki.dbpedia.org/.
15. The Open Definition and compliant licenses can be found under http://opendefinition.org.
16. Available under https://lod-cloud.net/.
17. Ranking criteria and number of Bird and Simons requirements per category.

- impossible with LOD 0/19

+ possible with/encouraged by LOD, but not required 8/19

++ required by LOD 3/19

+++ required in a more specific or stricter form by (L)LOD 8/19

18. http://datahub.io/, https://vlo.clarin.eu, http://metashare.elda.org/; for language documentation data, the Open Language Archives Community (OLAC, http://www.language-archives.org/) would be an option; it provides an RDF dump, but its metadata are not yet imported into LingHub.

19. Linguistic relevance is a requirement for Linguistic Linked Open Data, but of course not for LOD data.

20. https://linguistics.okfn.org/.

21. https://linguistics.okfn.org/, https://lists.okfn.org/mailman/listinfo/open-linguistics.

22. https://www.w3.org/community/ontolex.

23. https://www.w3.org/community/ld4lt/.

24. https://www.w3.org/community/bpmlod.

25. This includes, for example, the projects POSTDATA (on European poetry, 2015–2020, funded by the European Research Council), Linked Open Dictionaries (on language contact studies, 2015–2020, funded by the German Federal Ministry of Education and Science), Linking Latin (on Latin philology, 2018–2023, funded by the European Research Council), and the Horizon 2020 Research and Innovation Action Prêt-à-LLOD (2019–2021).

References

Aguado de Cea, G., I. Álvarez de Mon, A. Gómez-Pérez, and A. Pareja-Lora. 2004. "OntoTag's Linguistic Ontologies: Improving Semantic Web Annotations for a Better Language Understanding in Machines." In Proceedings of the International Conference on Information Technology: Coding and Computing, 2004 (ITCC 2004), Vol. 2, 124–128. Las Vegas, Nevada, USA.

Banisar, D. 2006. "Freedom of Information around the World 2006: A Global Survey of Access to Government Information Laws." Technical report, Privacy International. Version of September 20, 2006.

Banisar, D. 2018. "National Right to Information Laws, Regulations and Initiatives 2018." Technical report, Privacy International. Version of September 21, 2018.

Bellandi, A., E. Giovannetti, and S. Piccini. 2018. "Collaborative Editing of Lexical and Termino-ontological resources: A Quick Introduction to LexO." In the XVIII EURALEX International Congress. Lexicography in Global Contexts, 23–27. Ljubljana, Slovenia.

Berners-Lee, T. 2006. Design issues: Linked Data. http://www.w3.org/DesignIssues/LinkedData.html. Revised 2010.

Berners-Lee, T., R. Fielding, and L. Masinter. 2005. Request for Comments: 3986. Uniform Resource Identifier (URI): Generic syntax. Technical report, The Internet Society. Network Working Group. Version of January 2005.

Berners-Lee, T., L. Masinter, and M. McCahill. 1994. Request for Comments: 1738. Uniform Resource Locators (URL). Technical report, Internet Engineering Task Force (IETF). Network Working Group. Version of December 1994.

Bird, S., and G. Simons. 2003. "Seven Dimensions of Portability for Language Documentation and Description." *Language* 79:557–582.

Burchardt, A., S. Padó, D. Spohr, A. Frank, and U. Heid. 2008. "Formalising Multi-layer Corpora in OWL/DL—Lexicon Modelling, Querying and Consistency Control." In *Proceedings of the 3rd International Joint Conf on NLP* (IJCNLP 2008), 389–396. Hyderabad, India.

Chiarcos, C. 2008. "An Ontology of Linguistic Annotations." *LDV Forum* 23 (1): 1–16.

Chiarcos, C., C. Fäth, H. Renner-Westermann, F. Abromeit, and V. Dimitrova. 2016. "Lin|gu|is|tik: Building the Linguist's Pathway to Bibliographies, Libraries, Language Resources and Linked

Open Data." In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA).

Chiarcos, C., M. Ionov, M. Rind-Pawlowski, C. Fäth, J. W. Schreur, and I. Nevskaya. 2017. "LLOD-ifying Linguistic Glosses." In International Conference on Language, Data and Knowledge (LDK 2017), edited by J. Gracia, F. Bond, J. McCrae, P. Buitelaar, C. Chiarcos, and S. Hellmann, 89–103. Galway, Ireland. Cham: Springer. Lecture Notes in Computer Science, vol. 10318.

Chiarcos, C., I. Khait, É. Pagé-Perron, N. Schenk, C. Fäth, J. Steuer, W. Mcgrath, J. Wang, et al. 2018. "Annotating a Low-Resource Language with LLOD Technology: Sumerian Morphology and Syntax." *Information* 9 (11): 290.

Chiarcos, C., B. Kosmehl, C. Fäth, and M. Sukhareva. 2018. "Analyzing Middle High German Syntax with RDF and SPARQL." In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). Miyazaki, Japan, ELRA.

Chiarcos, C., S. Nordhoff, and S. Hellmann. 2012. "Linked Data in Linguistics." Berlin/Heidelberg: Springer.

Chiarcos, C. and M. Sukhareva. 2015. "OLiA—Ontologies of Linguistic Annotation." *Semantic Web Journal* 518:379–386.

Cimiano, P., J. McCrae, and P. Buitelaar. 2016. "Lexicon Model for Ontologies." Technical report, W3C Community Report, May 10, 2016.

de Melo, G. 2015. "Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud." *Semantic Web Journal* 6 (4): 393–400.

Duerst, M. and M. Suignard. 2005. Request for Comments: 3987. Internationalized Resource Identifiers (IRIs). Technical report, The Internet Society. Network Working Group. version of January 2005.

Herman, I., B. Adida, M. Sporny, and M. Birbeck. 2015. "RDFa 1.1 Primer." 3d ed. W3C working group note, World Wide Web Consortium.

Lehmann, J., C. Bizer, G. Kobilarov, et al. 2009. "DBpedia—A Crystallization Point for the Web of Data." *Journal of Web Semantics* 7 (3): 154–165.

McCrae, J. P., and P. Cimiano. 2015. "Linghub: A Linked Data-based Portal Supporting the Discovery of Language Resources." In Proceedings of the 11th International Conference on Semantic Systems (SEMANTiCS 2015), 88–91.Vienna, Austria.

Nordhoff, S., and H. Hammarström. 2011. "Glottolog/Langdoc: Defining Dialects, Languages, and Language Families as Collections of Resources." In First International Workshop on Linked Science (LISC-2011), held in conjunction with ISWC 2011. Bonn, Germany.

Pareja-Lora, A. 2012. *Providing Linked Linguistic and Semantic Web Annotations: The OntoTag Hybrid Annotation Model*. Saarbrücken: LAP–LAMBERT Academic Publishing.

Tittel, S., H. Bermúdez-Sabel, and C. Chiarcos. 2018. "Using RDFa to Link Text and Dictionary Data for Medieval French." In Proceedings of the Sixth Workshop on Linked Data in Linguistics (LDL-2018), 7–12. Miyazaki, Japan, ELRA.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (160018).