When Linguistics Meets Web Technologies. Recent advances in Modelling Linguistic Linked Data

Anas Fahad Khan [a], Christian Chiarcos [b], Thierry Declerck [c], Daniela Gifu [d], Elena González-Blanco García [e], Jorge Gracia [f], Maxim Ionov [b], Penny Labropoulou [h], Francesco Mambrini [i], John P. McCrae [j], Émilie Pagé-Perron [k], Marco Passarotti [i], Salvador Ros Muñoz [l], Ciprian-Octavian Truică [m]

[a] *Istituto di Linguistica Computazionale «A. Zampolli», Consiglio Nazionale delle Ricerche, Italy*
E-mail: fahad.khan@ilc.cnr.it
[b] *Applied Computational Linguistics Lab, Goethe-Universität Frankfurt am Main, Germany*
E-mail: chiarcos@informatik.uni-frankfurt.de,
E-mail: ionov@informatik.uni-frankfurt.de
[c] *DFKI GmbH, Multilinguality and Language Technology, Saarbrücken, Germany*
E-mail: declerck@dfki.de
[d] *Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Romania*
E-mail: daniela.gifu@info.uaic.ro
[e] *Laboratory of Innovation on Digital Humanities, IE University, Spain*
E-mail: egonzalezblanco@faculty.ie.edu
[f] *Aragon Institute of Engineering Research, University of Zaragoza, Spain*
E-mail: jogracia@unizar.es
[g] *Applied Computational Linguistics Lab, Goethe-Universität Frankfurt am Main, Germany*
E-mail: ionov@informatik.uni-frankfurt.de
[h] *Institute for Language and Speech Processing, Athena Research Center, Greece*
E-mail: penny@athenarc.gr
[i] *CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milan, Italy*
E-mail: francesco.mambrini@unicatt.it,
E-mail: marco.passarotti@unicatt.it
[j] *Insight SFI Research Centre for Data Analytics, Data Science Institute, National University of Ireland Galway, Ireland*
E-mail: john.mccrae@insight-centre.org
[k] *Wolfson College, University of Oxford, United Kingdom*
E-mail: emilie.page-perron@wolfson.ox.ac.uk
[l] *Laboratory of Innovation on Digital Humanities, National Distance Education University UNED, Spain*
E-mail: sros@scc.uned.es
[m] *Computer Science and Engineering Department, Faculty of Automatic Control and Computers, University Politehnica of Bucharest, Romania*
E-mail: ciprian.truica@upb.ro

Editor: Philipp Cimiano, Bielefeld University, Germany
Solicited reviews: Armando Stellato, University of Rome Tor Vergata, Italy; 3 Anonymous Reviewers

Abstract. This article provides a comprehensive and up-to-date survey of models and vocabularies for creating linguistic linked data (LLD) focusing on the latest developments in the area and both building upon and complementing previous works covering similar territory. The article begins with an overview of some recent trends which have had a significant impact on linked data models and vocabularies. Next, we give a general overview of existing vocabularies and models for different categories of LLD resource. After which we look at some of the latest developments in community standards and initiatives including descriptions of recent work on the OntoLex-Lemon model, a survey of recent initiatives in linguistic annotation and LLD, and a discussion of the LLD metadata vocabularies META-SHARE and *lime*. In the next part of the paper, we focus on the influence of projects on LLD models and vocabularies, starting with a general survey of relevant projects, before dedicating individual sections to a number of recent projects and their impact on LLD vocabularies and models. Finally, in the conclusion, we look ahead at some future challenges for LLD models and vocabularies. The appendix to the paper consists of a brief introduction to the OntoLex-Lemon model.

Keywords: linguistic linked data, FAIR, corpora, annotation, language resources, OntoLex-Lemon, Digital Humanities, metadata, models, lexicon, language identification

1. Introduction

The growing popularity of linked data, and especially of linked *open* data (that is, linked data that has an open license), as a means of publishing language resources (lexica, corpora, data category registers, etc.) calls for a greater emphasis on shared *models* and *vocabularies* for linguistic linked data (LLD), since these are key to making linked data resources more reusable and more interoperable (at a semantic level). The purpose of this article is to provide a comprehensive and up-to-date survey of such models, while also touching upon a number of other closely related topics. The article will focus on the latest developments in this area and will both build upon and attempt to complement previous works covering similar territory by avoiding too much repetition and overlap with the latter.

In the following section, Section 2, we give an overview of a number of trends from the last few years which have had/are having/are likely to have, a significant impact on the definition and use of LLD models. This overview is intended to help to locate the present work within a wider research context, something that is particularly useful in an area as active as linguistic linked data, as well as helping readers in navigating the rest of the article. Section 3 gives an overview of related work, and Section 4 an overview of the most widely used models in LLD. Next, in Section 5, we take a look at recent developments in community standards and initiatives: this includes a description of the latest extensions of the OntoLex-Lemon model, as well as a discussion of relevant work in the modelling of corpora and annotations and LLD metadata. Finally, the article contains a section dedicated to the use of models in LLD-centered projects, Section 6, and a concluding section, Section 7 in which we look at some potential future trends.

2. Setting the Scene: An Overview of Relevant Trends in LLD

We have decided to focus on three overarching trends in this overview. These are: the FAIRification of data in **Section 2.1**; the role of projects and community initiatives in **Section 2.2**; and, finally, the increasing influence of Digital Humanities use cases in **Section 2.3**. All three of these trends have arguably had a major impact on the development of and need for shared LLD models and vocabularies. The second of the themes listed above – the role of projects and community initiatives in the creation and maintenance of LLD models – has always been important for our topic and continues to be so; the other two, however, have really begun to taken on a marked relevance for LLD over the last few years.

FAIR data (defined below, in Section 2.1) plays a central role in a number of prominent initiatives which have recently been proposed for the promotion of open science and data on the part of numerous organisations and especially of research funding bodies. It would be useful to understand therefore how LLD models can contribute to the creation of FAIR language resources, and this is the topic of Section 2.1. Similarly, the Digital Humanities, an area of research which has rapidly gained ground over the last few years, have also become more and more significant as a both a producer and consumer of LLD, something which has inevitably had an impact on LLD vocabularies and models, see Section 2.3.

2.1. FAIR New World

It should come as no surprise, given the growing importance of Open Science initiatives and in particular those promoting the FAIR guidelines (where FAIR stands for Findable, Accessible, Interoperable and Reusable) for the modelling, creation and publication of data [1], that shared models and vocabularies have begun to take on an increasingly prominent role within numerous disciplines, and not least in the fields of linguistics and language resources. And although the linguistic linked data community has been active in advocating for the use of shared RDF-based vocabularies and models for quite some time now, this new emphasis on FAIR language resources is likely to have a considerable impact in several ways, not least in terms of the necessity for these models and vocabularies to demonstrate greater coverage with respect to the kinds of linguistic phenomena they can describe, and for them to be more interoperable with each other. We will look at one recent and influential series of FAIR related recommendations for models in Section 4 in order to see how they might be applied to the case of LLD. In the rest of this subsection, we will take a closer look at the FAIR principles themselves and show why their widespread adoption is likely to lead to a greater role for LLD models and vocabularies in the future.

In *The FAIR Guiding Principles for scientific data management and stewardship* [1], the article which first articulated the by-now ubiquitous FAIR principles, the authors state that the criteria proposed by those principles are intended both "for machines and people" and that they provide "'steps along a path' to machine actionability", where the latter is understood to describe structured data that would allow a "computational data explorer" to determine:

- The type of "digital research object"
- Its usefulness with respect to tasks to be carried out
- Its usability especially with respect to licensing issues, with this information represented in a way that would allow the agent to take "appropriate action".

The current popularity of the FAIR principles and, in particular, their promotion by governments, transnational organisations and research funding bodies, such as the European Commission,[1] reflects a wider recognition of the potential of structured, interoperable, machine actionable data to help effect a major shift in how research is carried out, and in particular, its potential to help underpin Open Science best practices. The FAIR ideal, in short, is to allow machines (non-human software agents) a greater level of autonomy in working with data by the expedient of rendering as much of the semantics of that data explicit (in the sense of machine actionable) as possible.

Publishing data using a standardised, general purpose, data model such as the Resource Description Framework[2] (RDF) goes a long way towards facilitating the publication of datasets as FAIR data. Indeed RDF, taken together with the other standards proposed in the Semantic Web stack and the technical infrastructure which has been developed to support it, was specifically intended to facilitate interoperability and interlinking between datasets. In order to ensure the interoperability and re-usability of datasets within a domain, however, it is vital that in addition to more generic data models such as RDF there also exist domain specific vocabularies/terminologies/models and data category registries (compatible with the former). Such resources serve to describe, ideally in a machine actionable way, the shared theoretical assumptions held by a community of domain experts as reflected in the terminology or terminologies in use within that community.

The following specific FAIR principles are especially salient here (emphasis ours):

- F2. data are described with *rich metadata*.
- I1. (meta)data use a formal, accessible, shared, and broadly applicable *language for knowledge representation*.
- I2. (meta)data use vocabularies that follow FAIR principles.

It is important to note here that the emphasis placed on machine actionability in FAIR resources (that is, recall, on enabling computational agents to find relevant datasets and resources and to take "appropriate action" when they find them) gives Semantic Web vocabularies/models/registries a substantial advantage over other (non-Semantic Web native) standards in the fields of linguistics and language resources, such as

[1] https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_0.pdf

[2] https://www.w3.org/TR/rdf-primer/

the Text Encoding Initiative (TEI) guidelines[3] [2], the Lexical Markup Framework (LMF) [3] or the Morpho-syntactic Annotation Framework (MAF) [4].

For a start, none of these other standards possess a 'native', widely-used, widely supported and broadly applicable formal knowledge representation (KR) language for describing the semantics of vocabulary terms in a machine-readable way, or at least nothing as powerful as the Web Ontology Language (OWL)[4] or the Semantic Web Rule Language (SWRL)[5]. This means that in effect there is no standardised way of, for instance, describing the meanings of terms such *morpheme*, or *lemma*, etc. in TEI in a machine actionable way. KR languages like OWL allow for precise, axiomatic definitions to be given to terms in a way that permits reasoning to be carried out on them (in the case of OWL there exist numerous, freely available reasoning engines such as Pellet [5]); more generally, they allow for much richer machine actionable metadata descriptions. Furthermore, the use of KR languages like OWL can be allied with already established conceptual modelling techniques and best practises – including the use of top level ontologies such as DOLCE[6] or BFO[7], both of which are available in OWL, and ontology validation methodologies such as OntoClean [6] which help to clarify what we mean when we say that one concept is a subtype of another – in order to define vocabularies and models which further enhance the interoperability and machine actionability of linguistic datasets.

Moreover, thanks to the use of a shared data model with a powerful native linking mechanism, LLD datasets can easily be integrated with, and therefore enriched by, linked data datasets belonging to other domains, for instance, geographical and historical datasets or gazetteers and authority lists. Indeed, OWL vocabularies, such as PROV-O,[8] make it straightforward to add complex, structured information describing when something happened or to make hypotheses explicit[9] – all of which contributes towards the creation of ever richer and more machine actionable metadata for linked data language resources.

[3]https://tei-c.org/guidelines/

[4]https://www.w3.org/TR/2012/REC-owl2-overview-20121211/

[5]https://www.w3.org/Submission/SWRL/

[6]http://www.loa.istc.cnr.it/dolce/overview.html

[7]https://basic-formal-ontology.org/

[8]https://www.w3.org/TR/prov-o/

[9]In the latter case for instance we could use the Semantic Web ontology CRMInfhttp://www.cidoc-crm.org/crminf/.

The pursuit of the FAIR ideal has in fact encouraged the definition of new ways of publishing linked data datasets, which offer additional opportunities for the re-use and integration of such datasets in an automatic or semi-automatic way. These include *nanopublications*, *cardinal assertions* and *knowlets*[10]. The potential of these new approaches for discovering new facts as well as for comparing different concepts together and tracking how single concepts change and evolve is well described in [7].

When it comes to language resources we are faced with a rich array of highly structured datasets arranged into different types (lexica, corpora, treebanks, etc) according to a series of widely shared conventions – something that would seem to lend itself well to making such resources FAIR in the machine-oriented spirit of the original description of those principles. However, in order to ensure the continued effectiveness of linked data and the Semantic Web in facilitating the creation of FAIR resources, it is critical that pre-existing vocabularies/models/data registries be re-used whenever possible in the modelling of language resources. In many instances, these models will not have sufficient coverage to capture numerous kinds of use cases, in which case we will have to define new extensions to these models (an ongoing process and one which is a major theme of this article, see for instance Section 5.1), in other cases it may be necessary to create training materials suitable for different groups of users. Part of the intention of this article, together with the foundational work carried out in [8], is to provide an overview of what exists in terms of LLD-focused models, to look at those areas and use-cases which have so far gained the most attention and to highlight those which are so far underrepresented.

2.2. The Importance of Projects and Community Initiatives in LLD

One significant indicator of the success which LLD has enjoyed in the last few years is the variety of newly

[10]*Nanopublications* are defined as the "smallest possible machine-readable graph-like structure that represents a meaningful assertion" [7] and consist of publishing a single subject-predicate-object triple with full provenance information; a generalisation of this idea is that of the *cardinal assertion* where a single assertion is associated with more than one provenance graph. A *knowlet* consists of a collection of multiple cardinal assertions, with the same subject concept [7] and can be viewed as locating that concept in a rich 'conceptual space'. For instance, this could be a cloud of predicates centered around a word or a sense.

funded projects which have emerged in this period, and which have included the publication of linguistic datasets as linked data as a core theme. These include projects both at a continental or transnational level – notably European H2020 projects[11], ERCs[12] and COST actions[13] – as well as at the national and regional levels. Arguably, this recent success in obtaining project funding reflects a wider recognition of the usefulness of linked data as a means of ensuring the interoperability and accessibility of language resources. It also demonstrates the ongoing maturation of the field, as LLD continues to be successfully applied to new domains and use cases within the context of such projects. In addition, these projects also offer us numerous examples of the application of some LLD vocabularies and models, which we look at in this article in the creation of medium to large-scale language resources.

We have therefore decided to dedicate a whole section of the present article, **Section 6**, to a detailed discussion of the current situation as regards research projects and LLD models and vocabularies. This includes a detailed overview of the area, **Section 6.1**, along with an extended description of a number of projects which we regard as the most significant from the point of view of LLD models and vocabularies. These are (in order of appearance): the **Linked Open Dictionaries (LiODi)** project (**Section 6.2.1**); the **Poetry Standardization and Linked Open Data (POSTDATA)** project (**Section 6.2.2**); the **European Lexicographic Infrastructure (ELEXIS)** project **Section 6.2.3**; the **LiLa: Linking Latin** ERC project (**Section 6.2.4**); the **Prêt-à-LLOD** project (**Section 6.2.5**); the **European network for Web-centred linguistic data science (NexusLinguarum)** COST action (**Section 6.2.6**). A list of all the projects described in Section 6 can be found in Table 3.

Note, however, that although the projects which we discuss in Section 6 have, in many cases, set the agenda for the development of LLD models and vocabularies, much of the actual work on the definition of these resources was carried out – and is being carried out – within community groups, such as the W3C OntoLex group. We therefore include an update on community standards and initiatives in **Section 5**. These include a subsection on the latest activities in the OntoLex group (**Section 5.1**); a discussion of recent work on LLD models for corpora and annotation (**Section 5.2**); and similarly for what concerns models and vocabularies for LLD resource metadata (**Section 5.3**). Section 6.1.2 features a discussion of the relationship between community initiatives and projects.

2.3. The Relationship of LLD to the Digital Humanities

Several of the projects discussed in this article fall under the umbrella of the Digital Humanities (DH). For this and other reasons this is the third major trend which we want to highlight here, since it represents a move away (or more precisely a branching off) from LLD's beginnings in computational linguistics and natural language processing (although these latter two still perhaps represent the majority of applications of LLD), and this we claim is something that is leading to a shift in emphasis in the definition and coverage of LLD models. The overlap between LLD and DH is especially apparent in the modelling of corpora annotation (**Section 5.2**) and in the context of linked data lexicographic use cases (see **Section 5.1.1** and **Section 6.2.3**).

One use case which clearly highlights these shared concerns is the publication of retro-digitised dictionaries as LLD lexica (a major theme of the ELEXIS project, see **Section 6.2.3**). This use case confronts us with the challenge of formally modelling both the *content* of a lexicographic work, that is, the linguistic descriptions which it contains, and those aspects which pertain to it as a *physical text* to be represented in digital form. In the latter case, this includes the representation of (elements of) the *form* of the text, i.e., its structural layout and overall visual appearance[14]; we may also wish to model different aspects of the *history* of the lexicographic work as a physical text.[15] In fact, as we touch upon in our description of the OntoLex-Lemon Lexicography module in (**Section 5.1.1**), the structural division of lexicographic works into textual units such as entries and senses is not always isomor-

[11] https://ec.europa.eu/programmes/horizon2020/what-horizon-2020

[12] https://erc.europa.eu/

[13] https://www.cost.eu/

[14] Encompassing what the TEI dictionary chapter guidelines call the typographical and editorial views. See https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html#DIMV

[15] For instance we might want to track the evolution of a historically significant lexicographic work over the course of a number of editions, in order to see, for example, how changes in entries reflected both linguistic and wider, non-linguistic trends. This was one of the motivations behind the Nénufar project [9], described in Section 6.1.1.

phic to the representation of the lexical content of those units using OntoLex-Lemon classes such as LexicalEntry and LexicalSense.

All of this calls for a much richer provision of metadata categories than has been considered up till now for LLD lexica: both at the level of the whole work and at the level of the individual entry. It also requires the capacity to model salient aspects of the same artefact or resource at different levels of description (something which is indeed offered by the OntoLex-Lemon Lexicography module, see **Section 5.1.1**). We discuss metadata challenges in humanities use cases more generally in **Section 5.3**. A related topic is the relationship between notions such as *word* taken from the lexical/linguistic and philological points of view and, more broadly, the relationship between linguistic and philologically motivated annotations of text. This latter topic which is just starting to gain attention within the context of LLD is being studied both at the level of community initiatives (see **Section 5.2**) and in projects such as LiLa (see **Section 6.2.4**) and POSTDATA (**Section 6.2.2**).

An additional series of challenges arises in the consideration of resources for classical and historical languages, or indeed, historical stages of modern languages. For instance in the case of lexical resources for historical languages we often come up against the necessity of having to model attestations (discussed in **Section 5.1.3**) and these can sometimes cite reconstructed texts, something that underscores the desirability of being able to represent different scholarly and philological hypotheses. This is a need which also arises in the context of modelling of word etymologies. The LiLa project [10] (**Section 6.2.4** for a more detailed description) provides a good example of the challenges and opportunities of adopting the LLD model to represent linguistic (meta)data for both lexical and textual resources for a classical language (Latin).

One extremely important (non RDF-based) standard for encoding documents in the Digital Humanities is **TEI/XML**. We discuss in this article the relationship between TEI and RDF-based annotation approaches (in Section 5.2.1), introduce the new lexicographic TEI-based standard **TEI Lex-0**, and describe current work on a crosswalk between OntoLex-Lemon and the latter (in **Section 6.2.3**).

Finally, see **Section 6.1.1** for an overview of a number of projects combining DH and LLD.

3. Related Work

This article is intended, among other things, to both complement and to update a previous general survey on models for representing LLD, published by Bosque-Gil et al. in 2018 [8]. Although we are now only four years on from the publication of that work, we feel that enough has happened in the intervening time period to justify a new survey article. In addition, our intention is to cover a much wider range of topics than the previous article. We also feel that our overall focus is quite different. Broadly speaking, that previous work offered a classification of various different LLD vocabularies according to the different levels of linguistic description that they covered. The current paper concentrates more on the use of LLD vocabularies in practise and on their availability (this is very much how we have approached the survey in **Section 4**). Moreover, the present article includes a detailed discussion of recent work in the use of LLD models and vocabularies in corpora and annotation, **Section 5.2**, as well as an extensive section on metadata, **Section 5.3**, neither of which were given the same detailed level of coverage in [8]. Additionally, we also cover the following initiatives which were not discussed in the previous article because they had not yet got underway:

- The development of new OntoLex-Lemon modules for morphology Section 5.1.2 and frequency, attestations, and corpus Information, described in **Section 5.1.3**
- An important new initiative in aligning LLD vocabularies for corpora and annotation, described in **Section 5.2.5**.

In what follows, we will assume that the reader already has some grounding in linked data in general – including a basic familiarity with the Resource Description Framework (RDF), RDF Schema (RDFS) and the Web Ontology Language (OWL) – and linguistic linked data in particular. In case the reader is missing this minimal background in linguistic linked data, the recently published *Linguistic linked data: representation, generation and applications* [11] should provide with a comprehensive introduction to and overview of the field, focusing on more established models and vocabularies and their application rather than on recent developments. Another important new book on the topic of LLD and which has relevance to the current work is the collected volume *Development of linguistic linked open data resources for collaborative data-*

intensive research in the language sciences [12] which aims to describe major developments since 2015. It consists mostly of position papers by linguists and researchers from the language resource communities.

4. LLD Models: An Overview

The current section gives an overview of some of the most well known and widely used models and vocabularies in LLD. A summary of the models discussed in the current section (and in the whole article) can be found in **Tables 1 and 2** (with Table 1 dealing with published LLD models/vocabularies and 2 with models/vocabularies that are currently unavailable or no longer updated). An account of some of the latest developments with regard to these models, on the other hand, can be found in Section 5. We classify each of the models described in this section according to the scheme given in the linguistic LOD cloud diagram[16] (the cloud itself is described in [13]). These are:

- Corpora (and Linguistic Annotations)(Section 4.1)
- Lexica and Dictionaries (Section 4.2)
- Terminologies, Thesauri and Knowledge Bases (Section 4.3)
- Linguistic Resource Metadata (Section 4.4)
- Linguistic Data Categories (Section 4.5)
- Typological Databases (Section 4.6)

For each category we list the most prominent and widely used LLD models/vocabularies belonging to that category (the relevant section is given in parentheses after the name of each category in the list above). These models were either originally designed to help encode that kind of dataset or have been widely appropriated for that end; in the case of the category *Linguistic Data Categories* we list LD linguistic data categories. For instance, the OntoLex-Lemon model falls under *Lexica and Dictionaries* since it was initially conceived as a means of enriching ontologies with lexical information, that is, of lexicalising ontological concepts, but subsequently gained popularity as a means of encoding linked data lexica with or without an associated ontology. Tables 1 and 2 give a summary of the LLD vocabularies and models covered in this paper (with the relevant sections of the article listed).

We describe our methodology for the rest of the section below. In Section 4.7 we discuss tools and platforms for the publication of LLD.

Our Approach to Classification

As this section is intended to be an overview we will not give detailed descriptions of single models or vocabularies here (several of these models and vocabularies are described in more detail in the rest of the article, or in the case of OntoLex-Lemon in the appendix, and others receive a more detailed treatment in [8] and [11]). Instead, we describe them on the basis of a number of criteria, many of which are related to their status as FAIR models and vocabularies. In doing so we refer to a recent survey on FAIR Semantics [14], the result of a dedicated brainstorming workshop and subsequently an evaluation session of the FAIRsFAIR project.[17] This report outlines a number of recommendations and best practices for FAIR *semantic artefacts* where the latter are defined as "machine-actionable and -readable formalisation[s] of a conceptualisation enabling sharing and reuse by humans and machines"; this term is intended to include taxonomies, thesauri and ontologies.

Even though all the recommendations listed in [14] are important, for reasons of space, we have selected the following subset on the basis of their salience to the set of models and vocabularies under discussion:

- *(P-Rec 2)* Globally Unique, Persistent, and Resolvable Identifiers must be used for Semantic Artefact Metadata Records. Metadata and data must be published separately, even if it is managed jointly;
- *(P-Rec 4)* Semantic Artefact and its content should be published in a trustworthy semantic repository;
- *(P-Rec 6)* Build semantic artefact search engines that operate across different semantic repositories;
- *(P-Rec 10)* Foundational Ontologies may be used to align semantic artefacts;
- *(P-Rec 13)* Crosswalks, mappings and bridging between semantic artefacts should be documented, published and curated;
- *(P-Rec 16)* The semantic artefact must be clearly licensed for use by machines and humans.

Neither of the recommendations (P-Rec 2) and (P-Rec 10) have been implemented by any of the mod-

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

[17] https://www.fairsfair.eu/

els/vocabularies which we look at below. Following them, however, greatly helps to make these resources (and the datasets which make use of them) more FAIR, and we regard their adoption as desirable future objectives for the models and vocabularies listed below.[18] In terms of the recommendation (P-Rec 13) at the time of writing, we can only mention ongoing efforts at developing a TEI Lex-0/OntoLex-Lemon crosswalk described in Section 6.2.3.

We use (P-Rec 16) as a guide in analysing the resources covered in the article. So that we point out cases where licensing information is available as machine actionable metadata, using properties like DCT:license and URI's such as https://creativecommons.org/publicdomain/zero/1.0/ as this practice enhances the re-usability of those resources. Recommendations (P-Rec 4) and (P-Rec 6), on the other hand, alert us to the value of being able to find models and vocabularies on specialised search engines/archives (findability being one of the pillars of FAIR). As we will see, several of the models discussed below are listed on the Linked Open Vocabulary (LOV)[19] search engine[20] [16] and the DBpedia archivo ontology archive.[21]

In addition to the textual descriptions of different LLD models given in the rest of this section, we also give a tabular summary of the most well-known/stable/widely available[22] of these models in Table 1; this table also refers, in relevant cases, to sections of the paper where more details about a model are given.

Every one of the models listed in the table uses the RDFS vocabulary, and each one of them is an OWL ontology. We also list the additional models/vocabularies which they make use of in the table on a case by case basis. These include the following well known ones: XML Schema Definition[23] (XSD); the Friend of a Friend Ontology[24] (FOAF); the Simple Knowledge Organisation System[25] (SKOS); Dublin Core[26] (DC); Dublin Core Metadata Initiative (DCMI) Metadata Terms;[27] the Data Catalog Vocabulary[28] (DCAT), described also in Section 5.3; and the PROV Ontology[29] (PROV-O).

In addition, the table also mentions the following vocabularies.

- Activity Streams(AS): a vocabulary for activity streams.[30]
- GOLD: an ontology for describing linguistic data, which is described in Section 4.5.
- MARL: a vocabulary for describing and annotating subjective opinions.[31]
- ITSRDF: an ontology used within the Internationalization Tag Set.[32]
- The Creative Commons vocabulary[33] (CC).
- VANN: a vocabulary for annotating vocabulary descriptions.[34]
- SKOS-XL: an extension of SKOS with extra support for "describing and linking lexical entities".[35] SKOS and SKOS-XL are, along with *lemon* and its successor OntoLex-Lemon, amongst the most well known ways of enriching linked data taxonomies and conceptual hierarchies with linguistic information. We will look at the use of a SKOS-XL vocabulary in the context of a project on the classification of folk tales in Section 6.

4.1. Vocabularies and Models for Corpora and Linguistic Annotations

Linguistic annotation for the purposes of creating digital editions, corpora, and linking texts with external resources etc, has long been a topic of interest within the context of RDF and linked data. Coexisting with relational databases, XML-based formats (most notably, TEI, see Section 5.2) or simply text-based formats, RDF-based annotation models have

[18] The adoption of foundational ontologies, for instance, would likely help to alleviate some problems raised by the proliferation of independently developed models as described in [8].

[19] https://lov.linkeddata.es/dataset/lov

[20] Note that the LOV site provides a list of criteria for inclusion on their search engine [15]: https://lov.linkeddata.es/Recommendations_Vocabulary_Design.pdf

[21] http://archivo.dbpedia.org/

[22] Several of the models which are described in the rest of the section and aren't available publicly but may be interesting for historical reasons.

[23] https://www.w3.org/TR/xmlschema-0/

[24] http://xmlns.com/foaf/spec/

[25] https://www.w3.org/2004/02/skos/

[26] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[27] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[28] https://www.w3.org/TR/vocab-dcat-2/)

[29] https://www.w3.org/TR/prov-o/

[30] https://www.w3.org/TR/activitystreams-vocabulary/

[31] http://www.gsi.dit.upm.es/ontologies/marl/

[32] https://www.w3.org/TR/its20/

[33] https://creativecommons.org/ns

[34] https://vocab.org/vann/

[35] https://www.w3.org/TR/skos-reference/skos-xl.html

Summary					
Name	**Other Vocabularies/Models Used**	**LLD Category**	**Licenses**	**Versions** (at time of writing 26/07/21)	**Extended Coverage in Current Article**
OntoLex-Lemon	CC, DC, FOAF, SKOS, XSD	Lexica and Dictionaries	CC0 1.0	Version 1.0, 2016 (but this is closely based on the prior *lemon* model [17])	Section 4.2, Section 5.3.3 and Appendix A
Lexicog (OntoLex-Lemon)	DC, LexInfo, SKOS, VOID, XSD	Lexica and Dictionaries	CC0	Version 1.0, (2019-03-08)	Section 4.2 and Section 5.1.1
MMoOn	DC, FOAF, GOLD, LexVo, OntoLex-Lemon, SKOS, XSD	Terminologies, Thesauri and KBs (Morphology)	CC-BY 4.0	Version 1.0, 2016	Section 4.3
Web Annotation Data Model (OA)	AS, FOAF, PROV, SKOS, XSD	Corpora and Linguistic Annotations	W3C Software and Document Notice and License	Version "2016-11-12T21:28:11Z"	Section 4.1 and Section 5.2.3
NLP Interchange Format (NIF Core)	DC, DCTERMS, ITSRDF, levont, MARL, OA, PROV, SKOS, VANN, XSD	Corpora and Linguistic Annotations	Apache 2.0 and CC-BY 3.0	Version 2.1.0	Section 4.1 and Section 5.2.2
POWLA	FOAF, DC, DCT,	Corpora and Linguistic Annotations	NA	Last Updated 2018-04-03	Section 5.2
CoNLL-RDF	DC, NIF Core, XSD	Corpora and Linguistic Annotations	Apache 2.0 and CC-BY 4.0	Last Updated 2020-05-26	Section 5.2.4
Ligt	DC, NIF Core, OA	Corpora and Linguistic Annotations	NA	Version 0.2 (2020-05-26)	Section 5.2.4
META-SHARE	CC, DC, DCAT, FOAF, SKOS, XSD	Linguistic Resource Metadata	CC-BY 4.0	Version 2.0 (pre-release)	Section 4.4 and Section 5.3.2
OLiA	DCT, FOAF, SKOS	Linguistic Data Categories	CC-BY-SA 3.0	Version last updated 27/02/20	Section 4.5
LexInfo	CC, Ontolex, TERMS, VANN	Linguistic Data Categories	CC-BY 4.0	Version 3.0, 14/06/2014	Section 4.5
LexVo	FOAF, SKOS, SKOSXL, XSD	Typological Databases	CC-BY-SA3.0	Version 2013-02-09	Section 4.6

Table 1
Summary of published LLD vocabularies

been steadily undergoing development and are increasingly being taken up in research and industry.

Currently there are two primary RDF vocabularies which are being widely used for annotating texts. These are **NLP Interchange Format** (NIF),[36] used mostly in the language technology sector and **Web Annotation**,[37] formerly known as *Open Annotation* (abbreviated here as OA), used in digital humanities, life sciences and bioinformatics. Each vocabulary has its own particular advantages and shortcomings, and a number of proposals to extend them have been proposed. Above all, however, there is a need for synchro-

[36] https://nif.readthedocs.io/en/latest/

[37] https://www.w3.org/TR/annotation-model/

Summary			
Name	**LLD Category**	**Status** (at time of writing 26/07/21)	**Extended Coverage in Current Article**
OntoLex-Lemon: FrAC	Lexica and Dictionaries	Under Development	Section 5.1.3
OntoLex-Lemon: Morphology	Lexica and Dictionaries	Under Development	Section 5.1.2
PHOIBLE	Terminologies, Thesauri and KBs	Unavailable	Section 4.3
FRED	Corpora and Linguistic Annotations	Project Specific Vocabulary	Section 5.2
NAF	Corpora and Linguistic Annotations	Project Specific Vocabulary	Section 5.2
GOLD	Linguistic Data Categories	No Longer Updated	Section 4.5

Table 2
Other LLD vocabularies discussed in this paper

nization between the two. Both are listed in LOV[38] and archivo[39] (the NIF core in the case of NIF[40]). The Web Annotation model, although it is covered by a W3C software and document notice and license, does not express this information as machine actionable metadata; while NIF does with its licensing information. More details about both models and their recent developments are given in Section 5.2.

Other vocabularies described in that section include POWLA, CoNLL-RDF and Ligt. The first of these, POWLA,[41] is available on archivo,[42] the only one of the three that has been made available in this way. CoNLL-RDF[43] expresses version info as a string using the owl:versionInfo property and is covered by a CC-BY 4.0 license as specified in the LICENSE.data page.[44]

[38]https://lov.linkeddata.es/dataset/lov/vocabs/nif and https://lov.linkeddata.es/dataset/lov/vocabs/oa

[39]http://archivo.dbpedia.org/info?o=http://www.w3.org/ns/oa

[40]http://archivo.dbpedia.org/info?o=http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core

[41]http://purl.org/powla/powla.owl

[42]https://archivo.dbpedia.org/info?o=http://purl.org/powla/powla.owl

[43]http://purl.org/acoli/conll#

[44]https://github.com/acoli-repo/conll-rdf/blob/master/LICENSE.data.txt

4.2. Lexica and Dictionaries

The most well known model for the creation and publication of lexica and dictionaries as linked data is **OntoLex-Lemon**[45] [18] (see Appendix A for an introduction to the model with examples, and Section 5.1 for extensions and further developments). This was an output of the W3C Ontology-Lexica working group (we will refer to this as the OntoLex group in what follows) which also manages its ongoing development along with the publication of further extensions. OntoLex-Lemon is based on a previous model, the **LExicon Model for ONtologies** (*lemon*) [17] and as was the case with its predecessor, it is intended as a model for enriching ontologies with linguistic information and not for modelling dictionaries and lexica *per se*. Thanks to its popularity, however, it has come to take on the status of a de facto standard for the modelling and codification of lexical resources in RDF in general. Resources which have been modelled using OntoLex-Lemon include the LLD version of the Princeton Wordnet,[46] DBnary (the linked data version of Wiktionary) [19], and the massive multilingual knowledge graph Babelnet [20].

The OntoLex-Lemon model is modular and consists of a core module along with modules for *Syntax*

[45]The URI for OntoLex-Lemon is: http://www.w3.org/ns/lemon/ontolex and the OntoLex-Lemon guidelines can be found at https://www.w3.org/2016/05/ontolex/.

[46]http://wordnet-rdf.princeton.edu/about

and Semantics,[47] *Decomposition*,[48] and *Variation and Translation*,[49] as well as a dedicated metadata module, *lime*[50] (all of which are described in Appendix A, except for *lime* which is described in Section 5.3.3).

OntoLex-Lemon is available on LOV as is its predecessor *lemon*.[51] All of its individual modules are listed separately: the core;[52] *lime*;[53] *vartrans*;[54] *synsem*;[55] the *decomp* module.[56] Three of its modules are available on archivo, the core:[57] the *lime* metadata module[58] and the Variation and Translation module.[59] All the OntoLex-Lemon modules have their licensing information (they are all CC0 1.0) described with RDF triples using the CC vocabulary[60] with a URI as an object. Version information is described using owl:versionInfo.

The OntoLex-Lemon Lexicography module[61], described in more detail in Section 5.1.1, was published separately from OntoLex-Lemon. It is not available on LOV yet, but it is available on archivo.[62] Its licensing information (CC-Zero) is described with RDF triples using the CC and DC vocabularies[63]. Version information is described using owl:versionInfo.

4.3. Vocabularies for Terminologies, Thesauri and Knowledge Bases

The **Simple Knowledge Organisation System (SKOS)** is a W3C recommendation for the creation of terminologies and thesauri, or more broadly speaking, knowledge organisation systems.[64] We will not discuss it in any depth here since it is a general purpose vocabulary which has applications well beyond the domain of language resources.

In terms of specialised vocabularies or models for the modelling of linguistic knowledge bases – and aside from linguistic data category registries, which will be discussed in Section 4.5 – we can list two prominent ones here. The first is **MMoOn ontology**[65] which was designed for the creation of detailed morphological inventories [21]. It does not currently seem to be available on any semantic repositories/archives/search engines, but it does have its own dedicated website[66] which offers a SPARQL endpoint.[67] Its licensing information (it has a CC-BY 4.0 license) is available as triples using dct:license with a URI as an object.

PHOIBLE is an RDF model for creating phonological inventories [8]. As of the time of writing, PHOIBLE data was no longer available as a complete RDF graph, but only in its native (XML) format from which RDF fragments are dynamically generated. The original data remains publicly available,[68] but on the PHOIBLE website, it is only possible to browse and export selected content into RDF/XML.[69] Since it no longer provides resolvable URIs for its components, PHOIBLE data does not fit within the narrower scope of LLD vocabularies anymore. It does, however, maintain a non-standard way of linking, as it has been absorbed into the Cross-Linguistic Linked Data infrastructure [22, CLLD] (along with other resources from the typology domain). CLLD datasets and their RDF exports continue to be available as open data under https://clld.org/.[70]

4.4. Linguistic Resource Metadata

Due to the importance of this topic, we give a more detailed overview in Section 5.3. Here, we consider only accessibility issues for the two models for language resource metadata, which are described in Section 5.3: The METASHARE ontology[71] and *lime*. The latter has been previously introduced and is described

[47] http://www.w3.org/ns/lemon/synsem
[48] http://www.w3.org/ns/lemon/decomp
[49] http://www.w3.org/ns/lemon/vartrans
[50] http://www.w3.org/ns/lemon/lime
[51] https://lov.linkeddata.es/dataset/lov/vocabs/lemon
[52] https://lov.linkeddata.es/dataset/lov/vocabs/ontolex
[53] https://lov.linkeddata.es/dataset/lov/vocabs/lime
[54] https://lov.linkeddata.es/dataset/lov/vocabs/vartrans
[55] https://lov.linkeddata.es/dataset/lov/vocabs/synsem
[56] https://lov.linkeddata.es/dataset/lov/vocabs/lexdcp
[57] https://archivo.dbpedia.org/info?o=http://www.w3.org/ns/lemon/ontolex
[58] http://archivo.dbpedia.org/info?o=http://www.w3.org/ns/lemon/lime
[59] http://archivo.dbpedia.org/info?o=http://www.w3.org/ns/lemon/vartrans
[60] Using the cc:license property
[61] The guidelines for the module can be found at https://www.w3.org/2019/09/lexicog/, the URL for the module is at http://www.w3.org/ns/lemon/lexicog#
[62] https://archivo.dbpedia.org/info?o=http://www.w3.org/ns/lemon/lexicog
[63] using cc:license in the former and dc:rights in the latter case
[64] https://www.w3.org/2004/02/skos/
[65] https://github.com/MMoOn-Project/MMoOn/blob/master/core.ttl
[66] https://mmoon.org/
[67] Although this was down at the time of writing.
[68] https://github.com/clld/phoible/tree/master/phoible/static/data
[69] See, for example, https://phoible.org/inventories/view/161.
[70] See Section 4.6 below for additional details.
[71] http://www.meta-share.org/ontologies/meta-share/meta-share-ontology.owl/documentation/index-en.html

in more detail in Section 5.3.3. The former is currently in its pre-release version 2.0 (the last update being 2020-03-20). Its license information (it has a CC-BY 4.0 license) is available as triples using dct:license with a URI as an object.

4.5. Linguistic Data Categories

History

Looking back to 2010, two major registries were in widespread use by different communities for addressing the harmonization and linking of linguistic resources via their data categories.

In computational lexicography and language technology, the most widely applied terminology repository was **ISOcat** [23] which provided human-readable and XML-encoded information about linguistic data categories that were applicable to tasks such as linguistic annotation, the encoding of electronic dictionaries and the encoding of language resource metadata via persistent URIs.

In the field of language documentation and typology, the **General Ontology of Linguistic Description (GOLD)** emerged in the early 2000s [24], having been originally developed in the context of the project Endangered Metadata for Endangered Languages Data (E-MELD, 2002-2007).[72] GOLD stood out in particular because of its excellent coverage of low resource languages. In the RELISH project, a curated mirror of GOLD-2010 was incorporated into ISOcat [25]. Unfortunately, since then, GOLD development has stalled and, while the resource is still being maintained by the LinguistList (along with the data from related projects) and still remains accessible,[73] it has not been updated since [26] (and for this reason we have not included it in our summary table). In parts, its function seems to have been taken over by ISOcat, but it is worth pointing out here that the ISOcat registry exists only as a static, archived resource, and is no longer an operational system.

The Current Situation

The 'official' successor of ISOcat, the CLARIN Concept Registry is briefly discussed in Section 5.3 below, but it is not strictly speaking a linked data vocabulary. Another successor of ISOcat is the **LexInfo ontology**,[74] the data category register used in OntoLex-Lemon and which has re-appropriated many of the concepts contained in ISOcat for use within the domain of lexical resources. Currently in its third version, LexInfo can be found both on the LOV search engine[75] and on archivo,[76] it appears both times however in its second version. Version 3.0 has been under development since late 2019 in a community-guided process via GitHub,[77] and is not registered with either service, yet. LexInfo has a (CC-BY 4.0) license, which is described with RDF triples using the CC vocabulary and DCT, with a URI as an object in both cases. Version information is described using owl:versionInfo.

A separate terminology repository for linguistic data categories in linguistic annotation exists: the **Ontologies of Linguistic Annotation** [27, OLiA].[78] OLiA has been in development since 2005 in an effort to link community-maintained terminology repositories such as GOLD, ISOcat or the CLARIN Concept Registry with annotation schemes and domain- or community-specific models such as LexInfo or the Universal Dependencies specifications by means of an intermediate "Reference Model". OLiA consists of a set of modular, interlinked ontologies and is designed as a native linked data resource. Its primary contributions are to provide machine-readable documentation of annotation guidelines and to link together other terminology repositories. It has been suggested that such a collection of linking models, developed in an open source process via GitHub, may be capable of circumventing some of the pitfalls of earlier, monolithic solutions of the ISOcat era [28]. At the moment, OLiA covers annotation schemes for more than 100 languages, for morpho-syntax, syntax, discourse and aspects of semantics and morphology. OLiA has a (CC-BY 4.0) license; this is described using the Dublin Core property license with a URI as an object.

4.6. Vocabularies for Typological Datasets

Relevant Resources and Initiatives

Linguistic typology is commonly defined as the field of linguistics that studies and classifies languages based on their structural features [29]. The field of linguistic typology has natural ties with language docu-

[72] http://emeld.org/
[73] https://linguistlist.org/projects/gold.cfm
[74] https://lexinfo.net/
[75] https://lov.linkeddata.es/dataset/lov/vocabs/lexinfo
[76] http://archivo.dbpedia.org/info?o=http://www.lexinfo.net/ontology/2.0/lexinfo
[77] It will be the first version that is compliant with OntoLex-Lemon.
[78] http://purl.org/olia

mentation, and accordingly, considerable work on linguistic typology and linked data has been conducted in the context of the GOLD ontology (see above, Section 4.5). We can identify the following relevant datasets.

One of the main contributors and advisors to the scientific study of typology is the **Association for Linguistic Typology (ALT)**.[79] They facilitate the description of the typological patterns underlying datasets. One of the most well-known resources that ALT makes available is the **World Atlas of Language Structures (WALS)**[80] [30, 31] which is a large database of phonological, grammatical, and lexical properties of languages gathered together from various descriptive materials. This resource can both be used interactively online and is also downloadable. The **CLLD**[81] **(Cross-Linguistic Linked Data)** project integrates WALS, thus, offering a framework that structures this typological dataset using the Linked Data principles.

Another collection that provides web-based access to a large collection of typological datasets is the **Typological Database System (TDS)** [32–34]. The main goals of TDS are to offer users a linguistic knowledge base and content metadata. The knowledge base includes a general ontology and dictionary of linguistic terminology, while the metadata describes the content of the term ontology databases. TDS supports a unified querying across all the typological resources hosted with the help of an integrated ontology. The **Clarin Virtual Language Observatory (VLO)**[82] incorporates TDS among its repositories.

Finally, another group of datasets relevant for typological research include large-scale collections of lexical data, as provided, for example, by **PanLex**[83] and **Starling**.[84] An early RDF edition of PanLex was described by [35] and was incorporated in the initial version of the Linguistic Linked Open Data cloud diagram; at the time of writing, however, this version does not seem to be accessible anymore. Instead, CSV and JSON dumps are being provided from the PanLex website. On this basis, [36] describe a fresh OntoLex-Lemon edition of PanLex (and other) data as part of the **ACoLi Dictionary Graph**.[85] However, they currently do not provide resolvable URIs, but rather redirect to the original PanLex page. The authors mention that linking would be a future direction, and in preparation for this, they provide a **TIAD-TSV** edition of the data along with the OntoLex-Lemon edition, with the goal to adapt techniques for lexical linking developed in the context of, for example, the ongoing series of shared tasks on translation inference across dictionaries (TIAD).[86] As for the specific modelling requirements of lexical datasets in linguistic typology, these are not fundamentally different from other forms of lexical data. They do, however, require greater depth with respect to identifying and distinguishing language varieties. This was one of the driving forces behind the development of Glottolog (see Section 5.3.4 below).

Vocabularies for Typological Datasets

In terms of linked data vocabularies and models which are relevant for the creation of typological databases, we can identify **LexVo**[87] [37]. This vocabulary bridges the gap between linguistic typology and the LOD community and brings together language resources and linked data entity relationships. Indeed, the project behind LexVo has managed to link a large variety of resources on the Web, besides providing global IDs (URIs) for language-related objects. LexVo is available on archivo[88] but is not yet available on LOV. Further discussion of this vocabulary can be found in Section 5.3.4

4.7. Excursus: Tools and Platforms for the Publishing of LLD

The availability of tools and platforms for the editing, conversion and publication of LLD resources, on the basis of the models which we discuss in this article, is critical for the adoption of those models amongst a wider community of end users. It can be especially important for users who are unfamiliar with the technical details of linked data and the Semantic Web, and yet who are highly motivated to create and/or make use of linked data resources. Such tools/platforms are helpful, for instance, when it comes to the validation and post-editing by domain experts of language resources which have been generated automatically or semi-automatically.

In terms of existing tools or software which offer dedicated provision for the models which we look at

[79] https://linguistic-typology.org/
[80] https://wals.info/
[81] https://clld.org/
[82] https://vlo.clarin.eu/
[83] http://panlex.org
[84] https://starling.rinet.ru/
[85] Data available under https://github.com/acoli-repo/acoli-dicts.
[86] https://tiad2021.unizar.es/
[87] http://lexvo.org/
[88] http://archivo.dbpedia.org/info?o=http://lexvo.org/ontology

in this article, we can mention **VocBench** and **LexO** for OntoLex-Lemon. Both of these are web-based platforms which allow for the collaborative development of computational lexical resources by a number of users. In the case of the **VocBench** platform, currently in its third release [38], users can also develop OWL ontologies and SKOS thesauri as well as OntoLex-Lemon lexica. **LexO** focuses on assisting users in the creation of OntoLex-Lemon lexical resources and was originally developed in the context of DitMaO a project on the medico-botanical terminology of the Old Occitan language [39]. A first generic (i.e., non-project specific) version of LexO, LexO-lite, is available at https://github.com/andreabellandi/LexO-lite.

Finally, we should mention **LLODifier**[89] a suite of tools for creating and working with LLD which is currently being developed by the Applied Computational Linguistics Lab of the Goethe University Frankfurt. These include the **vis** visualization routines[90] for working with NIF and **unimorph** which works with CoNLL-RDF.

5. An Overview of Developments in LLD Community Standards and Initiatives

Summary and Overview The current section comprises an extensive overview of recent developments in various different LLD community standards and initiatives as they relate to LLD models and vocabularies. In particular, it focuses on three areas that we believe have either been the most active or most prominent over the last few years. These are lexical resources (Section 5.1), annotation and corpora (Section 5.2), and metadata (Section 5.3). We have referred to these as community standards/initiatives because they have been pursued or developed as community efforts rather than within a single research group or project. Membership in these communities is (often) open to all, rather than being limited to members of a specific project or to experts nominated by a standards body. The intention being to allow for the participation of a wider range of stakeholders, as well as encouraging the collection of a wider variety of use-cases than might otherwise be possible.

One of the most notable community efforts in the context of LL**O**D is the Open Linguistics Working Group (OWLG) of Open Knowledge International[91]. It was OWLG which first introduced the vision of a Linguistic Linked Open Data cloud in 2011 [40], and it was OWLG's activities, most notably the organization of the long-standing series of international Workshops on Linked Data in Linguistics (LDL, since 2012), as well as the publication of the first collected volume on the topic of Linked Data in Linguistics [41], which ultimately led to the implementation of LLOD cloud in 2012 (something which was celebrated with a special issue of the Semantic Web Journal published in 2015 [42]). The LLOD cloud, now hosted under http://linguistic-lod.org/, has been enthusiastically embraced, with the *Linguistics* category becoming a top-level category in the 2014 LOD cloud diagram, and since 2018, it has represented the first LOD domain sub-cloud.

Around the same time, a number of more specialized initiatives emerged for which the Open Linguistics Working Group acted and continues to act as an umbrella organisation, facilitating information exchange among them and between these initiatives and the broader circles of linguists interested in linked data technologies and knowledge engineers interested in language. Currently, the main activities of the OWLG are the organization of workshops on Linked Data in Linguistics (LDL), the coordination of datathons such as Multilingual Linked Open Data for Enterprises (MLODE 2012, 2013) and the Summer Datathon in Linguistic Linked Open Data (SD-LLOD, 2015, 2017, 2019), maintaining the Linguistic Linked Open Data (LLOD) cloud diagram[92] and continued information exchange via a shared mailing list[93]

Over the years, the focus of discussion has shifted from the OWLG to more specialized mailing lists and communities. At the time of writing, particularly active community groups concerned with data modelling include

- the W3C Community Group Ontology-Lexica,[94] originally working on ontology lexicalization, the group extended their activities after the publication of the OntoLex-Lemon vocabulary (May 2016) and now represents the main locus for discussing the modelling of lexical resources with web standards and in LL(O)D. See Section 5.1.

[89] https://github.com/acoli-repo/LLODifier
[90] https://github.com/acoli-repo/LLODifier/tree/master/vis
[91] https://linguistics.okfn.org/
[92] http://linguistic-lod.org/
[93] Since early 2020, the mailing list operates via https://groups.google.com/g/open-linguistics. Earlier messages are archived under https://lists-archive.okfn.org/pipermail/open-linguistics/.
[94] https://www.w3.org/community/ontolex/

– the W3C Community Group Linked Data for Language Technology,[95] with a focus on language resource metadata and linguistic annotation with W3C standards

Most recently, these activities have converged in funded networks, especially, the Cost Action NexusLinguarum, see Section 6.2.6. We take the standards and initiatives proposed by these communities as our basis of the topics in this section, but in the interests of completeness and to understand current trends we will also look at significant developments respecting these standards and initiatives outside and independent of these groups (see Section 5.1.4).

A discussion of the relationship between community initiatives and projects can be found in Section 6.1.2 below.

5.1. Lexical Resources: OntoLex-Lemon and its Extensions

Summary In this section we describe some of the most recent work that has been carried out on the OntoLex-Lemon model,[96] both within and outside of the ambit of the W3C OntoLex group. With regard to the former case, we discuss three of the latest extensions to the model (the first of which has been published with the other two are still currently under development) in Sections 5.1.1, 5.1.2, and 5.1.3. In Section 5.1.4 we look at a number of new extensions to OntoLex-Lemon which have emerged independently of the W3C OntoLex group over the last two years and which moreover have not been discussed in [8] (for an in-depth discussion of such developments prior to 2018 please refer to the latter paper).

Note that the use of OntoLex-Lemon in a number of different projects is described in Section 6.

5.1.1. The OntoLex-Lemon Lexicography Module (lexicog)

As mentioned previously, *lemon* and its successor OntoLex-Lemon have been widely adopted for the modelling and publishing of lexica and dictionaries as linked data. Both of them have proven to be reasonably effective in capturing many of the most typical kinds of *lexical* information contained in dictionaries and in lexical resources in general (e.g., [43–47]). However, there are some fairly common situations in which the model falls short, and most notably in the representation of certain specific elements of dictionaries and other lexicographic datasets [48]. This is not surprising, given that (as we have mentioned above) *lemon* was initially conceived as a model for a somewhat different use case (grounding ontologies with linguistic information).

In order to adapt OntoLex-Lemon to the modelling necessities and particularities of dictionaries and other lexicographic resources, the W3C OntoLex community group developed a new **OntoLex-Lemon Lexicography Module (*lexicog*)**.[97] This module was the result of collaborative work with contributions from lexicographers, computer scientists, dictionary industry practitioners, and other stakeholders and was first released in September 2019. As stated in the specification, the *lexicog* module "overcome[s] the limitations of *lemon* when modelling lexicographic information as linked data in a way that is agnostic to the underlying lexicographic view and minimises information loss".

The idea is to keep purely lexical content separate from lexicographic (textual) content. For that purpose, new ontology elements have been added that reflect the dictionary structure (e.g., sense ordering, entry hierarchies, etc.) and complement the OntoLex-Lemon model. The *lexicog* module have been validated with real enterprise-level dictionary data [49] and a final set of guidelines have been published as an output of the W3C OntoLex group. We give a description of the main classes and properties of the model below[98]

In *lexicog* the structural organisation of a lexicographic resource is now associated with the class LexicographicResource (a subclass of the VoID[99] class Dataset) whereas the lexical content is (as previously) associated with the *lime* class Lexicon (see Section 5.3.3). The former is described as representing "a collection of lexicographic entries[...]in accord with the lexicographic criteria followed in the development of that resource".[100]

These lexicographic entries are represented in their turn by another new *lexicog* class, namely, the class Entry, which is defined as being a "structural element that represents a lexicographic article or record *as it arranged in a source lexicographic resource*"[101] (empha-

[95] https://www.w3.org/community/ld4lt

[96] An introduction to the model is given in Appendix A

[97] https://www.w3.org/2019/09/lexicog/

[98] Please see the guidelines for a comprehensive description with examples.

[99] https://www.w3.org/TR/void/

[100] https://www.w3.org/2019/09/lexicog/#lexicographic-resource

[101] https://www.w3.org/2019/09/lexicog/#Entry

sis ours). An Entry furthermore is related to its source Lexicographic Resource via the object property entry.

The class Entry is a subclass of the more general class Lexicographic Component, defined as "a structural element that represents the (sub-)structures of lexicographic articles providing information about entries, senses or sub-entries", members of this class "can be arranged in a specific order and/or hierarchy".[102] That is, Lexicographic Component allows for the representation of the ordering of senses in an entry (and even potentially entries if this is required), the arrangement of senses and sub-senses in a hierarchy, etc. in a published lexicographic resource (by making use of the classes and properties we have looked at above, along with the *lexicog* object property subComponent) separately from the representation of the same resource as lexical content.

Finally, we need some way of linking together these two levels of representation. This is provided by the *lexicog* object property describes which relates individuals of class Lexicographic Component, which belong to a specific lexicographic resource, "to an element that represents the latest information provided by that component in the lexicographic resource".[103] These elements are described in 1.

As an example, let's look a *lexicog* encoding for the entry for the Italian word *chiaro* 'clear' in the popular Italian language dictionary *Treccani*.[104] This latter lists the word an adjective, a masculine noun and an adverb. It also lists the adverb *chiaramente* 'clearly' as a related entry, along with the diminutive *chiaretto*.

More precisely, the first two of the (four) subsenses of the entry are classed as adjectives, the third as a noun, and the fourth as an adverb. We will simplify this for the purposes of exposition by assuming that the first subsense is an adjective, the second a noun, and the third an adverb. This can be represented as follows. First, we represent the encoding of the Treccani dictionary structure itself, and the different sub-components of the entry for *chiaro*:

```
:treccaniRDF a lexicog:LexicographicResource;
     dc:language "it" ;
     lexicog:entry :chiaro_entry .

:chiaro_entry a lexicog:Entry ;
     rdfs:member :chiaro_1_comp,
     :chiaro_2_comp,
     :chiaro_3_comp.

:chiaro1_comp a lexicog:LexicographicComponent;
:chiaro2_comp a lexicog:LexicographicComponent;
:chiaro3_comp a lexicog:LexicographicComponent.
```

Next we encode a lexicon which represents the content of the resource in the last listing.

```
:myItLexicon a lime:Lexicon;
     lime:language "it" ;
     lime:entry :chiaro1_adj, :chiaro2_n,
     :chiaro3_adv .

:chiaro1_adj a ontolex:LexicalEntry .
:chiaro2_n a ontolex:LexicalEntry .
:chiaro3_adv a ontolex:LexicalEntry .
```

Finally, we bring the two resources together using the describes property.

```
:chiaro1_comp lexicog:describes :chiaro1_adj .
:chiaro2_comp lexicog:describes :chiaro2_n.
:chiaro3_comp lexicog:describes :chiaro3_adv
```

5.1.2. OntoLex-Lemon Morphology Module

Morphology often an important role in the description of languages in lexical resources, even if the extent of its presence in can often vary, ranging from the sporadic indication of certain specific forms in a dictionary (e.g. plural form for some nouns) to electronic resources which provide tables with entire inflectional paradigms for every word.[105] Consequently, the W3C OntoLex community group since November 2018 has been developing another extension of the original model that would allow for better representation of morphology in lexical resources.

The original OntoLex-Lemon model, together with LexInfo (see Section 4.5), provides the means of encoding basic morphological information. For lexical entries, morpho-syntactic categories such as part of speech can be provided and basic inflection information (i.e., the morphological relationship between a lexical entry and its forms) can be modelled by creating additional inflected forms with corresponding morpho-syntactic features (e.g. case, number, etc.). However, this only covers a small portion of the morphological data to be modelled in many lexical resources. Neither derivation (i.e. morphological relationships between lexical entries) nor additional inflectional information (e.g. declension type for Latin nouns) can be properly modelled with the original model. The new **OntoLex-Lemon Morphology** mod-

[102] https://www.w3.org/2019/09/lexicog/#lexicographic-component

[103] https://www.w3.org/2019/09/lexicog/#describes-0

[104] https://www.treccani.it/vocabolario/chiaro/

[105] For example, *Wiktionary*, https://en.wiktionary.org/wiki/Buch#Declension.

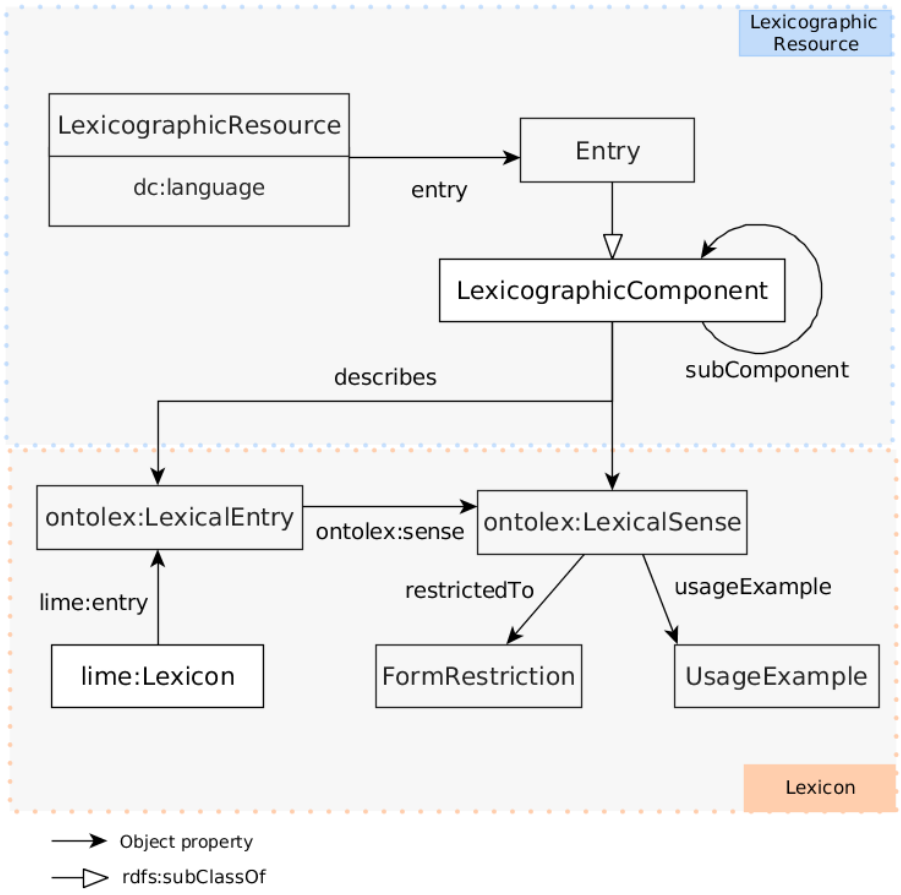

Fig. 1. The *lexicog* module (taken from the guidelines).

ule has been proposed to address these limitations. The scope of the module is threefold:

- *Representing derivation*: for a more sophisticated description of the decomposition of lexical entries;
- *Representing inflection*: introducing new elements to represent paradigms and wordform-building patterns;
- Providing means to *create wordforms automatically* based on lexical entries, their paradigms and inflection patterns.

Figure 2 presents a diagram for the module.

The central class of the module, used in the representation of both derivation and inflection, is Morph with subclasses for different types of morphemes.

For derivation, elements from the *decomp* module are reused. A derived lexical entry has Components for each of the morphemes of which it consists. A *stem* corresponds to a different lexical entry whereas morphemes, which do not correspond to any headwords, correspond to an object of a Morph class. A derived lexical entry has constituent properties pointing to objects of the Component class:

```
:lex_drive_v a ontolex:LexicalEntry .
:lex_driver_n a ontolex:LexicalEntry ;
              decomp:constituent :component_drive,
                                 :component_er .

:component_drive a decomp:Component ;
                 decomp:correspondsTo :lex_drive_v .
:component_er a decomp:Component ;
              decomp:correspondsTo :suffix_er .

:suffix_er a morph:AffixMorph .
```

Inflection is modelled as follows: every instance of Form has properties morph:consistsOf which point to

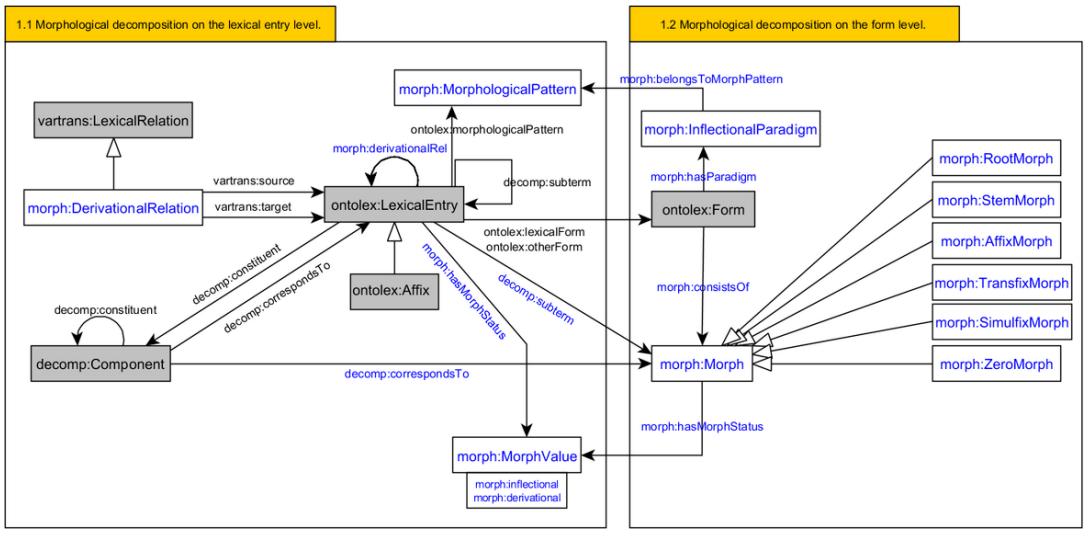

Fig. 2. Preliminary diagram for the Morphology Module.

instances of morph:Morph.[106] These instances can have morpho-syntactic properties expressed by linking to an external vocabulary, e.g. LexInfo:

```
:lex_drive_v a ontolex:LexicalEntry ;
             ontolex:otherForm :form_drives .

:form_drives a ontolex:Form ;
             consistsOf :stem_drive_v, :suff_s .

:suff_s a morph:AffixMorph ;
        lexinfo:number lexinfo:Plural .
```

The module[107] has not yet been published and is still very much under development by the W3C group. At the time of writing, a consensus was reached on the first two parts of the module, and an overview of these has been published in [50]. The third part, which concerns the automatic generation of forms, is currently being discussed, and the next step will be validating the model by creating resources using the module.

5.1.3. *OntoLex-FrAC: Frequency, Attestations, Corpus Information*

In parallel with the development of the Morphology Module, the OntoLex W3C group has also started developing a separate module that would allow for the enrichment of lexical resources with information drawn from corpora. Most notably, this includes the representation of attestations (often used as illustrative examples in a dictionary). These latter were originally discussed within *lexicog* (See 5.1.1), but this discussion quickly outgrew the confines of computational lexicography/e-lexicography alone. Furthermore, it was observed that OntoLex-Lemon lacked any support for corpus-based statistics, a cornerstone not only of empirical lexicography, but also of computational philology, corpus linguistics and language technology, and thus, again, beyond the scope of the *lexicog* module. Finally, the OntoLex community group felt the need to specifically address the requirements of modern language technology by extending its expressive power to corpus-based metrics and data structures like word embeddings, collocations, similarity scores and clusters, etc.

The development of the module has been use-case-based, which has dictated the order and development for various parts of the FRaC module. The stable parts of the module include the representation of (absolute) frequencies and attestations, and, by analogy, any use case that requires pointing from a lexical resource into an annotated corpus or other forms of external empirical evidence [51]. We will limit ourselves to describing these stable parts in what follows.

The central element which has been introduced in FRaC e is frac:Observable defined as "an abstract superclass for any element of a lexical resource that frequency, attestation or corpus-derived information can

[106] One of the problems with this approach is that the order of the affixes is undefined, there are several possible solutions for this, e.g. a property next between two morphs, but currently there is no consensus in the community on how to model the order.

[107] https://www.w3.org/community/ontolex/wiki/Morphology

be expressed about"[108]. Since the type of elements for which corpus-based information can be provided is not limited to an entry, form, sense, or concept but can be any of these, Observable was introduced as a superclass for all these classes, among others to be potentially defined by the user.

The module provides means to model only absolute frequency, because "relative frequencies can be derived if absolute frequencies and totals are known" [51, p. 2]. To represent frequency, a property frequency with an instance of CorpusFrequency as an object should be defined. This instance must implement the properties corpus and rdf:value:[109]

```
epsd:kalag_strong_v a ontolex:LexicalEntry;
   frac:frequency [
      a frac:CorpusFrequency;
      rdf:value "2398"^^xsd:int;
      frac:corpus
         <http://oracc.museum.upenn.edu/epsd2/pager>
] .
```

The usage recommendation is to define a subclass of CorpusFrequency for a specific corpus when representing frequency information for many elements in the same corpus.

In FRAC corpus attestations, i.e. corpus evidence in FrAC, are defined as "a special form of citation that provides evidence for the existence of a certain lexical phenomenon; they can elucidate meaning or illustrate various linguistic features"[110]. As with frequency, there is a class Attestation, an instance of which should be an object of a property attestation. This class is associated with two properties: attestationGloss – the text of the attestation – and locus – the location where the attestation can be found:

```
diamant:sense_1 a ontolex:LexicalSense;
  frac:attestation diamant:attestation_1 ;
  diamant:attestation_1 a frac:Attestation ;
  cito:hasCitedEntity diamant:cited_document_1 ;
  cito:hasCitingEntity diamant:sense_1;
  frac:locus diamant:locus_1 ;
  frac:quotation "... dat men licht yemant de cat
              aen het been kan werpen," .
```

The FrAC module does not provide an exhaustive vocabulary and instead promotes reuse of external vocabularies, such as CITO [52] for a citation object and NIF or WebAnnotation (see 5.2) to define a locus.

Another, more recent paper focused on representing embeddings in lexical resources is [53]. It should be noted that the term *embedding* is used here in a broader sense than is usual in the field of natural language processing, namely as a morphism Y ($f : X \to Y$).[111] Therefore, the class Embedding has subclasses for modelling bags of words and time series.

The main motivation to model embeddings as a part of this module is to provide metadata as RDF for precomputed embeddings, therefore a word vector itself is stored as a string with an embedding vector:

```
:embedding a
   frac:Embedding;
   dc:extent "300"^^xsd:int;
   rdf:value "0.145246 0.38873 ...";
```

As with modelling frequency, the recommendation is to define a subclass for the specific type of embedding concerned in order to make the RDF less verbose.

Figure 3 presents a diagram of the latest version of the module. Note that we will not go into detail on the classes Similarity, Collocation and ContextualRelation here, since the definitions of these classes and their related properties is still under discussion. However, we leave them in the diagram to give the reader an idea of the current progress of the model.

At the time of writing, module development is focused on collecting and modelling various use-cases. Among the many use-cases that were proposed during this phase, one stood out in particular and seemed to be more challenging than the others: this was related to the modelling of sign language data. Given the nature of the data (video clips with signs and/or time series of key coordinates for preprocessed data), it was decided that although the use-case was out of the scope of the FrAC module, it did indeed raise serious interest within the community, and therefore discussion on whether it will be developed as a separate module in the future, is now underway. The question of the scope of this new module and, more generally, its connection to OntoLex-Lemon, is currently subject to discussion.

5.1.4. Selected individual contributions

'Unofficial' OntoLex-Lemon extensions developed outside the W3C OntoLex Community Group are manifold, and while these are not yet being pursued as candidates for future OntoLex-Lemon modules by the group, they may represent a nucleus and a cumulation point for future directions.

[108] https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md (Accessed 20/01/2022)

[109] Examples in this section are based on those in [51].

[110] https://github.com/ontolex/frequency-attestation-corpus-information/blob/master/index.md (Accessed 20/01/2022)

[111] An injective structure-preserving map.

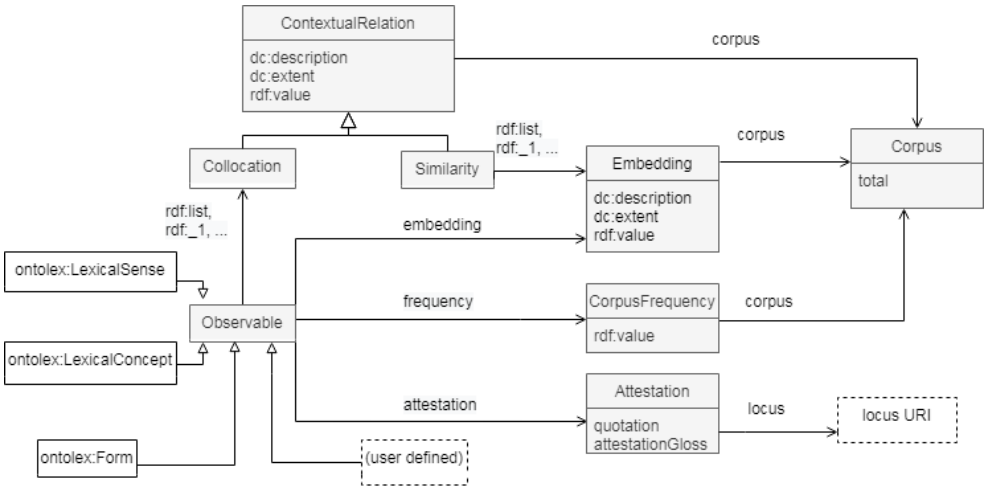

Fig. 3. Preliminary diagram for the FrAC Module.

Selected recent extensions include ***lemon-tree*** [54], an OntoLex-Lemon and SKOS based model for publishing *topical thesauri*, where the latter are defined as lexical resources which are organised on the basis of meanings or topics.[112] The use of the lemon-tree model to publish the Thesaurus of Old English [55] reveals the flexibility of the OntoLex-Lemon/LLD approach in modelling more specialised kinds of linguistic information. As indeed does ***lemonEty*** [56] another 'unofficial' extension of the OntoLex-Lemon model, which has been proposed as a means of encoding etymological information contained both in lexica and dictionaries as well in other kinds of resources (such as articles or monographs). The *lemonEty* model does this by exploiting the graph-based structure of RDF data and by rendering explicit the status of etymologies as linguistic hypotheses.

In both of these cases, the RDF data model together with the various different standards and technologies which make up the Semantic Web stack as a whole, allows for the structuring of data that is strongly heterogeneous and integrates together temporal[113], geographical and historical information.

5.2. Annotation and Corpora

Summary In this section, we give an overview of a number of LLD vocabularies for the annotation of texts. Section 5.2.1 constitutes a detailed introduction and general overview of this topic. Then we focus on the two most popular LLD vocabularies for text annotation, *the NLP Interchange Format* (Section 5.2.2) and *Web Annotation* (Section 5.2.3). Next, in Section 5.2.4 we look at two domain specific vocabularies, Ligt and CoNLL-RDF. Finally, in Section 5.2.5, we look at the prospects of a convergence between the vocabularies which we have discussed. (Note that in this section, we only discuss vocabularies that define *data structures* for linguistic annotation by NLP tools and in annotated corpora. Linguistic categories and grammatical features, as well as other information that represents the content of an annotation, are assumed to be provided by a(ny) repository of linguistic data categories (see above))

5.2.1. Introduction and Overview

Linguistic annotation of corpora by NLP tools in a way that integrates Semantic Web standards and technologies has long been a topic of discussion within LLD circles, with different proposals grounded in traditions from natural language processing [58], web technologies [59], knowledge extraction [60], but also from linguistics [61], philology [62], and the development of corpus management systems [63, 64].

A practical introduction to the various different vocabularies used (by various different communities, for different purposes and according to different capabilities) for linguistic annotation in RDF today is given over the course of several chapters in [11]. In brief, the RDF vocabularies which are most widely used for this purpose are the **NLP Interchange Format** (NIF,

[112] The lemon-tree specifications can be found here https://ssstolk.github.io/onto/lemon-tree/

[113] For a discussion of the possibilities of integrating temporal information in OntoLex-Lemon see [57]

in language technology) and **Web Annotation** (OA, in bioinformatics and digital humanities), as well as customizations of these. We describe NIF in Section 5.2.2 and Web Annotation in Section 5.2.3.

In the current section we give an overview of the relationship between RDF and two other pre-RDF vocabularies, then we will touch upon some platform specific RDF vocabularies for annotations that have been developed over the years. Aside from software- or platform-specific formats, a number of vocabularies has been developed that address specific problems or user communities.

Pre-RDF Vocabularies

Developed by the ISO TC37/SC4 Language Resource Management group, the **Linguistic Annotation Framework (LAF)** vocabulary represents "universal" data structures shared by the various, domain- and application specific ISO standards [65]. Following the earlier insight that a labelled directed multigraph can represent any kind of linguistic annotation, LAF produces concepts and definitions for four main aspects of linguistic annotation: **anchors and regions** elements in the primary data that annotations refer to; **markables (nodes)** elements that constitute and define the scope of the annotation by reference to anchors and regions; **values (labels)** elements that represent the content of a particular annotation; and **relations (edges)** links (directed relations) that hold between two nodes and can be annotated in the same was as markables.

Note that in relation to Web Annotation *anchors* roughly correspond to Web Annotation selectors (or target URIs); *markables* roughly correspond to annotation elements; *values* to the body objects of Web Annotation. In Web Annotation, relations as data structures are not foreseen.[114] As for NIF, its relation with LAF is more complex. Like Web Annotation, NIF does not provide a counterpart of LAF relations, but more importantly, the roles of regions and markables are conflated in NIF: Every markable must be a string (character span), and for every character span, there exists exactly one potential markable (URI, or, a number of URIs with different schemes that are owl:sameAs).

At the moment, direct RDF serializations of LAF do not seem to be widely used in an LLOD context. The reason is certainly that the dominant RDF vocabularies for annotations, despite their deficiencies, cover the large majority of use cases. One notable RDF serialisation of LAF however is **POWLA** [66], an OWL2/DL serialization of PAULA, a standoff-XML format that implemented the LAF as originally described by [67]. POWLA complements LAF core data structures with formal axioms and slightly more refined data structures that support, for example, effective navigation of tree annotations. On current applications of POWLA see the CoNLL-RDF Tree Extension below.[115]

It is also worth mentioning **TEI/XML** in the context of this discussion. The standard, widely used in the digital humanities and in computational philology, only comes with partial support for RDF and does not represent a publication format for Linked Data. Traditionally there has been an acknowledgement on the part of the TEI community of the value in being able to link from a digital edition (or another TEI/XML document) to a knowledge graph.[116] Interlinking between (elements of) electronic editions created with TEI was addressed by means of specialised XML attributes with narrowly defined semantics. Accordingly, electronic editions in TEI/XML do not normally qualify as Linked Data, even if they use and provide resolvable URIs (TEI pointers).[117]

The annotation *of* rather than *within* TEI documents, however, has been pursued by Pelagios/-Pleiades, a community interested in the annotation of historical documents and maps with geographical identifiers and other forms of geoinformation (though

[114] Although Web Annotation lacks any formal counterpart of edges or relations as defined by LAF there have been attempts to define a vocabulary that extends Web Annotation with LAF data categories [59], but this has apparently never been applied in practice.

[115] Others include [63] utilised an RDF graph, with an RDF vocabulary for nodes, labels and edges to express linguistic data structures over a corpus backend natively based on an RDBMS; a prototypical extension of Web Annotation with an RDF interpretation of the LAF described by [59], which and the LAPPS Interchange Format, conceptually and historically an instance of LAF, which has see the discussion below on platform-specific vocabularies.

[116] This is useful for instance for managing prosopographical, bibliographical or geographical information

[117] This may not be considered to be drastic for electronic editions of historical manuscripts which one could conceivably complement with information drawn from the LLOD cloud. The situation is quite different for dictionaries whose content could easily be made accessible and integrated with other lexical resources on the LLOD cloud, e.g., for future linking. The situation has begun to change over the last few years, and long-standing efforts to develop technological bridges between both TEI and LOD are beginning to yield concrete results. For instance, different tools for the conversion of lexical resources in different TEI dialects to OntoLex-Lemon have been presented in the last years. Among others, this includes a converter for TEI Dict/FreeDict dialect, https://github.com/acoli-repo/acoli-dicts/tree/master/stable/freedict [36]. For ELEXIS related developments, see Section 6.2.3.

this does not yet run to linguistic annotations). One result of these efforts is the development of a specialised editor called Recogito, and its extension to TEI/XML. In this case the annotation is not part of the TEI document, but stored as standoff annotation in a JSON-LD format, and thus, is in compliance with established web standards and re-usable by external tools and addressable as Linked Data. However, this approach is restricted to cases in which the underlying TEI document is static and no longer changes.[118] Therefore, there is a need for encoding RDF triples directly inline in a TEI document. Happily, it has been demonstrated that this can be done in a W3C- and XML-compliant way by incorporating RDFa attributes into TEI [68, 69]. As a result and after more than a decade of discussions, the TEI started in May 2020 to work on a customization that allowed the use of RDFa in TEI documents.[119]

Platform Specific RDF Vocabularies

Over the years, several platforms, projects and tools have come up with their own approaches for modelling annotations and corpora as linked data. Notable examples include the RDF output of machine reading and NLP systems such as FRED [70], NewsReader [71] or the LAPPS Grid [72]. We discuss these below.

FRED provides output based on NIF or EARMARK [73], with annotations partially grounded in DOLCE [74], but enriched with lexicalized ad hoc properties for aspects of annotation covered by these.[120] The **NewsReader Annotation Format (or NLP Annotation Format) NAF**, is an XML-standoff format for which an NIF-inspired RDF export has been described [75], and LIF, the LAPPS Interchange Format [76], a JSON-LD format used for NLP workflows by the LAPPS Grid Galaxy Workflow Engine [77][121].

Both LIF and NAF-RDF are, however, not generic formats for linguistic annotations but rather, provide (relatively rich) inventories of vocabulary items for specific NLP tasks.[122] Neither seem to have been used as a format for data publication, and we are not aware of their use independently of the software they have originally been created for or are being created by.

5.2.2. NLP Interchange Format

The NLP Interchange Format (NIF),[123] developed at AKSW Leipzig, was designed to facilitate the integration of NLP tools in knowledge extraction pipelines, as part of the building of a Semantic Web tool chain and a technology stack for language technology on the web [60]. NIF provides support for a broad range of frequently occurring NLP tasks such as part of speech tagging, lemmatization, entity linking, coreference resolution, sentiment analysis, and, to a limited extent, syntactic and semantic parsing. In addition to providing a technological solution for integrating NLP tools in semantic web annotations, NIF also provides specifications for web services.

A core feature of NIF is that it is grounded in a formal model of strings and that it makes the use of String URIs as fragment identifiers obligatory for anything annotable by NIF. Every element that can be annotated in NIF has to be a string.[124] NIF does support different fragment identifier schemes, e.g., the offset-based scheme defined by RFC 5147. [79] As a consequence, any two annotations that cover the same string are bound to the same (or owl:sameAs) URI. While this has the advantage of being able to implicitly merge the output of different annotation tools, this limits the applicability of NIF to linguistically annotated corpora.

As an example, NIF does not allow us to distinguish multiple syntactic phrases that cover the same token. Consider the sentence “Stay, they said.”[125] The Stanford PCFG parser[126] analyzes *Stay* as a verb phrase contained in (and only constituent of) a sentence. In NIF, both would be conflated. Likewise, zero elements

[118] Otherwise, the efforts for synchronization will by far outweigh any benefit that the use of W3C standards for encoding the annotation brings

[119] For the current status of the discussion, cf. https://github.com/TEIC/TEI/issues/311 and https://github.com/TEIC/TEI/issues/1860

[120] For the rendering of discourse relations, for example, it produces properties such as fred:becauseOf (apparently extrapolated from the surface string, so, not ontologically defined).

[121] A more recent development in this regard is that efforts have been undertaken to establish a clear relation between LIF and pre-RDF formats currently used by CLARIN [78].

[122] Historically, LIF is grounded in LAF concepts and has been developed by the same group of people, but no attempt seems to have been made to maintain the level of genericity of the LAF. Instead, application-specific aspects seem to have driven LIF design.

[123] https://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html

[124] In particular, this includes the classes nif:Phrase and nif:Word. With the introduction of support for provenance annotations, NIF 2.0 also introduced nif:Annotation which can be attached as a property to a NIF string. However, it is to be noted that the linguistic data structures defined by NIF 2.0 are *not* subclasses of nif:Annotation, but of nif:String.

[125] From Stephen Dunn (2009), ‘Dont Do That’, poem published in the New Yorker, June 8, 2009.

[126] http://nlp.stanford.edu:8080/parser/index.jsp

in syntactic and semantic annotation cannot be expressed. Another limitation of NIF is its insufficient support for annotating the internal structure of words. It is thus largely inapplicable to the annotation of morphologically rich languages.

Overall, NIF fulfills its goals to provide RDF wrappers for off-the-shelf NLP tools, but it is not sufficient for richer annotations such as are frequently found in linguistically annotated corpora. Nevertheless, NIF has been used as a publication format for corpora with entity annotations.[127] It also continues to be a popular component of the DBpedia technology stack. At the same time, active development of NIF seems to have slowed down since the mid-2010s, whereas limited progress on NIF standardization has been achieved. A notable exception in this regard is the development of the Internationalization Tag Set [80, ITS] that aims to facilitate the integration of automated processing of human language into core Web technologies. A major contribution of ITS 2.0 has been to add an RDF serialization into NIF as part of the standard.

More recent developments of NIF include extensions for provenance (NIF 2.1, 2016) and the development of novel NIF-based infrastructures around DBpedia and Wikidata [81]. In parallel to this, NIF has been the basis for the development of more specialised vocabularies, e.g., CoNLL-RDF for linguistic annotations originally provided in tabular formats, see Section 5.2.4.

5.2.3. Web Annotation

The Web Annotation Data Model is an RDF-based approach to standoff annotations (in which annotations and the material to be annotated are stored separately) proposed by the Open Annotation community.[128] It is a flexible means of representing standoff annotation for any kind of document on the web. Although the most common use case of Web Annotation is the attaching of a piece of text to a single web resource, it is intended to be applicable across different media formats. So far, Web Annotation has been primarily applied to linguistic annotations in the biomedical domain, although other notable applications include NLP [59] or Digital Humanities [84]. Web Annotation recommends the use of JSON-LD to add a layer of standoff annotations to documents and other resources accessible over the web, with primary data structures defined by the Web Annotation Data Model, formalised as an OWL ontology.

The core data structure of the Web Annotation Data Model is the annotation, i.e., instances of oa:Annotation that have an oa:hasTarget property that identifies the element that carries the annotation, and the oa:hasSource property that – optionally – provides a value for the annotation, e.g., as a literal. The target can be a URI (IRI) or a selector, i.e., a resource that identifies the annotated element in terms of its contextual properties, formalised in RDF, e.g., its offset or characteristics of the target format. By supporting user-defined selectors and a broad pool of pre-defined selectors for several media types, Web Annotation is applicable to any kind of media on the web. Targets can also be more compact string URIs, as introduced, for example, by NIF. NIF data structures can thus be used to complement Web Annotation [60].

Web Annotation can be used for any labelling or linking task, e.g., POS tagging, lemmatization, entity linking. It does, however, not support relational annotations such as syntax and semantics, nor (like NIF) the annotation of empty elements. The addition of such elements from LAF has been suggested [59], but does not seem to have been adopted, as labelling tasks dominate the current usage scenarios of Web Annotation.

Unlike NIF, Web Annotation is ideally suited for the annotation of multimedia content or entities that are manifested in different media simultaneously (e.g., in audio and transcript). As a result, it has become popular in the digital humanities, e.g., for the annotation of geographical entities with tools such as Recogito [85], especially since support for creating standoff annotations for static TEI/XML documents was added (around March 2018 [86, p.247]).

5.2.4. Domain-specific solutions: Ligt and CoNLL-RDF

Interlinear glossed text (IGT) is a notation where annotations are placed, as the name suggests, between the lines of a text with the purpose of helping readers to understand and interpret linguistic phenomena. The notation is frequently used in education and various language sciences such as language documentation, linguistic typology, and philological studies (for instance, it is commonly used to gloss linguistic examples). Moreover, IGT data can consist of different layers, including translation and transliteration layers, and usually contains layers for ensuring morpheme-

[127] The most prominent example, the NIF edition of the Brown corpus published in 2015, formerly available from http://brown.nlp2rdf.org/, does not seem to be accessible anymore. Attempted to access on Jan 23, 2021.

[128] The Web Annotation data model and vocabulary were published as W3C recommendations in 2017 [82, 83].

level alignment. IGT is not supported by any established vocabularies for representing annotations on linguistic corpora. And although there exist several specialised formats which are specifically designed for the storage and exchange of IGT, these formats are not reused across different tools, limiting the reusability of annotated data.

In order to help overcome this situation and improve data interoperability, the RDF vocabulary **Ligt** [87] has been proposed for representing IGT as linked data. Ligt is a tool-agnostic representation model for IGT which in addition to structural interoperability also enables the use of LLD vocabularies and terminology repositories.

The Ligt vocabulary was developed as a generalisation over the data structures employed by established tools for creating IGT annotations, most notably Toolbox [88], FLEx [89] and Xigt [90].[129] Ligt is intended to facilitate a pivot format that faithfully captures the linguistic information produced by these tools in a uniform way for subsequent processing. Notably, since its publication, Ligt has been adopted by third party users to model and annotate IGT from 280 endangered languages and their publication as Linked Open Data [91].

Although Ligt was designed for a very specific set of domain requirements, it can be considered a useful contribution to LLD vocabularies for textual annotation. This is because it provides data structures that are relevant for low-resource and morphologically rich languages but which had been neglected by earlier RDF vocabularies for linguistic annotation on the web, in particular, by NIF and Web Annotation.[130]

Another domain specific RDF-based vocabulary which aims to provide a serialisation-independent way of dealing with textual annotations is **CoNLL-RDF** [92]. This latter vocabulary is based on the so-called "CoNLL formats", a family of a tab-separated values (TSV) based-formalisms used to represent linguistically annotated natural language in fields such as NLP,[131] corpus linguistics, and more generally in the language sciences. CoNLL-RDF [92] provides a data model and a programming library that aim to facilitate the processing and transformation of such data regardless of the original order and number of columns, whether the source format used fixed-size tables (as for most CoNLL dialects) or variable size tables (such as all CoNLL formats that contain semantic role annotations). Sentences are sequentially converted to an RDF graph in accordance to the label information provided by the user. The listing below provides a slightly simplified annotation from the 2005 edition of the Shared Task of the SIGNLL Conference on Computational Natural Language Learning (CoNLL-05):

```
#  WORD          POS  PARSE
   The           DT    (S (NP   *
   spacecraft    NN             *)
   ...
```

Here, the wordform is provided in the first column, the second column provides a part-of-speech tag. The PARSE column contains a full parse in accordance with the Penn Treebank [93].The CoNLL-RDF library reads such data as a continuous stream; every sequence of rows enclosed in empty lines is processed as a block, assigned a URI and the type nif:Sentence, every row is assigned a URI and the type nif:Word, and the annotation of every column stored as value of a property in the conll namespace that is generated from the column label.[132] Links between and among sentences and words are encoded in accordance with NIF:

```
:s1_1 a nif:Word; nif:nextWord :s1_2;
      conll:WORD "The"; conll:POS "DT";
      conll:PARSE "(S (NP *".
```

Among other things, a CoNLL-RDF edition of the Universal Dependencies corpora[133] is available in the LLOD cloud diagram. The corpora are linked with the OLiA ontologies; further linking with additional LLOD resources, in particular, lexical resources, has not been explored at the time of writing. CoNLL-RDF has also been applied to the linking of corpora to dictionaries [94] and knowledge graphs [95]. It has also formed the basis of work on the syntactic parsing of historical languages [96, 97], the consolidation of syntactic and semantic annotations [98], corpus querying [99], and language contact studies [100]. In addition to the storing of syntactic parses as plain strings,

[129] One should note that these tools are currently incompatible with each other and information can only be exchanged between them if manual corrections are applied.

[130] However, it would be possible to encode Ligt information with a generic LAF-based vocabulary such as POWLA

[131] Indeed in NLP the CoNLL formats have become de-facto standards for the most frequently used types of annotations having been popularised in a long-standing series of shared tasks over the last two decades

[132] The columns HEAD (for dependency annotation) and PRED-ARGS (for semantic role annotations) are treated differently as they produce object properties, i.e., links, rather than datatype properties. Similarly, the column ID receives special handling. If provided as column label, as its value is used to overwrite the offsets that CoNLL-RDF normally adopts for creating word (row) URIs.

[133] https://universaldependencies.org/

a further extension of CoNLL-RDF adds native support for tree structure [101], extending NIF/CoNLL-RDF data structures with POWLA [66]. As a result, the phrase structure of the example above can now be represented as:

```
:s1_1 a nif:Word; nif:nextWord :s1_2;
      conll:WORD "The"; conll:POS "DT";
      powla:hasParent _:np.
_:np a conll:PARSE; rdf:value "NP";
      powla:next _:vp;
      powla:hasParent _:s.
_:s a conll:PARSE; rdf:value "S".
...
```

The CoNLL-RDF tree extension uses a minimal fragment of POWLA, the properties powla:hasParent (pointing to the parent node in a DAG) and powla:next (pointing to the following sibling in a tree). The class powla:Node, implicit in the listing above, can be inferred (using RDFS) from the use of these properties.

5.2.5. Towards a Convergence

The large number of vocabularies mentioned above already reveals something of a problem, that is, that applications and data providers may choose from a broad range of options, and depending on the expectations and requirements of their users, they may even need to support multiple different output formats, protocols and service specifications that could potentially be mutually incompatible. So far, no clear consensus on a single Semantic Web vocabulary for linguistic annotations has emerged, albeit NIF and Web Annotation appear to enjoy relatively high popularity in their respective user communities. However, they are not compatible with each other and neither do they support linguistic annotation to the same (or even, what the authors would consider a sufficient) extent, thus motivating the continuous development of novel, more specialised vocabularies. Synergies between Web Annotation and NIF were explored relatively early on [60], and Cimiano et al. [102, p.89-122] describe how they can be used in combination with each other, in conjunction with more specialised vocabularies such as CoNLL-RDF, and more general vocabularies such as POWLA to model data in a way that suits the following criteria:

- it is applicable to any kind of primary data, including non-textual data (via Web Annotation selectors);
- it can also express reference to primary data in a compact fashion (via NIF String URIs);
- it permits round-tripping between RDF graphs and conventional formats (via CoNLL-RDF and the CoNLL-RDF library);
- it supports generic linguistic data structures (via POWLA, resp., the underlying LAF model).

However, while the combination of these various components is possible and in principle operational, this also means that a user or provider of data needs to understand and develop a coherent vision of at least five different data models: Web Annotation, NIF, CoNLL-RDF, POWLA and the original or conventional structure of the data. Moreover, the data structures of these formats are parallel, in parts, and then, a principled and consistent choice between, say, a oa:Annotation (from Web Annotation), a powla:Node (from POWLA), a nif:String and a nif:Annotation, has to be made.

Generally speaking, this situation is intractable, and thus, **the W3C Community Group Linked Data for Language Technology (LD4LT)** is currently engaged in a process to develop a harmonisation of these vocabularies. While this has been under development since about mid-2018, regular discussions via LD4LT only began in early 2020. Concrete results so far include a survey of requirements that any vocabulary for linguistic annotation on the web should have and the degree to which NIF, Web Annotation and other vocabularies support these at the moment.[134] So far, 51 requirements have been identified, clustered in 6 groups:

1. LLOD compliance (adherence to web standards, compatibility with community standards for linguistic annotation)
2. expressiveness (necessary data structures to represent and navigate linguistic annotations)
3. units of annotation (addressing primary data and annotations attached to it)
4. sequential data structures (preserving and navigating sequential order)
5. relations (annotated links between different units of annotation)
6. support for/requirements from specific applications and use cases (e.g., intertextual relations, linking with lexical resources, alignment, dialogue annotation).

So far, this is still work in progress, but if these challenges can indeed be resolved at some point in the future, and a coherent vocabulary for linguistic annotations emerge, we expect a similar rise in popularity for

[134] The survey can be accessed via https://github.com/ld4lt/linguistic-annotation/blob/master/survey/required-features.md, also compare the tabular view under https://github.com/ld4lt/linguistic-annotation/blob/master/survey/required-features-tab.md.

the adoption of the Linked Data paradigm for encoding linguistic annotations as we have seen in the last years for lexical resources. This latter was largely driven by the existence of a coherent and generic vocabulary, and indeed, the drift in applications that the OntoLex-Lemon model has recently experienced very much reflects the need for consistent, generic data models.

A question at this point may be what the general benefit of modelling annotations as linked data may be in comparison to other conventional solutions, and different user communities may have different answers to that. It does seem, though, that one potential killer application can be seen in the capacity to integrate, use and re-use pieces of information from different sources. A still largely unsolved problem in linguistic annotation is how to efficiently process standoff annotation, and indeed, the application of RDF and/or Linked Data has long been suggested as a possible solution [58, 61, 63, 66], but only recently, have systems that support RDF as an output format emerged [64]. While it is clear that standoff is a solution, it is also true that the different communities involved have not agreed on commonly used standards to encode and exchange their respective data. In DH and BioNLP, Web Annotation and JSON-LD seems to dominate; in knowledge extraction and language technology, NIF (serialised in JSON-LD or Turtle) seem to be more popular; for digital humanities, the TEI is currently revising XML standoff specifications,[135] and support for RDF serializations (RDFa) or standoff (Web Annotation in JSON-LD) also seems to be growing, as mentioned above.

5.3. Metadata

Summary In the first subsection of the current section, Section 5.3.1, we give an introduction and overview of metadata trends in LLD and other related areas. Next, we give a detailed description of two important metadata resources for LLD. These are META-SHARE, described in Section 5.3.2, and the OntoLex-Lemon *lime* module, described in Section 5.3.3. The latter section also features a discussion of future metadata challenges for LLD language resources. Finally, in Section 5.3.4 we address the ongoing challenge of language identification, which is an essential part of the metadata of a language resource.

[135] See https://github.com/TEIC/TEI/issues/1745 for pointers.

5.3.1. Introduction

The rise of data-driven approaches that use Machine Learning, and in particular recent breakthroughs in the field of Deep Learning, have secured a central place for data in all scientific and technological areas. Cross-disciplinary research has also boosted the sharing of data within and across different communities. Moreover, a huge volume of data has become available through various repositories, but also via aggregating catalogues, such as the European Open Science Cloud[136] and the Google dataset search service[137]. Metadata play an instrumental role in the discovery, interoperability and hence (re-)use of digital objects, and indeed act as an intermediary between consumers (humans and machines) and digital objects. For this reason, the FAIR principles [1] include specific recommendations for metadata (see also Section 1). Of particular relevance to this section is principle R1.3 which recommends that "(Meta)data meet domain-relevant community standards". According to this principle, the adoption of community standards or best practices for data archiving and sharing, including "documentation (metadata) following a common template and using common vocabulary" facilitates the re-use of data. In this section we therefore take a closer look at metadata models commonly used for language resources in the linguistics, digital humanities and language technology communities.

Although the focus of this section is on community models, we cannot leave the most popular general purpose models for dataset description out of this overview. Language is an essential part of human cognition and is thus present in all types of data; research on language and language-mediated research is carried out on data from all domains and human activities. All of this obviously extends the search space for data to catalogues other than the purely linguistic ones. The three models that currently dominate the description of datasets are DCAT[138], schema.org[139] and DataCite[140].

DCAT profiles are used in various open data catalogues, such as the EU Open Data portal[141], while schema.org is used for the Google dataset search engine; finally, DataCite, a leading provider of persistent identifiers (namely DOIs), has developed a schema

[136] https://www.eosc-portal.eu
[137] https://toolbox.google.com/datasetsearch
[138] https://www.w3.org/TR/vocab-dcat-2/
[139] https://schema.org/
[140] https://schema.datacite.org/
[141] https://data.europa.eu/euodp/en/data/

with a small set of core properties which have been selected for the accurate and consistent identification of resources for citation and retrieval purposes.

There are various initiatives for the collection of crosswalks of community-specific metadata models with these models[142], as well as recommendations for extensions for specific data types (e.g., CodeMeta[143] and Bioschemas[144] for source code software and life science resources respectively). Of course, these models are not intended to capture all the specificities required for the description of linguistic features and, thus, we do not go into further details on them in this paper.

Among models for the description of language resources in general (and not just LLD resources), the **Component Metadata Infrastructure** (CMDI) profiles [103, 104], and the TEI guidelines (introduced above) stand out. CMDI is a framework designed to describe and re-use metadata; "profiles" can be constructed on the basis of building blocks ("components") that group together semantically related metadata elements (e.g., address, identity, etc.) and can be used as ready-made templates catering for specific use cases (e.g., for lexica, for linguistic corpora, for audio corpora, etc.). CMDI profiles are used by various humanities and social sciences communities within the CLARIN[145] research infrastructure. The TEI standard specifies an encoding scheme for the representation of texts in digital form, chiefly in the humanities, social sciences and linguistics; it includes specific elements for the description of texts at both the collection and individual text levels. Both CMDI and TEI, however, are XSD-based[146], and therefore not discussed further in this section.

We should also mention the **CLARIN Concept Registry** (CCR)[147], which is a collection of linguistic concepts [106, 107]. It is the successor to the ISOcat data category registry (described in Section 4.5) and is currently maintained by CLARIN. The CCR is implemented in SKOS and includes a concept scheme for metadata, but this is a structured list without ontological relations, either internally or externally to other vocabularies. It mainly serves as the semantic interoperability layer of CLARIN; such interoperability is achieved by linking metadata fields included in CMDI profiles to concepts from the CCR.

[142] See, for instance, https://rd-alliance.github.io/Research-Metadata-Schemas-WG/

[143] https://codemeta.github.io/

[144] https://bioschemas.org/

[145] https://www.clarin.eu

[146] The conversion of CMDI metadata records offered in CLARIN into RDF [105] should not be confused with the construction of an RDF model for CMDI profiles

[147] https://concepts.clarin.eu/ccr/browser/

5.3.2. *Language Resource Metadata: The META-SHARE ontology*

The **META-SHARE**[148] model [108] (known as **MS-OWL** for short in its implementation as an OWL ontology) has been designed specifically for language resources, including data resources (structured or unstructured datasets, lexica, language models, etc.) and technologies used for language processing [109]. The first version of MS-OWL was (semi-)automatically created from the META-SHARE XSD schema [109, 110] (originally designed to support the META-SHARE infrastructure [111]) and developed within the framework of the LD4LT group mentioned above. The second version of MS-OWL, which is described here, evolved from the first version by taking into account advancements in the Language Technology domain and related metadata requirements (such as the necessity for the description of workflows, interoperability issues between language processing tools and processing resources, etc.) as well as current trends in the overall metadata landscape [108].

MS-OWL has been constructed by taking three key concepts into consideration: *resource type*, *media type* and *distribution*. These give rise to the following basic classes:

- LanguageResource, with four subclasses derived from the notion of resource type:
  - Corpus: for structured collections of pieces of language data, typically of considerable size and which have been selected according to criteria external to the data (e.g., size, language, domain, etc.) with the aim of representing as comprehensively as possible a specific object of study;
  - LexicalConceptualResource: covering resources such as term glossaries, word lists, semantic lexica, ontologies, etc., organised on the basis of lexical or conceptual units (lexical items, terms, concepts, phrases, etc.) along with supplementary information (e.g., grammatical, semantic, statistical information, etc.);
  - LanguageDescription: for resources which are intended to model a language or some aspect(s)

[148] http://w3id.org/meta-share/meta-share

of a language via a systematic documentation of linguistic structures; members of this class are typically: statistical and machine learning-computed language models and computational grammars;
  - ToolService: for any type of software that performs language processing and/or related operations (e.g., annotation, machine translation, speech recognition, speech-to-text synthesis, visualization of annotated datasets, training of corpora, etc.);

- MediaPart: this is a parent class for a number of other subclasses, combining the notions of resource and media type; it is not meant to be used directly in the description of language resources. The media type refers to the form/physical medium of a data resource (i.e., member of one of the first three subclasses under LanguageResource above) and it can take the values text, audio, image, or video. To cater for multimedia/multimodal language resources (e.g. a corpus of videos and their subtitles, or corpora of audio recordings and their transcripts), language resources are represented as *consisting* of at least one media part, e.g., the mediaPart property is used to link an instance of the class Corpus to instances of CorpusTextPart, CorpusAudioPart, and so on; similarly, LexicalConceptualResource is linked to LCRTextPart, LCRVideoPart, etc.
- DatasetDistribution and SoftwareDistribution: these are conceived as subclasses of dcat:Distribution, which represents the accessible form(s) of a resource. For instance, software resources may be distributed as web services, executable files or source code files, while data resources as PDF, CSV or plain text files or through a user interface.

MS-OWL caters for the description of the full lifecycle of language resources, from conception and creation to integration in applications and usage in projects as well as for recording relations with other resources (e.g., raw and annotated versions of corpora, tools used for their processing, models integrated in tools, etc.) and related/satellite entities[149].

The properties recommended for the description of language resources are assigned to the most relevant class. Thus, the LanguageResource class groups properties common to all resource/media types, such as those used for identification purposes (title, description, etc.), recording provenance (creation, publication dates, creators, providers, etc.), contact points, etc. More technical features and classification elements, that depend on resource/media types, as well as instances of MediaPart and Distribution are attached to the respective LanguageResource subclasses. Thus, properties for LexicalConceptualResource encode the subtype (e.g. computational lexicon, ontology, dictionary, etc.), and the contents of the resource (unit of description, types of accompanying linguistic and extralinguistic information, etc.); properties for Corpus include corpus subclass (raw, annotated corpus, annotations), and information on corpus contents. It should be noted that the language of the resource's contents, a piece of metadata of particular relevance to all language resources, is encoded in the media part subclasses rather than the top LanguageResouce class; this is in line with the principles adopted for the representation of multimedia/multimodal resources consisting of parts with different languages (e.g. a corpus of video recordings in one language, its subtitles in the same language and their translations in another language). Finally, the two distribution classes (DatasetDistribution and SoftwareDistribution) provide information on how to access the resource (i.e., how and where it can be accessed), technical features of the physical files (such as size, format, character encoding) and licensing terms and conditions. A dedicated module has been devised for the structured representation of licenses commonly used for language resources, reusing existing vocabularies and extending the Open Digital Rights Language[150] core model [112].

To better illustrate the structure of the MS-OWL, Figure 4 depicts a subset of the mandatory and recommended properties for the description of a corpus.

Amongst the additions made between the two versions of the MS ontology is the development of an additional vocabulary, again implemented as an OWL ontology, **OMTD-SHARE**[151] [113]. OMTD-SHARE can be considered to be complementary to MS-OWL. It covers *functions* (tasks performed by software components), *annotation types* (types of information extracted or annotated by such software), *methods* (clas-

[149] The current work discusses only the core part of MS-OWL targeting the description of language resources and leaves aside the representation of satellite entities (persons, organizations, projects, etc.)

[150] https://www.w3.org/ns/odrl/2/

[151] http://w3id.org/meta-share/omtd-share/

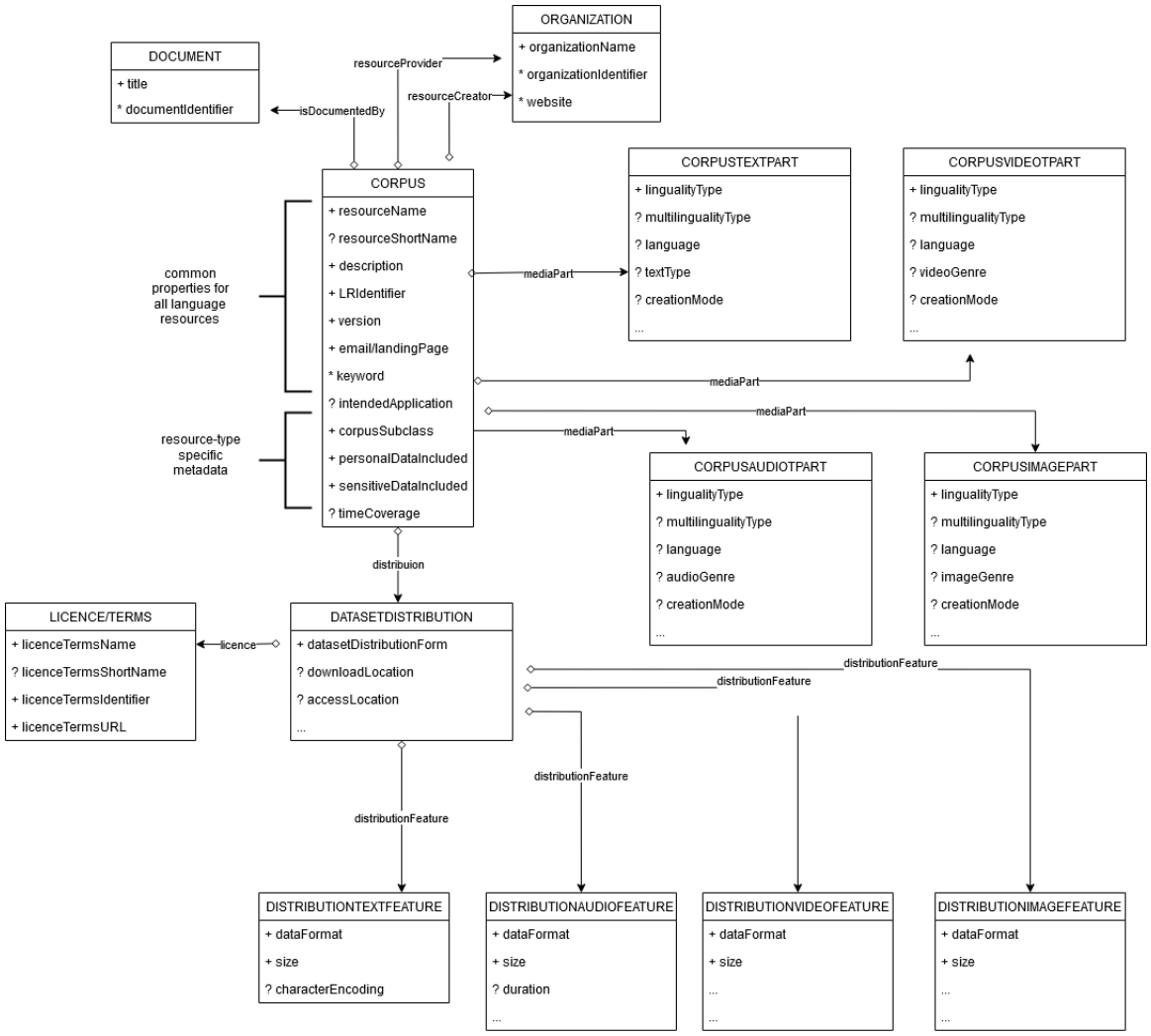

Fig. 4. Simplified subset of the MS-OWL for corpora.

sification of the theoretical method used in the algorithm), and *data formats* of the resources that can be processed by such software. The ontology was begun within the framework of the OpenMinTeD project[152], which focused on Text and Data Mining resources, and which has been enriched afterwards. The class Operation has been extended to cover Language Technology (LT) operations at large (now also referred to as "LT taxonomy"). Specific properties of MS-OWL make reference to the relevant OMTD-SHARE classes. Operation is used for describing the function of tools/services, as well as for applications for which a data resource can be used or has already been used. annotationType for annotated corpora takes values from the AnnotationType class; linguistic annotation types are linked to the OLIA ontology (work in progress), while domain-specific annotation types for neighbouring domains are also foreseen (e.g., for elements in the document structure of publications, biomedical entities, etc.).

Both the MS-OWL and OMTD-SHARE ontologies have been published and are currently undergoing evaluation and improvements. They are deployed in the description of language resources in catalogues of language resources. More specifically,

[152]https://www.openminted.eu

the first version of MS-OWL is used in **LingHub**[153], a data portal aggregating metadata records for language resources hosted in various repositories and catalogues [114, 115], while the second version, the one described here, is used in the European Language Grid[154], which is a platform for language resources with a focus on industry-relevant Language Technology in Europe [116]. Amongst the immediate plans, crosswalks with DCAT and schema.org are a priority, to ensure wider uptake and interoperability with (meta)data from other communities.

5.3.3. Linguistic Metadata for Lexical Resources: *lime*

Another metadata model that is deeply relevant to the current discussion is OntoLex-Lemon's own dedicated metadata module. The latter, in keeping with the overall citric theme, is called *lime*, which is short for the *LInguistic MEtadata* module [117][155]. A diagram for the module is given in Figure 5.

Before we go onto describing *lime* in more detail, it is worth pointing out that the module focuses on providing metadata descriptions at the "level of lexicon-ontology interface"[156]. That is, it concentrates on how ontological concepts in a so-called *reference dataset*[157] are *lexicalised* or given a linguistic grounding in a lexicon[158] (Figure 5 also makes reference to a ConceptSet which is defined in the OntoLex-Lemon guidelines as a set of individuals of class Lexical Concept described as potentially "bearing a conceptual backbone to a lexicon").

The aim of the *lime* module then is to provide quantitative and qualitative (metadata) information about the relations between the aforementioned kinds of resource. In other words many (though as we will see below not all) of its classes and properties will not apply in cases where OntoLex-Lemon is *only* used to encode a lexicon, and where entries and their senses aren't linked to either Lexical Concept individuals or to ontology entities (such as is true of an increasing number of lexicon-centric use cases, as we discuss elsewhere in the current article).

More generally, useful classes and properties include the lime:Lexicon class which is defined as a subclass of void:Dataset[159]) and represents a set of individuals of the class Lexical Entry which are related to lime:Lexicon via the property lime:entry. The whole lexicon, as well as individual entries, can be assigned to a certain language, as specified by the datatype property lime:language (the OntoLex-Lemon guidelines also recommend the use of the Dublin Core property and the use of either LexVo or Library of Congress language tags, see Section 5.3.4 for an extended discussion of both of these and of language tags in general). In addition, the property lime:linguisticCatalog specifies the linguistic model, i.e. the catalogue of linguistic categories used for the annotation of the lexical entries; this could be, for instance, LexInfo (see Section 4.5 above).

In order to show the use of these more general *lime* classes to relate a lexicon together with its entries, we will look at a very simple example taken from the W3C guidelines[160]. This can be seen in diagrammatic form in Figure 6. The diagram corresponds to the following listing.

```
:lexicon a lime:Lexicon;
  lime:language "en";
  lime:entry :lex_high;
  lime:entry :lex_cat;
  lime:entry :lex_marry;
  lime:entry :lex_intangible_assets.
] .
```

As the example demonstrates, *lime* properties and classes allow for the description of some of the most fundamental lexicon-specific metadata categories of a lexical resource. In addition, we of course use Dublin Core properties such as description and creator to further flesh out the metadata description of a lexical resource. We now look at some other classes in *lime*.

The lime:LexicalizationSet class (once again a subclass of void:Dataset) represents a collection of *lexicalizations*, each of which is a pair consisting of a lexical entry and an associated entry in the reference dataset (this might be an OWL ontology but could also be any "RDF dataset which contains references to objects of a domain of discourse"). The metadata properties associated with lime:LexicalizationSet enable us to describe, amongst other things[161]: *how many entities have been lexicalised (by at least one entry), how many pairs of entries and ontology elements there are, as well as how*

[153] http://linghub.org/

[154] https://live.european-language-grid.eu/

[155] The rest of this section assumes some familiarity with OntoLex-Lemon; an introduction to the model is given in Appendix A

[156] See https://www.w3.org/2016/05/ontolex/#metadata-lime

[157] Here defined as an ontology that describes "the semantics of the domain" [117]

[158] Here viewed as a collection of lexical entries

[159] See https://www.w3.org/TR/void/

[160] https://www.w3.org/2016/05/ontolex/#metadata-lime

[161] see https://www.w3.org/2016/05/ontolex/ for a full description.

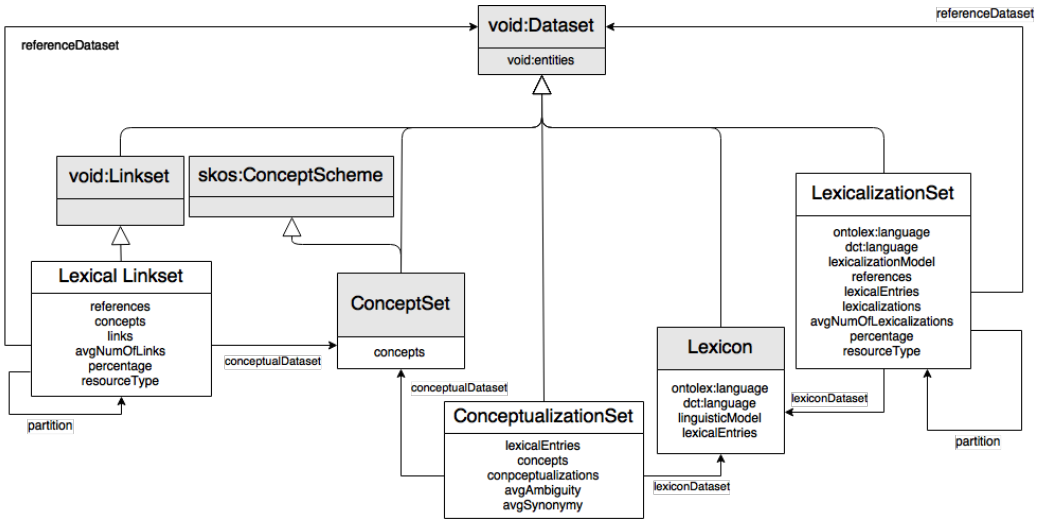

Fig. 5. The *lime* module.

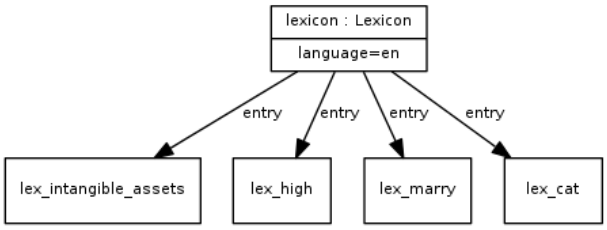

Fig. 6. *lime* Example (diagram taken from OntoLex-Lemon guidelines).

many ontology elements have been lexicalised on average.

lime also defines the class lime:LexicalLinkSet (a subclass of void:Dataset), individuals of which are links between a set of lexical concepts (i.e., members of the class ontolex:LexicalConcept) and the reference dataset. For this class, *lime* defines properties describing, for example, the number of links between the two resources in question.

Lastly, the lime:ConceptualizationSet class is analogous to lime:LexicalizationSet but describes the links between the lexicon and the concept set.

Metadata for Heterogeneous Use Cases

Language Resources are often complex informational objects and as such require their description requires the use of specialised vocabularies in addition to, and in combination with, the general LLD metadata vocabularies we have mentioned above. Take for instance the case of the publication of retrodigitised dictionaries and/or the modelling of historical and scholarly lexical resources as LLD. Here there is a need for extensive metadata provision at both the lexical and the individual entry levels in order, e.g., to encode historic and bibliographic information as well as to explicitly represent scholarly hypotheses as such (as in the case of etymologies, see for instance [56]). In addition, and as mentioned in Section 4.4 above metadata for retrodigitised resources should ideally feature information on the original physical work as mentioned.

In this case, we can use classes and properties belonging to a number of other vocabularies from outside the language resource/linguistic domain. These include the Semantic Publishing and Referencing suite of ontologies for bibliographic information[162], and the CIDOC-CRM family of ontologies for dealing with hybrid informational artefacts. The challenge lies in combining these vocabularies and others together with META-SHARE and *lime* in creating metadata so-

[162]http://www.sparontologies.net/

lutions, and potentially application profiles, each of which is targeted to an individual such kind of use-case. Here, the use of a top level ontology for integrating the disparate kinds of data together can be particularly useful. This however is still a fairly new area of research. A first proposal in dealing with the use-case of retrodigitised dictionaries and using CIDOC-CRM as a framework for bringing together different kinds of information in one complex hybrid object can be found in [118].

5.3.4. Language Identification

The reliable identification of languages and language varieties is of the utmost importance for language resources. For applications in linguistics and lexicography it defines the very scope of investigation of the data provided by a language resource; for applications in language technology and knowledge extraction, language identifiers define the suitability of training data or the applicability of a particular tool to the data at hand.

There are two different ways of encoding language identification information currently in use in RDF datasets. The first is via a URI-based mechanism that uses terminology repositories, the other is by attaching a language tag to a literal to indicate its language.

In the latter case, the language tag is treated similarly to a data type. Language information provided in this way does not entail an additional RDF statement and allows for a compact, readable and efficient identification of language information with minimal overhead on data modelling. Note that the original RDF specifications [119] already included provision for the use of language identification via the attachment of language tags to strings. In the former case , the URI-based mechanism, there exist a number of RDF vocabularies which provide the means to mark the language of a resource explicitly using RDF triples, i.e., using properties such as dc:language (for language URIs or string representations) or lime:language (for string representations). We elaborate on the differences in practice below.

RDF language codes are defined by BCP47[163] and the IANA[164] registry on the basis of the ISO 639 standard for language codes[165]. For application to RDF data, ISO provides three relevant subsets of language tags: **ISO 639-1**, maintained by the Library of Congress and available as plain text or RDF data,[166] provides an extensive set of two-letter codes for major languages that date back to the beginning of the modern-age of computing, but long before the emergence of the internet. ISO 639-1 codes are composed of two lower-case letters with values from *a* to z each. In theory, such a system is sufficient to identify up to 676 languages.

Yet, with language technology developing into a truly global phenomenon, it became clear that two-letter codes were not sufficient to reflect the linguistic diversity of the world both past and present – and in the present case this diversity is estimated to comprise more than 6,000 language varieties. As a response to this, **ISO 639-2** provides a set of three-letter codes for (theoretically) up to 17,576 languages. Again, the Library of Congress acts as maintainer and provides the data both in human-readable form and as RDF[167]. However, it should be pointed out that the primary use case for ISO 639-2 was a library-based one and focused on languages with an extensive literature, whereas the demands of linguistics and lexicography, especially historical linguistics and language documentation, exceed far beyond this. Indeed, they comprise languages that are primarily spoken, not written, but for which field recordings, text books, grammars or word lists must nevertheless be identifiable in order to be retrieved from metadata portals such as, as an example, the Open Language Archives Community (OLAC)[168].

For applications in linguistics, SIL International acts as maintainer of **ISO 639-3**, which is another, and more extensive, set of three-letter codes. In distinction to ISO 639-1/2 codes, which are meant to be stable and develop at a slow pace, if at all,[169] ISO 639-

[163] https://tools.ietf.org/rfc/bcp/bcp47.txt

[164] https://www.iana.org/

[165] The need for the provision of machine-readable identifiers for single languages or language varieties is clear from instances where a language has more than one name. For instance, the Manding language *Bamanakan* (bm) which is also known as *Bambara*. It is also essential for dealing with cases where the same language name is used to refer to what are quite different varieties. Take, for instance, the case of *Saxon* which as well as being an English heavy metal band has also been used to designate both Old English (Anglo-Saxon, ISO 639-3 ang) and a number of varieties of *Low German*, both historical and modern (Old Saxon, osx; Low Saxon, nds), along with various different dialects of *High German* (Upper Saxon, sxu; Transylvanian Saxon [currently no ISO language code]).

[166] https://id.loc.gov/vocabulary/iso639-1.html

[167] https://id.loc.gov/vocabulary/iso639-2

[168] http://www.language-archives.org/

[169] Changes in ISO 639-1 and 639-2 codes are very rare and occur mostly as a result of political changes, e.g., after the split of Yugoslavia, Serbian (sr, srp) and Croatian (hr, hrv) were to be considered independent languages (with two tags) whereas they were

3 codes are actively maintained by the research community and a continuous process of monitoring, approval (or rejection) of updates, additions and deprecation requests is in place. At the moment, ISO 639-3 codes are published by means of human-readable code tables only,[170] along with their history and associated documentation, but not in any machine-readable form. Within the LLOD community, it is a common practice to apply the ISO 639-3 codes provided as part of LexVo [37] whenever language URIs are required and ISO 639-3 codes are sufficient. However, it is to be noted that, unlike SIL code tables, LexVo identifiers are not authoritative and may not be up-to-date with the latest version of SIL.

But ISO 639-3 only represents the basis for language tags as specified by BCP47 [120, Best Common Practices 47, also referred to as IETF language tags or RFC 4646] as incorporated into the RDF specifications. BCP47 defines how ISO 639 language tags can be extended with information regarding geographical use, script, among other variables, as follows:

```
language(-script)(-region)(-variant)*
   (-extension)*(-x-privateuse)
```

where:

- **language**: this is an ISO 639-1 tag if this is available or an ISO 639-3 tag otherwise;
- **Script** (optional): an **ISO 15924** 4-letter code, for instance the code for Latin is `Latn`;
- **region** (optional): this is an **ISO 3166** 2-letter region code or a **UN M.49** 3-number) code, for instance either `US` or `840` for the United States of America
- **variant**: zero or more registered variants taken from the current list of registered variants provided by IANA[171].
- **extension**: zero or more extensions in one or more custom schemes
- **private use** (optional): for use for internal notes about identification in a single application.

The W3C provides means for validating BCP 47 language tags, part of the specification is also that language tags should be registered at the Internet Assigned Numbers Authority. The IANA language subtag registry[172] currently provides registered language tags in XML, HTML and plain text. As of 2020, discussions about the provision of a machine-readable view in RDF and by means of resolvable URIs are in progress and are expected to bear fruit in the coming years. We expect that, by then, the IANA registry will supersede LexVo as a default provider of ISO 639(-3) language URIs[173]. However, it should be noted that the very notion of language tags has been criticised as being both too inflexible and unable to address the needs of linguistics, e.g., recently by [121, 122], and alternatives are being explored [123].

URI-based language identification represents a natural alternative in such cases, as these are not tied to any single standardization body or maintainer, but allow the marking of both the respective organization or maintainer of the resource (as part of the namespace) and the individual language (in the local name). As a consequence, they would naturally support the shift from one provider to another, if this were required for a particular task.

Finally, another provider of language identifiers relevant to the current discussion is **Glottolog** [124],[174]. This is a repository of identifiers for language varieties with a specific focus on (although by no means restricted to) low-resource languages and with an eye to applications in linguistic typology and language documentation. Glottolog maintains an independent set of language variety identifiers accessible in human- and machine-readable (RDF) form via resolvable URIs, along with additional metadata, an associated bibliography along with data on the phylogenetic structure of specific varieties[175]. In order to avoid any of the unintended political connotations that inevitably arise from the use of the term 'language'[176], Glottolog uses the more neutral (though rather uglier) term *languoid*, where this latter is defined as a language variety about

previously considered dialects of a single language, Serbo-Croatian (language tag sh, deprecated in 2000).

[170] https://iso639-3.sil.org/

[171] https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry (accessed 10-07-2019)

[172] https://www.iana.org/assignments/lang-subtags-templates/lang-subtags-templates.xhtml

[173] Cf. https://github.com/w3c/i18n-discuss/issues/13 .

[174] https://glottolog.org/

[175] That is, Glottolog allows for the specification of the phylogenetic relationships between different varieties, specifying English, for instance, as a subconcept of the category 'Macro-English' (macr1271), which groups together Modern Standard English and a number of English Pidgins: and relating it in its turn to narrower subconcepts such as Indian English (`indi1255`) and New Zealand English (`newz1240`).

[176] Recall Max Weinreich's famous observation that "a language is a dialect with an army and a navy".

which, or in which, there is exists some kind of literature.

A Glottolog ID for a languoid, then, consists of a 4-letter alphabetic code followed by a 4-character numerical code; for instance, the Glottolog ID for standard English is `stan1293`. These codes are the basis of resolvable URIs, for instance http://glottolog.org/resource/languoid/id/stan1293, which once resolved provide links to other relevant resources such as ISO 639. Note that Glottolog maintains a certain bias towards endangered modern languages, and therefore remains rather sketchy for what concerns the historical dimension. Yet the popularity of Glottolog and the fact that it has already had a wide uptake beyond the language documentation community, and including on Wikipedia, would suggest that the provision of identifiers for historical varieties is (and pardon the pun) only a matter of time.

6. Projects

Summary In this section, we give an overview of a range of different projects that have had an impact (or which are currently having an impact) on the use and/or definition of LLD vocabularies; see Table 3 for a summary of the projects discussed in the section. In Section 6.1 we give a detailed overview of this topic; this overview includes a subsection on recent projects which combine LLD and DH (Section 6.1.1) and introduction and description of an LLD project matrix given as Figure 7 (Section 6.1.2). Next, we describe a series of selected projects in detail. These are (in order of appearance):

- LiODi (Section 6.2.1)
- POSTDATA (Section 6.2.2)
- ELEXIS (Section 6.2.3)
- Prêt-à-LLOD (Section 6.2.5)
- NexusLinguaram (Section 6.2.6)

6.1. An Overview

As mentioned in the introduction to this paper, we take the funding, at a transnational (including European), national, and regional level, of an ever-increasing number of projects in which LLD plays a key role as evidence of the success of the latter as a means of publishing language resources. These projects also offer us a crucial snapshot of the application of LLD models and vocabularies across different disciplines and use cases, as well as indicating where future challenges may lie. Therefore, in conjunction with an information gathering task being undertaken as part of the NexusLinguarum COST action (see Section 6.2.6), we decided to carry out a survey of research projects in which a significant part of the project was dedicated to making language resources available using linked data or which had LLD as one of its main themes.

The survey has so far been carried out via queries on **CORDIS**[177] and the **OpenAIRE explorer site**[178], as well as through a study of the literature and by soliciting input from other participants of the NexusLinguarum COST action.[179]

Our project survey also included an analysis of influential survey articles as well as anthologies dealing with linguistic linked data (such as [11, 12]) along with a study of the programs of the major conferences in the sector of language resources[180]. This may of course have, inadvertently, led us towards a natural selection bias in the project overview, namely, towards projects that tended to publish their results at these venues. Moreover, it should also be noted that since our most important sources of project information were the CORDIS and OpenAIRE project platforms, both of which have a severely limited coverage of national and non-European projects, we were also at a disadvantage with respect to information with regard to these categories of project. We were however able to partially compensate for this by information retrieved via the active consultation of our respective networks.

Based on this exploratory work we were able to make a number of observations. Probably the most important of these is that the effort towards the definition

[177] https://cordis.europa.eu/projects

[178] https://explore.openaire.eu/

[179] As part of the preparation for the survey, we set up a Wikipedia page on OntoLex, (https://en.wikipedia.org/wiki/OntoLex) and extended another Wikipedia page on Linguistic Linked Open Data (https://en.wikipedia.org/wiki/Linguistic_Linked_Open_Data. We also encouraged partners from our respective networks to contribute and extend those pages, especially with respect to applications of OntoLex-Lemon and LLOD in general. Information retrieved as part of this process was used to complement the survey described above.

[180] In particular, the Language Resource and Evaluation Conference (LREC) series and associated workshops as well as domain-specific events (workshops on Linked Data in Linguistics (LDL), conferences on Language, Data and Knowledge (LDK), lexicographic events such as EURALEX, ASIALEX, and GLOBALEX as well as the eLex series of electronic lexicography conferences, and associated workshops.

Summary			
Project Name	**Duration**	**Type**	**Coverage in Current Article**
EAGLES	1993-1995	European Project (FP3)	Section 6.1
ISLE	2000-2002	European Project (FP5)	Section 6.1
E-MELD	2007-2012	American National Project (NSF)	Section 6.1.2
MONNET	2010-2013	European Project (FP7)	Section 6.1
SemaGrow	2012-2015	European Project (FP7)	Section 6.1
CLLD	2013-2016	German Project (Max Planck)	Section 5.2
LIDER	2013-2015	European Project (FP7)	Section 6.1
QTLeap	2013-2016	European Project (H2020)	Section 6.1
TDWM	2014-2029	German Regional Project	Section 6.1.1
FREME	2015-2017	European Project (H2020)	Section 6.1
LiODi	2015-2022	German Project	Section 6.2.1
Lynx	2017-2021	European Project (H2020)	Section 6.1
DiTMAO	2016-2019	German-Italian (funded by Deutsche Forschungsgemeinschaft (DFG))	Section 6.1.1
POSTDATA	2016-2022	European Project (H2020-ERC)	Section 6.2.2
MTAAC	2017-2020	International (funding from DFG, SSHRC and NEH)	Section 6.1.1
Nénufar	2017-	French Project (mixed funds)	Section 6.1
ELEXIS	2018-2022	European Project (H2020-ERC)	Section 6.2.3
LiLa	2018-2023	European Project (H2020-ERC-CoG)	Section 6.2.4
Prêt-à-LLOD	2019-2022	European Project (H2020-ERC)	Section 6.2.5
NexusLinguaram	2019-2023	EU Cost Action	Section 6.2.6
ItAnt	2020-2023	Italian National Project (PRIN)	Section 6.1.1
MORdigital	2021-2024	Portuguese National Project	Section 6.1.1

Table 3

Projects Discussed in the Current Article

of common models for linguistic linked data has never been dependent on any single, large-scale project, but has largely conducted within the confines of a much broader community: a broader community whose initiatives and activities did however overlap with a number of funded projects, often carried out in parallel. Over and above this, the community was also maintained by other kinds of networks and initiatives. What also came through quite strongly, however, both from the research carried out as part of the survey and from the authors' personal experiences is that international (and especially European level) projects played a crucial role in **supporting and sustaining** LLD models and vocabularies, *after they had already been proposed*. This can be demonstrated by looking at the development history of OntoLex-Lemon, probably the most popular of the LLD models featured in this article.

The original inspiration of this model can ultimately be traced back to the Lexical Markup Framework (LMF) [125], a conceptual Uniform Markup Language (UML)-based model[181] for representing NLP-lexica and machine-readable dictionaries. LMF was developed over the course of a number of projects which carried out early, pioneering work on lexical resources in NLP and related use cases, the most notable of these being **Expert Advisory Group on Language Engineering Standards (EAGLES, 1993-1995)**[182], and **International Standards for Language Engineering (ISLE, 2000-2002)**[183]. LMF then underwent further development within the ISO committee ISO TC37 and indeed the standard continues to be developed under the auspices of this organisation to this day, with the

[181] LMF also had an official XML serialization was included as part of the standard. Attempts towards a RDF/OWL serialization were made by Gil Francopoulo and can be found linked under http://www.lexicalmarkupframework.org/, but have not been otherwise published.

[182] http://www.ilc.cnr.it/EAGLES/home.html

[183] http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm

latest version of LMF being a multi-part standard, serialised in TEI, of which the first five parts have been published at the time of writing [126]. The **Multilingual Ontologies for Networked Knowledge (MONNET, 2010-2013)**[184] project subsequently proposed the original *lemon* model largely on the (conceptual) basis of LMF, in order to meet the need for a RDF-based model for giving lexical grounding to Semantic Web ontologies. In 2011, MONNET project members initiated the formation of W3C Community Group Ontology-Lexica. This led to OntoLex-Lemon, a revision of the original *lemon* model, whose development which was initially carried out within the ambit of this community group. OntoLex-Lemon was then further developed in the **LIDER project (2013-2015)**.[185] The latter project contributed to the founding of numerous W3C community groups as a means of assuring the long-term sustainability of it its activities. As far as lexical resources are concerned, this included the a W3C Community Group on **Best Practices for Multilingual Linked Open Data (BP-MLOD)** which, among other contributions, developed guidelines for the application of OntoLex-Lemon for modelling lexical resources (dictionaries and terminologies) independently from ontologies. This latter represents the basis for most modern uses of OntoLex-Lemon, and its development towards a general-purpose community standard for publishing lexical resources on the Semantic Web.

Monnet and LIDER were seminal in their impact on the development of LLD models and vocabularies. Other important (European) projects in this regard include the FP7 project **Eurosentiment**[186] which leveraged *lemon* to model language resources for sentiment analysis; **FREME**[187] which explored the application of the NIF and *lemon*; and **SemaGrow**[188] which, along with the LIDER project, helped to support the development of the *lime* metadata module).

Additional projects with a significant recent impact on the application of LLD vocabularies include: the Horizon 2020 project **Lynx: Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe** (2017-2021) [127] for its use of NIF for the annotations produced by NLP service.and the project **Quality Translation by Deep Language Engineering Approaches (QTLeap)**, (2013-2016) which has primarily focused on Natural Language Processing. Due to their more recent impact on the definition and use of LLD models and vocabularies we will dedicate specific sections to the following European H2020 projects **ELEXIS** (Section 6.2.3), **Prêt-à-LLOD** (Section 6.2.5), and the ERC projects **LiLa** (Section 6.2.4) and **POSTDATA** (Section 6.2.2) below.

6.1.1. Recent Projects combining LLD and DH

The projects which we describe in this section, along with ELEXIS, LiLa and POSTDATA described in their own sections below, are notable for bringing together DH and LLD. As is so often the case with DH projects, they aim to engage with a wide and diverse scholarly community, which includes linguists, philologists, historians, and archaeologists; in the case of the Classics (the case of LiLa in particular Section 6.2.4), there is also a reliance on, and a necessity to engage with, an extensive tradition of past scholarship. However by making it easy to structure data in a way which highlights different kinds of relationships both within and between different past civilisations, their languages and cultures, LLD offers a powerful and effective solution to the challenges of modelling heterogeneous humanities data, making it both findable and interoperable. In particular LLD is well placed to facilitate the integration of historical and geographical with lexicographic and linguistic information as the use of linked data in DH projects such as Pelagios [84], Mapping Manuscript Migrations [128] and in the Finnish Sampo datasets [129], among others, very clearly demonstrates. In the rest of this section we will provide summaries of a number of small and medium scale projects that are at the overlap of LLD and DH.

At a national level, we can list the French project **Nénufar**, already mentioned above, which aims towards the creation of successive early 20th century editions of the French language **Le Petit Larousse Illustré** dictionary in both TEI/XML and in RDF using OntoLex-Lemon [130][189], along with the German project **Linked Open Dictionaries** which is described in detail in Section 6.2.1. In addition we can also mention the Italian project (part of the Progetti di Rilevante Interesse Nazionale or PRIN program) **Languages and Cultures of Ancient Italy. Histori-**

[184] https://cordis.europa.eu/project/id/248458
[185] http://lider-project.eu/lider-project.eu/index.html
[186] https://cordis.europa.eu/project/id/296277
[187] https://cordis.europa.eu/project/id/644771
[188] https://cordis.europa.eu/project/id/318497

[189] Despite the best of intentions however the RDF part isn't currently very well developed.

cal Linguistics and Digital Models (ItAnt) (currently ongoing) which aims to publish a linked data lexicon of the ancient Italic languages[190] using the OntoLex-Lemon model and its extensions. Also relevant here is the Italo-German project **DiTMAO**, funded by the DFG (Deutsche Forschungsgemeinschaft), (completed) which produced a lexicon of Old Occitan medical terminology for which it also proposed an extension of *lemon*[191] [131]. Yet another national project worth mentioning here is the recently initiated Portuguese project **MORdigital** [132] which has the aim of digitising the historically significant 18th-century Portuguese language dictionary *O Diccionario da Lingua Portugueza* by António de Morais Silva with the intention of producing digital editions of this important lexicographic work both in TEI Lex-0 and OntoLex-Lemon. The MORdigital project will be an important test case for understanding both the coverage of already existing LLD vocabularies when it comes to retrodigitized dictionaries, and the advantages and disadvantages of using linked data as a means of publishing such data when compared with TEI.

Many of the projects we have mentioned have used OntoLex-Lemon or its predecessor *lemon*. However, a lightweight alternative to these vocabularies, and one which enables for the multilingual annotation of conceptual hierarchies, is SKOS-XL. This latter has been used, for instance, in several related projects at the Computational Linguistics lab of the University of Saarland, as part of a major effort towards the transformation of a number of influential classification schemes in the field of folk literature[192] (including among others folktales, ballads, myths, fables) into Semantic Web representation languages in order to support interoperability between those schemes; see [136]. The terms used in the original classification schemes were transformed into (multilingual) SKOS-XL labels; these were then used for encoding folktale text sequences, extracted from a manually annotated multilingual folktale corpus having been identified as representing motifs listed in [134]. The use of SKOS-XL meant that motifs could be annotated in different multilingual versions of tales.

Another project worth mentioning here, and one that also uses a range of different (L)LD vocabularies is **Text Database and Dictionary of Classic Mayan (TDWM)**[193] (2014-2029). TDWM aims to develop a corpus-based dictionary of Mayan hieroglyphic writing alongside a near-exhaustive corpus of Classic Mayan something which would allow for the verification of different textual interpretations and aid in arriving at the complete decipherment of Maya writing. The project faces the problem, typical of ancient languages, of the necessity of representing multiple interpretations of characters and texts(in part due to damaged sources) and in concomitance with the need to update this data with the inclusion of new data during dictionary development; in the case of the Mayan the situation is even more difficult due to the signs not yet having been fully deciphered. In order then to deal with the challenges which arise from the existence of different sign catalogues (which might cluster different signs into meanings differently) and the necessity of linking with other catalogues which have been developed in the field, the project's sign catalogue has been formalised in SKOS in addition to using properties and concepts from the CIDOC-CRM vocabulary[194] and GOLD (mentioned above in Section 4.5). The TDWM project is also developing its own vocabulary for identifying signs, linking them to different sign catalogues, possible readings, graphical variants, etc. At the time of writing, neither the sign catalog nor any texts are publicly available, but Diehr et al. [137] provide a detailed description[195].

Finally, another recent project which exploited a range of different LLD vocabularies is **Machine Translation and Automated Analysis of Cuneiform Languages (MTAAC)**. This was a *Data international* funded project which saw the collaboration of specialists of cuneiform languages and computational lin-

[190] https://www.prin-italia-antica.unifi.it/

[191] https://www.uni-goettingen.de/en/ditmao/487498.html

[192] The classification schemes in question were those proposed by Vladimir Propp [133], Stith Thompson [134] and Anti Aarne, Stith Thompson and Hans-J. Uther [135].

[193] Based at the University of Bonn, Germany.

[194] http://www.cidoc-crm.org/

[195] It is interesting to note that TDWM stands in a longer tradition of projects in the Digital Humanities that aim to complement a TEI/XML edition with terminology management using an ontology. Similar ideas have already been driving force behind the project *Sharing Ancient Wisdoms (SAWS, 2010-2013)*(http://www.ancientwisdoms.ac.uk/), a joint project at King's College London, UK, the Newman Institute in Uppsala, Sweden, and the University of Vienna, Austria, funded in the context of the Humanities in the European Research Area (HERA) program to facilitate the study and electronic edition of ancient wisdom literature. Both projects employ resolvable URIs, but the linking is expressed by means of narrowly defined TEI/XML attributes rather in terms of RDF semantics. In that regard, the data published in accordance with these guidelines does not qualify as Linked Data, but can still be converted to Linked Data with moderate effort.

guists in the development of cutting edge tools for the annotation and distribution of linguistic data relating to the cuneiform corpus. Although the project's overall objective was to open the way to the development of tools and the production of richer linguistic data for all cuneiform languages, its specific focus was on a group of unannotated Sumerian texts issued from the bureaucratic apparatus of the Ur III period (21th century BC)[196]; in addition, another corpus composed of royal inscriptions in the Sumerian language [138], annotated with morphology, was also employed. Amongst the several objectives of the project [139] was the aim of formalising the new data produced by the project by utilising (L)LOD vocabularies, and fostering the practices/technologies of standardisation, open data and LOD as integral to projects in digital humanities and computational philology. The project, which ended in 2020 made significant headway towards these aims, including making new data in the form of linguistic annotations and translations available under open licenses[197]; this will soon be accessible through the new web platform of the Cuneiform Digital Library Initiative (CDLI https://cdli.ucla.edu) in many forms, including (L)LOD.

CoNLL was chosen, due to its flexibility and robustness, as the storage format for the multi-layer annotations which were produced and worked on as part of the project.[198] However CoNLL-RDF was also employed in the project in order to ensure integration with LLOD, as well as for easier querying and transformation, and was used to link annotations, lexical information, and metadata. The ETCSRI morphological annotations[199] were mapped to Unimorph[200] using Turtle-RDF[201], rendering Sumerian material accessible for cross-linguistic queries. SPARQL was leveraged through CoNLL-RDF for syntactic annotation which was mapped to Universal Dependencies for POS and dependency labels. Lexical data was linked to guide word entries through the employment of an OntoLex-Lemon compliant index. Metadata concerning the analysis of the medium of the text and other meta classifications of the texts were mapped to the CIDOC-CRM. Overall, MTAAC succeeded in preparing a (L)LOD edition and linking of Sumerian language corpora. The model can be extended in part to other cuneiform languages. Various Assyriological resources had been integrated using (L)LOD [100]: The CDLI data, (CoNLL-RDF plus CIDOC-CRM), ORACC:ETSCRI (by conversion; CoNLL-RDF), ePSD (by conversion and links to HTML; lemon) and ModRef & BM (by federation; CIDOC-CRM). Other vocabularies are planned to be added in the future (Pleiades, perio.do, etc.). The model developed is currently being integrated into the CDLI platform.

[196] These texts were extracted from CDLI

[197] https://gitlab.com/cdli/framework; https://github.com/cdli-gh

[198] A derivative internal format, called CDLI-CoNLL is employed to store the data locally – this was an essential step to support the preservation of domain specific annotation which are richer than their counterparts found in linguistic all-encompassing models. But this can be exported in CoNLL-U format, as well in Brat Standalone format, for better compatibility.

[199] http://oracc.museum.upenn.edu/etcsri/parsing/index.html

[200] http://unimorph.org/.

[201] https://github.com/cdli-gh/mtaac_work/blob/master/lod/annotations/um-link.ttl.

6.1.2. An LLD Project Matrix; The Relationship between Projects and Community Initiatives

Figure 7 provides an overview in the form of a matrix of the contribution made by various different funded projects to a number of LLD vocabularies. We distinguish three kinds of contribution: namely, a project is said to have:

developed (deep green) a vocabulary if the development of that vocabulary was a designated project goal,

contributed (light green) to a standard if vocabulary development was not a designated project goal, but the project provided a use case or application that was discussed in the process of its development,

used (yellow) a vocabulary if they applied an existing vocabulary, worked with or produced data of that type

Note that this survey, and indeed any survey which focuses on projects, will provide a partial view only. In particular, contributions by community groups are not explicitly covered in this section (although they are described in some depth in Section 5 and their contribution is also discussed in Sections 2.2 and 6.1). For instance the reader will notice that very few of the projects in Fig. 7 address the area of LLD for linguistic typology. In fact the interaction between linguistic typology and language technology operates primarily on the basis of informal contacts on mailing lists and via workshops and less in terms of large-scale infrastructural projects, and that, thus, the development of standard (computational) models and vocabularies

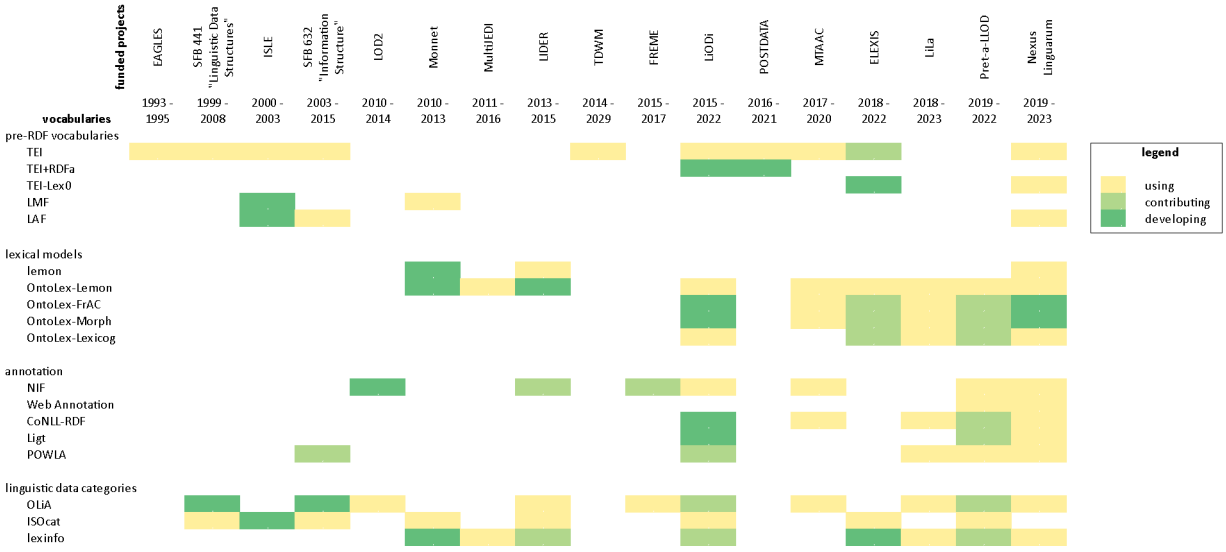

Fig. 7. Usage of and contribution to major LLOD vocabularies by selected research projects

has only rarely a priority in typological projects[202]. For such discussions, these more informal networks present critical opportunities (and act as a driver) for experts to participate in the Linguistic Linked Open Data movement, whereas the chances for acquiring substantial funding directed towards vocabulary development and community participation are rather unreliable (if the past experience of the authors is anything to go).

Note also that in this section, we have concentrated on research projects with a specific focus on linguistic linked (open) data – several of them, indeed, featuring the involvement of industrial partners – but which do not, for the most part, directly target industrial applications. More industry-focused LLD projects do exist, however, and are the basis for businesses specialising in text analytics [140], terminology and knowledge management [141] or lexicography [142]. But linked data in these contexts tends to be viewed as a technical facet that has an impact on interoperability, (re)usability and information aggregation rather than being fundamental for to the existing business model. With the increasing maturity of the technology, however, this may change over the longer term, especially in the area of establishing interoperability between AI platforms [143], their providers and users and data provided and exchanged between them [144].

To conclude then, it really has been the *combination* of open community initiatives and projects that determined the success and then the subsequent maintenance of the LLD models and vocabularies. The importance of funded projects is clear for the development of tools and hosting solutions for Linguistic Linked (Open) Data which are not yet in place; open community initiatives have also proven themselves vital for dissemination and wider community engagement. With the increasing maturity of OntoLex-Lemon and the convergence between existing solutions in linguistic annotation, the necessary requirements for developing large-scale Linguistic Linked (Open) Data infrastructures and their respective linking are in place, now. Note that in Section 7.1.1 below we take a brief look at the prospects for the involvement of research infrastructures in the kinds of initiatives mentioned in this section.

In what follows we will give extended descriptions of six ongoing projects. We have chosen these projects

[202] There are notable exceptions here the **E-MELD project** (http://emeld.org/), for example, developed the GOLD ontology as part of an attempt to improve interoperability and sustainability of language documentation material. But while several typological projects developed ontologies and RDF vocabularies, and have been actively contributing to the community, esp., in the Open Linguistics working group, we see a very limited degree of linking between such resources. The **Cross-Linguistic Linked Data project (CLLD, https://clld.org/)**, for example, does provide an RDF view on their data, but linking is primarily internal, and neither complete data dumps nor a SPARQL end point or any form of an API is provided. Instead, their RDF data seems to be generated on the fly, without any links to external resources. We take this to reflect the fact that for this community, interoperability is a priority, but also, to maintain control over internal data and independence from external contributions.

on the basis of their importance in the development of well known LLD models and vocabularies and/or in their innovative use of such. These are **LiODi** in Section 6.2.1; **POSTDATA**, in Section 6.2.2; **Prêt-à-LLOD** in Section 6.2.5; **ELEXIS** in Section 6.2.3 and finally **NexusLinguarum** in Section 6.2.6. Please note that the length of the following project descriptions will vary on the basis of their relevance to the models and vocabularies discussed in the rest of this paper.

6.2. Innovative Projects

6.2.1. LiODi (2015-2022)

The **Linked Open Dictionaries project (LiODi)**[203] aims to develop LLOD-enabled methodologies and infrastructures to facilitate language research for low-resource languages, validating these developments for the most part on the languages of the Caucasus. As part of the project, a set of loosely connected tools are being created with the aim of facilitating language contact studies over lexical and corpus data. One of the primary development goals of the project is the creation of an environment for detecting semantically and phonologically similar words across different languages as a means of facilitating the detection of possible cognates. Other tools include interfaces for converting, validating, and exploring linguistic data to aid in linguistic research both within and outside of the project. Tool development and linguistic research are both integral parts of LiODi and the tools and pipelines implemented within it are also tested on the data generated and used in the project [100, 145].

The most important contributions of LiODi from a modelling perspective relate to the fact that its members have developed, and are in the course of developing, LLD vocabularies for a wide-range of applications in the language sciences: in particular, vocabularies with an emphasis on the requirements of low-resource languages and especially morphologically rich languages which have so far not been well served by existing formats. These vocabularies include individual, task-specific vocabularies such as Ligt and CoNLL-RDF (see 5.2.4), but also an extension of OntoLex-Lemon for diachronic relations (cognate and loan relations) [46]. In addition to that, the LiODi project (along with Prêt-à-LLOD, see 6.2.5) is the main contributor to the **ACoLi Dictionary Graph** [36][204] which, at the time of writing and to the best of our knowledge, represents the most extensive collection of machine-readable bilingual open source dictionaries available: it currently features more than 3000 substantial data sets for more than 430 ISO 639-3 languages (including full OntoLex-Lemon editions of **PanLex**[205], **Apertium**[206], **FreeDict**[207], **MUSE**[208], **Wikidata**[209], **the Open Multilingual WordNets**[210], **the Intercontinental Dictionary Series**, **XDXF**[211] and **StarDict**[212] (the latter only to the extent that the copyright could be clarified and an open license was confirmed).

More significant than lexical resources and novel vocabularies, however, are the contributions of LiODi to the development of community standards for LLD vocabularies. This includes, among other aspects, significant contributions to the emerging OntoLex-Lemon Morphology module (Section 5.1.2), initiating and moderating the development of the OntoLex-Lemon FrAC module (Section 5.1.3) and the LD4LT initiative on harmonizing vocabularies for linguistic annotation on the web.

Furthermore, LiODi has a strong commitment to the dissemination and promotion of linked data approaches to linguistics. As a demonstration of this, the project co-organised two summer schools, **SD-LLOD 2017** and **SD-LLOD 2019**; two conferences **LDK 2017** and **LDK 2019**; three workshops **LDL 2016**, **LDL 2018**, and **LDL 2020**; and collaborated with international partners and the Prêt-à-LLOD project (see Section 6.2.5) in the publication of the first monograph on the topic [11] along with a number of edited volumes (not counting the five volumes of proceedings which resulted from the aforementioned events, including a collection on linked data for collaborative, data-intense research in the language sciences [12]).

Outside of conjoined activities at summer schools and datathons, the project supports numerous external partners in expertise with data modelling and language resource management. Indeed LiODi has close ties with most of the projects listed here. To mention one notable example here, a collaboration with the POSTDATA project (see the next section) and the Academy of Sciences in Heidelberg, Germany, led to the first

[203] https://acoli-repo.github.io/liodi/
[204] https://github.com/acoli-repo/acoli-dicts
[205] https://panlex.org/
[206] https://www.apertium.org/
[207] https://freedict.org
[208] https://github.com/facebookresearch/MUSE
[209] https://www.wikidata.org/
[210] http://compling.hss.ntu.edu.sg/omw/
[211] https://sourceforge.net/projects/xdxf/
[212] http://stardict.sourceforge.net/

practical applications of RDFa within TEI editions in the Digital Humanities [69, 146], and ultimately to the development of an official TEI+RDFa customization (see above).

6.2.2. POSTDATA (2016-2021)

The **Poetry Standardization and Linked Open Data (POSTDATA)** project[213], seeks to bridge the digital gap between traditional cultural assets and the growing sophistication of data modelling and publication practices in the field of the Digital Humanities. It focuses on poetry analysis, bringing Semantic Web standards and technologies to bear on a variety of different poetry-related resources. The project is founded upon two central pillars. The first is the use of linked open data; in fact one of the key aims of the project is to share scholarly knowledge about the domain of poetry and publish literary works on the linked open data cloud. And the second is the implementation and utilisation of a set of dedicated Natural Language Processing (NLP) tools, **PoetryLAb**.

As part of its focus on the Semantic Web, POSTDATA is developing a poetry ontology in OWL. This ontology is based on the analysis and comparison of different data structures and metadata arising from eighteen projects and databases devoted to poetry in different languages at the European level [147–150]. The POSTDATA ontology is *an encapsulated ontology model*, where domain knowledge is implemented in 3 layers: **Postdata-core**, **Postdata metrical and literary analysis** and **Postdata-transmission**. It re-uses other ontologies relevant to the project's domain of interest and covers different levels of description from the abstract concept of the poetry work to its bibliographic representation [151–155]. The model is intended to support tasks associated with the analysis of poetry and which fall under the categories of close reading, distant reading or critical analysis. All of these ontologies will be exposed via SPARQL endpoints.

The POSTDATA metrical layer encapsulates knowledge pertaining to the poetical structure and prosody of a poem by making use of the salient (general) linguistic, phonetic and metrical concepts. From the metrical point of view, a poem is formed by *stanzas* that contain *lines*, where individuals of the latter category are understood as a list of *words*. Although the concept of *word* is present in OntoLex-Lemon and NIF, in both of these cases its definition is insufficient for capturing all of the knowledge needed for the analysis and description of a word from a metrical point of view. Indeed according to this latter the concept *word* should be associated both with more general linguistic information (such as its *lemma*) – information which is captured by the former models – as well as more specific phonetic features such as *syllable, foot, feet type onset or coda* along other types of metrical information. This led to the definition of a class Word in the POSTDATA metrical ontology. However, the intention is to link this class with the OntoLex-Lemon class Word through the property wordsense, allowing us to capture the range of meanings of the concept. Furthermore, the POSTDATA Word class will also be linked to the NIF Word class due to the shared relationship of both of them to NLP operations.

The second pillar of POSTDATA, the use of NLP tools, is represented by PoetryLab[214], encompasses the several different levels of poetry scholarship, from the most formal analyses relating to scansion, to more cognitive levels which concern the understanding of metaphor as well as others related to knowledge and subjective perception involving AI techniques. POSTDATA has already implemented the first level of NLP algorithms for poem analysis. These allow for the automated extraction of information from poems at different levels of description and include an Name-Entity Recognition system (NER) for medieval place names and organizations, [156] as well as automatic enjambment analysis and basic metrical scansion tools (which allow for lexical syllabification and the recognition of stressed and unstressed syllables) testing different approaches. These latter range from traditional ruled-based systems to the latest deep learning based techniques, [157–159]. The goal in this case is to use the results of these tools in order to build an RDF knowledge graph that is compliant with Postdada ontology.

6.2.3. ELEXIS (2018-2022)

A follow-up to the European Network for e-Lexicography COST Action[215], the ELEXIS project is in the process of undertaking the construction of a European infrastructure for electronic lexicography [160]. LLD will play a key role in this infrastructure, namely, as means of connecting dictionaries and other lexicographic resources both within and across language boundaries. In fact, the idea of ELEXIS is to eventually construct a network of interlinked electronic lexica and other lexicographic and language resources in several lan-

[213] http://postdata.linhd.uned.es

[214] http://postdata.uned.es/poetrylab/

[215] https://www.elexicography.eu/

guages, a network that the project calls a **Matrix Dictionary**. Another relevant aspect of the project concerns the conversion of legacy lexicographic resources into structured data, and potentially, linked data in order to feed into the Matrix Dictionary.

The main models being used in the project are OntoLex-Lemon and the TEI Lex-0 model mentioned above [161]. And here it will perhaps be useful to give a brief description of the latter.

TEI Lex-0 and ELEXIS TEI Lex-0 is a custmomization of the TEI schema[216] adapted to the encoding of lexical resources. More precisely, it was designed to enhance the interoperability of such datasets by limiting the range of encoding possibilities (offered by the current TEI guidelines) in the representation of lexical content (for instance, TEI Lex-0 has deprecated elements such as `superEntry` or `entryFree`). This makes the possibility of a crosswalk from (at least a subset of) TEI Lex-0 to OntoLex-Lemon more feasible than, say, a crosswalk from say, a minimal customisation of TEI based on the TEI dictionary guidelines to OntoLex-Lemon[217].

TEI Lex-0 is being developed by a special working group which (pre-Covid) organised regular in-person training schools with support from ELEXIS. Both OntoLex-Lemon and TEI Lex-0 have been previously used for smaller lexicography projects, but never in a project with such wide coverage in terms of the languages and kinds of lexicographic resource under consideration. ELEXIS has provided support to the development of both OntoLex-Lemon as well as TEI Lex-0 and a joint workshop was held between these projects at the 2019 edition of the e-lexicography convention eLex.

The project is also promoting the standardisation of OntoLex-Lemon and TEI Lex-0 through the OASIS working group on Lexicographic Infrastructure Data Model and API (LEXIDMA)[218], where the intention is to produce a new unifying standard for lexicographic data that will be serialised in both OntoLex-Lemon and TEI Lex-0.

The Impact of ELEXIS on the Use of OntoLex-Lemon ELEXIS aims to provide support for the creation and editing of dictionary resources using OntoLex-Lemon. To this end extensive teaching materials are also being developed as part of the project with the aim of introducing lexicographers to linked data and the OntoLex-Lemon model. It should be noted that the availability of manuals and targeted teaching materials plays an important factor in increasing the uptake of models such as OntoLex-Lemon and technologies such as linked data, (as of course is the case with new technologies and new technological approaches in general), especially amongst users who haven't had much previous exposure to linked data or conceptual modelling. The original designers of such models are usually unable to take into consideration of every kind of use-case for which the model might be used. Such targeted training materials can help to bridge the gap between a general purpose model as it is presented in some final set of guidelines, and its use or appropriation (along with other pertinent models and vocabularies) in a specialist domain or task (the use of design patterns can also help in this respect, see Section 7.1.2). This is also one of the motivations behind the strong emphasis on training in NexusLinguaram (see Section 6.2.6).

Both the production of training materials and the push to promote OntoLex-Lemon as a common serialisation format for a standard for e-lexicography seems to promise much in terms of the future use of linked data in this domain. It is inevitable that the experiences of lexicographers and linguists in using OntoLex-Lemon (and its lexicographic extension, see Section 5.1.1) both within and outside of the ELEXIS project to create and edit lexicographic resources will have an important impact on the use of the model and also, potentially, on future extensions and/or versions of OntoLex-Lemon.

6.2.4. LiLa (2018-2023)

The **LiLa: Linking Latin** ERC project[219] aims to connect language resources developed for the study of Latin, bringing the worlds of textual corpora, digital libraries, lexica and tools for NLP together. To this end, LiLa makes use of the LOD paradigm and of a set of ontologies from the LLOD cloud to build an interoperable network of resources. LiLa's ambition is to create an infrastructure where researchers and students of Latin can find answers to complex questions

[216] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

[217] Work is also underway on a crosswalk between TEI Lex-0 and OntoLex-Lemon. The latest version of a proposed TEI Lex-0 to OntoLex-Lemon converter can be found at https://github.com/elexis-eu/tei2ontolex.

[218] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=lexidma

[219] http://lila-erc.eu

that involve multiple layers of linguistic annotations and knowledge, such as: *what subjects are constructed with verbs formed with a certain prefix* [162]? and *What WordNet synsets do they belong to?*

As Latin is characterized by a very rich morphology (where, for instance, a single verb can potentially yield more than 100 forms, excluding the nominal inflection of participles), LiLa focuses on lemmatization as the key task that allows for a meaningful and functional connection between the different layers of annotation and information involved in the project. Indeed, while lemmas are used by lexica to label entries, lemmatization is often performed in digital libraries of Latin texts to index words and is included in most NLP pipelines (like e.g. UDPipe)[220] as a preliminary step for more advanced forms of analysis[221].

LLD standards such as OntoLex-Lemon (see Section 5.1) provide an adequate framework to model the relations between the different classes of resources via lemmatization, while also offering a robust solution for modelling the information contained in most lexica. The central component in LiLa's framework, the gateway between different projects, is the collection of canonical forms that are used to lemmatize texts (called the **lemma bank**). This collection was created starting from the lexical database of the morphological analyzer **Lemlat** [164], and currently includes a set of about 190,000 forms that can potentially be used as lemmas in corpora or lexica[222].

The forms in the lemma bank are described in an OWL ontology that reuses several concepts from the LLD standards discussed in the previous sections. The canonical forms are instances of the class Lemma, which is defined as a subclass of the Form from the OntoLex-Lemon vocabulary. The part-of-speech and morphological annotations in the Lemlat database have been included in the ontology and linked to the OLiA reference model (see Section 4.5). For a selection of circa 36,000 lemmas, the lemma bank also includes derivational information, listing the morphemes (i.e. the prefixes, affixes and lexical bases) that can be identified in each lemma [165].

The fact that OntoLex-Lemon forms are allowed to have multiple written representations is a particularly helpful feature for a language which is attested across circa 25 centuries and in a wide spectrum of genres, and which is, moreover, characterised by a substantial amount of spelling variation. Harmonising different lemmatisation solutions adopted by corpora and NLP tools, however, requires practitioners to deal with other kinds of variation as well [166]. In the case of words with multiple inflectional paradigms or forms which may be interpreted as either autonomous words or inflected forms of a main lemma (such as participles, or adverbs built from adjectives: see e.g. English "quickly" from "quick"), different projects may vary considerably in the adopted strategies. For these reasons, the LiLa ontology introduces one sub-class of the Lemma class and two new object properties that connect forms to forms. The property lemma variant connects two lemmas that can be alternatively used to lemmatise forms of the same words. Hypolemma is a new sub-class of Lemma that groups forms (e.g. participles) that can be either promoted to canonical or be lemmatised under a hyperlemma (e.g. the main verb); hypolemmas are connected to their hyperlemma via the is hypolemma property.

Currently, the canonical forms in the LiLa lemma bank connect lexical entries of four lexical resources. Two lexica provide etymological information, modelled using the OntoLex-Lemon extension *lemonEty* [56], respectively dealing with the lexicon inherited from proto-Indo-european[223] [167] and loans from Greek[224] [168]. The polarity lexicon *LatinAffectus* connects a polarity value (expressed using the **Marl** ontology[225]) to a general sense for 1,998 entries[226] [169]. Finally, 1,421 verbs from the *Latin WordNet* have been manually revised and published as LOD[227] [170].

In addition to lexica, two annotated corpora are currently linked to the LiLa lemma bank. The *Index Thomisticus* Treebank[228] provides morpho-syntactic annotation for 375,000 tokens from the Latin works of

[220] https://ufal.mff.cuni.cz/udpipe.

[221] For the state of the art in automatic lemmatization and PoS tagging for Latin, see the results of the first edition of **EvaLatin**, a campaign devoted to the evaluation of NLP tools for Latin [163]. The first edition of EvaLatin focused on two shared tasks (i.e. lemmatization and PoS tagging), each featuring three sub-tasks (i.e. Classical, Cross-Genre, Cross-Time). These sub-tasks were specifically designed to measure the impact of genre variation and diachrony on NLP tool performances.

[222] The lemma bank can be queried using the lemmaBank SPARQL endpoint of the project: https://lila-erc.eu/sparql/.

[223] https://lila-erc.eu/data/lexicalResources/BrillEDL/Lexicon

[224] https://lila-erc.eu/data/lexicalResources/IGVLL/Lexicon

[225] http://www.gsi.upm.es:9080/ontologies/marl/

[226] https://lila-erc.eu/data/lexicalResources/LatinAffectus/Lexicon

[227] http://lila-erc.eu/data/lexicalResources/LatinWordNet/Lexicon

[228] http://lila-erc.eu/data/corpora/ITTB/id/corpus

Thomas Aquinas (13th century CE), while the *Dante Search* corpus[229] includes the lemmatized text of four Latin works of Dante Alighieri (14th century), which are currently undergoing a process of syntactic annotation following the Universal Dependencies annotation style [171][230]. The POWLA ontology was used to represent texts and annotations for both corpora. However, the link between a corpus token and a lemma of the LiLa collection was expressed using a custom property has lemma defined in the LiLa ontology[231], which takes an instance of the Lemma class as its range, since no existing vocabulary provided a suitable way to express such relation.

6.2.5. Prêt-à-LLOD (2019-2022)

The goal of the **Prêt-à-LLOD** project is to make linguistic linked open data 'ready-to-use' and part of this mission is to contribute to the development of new vocabularies for linguistic linked data in application scenarios that facilitate the development of a next-generation multilingual internet. Several aspects of linked data technology are being pursued in this context. This includes, without being restricted to

linking In its linking aspect, Prêt-à-LLOD explores technologies to facilitate the linking between and among lexical, terminological and ontological resources. In this context, it has provided significant support to the development of OntoLex-Lemon, including the development of a module for lexicography, a module for morphology, and corpus information (all of which are discussed in Section 5.1). Further extensions for terminologies and linking metadata (Fuzzy Lemon) have been proposed in the context of the project, as well. In addition, the project is contributing models for dataset linking to the Naisc project [232] that provides a toolkit for generic dataset linking.

transformation Prêt-à-LLOD provides a generic framework for transforming, enriching and manipulating language resources by means of RDF technology [172]. The idea here is to transform a language resource into an equivalent RDF representation, to manipulate and enrich it with a SPARQL transformation and external knowledge, and to serialize the result in RDF or non-RDF formats. To the extent that different formats can be mapped to or generated from the same RDF representation, they can be transformed one into another. For lexical data, the OntoLex-Lemon model and its aforementioned extensions represent a de facto standard and are being used as such. For linguistic annotations, several competing standards exist, and Prêt-à-LLOD contributes to on-going consolidation efforts within the W3C CG Linked Data for Language Technology with case studies on and support for CoNLL-RDF, NIF, Ligt, POWLA, and OLiA (see Sect. 5.2).

metadata Prêt-à-LLOD provides a workflow management system, a metadata repository for language resources, and machine-readable license information. In that regard, it also contributes to the development of metadata standards. This work is leading to a new version of the Linghub site [114][233], that is based around the DSpace open source software repository as well as the linking technologies to provide a single authoritative source of information about language resources across a wide range of languages.

The key priority of Prêt-à-LLOD, however, is less to develop novel vocabularies, than to develop technical solutions on that basis. Accordingly, Prêt-à-LLOD involves four industry-led pilot projects that are designed to demonstrate the relevance, transferability and applicability of the methods and techniques under development in the project to concrete problems in the language technology industry. The pilots showcase potentials in the context of various sectors: technology companies, open government services, pharmaceutical industry, and finance, details of which are described in [173] As overarching challenges, all pilots are addressing facets of *cross-language transfer* or *domain adaptation* to varying degrees. Particularly relevant to LLOD, the project is developing tools that are helpful to practical lexicographic applications, including for the Oxford Dictionaries [174].

Notable project results in the context of this paper are **a Report on Vocabularies for Interoperable Language Resources and Services** that gives a brief overview over standards for language resources as of 2019[234] and the publication of the first mono-

[229] http://lila-erc.eu/data/corpora/DanteSearch/id/corpus
[230] https://universaldependencies.org/guidelines.html
[231] https://lila-erc.eu/lodview/ontologies/lila/
[232] https://github.com/insight-centre/naisc
[233] https://linghub.org
[234] Christian Chiarcos, Philipp Cimiano, Julia Bosque-Gil, Thierry Declerck, Christian Fäth, Jorge Gracia, Maxim Ionov, John McCrae, Elena Montiel-Ponsoda, Maria Pia di Buono, Roser Saurí, Fernando Bobillo, Mohammad Fazleh Elahi (2020), Report on

graph on LLOD technologies [11]. Whereas the latter builds on long-standing collaborations between its authors in previous projects and community groups, it was finalized with support from the Prêt-à-LLOD project.

6.2.6. NexusLinguarum (2019-2023)

The **European network for Web-centred linguistic data science (NexusLinguarum)**[235] is a COST Action project involving researchers from 42 countries. The network started in October 2019 and will last a total of four years. The COST Action promotes synergies across Europe between linguists, computer scientists, terminology experts, language professionals, and other stakeholders from both industry and society, in order to investigate into and to extend the areas of applicability of **linguistic data science** in a Web-centred context. Linguistic data science is concerned with providing a formal basis for the analysis, representation, integration and exploitation of linguistic data for language analysis (e.g. syntax, morphology, terminology, etc.) and language applications (e.g. machine translation, speech recognition, sentiment analysis, etc.). NexusLinguarum seeks to identify several key technologies to support such a study, including language resources, data analysis, NLP, and LLD. The latter is considered to be a cornerstone for the building of an ecosystem of multilingual and semantically interoperable linguistic data technologies and resources at a Web scale. Such an ecosystem is needed to foster the systematic cross-lingual discovery, exploitation, extension, curation and quality control of linguistic data.

One of the main research coordination objectives of NexusLinguarum is to propose, agree upon and disseminate best practices and standards for linking data and services across languages. In that regard, an active collaboration has been established with W3C community groups for the extension of existing standards such as OntoLex-Lemon as well as for the convergence of standards in language annotation (see Section 5). Several surveys of the state of the art are also being drafted by the NexusLinguarum community covering different salient aspects of the domain (e.g., multilingual linking across different linguistic description levels). A number of activities organised by NexusLinguarum have been planned with the aim of fostering collaboration and communication across communities. These include scientific conferences (e.g., LDK 2021[236]), and training schools (e.g., EuroLAN 2021[237]), where linguistic linked data will take on a central role. Finally, NexusLinguarum is also devoted to the collection and analysis of relevant use cases for linguistic data science and to developing prototypes and demonstrators that will address a selection of prototypical cases. In an initial phase, the definition of use cases will cover Humanities and Social Sciences, Linguistics (Media and Social Media, and Language Acquisition), Life Sciences, and Technology (Cybersecurity and FinTech). The COST action also places a strong emphasis on lesser resourced languages.

A NexusLinguarum Use Case: ReTeRom

As an example of the kinds of complex, heterogeneous resources which have been proposed by consortium members as candidates for modelling and publication as linked data with the support of members of the COST action, we will look at the corpora being produced in a Romanian language project.

The ReTeRom (*Resources and Technologies for Developing Human-Machine Interfaces in Romanian*) project[238] is working towards adding the Romanian language to the multilingual Linguistic Linked Open Data cloud[239]. There are four different ReTeRom components. These are CoBiLiRo, SINTERO[240], TEPRO-

Vocabularies for Interoperable Language Resources and Services, available from https://cordis.europa.eu/project/id/825182/results

[235] https://nexuslinguarum.eu/

[236] http://2021.ldk-conf.org/

[237] http://eurolan.info.uaic.ro/2021

[238] https://www.racai.ro/p/reterom/index_en.html/

[239] Note that several Romanian language resources (e.g. Romanian WordNet (RoWN), Romanian Reference Treebank (RoRefTrees or RRT), Corpus-driven linguistic data, etc.) are currently in the process of conversion to LLD. The converter implementation is open source (https://github.com/racai-ai/RoLLOD/)

[240] SINTERO (Technologies for the Realization of Human-Machine Interfaces for Text-to-Speech Synthesis with Expressivity), coordinated by Technical University of Cluj-Napoca (UTCN), primarily aims to implement a text-speech synthesis system in Romanian that allows the modelling and control of prosody (intonation in speech) in an appropriate way of natural speech. Secondly, SINTERO aims is to create as many voices synthesised in Romanian as possible (in this project at least 10 voices), so that they too can be used by an extended community, including in commercial applications [175]

LIN[241] and TADARAV[242]; we will focus on the first of these, CoBiLiRo, in the rest of the section.

CoBiLiRo (*Bimodal Corpus for Romanian Language*), coordinated by the "Alexandru Ioan Cuza" University of Iași (UAIC), is working with a large collection of parallel speech/text data [178]. This collection is annotated at different levels of both the acoustic and the linguistic components [179], something which greatly facilitates querying, editing and the carrying out of statistical analysis. Three types of formats pairing speech and text components were identified in the building of the CoBiLiRo repository: (1) PHS/LAB, a format which separates text, speech and alignment in different files; (2) MULTEXT/TEI, a format described initially in the MULTEXT project and later used in the building of various language resources; (3) TEXTGRID, a format supported by a large community of European developers and used in a large set of existing resources. In order to share and distribute these bimodal resources, a standard format for CoBiLiRo has been proposed, inspired by the TEI-P5.10 standard [2] and based on the idea of alignment between the speech and text components, taking into consideration several annotation conventions proposed in 2007 by Li and Zhi-gang [180]. At present, the header of this format includes the following metadata: *source of the object stored*; *speakers gender*; *speakers identity (if she/he agreed to this)*; *vocal type (spontaneous or in-reading)*; *recording conditions*; *duration*; *speech file type*; *speech-text alignment level*, etc. Moreover, the CoBiLiRo format allows for three types of segmentation ("file - adequate for resources held in multiple files, "startstop - adequate for resources that include only one speech file, and "file-start-stop a combination of the two types described before) and speech-text alignment, marked using <unit> tags. A <unit> tag includes two child nodes: the <speech> that names the file containing the speech component and the <text> that points to the corresponding textual transcription file.

As we hope the preceding example has demonstrated (and it is only one of numerous case studies within the project straddling several different disciplines, media and technical domains) the NexusLinguaram COST action has enormous potential as a testing ground for many of the new vocabularies and modules mentioned above.

7. Conclusions and Discussions of Future Challenges

We have attempted, in the present article to give a comprehensive survey and a near-exhaustive[243] description of the current state of affairs with respect to the use, definition and availability of models and vocabularies for Linguistic Linked (Open) Data. We have also gone into some detail as to the role of these models in various different initiatives, both past and present.

As we hope that the article has demonstrated, LLD is an extremely active and dynamic an area of research, with numerous projects and initiatives underway, or due to commence in the short term, which promise to bring further updates and improvements in coverage and expressivity in addition to what we have described here. For this reason, and in a vain attempt to stave off the risk of rapid obsolescence, we have attempted throughout this article to situate our descriptions of recent advances in the field within a discussion of more general, ongoing trends. Indeed this was our specific intention with Section 2 and in many other parts of the article: we want this survey to give the reader a good idea both of the future challenges which have yet to be fully confronted in LLD as well as the areas of immense opportunity which currently remain untapped.

In this rest of this section we will summarise the future prospects/challenges described in this paper. In the next and final subsection, Section 7.1, we focus on two particular areas and suggest a possible future trend and a proposal for a further direction of research.

[241] TEPROLIN (*Technologies for Processing Natural Language - Text*) which is coordinated by the Research Institute for Artificial Intelligence Mircea Drgnescu (ICIA), aims to create Romanian text processing technologies that can be readily used by the other component-projects of ReTeRom. For instance, higher layers of annotation may be performed using TEPROLIN services: on the speech component - the prosodic annotation (e.g. decrease of the fundamental frequency) and on the textual component sub-syntactic (e.g. clauses) and syntactic annotation (e.g. parsing trees). TEPROLIN works inside a major language processing and text mining platform such as UIMA, GATE or TextFlows [176]

[242] TADARAV *(Technologies for automatic annotation of audio data and for the creation of automatic speech recognition interfaces*), coordinated by the University Politehnica of Bucharest (UPB), primarily aims to develop a set of advanced technologies for generating transcripts aligned correctly with the voice signal from the body collected in the CoBiLiRo component project. Secondly, TADARAV aims to increase the accuracy of the current SpeeD automatic speech recognition system [177] by requalifying its acoustic model based on the entire body of speech collected and using more powerful language models generated in the TEPROLIN component project.

[243] We were certainly exhausted after writing it.

A Summary of the Present Article

In Section 4, we gave an overview of the most well known and widely used models for linguistic linked data, emphasising their FAIR-ness, and in particular: their accessibility via ontology search engines, whether and how licensing information is made available, and how versioning is handled; we saw how, in many cases, there still remained work to be done in these areas. We classified these models into different kinds of resource based on the LLOD cloud; this helped us to show how some areas were better served than others. We also briefly discussed the provision of dedicated tools for LLD models; again this is an area which is still very much under development.

Next, in Section 5 we looked at the latest developments in LLD community standards. This section was divided into a subsection discussing OntoLex-Lemon related developments (Section 5.1), a section on the latest developments regarding LLD models for annotation (Section 5.2), and a section on metadata (Section 5.3). Each of these sections features a detailed description of different initiatives in their respective areas (including those still in progress), including in the case of Section 5.2 and Section 5.3 discussions of future trends and prospects (Section 5.2.5 and Section 5.3.3 respectively). The main challenge in the case of LLD vocabularies for annotation is to respond to the need for a convergence of vocabularies. In the case of metadata vocabularies we looked at coverage issues, especially with regard to language identification.

Then in Section 6 we presented an overview of the impact of projects on the definition and use of LLD models and vocabularies. We focused on a number of ongoing projects and looked at their current and potential future contributions to LLD models and vocabularies. In the rest of this concluding section we will look at one important potential future trend, the involvement of research infrastructures alongside community groups and projects in the definition and ongoing development of models and vocabularies (Section 7.1.1. We will also make a proposal for handling the increasing complexity of LLD vocabularies (especially in the domain of language resources), namely, the recourse to ontology design patterns (Section 7.1.2).

7.1. Discussion of Future Trends and Challenges

7.1.1. Linguistic Linked Data, Projects, and Research Infrastructures

Throughout this article we have sought to underline the role of research projects alongside that of community groups such as the Open Linguistics Working Group or the W3C Ontology-Lexicon Community Group in driving the development of LLD vocabularies and models. Moving ahead however, the role of SSH research infrastructures (RI) could also begin to play an important role by helping to ensure longer term hosting solutions and the greater sustainability of resources and tools based on these models. RIs could also help to give long term support to the community groups which are developing such models and vocabularies: in addition to and in a complimentary way to the support received from projects and COST actions in the short-to-medium term. In this, inspiration can be taken from cases such as that of TEI Lex-0 (described in Section 6.2.3) an initiative which has been supported both by a number of funded projects and COST actions as well as by the DARIAH "Lexical Resources" Working Group[244].

Related to this, RIs could also assist in the dissemination of LLD vocabularies and models, making them more accessible to wider numbers of users and cutting across different disciplinary boundaries via the kinds of training and teaching activities in which they have already established expertise. In other words, the Linguistic Linked Data community could exploit both the technical and the knowledge infrastructures provided by such **European Resource Infrastructure Consortia** (ERICs) as CLARIN, DARIAH in order to further sustain the work carried out in individual research projects and via open community groups. In this connection we should mention a recent CLARIN event which brought members of these two communities together in order to initiate a dialogue on future collaboration between the two[245]. The event was well received and would seen to be a promising start for future collaborations.[246]

7.1.2. A Proposal: The Use of Design Patterns

The OntoLex-Lemon model has come to be used in (or has at least been proposed for) a wide range of use cases belonging to an increasing number of different

[244] See https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

[245] A textual summary of the virtual event and recordings of the presentations and discussion can be found here https://www.clarin.eu/event/2021/clarin-cafe-linguistic-linked-data

[246] Note that although we do not discuss it here (as it would have shifted us too far into the realms of research policy), the role played by the European Open Science Cloud (https://eosc-portal.eu/)will also be crucial here (at the very least for projects and initiatives taking place in Europe) and especially for its promotion of FAIR data.

disciplines and types of resource. As we have seen, the original model is currently being extended to cover new kinds of use cases by the W3C OntoLex-Lemon group through the definition and publication of new extensions each of which carries its own supplementary guidelines. In the long term, however, this has the potential to become very complicated very quickly.

As an example, take the modelling of specialised language resources for such areas as the study of morpho-syntax or historical linguistics (in the former case these are dealt with in in part in the original guidelines and in the new morphology module). In both of these cases, there are so many different types (and sub-types) of resource as well as varieties of theoretical approach and diversities of schools of thought (not to mention language-specific modelling requirements) that it would be difficult to produce guidelines with detailed enough provision for any and all of the exigencies that might potentially arise. Or instead, take the modelling of lexicographic resources (something which falls within the compass of the lexicog extension, Section 5.1.1). This could encompass numerous different kinds of sub-cases – e.g., etymological dictionaries, philological dictionaries, rhyming dictionaries – each of which brings its own specific varieties of modelling challenges. And moreover there often exist distinct technical solutions to given modelling problems without a strong enough consensus on any single one of these to make it the default. Such, for instance, is the case with modelling ordered sequences in RDF.

One way of handling this potential modelling complexity that avoids the drafting of ever more elaborate guidelines in conjunction with the definition of ever more specialised modules is via the publication and maintenance of a repository of **ontology design patterns** (ODP). ODP's are modelling solutions for recurring problems in the field of conceptual modelling and are intended as a means of enhancing resuability in knowledge base design. As the name suggests, they are based on previous work on design patterns in software engineering. ODPs are arranged in six types [181]. These range from so called **Logical ODPs**, i.e., patterns that deal with problems in expressivity of formal knowledge engineering languages such as OWL (such as the representation of n-ary relations), and **Architectural ODPs** which are compositions of Logical ODPs, to **Reasoning ODPs** which propose procedures for automatic inference (for a full list see [181]). In our case the most relevant of these types are the **Content ODPs**, which are described as solving domain specific problems.

The idea would be to define, promote, and collect OntoLex-Lemon specific design patterns (as well as those pertaining to other similar vocabularies) within the LLD community and beyond. This is not a completely new idea and design patterns had been created for OntoLex-Lemon's predecessor *lemon* in the past. These previous patterns are currently available on github[247] and offer templates for the creation of nominal, verbal and adjectival lexical entries as well as more specific kinds of these such as Relational Nouns, State Verbs and Intersective Adjectives. They are fairly limited in scope however and so our proposal would be for the creation of patterns covering a far wider variety of different areas/kinds of use cases. These new patterns would deal with the various sections of the W3C OntoLex-Lemon guidelines, such as for example the syntax and semantics and the decomposition sections, along with the lexicography module and the forthcoming Morphology and Frequency Attestation and Corpus (FrAC) modules. Each new pattern would follow the set of criteria proposed in for instance [181] for Content ODPs and would be based on competency questions, e.g., potential SPARQL queries.

These OntoLex-Lemon ODPs could then either be hosted on the ontology design patterns site[248], or a special repository, or both. They would provide a bridge between the OntoLex-Lemon guidelines and concrete applications; they would help to prevent those guidelines from becoming overly-complicated and unwieldy and would keep the extensions themselves as simple (and hopefully uncontroversial) as possible[249]. They would make models such as OntoLex-Lemon, and indeed several of the other models featured in this article, more accessible. Furthermore, they would allow us to recommend the re-use of other vocabularies without having to include them 'officially' within the OntoLex-Lemon guidelines themselves, ensuring the decoupling of the OntoLex-Lemon guidelines from these other vocabularies.

Acknowledgments

The authors thank Milan Dojchinovski and Francesca Frontini for several very helpful suggestions. This arti-

[247] https://github.com/jmccrae/lemon.patterns

[248] http://ontologydesignpatterns.org/wiki/Main_Page

[249] Although of course the original modules would still need to be revised and extended on the basis of new kinds of use-cases/modelling needs; ODPs would help to keep these to a minimal.

cle is based upon work from COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology). The article is also supported by the Horizon 2020 research and innovation programme with the projects Prêt-à-LLOD (grant agreement no. 825182) and ELEXIS (grant agreeement no. 731015). It has been also partially supported by the Spanish project PID2020-113903RB-I00 (AEI/FEDER, UE), by DGA/FEDER, and by the *Agencia Estatal de Investigación* of the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the "Ramón y Cajal" program (RYC2019-028112-I).

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, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* **3**(1) (2016), 160018. doi:10.1038/sdata.2016.18. https://www.nature.com/articles/sdata201618.

[2] TEI Consortium, TEI P5: Guidelines for Electronic Text Encoding and Interchange, Zenodo, 2020. doi:10.5281/zenodo.3992514.

[3] G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet and C. Soria, Multilingual resources for NLP in the lexical markup framework (LMF), *Lang. Resour. Evaluation* **43**(1) (2009), 57–70. doi:10.1007/s10579-008-9077-5.

[4] L. Clément and É. Villemonte de La Clergerie, MAF: a Morphosyntactic Annotation Framework, in: *2nd Language & Technology Conference (LTC'05)*, Z. Vetulani, ed., 2nd Language & Technology Conference (LTC'05), Poznan, Poland, 2005, pp. 90–94. https://hal.archives-ouvertes.fr/hal-01104466.

[5] E. Sirin, B. Parsia, B.C. Grau, A. Kalyanpur and Y. Katz, Pellet: A practical OWL-DL reasoner, *Web Semantics: Science, Services and Agents on the World Wide Web* **5**(2) (2007), 51–53, Software Engineering and the Semantic Web. doi:http://dx.doi.org/10.1016/j.websem.2007.03.004. http://www.sciencedirect.com/science/article/pii/S1570826807000169.

[6] N. Guarino and C.A. Welty, An Overview of OntoClean, in: *Handbook on Ontologies*, Springer Berlin Heidelberg, 2004, pp. 151–171. doi:https://doi.org/10.1007/978-3-540-24750-0_8.

[7] B. Mons, FAIR Science for Social Machines: Let's Share Metadata Knowlets in the Internet of FAIR Data and Services, *Data Intelligence* **1**(1) (2019), 22–42. doi:https://doi.org/10.1162/dint_a_00002.

[8] J. Bosque-Gil, J. Gracia, E. Montiel-Ponsoda and A. Gómez-Pérez, Models to represent linguistic linked data, *Natural Language Engineering* **24**(6) (2018), 811–859. doi:10.1017/S1351324918000347. https://www.cambridge.org/core/journals/natural-language-engineering/article/models-to-represent-linguistic-linked-data/805F3E46882414B9144E43E34E89457D.

[9] H. Bohbot, F. Frontini, F. Khan, M. Khemakhem and L. Romary, Nénufar: Modelling a Diachronic Collection of Dictionary Editions as a Computational Lexical Resource, in: *Electronic lexicography in the 21st century. Proceedings of eLex 2019*, 2019.

[10] M. Passarotti, F. Mambrini, G. Franzini, F.M. Cecchini, E. Litta, G. Moretti, P. Ruffolo and R. Sprugnoli, Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin, *Studi e Saggi Linguistici* **58**(1) (2020), 177–212. doi:10.4454/ssl.v58i1.277.

[11] P. Cimiano, C. Chiarcos, J.P. McCrae and J. Gracia, *Linguistic Linked Data: Representation, Generation and Applications*, Springer International Publishing, 2020. doi:10.1007/978-3-030-30225-2.

[12] A. Pareja-Lora, M. Blume, B.C. Lust and C. Chiarcos, *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*, MIT Press, 2020. doi:10.7551/mitpress/10990.003.0003.

[13] C. Chiarcos, S. Hellmann and S. Nordhoff, Linking linguistic resources: Examples from the open linguistics working group, in: *Linked Data in Linguistics*, Springer, 2012, pp. 201–216. doi:https://doi.org/10.1007/978-3-642-28249-2_19.

[14] W. Hugo, Y. Le Franc, G. Coen, J. Parland-von Essen and L. Bonino, D2.5 FAIR Semantics Recommendations Second Iteration, Zenodo, 2020. doi:10.5281/zenodo.5362010.

[15] P.-Y. Vandenbussche and B. Vatant, Metadata recommendations for linked open data vocabularies, Technical Report, 2012.

[16] P.-Y. Vandenbussche, G.A. Atemezing, M. Poveda-Villalón and B. Vatant, Linked Open Vocabularies (LOV): a gateway to reusable semantic vocabularies on the Web, *Semantic Web* **8**(3) (2017), 437–452. doi:10.3233/SW-160213.

[17] J. McCrae, G. Aguado-de-Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda and D. Spohr, Interchanging lexical resources on the semantic web, *Language Resources and Evaluation* **46**(4) (2012), 701–719, Publisher: Springer. doi:10.1007/s10579-012-9182-3.

[18] P. Cimiano, J.P. McCrae and P. Buitelaar, Lexicon Model for Ontologies: Community Report, W3C, 2016. https://www.w3.org/2016/05/ontolex/.

[19] G. Sérasset, DBnary: Wiktionary as a Lemon-based multilingual lexical resource in RDF, *Semantic Web* **6**(4) (2015), 355–361, Publisher: IOS Press. doi:10.3233/SW-140147.

[20] M. Ehrmann, F. Cecconi, D. Vannella, J.P. Mccrae, P. Cimiano and R. Navigli, Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0, in: *Proceedings of the Ninth International Conference on Language Resources and Evalu-*

ation (LREC'14), European Language Resources Association (ELRA), 2014.

[21] B. Klimek, N. Arndt, S. Krause and T. Arndt, Creating linked data morphological language resources with mmoon-the hebrew morpheme inventory, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), 2016, pp. 892–899.

[22] R. Forkel, The cross-linguistic linked data project, in: *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL-2014): Multilingual Knowledge Resources and Natural Language Processing*, 2014, p. 61.

[23] M. Kemps-Snijders, M. Windhouwer, P. Wittenburg and S.E. Wright, ISOcat: Corralling data categories in the wild, in: *6th International Conference on Language Resources and Evaluation (LREC 2008)*, European Language Resources Association (ELRA), 2008. https://aclanthology.org/L08-1431/.

[24] S. Farrar and D.T. Langendoen, A linguistic ontology for the semantic web, *GLOT international* **7**(3) (2003), 97–100.

[25] H. Aristar-Dry, S. Drude, M. Windhouwer, J. Gippert and I. Nevskaya, Rendering endangered lexicons interoperable through standards harmonization: the relish project, in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association (ELRA), 2012, pp. 766–770.

[26] D.T. Langendoen, Whither GOLD?, in: *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*, A. Pareja-Lora, B. Lust, M. Blume and C. Chiarcos, eds, MIT Press, 2019. doi:10.7551/mitpress/10990.003.0003.

[27] C. Chiarcos and M. Sukhareva, Olia – ontologies of linguistic annotation, *Semantic Web* **6**(4) (2015), 379–386. doi:10.3233/SW-140167.

[28] C. Chiarcos, C. Fäth and F. Abromeit, Annotation Interoperability for the Post-ISOCat Era, in: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, European Language Resources Association (ELRA), 2020.

[29] C.A. Ferguson, Diglossia, *WORD* **15**(2) (1959), 325–340. doi:10.1080/00437956.1959.11659702.

[30] D.G. Martin Haspelmath Matthew S. Dryer and B. Comrie, *The world atlas of language structures*, Oxford University Press, 2005. ISBN 9780199255917.

[31] M.S. Dryer and M. Haspelmath (eds), *WALS Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. doi:10.5281/zenodo.4683137. https://wals.info/.

[32] P. Monachesi, A. Dimitriadis, R. Goedemans, A.-M. Mineur and M. Pinto, The typological database system, in: *Proceedings of the IRCS workshop on linguistic databases*, 2001, pp. 181–186.

[33] P. Monachesi, A. Dimitriadis, R. Goedemans and A.-M. Mineur, A unified system for accessing typological databases, in: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain, 2002. http://www.lrec-conf.org/proceedings/lrec2002/pdf/279.pdf.

[34] A. Dimitriadis, M. Windhouwer, A. Saulwick, R. Goedemans and T. Bíró, How to integrate databases without starting a typology war: The Typological Database System, *The Use of Databases in Cross-Linguistic Studies, Mouton de Gruyter, Berlin* (2009), 155–207. doi:10.1515/9783110198744.155.

[35] P. Westphal, C. Stadler and J. Pool, Countering language attrition with PanLex and the Web of Data, *Semantic Web* **6**(4) (2015), 347–353. doi:10.3233/SW-140138.

[36] C. Chiarcos, C. Fäth and M. Ionov, The ACoLi dictionary graph, in: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, European Language Resources Association (ELRA), 2020, pp. 3281–3290.

[37] G. De Melo, Lexvo. org: Language-related information for the linguistic linked data cloud, *Semantic Web* **6**(4) (2015), 393–400, Publisher: IOS Press. doi:10.3233/SW-150171.

[38] A. Stellato, M. Fiorelli, A. Turbati, T. Lorenzetti, W. van Gemert, D. Dechandon, C. Laaboudi-Spoiden, A. Gerencsér, A. Waniart, E. Costetchi and et al., VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons, *Semantic Web* **11**(5) (2020), 855–881. doi:10.3233/SW-200370.

[39] A. Bellandi, E. Giovannetti and A. Weingart, Multilingual and multiword phenomena in a lemon old occitan medico-botanical lexicon, *Information* **9**(3) (2018), 52. doi:https://-doi.org/10.3390/info9030052.

[40] C. Chiarcos, S. Hellmann and S. Nordhoff, Towards a Linguistic Linked Open Data cloud: The Open Linguistics Working Group, *TAL Traitement Automatique des Langues* **52**(3) (2011), 245–275. doi:10.1.1.377.2076.

[41] C. Chiarcos, S. Nordhoff and S. Hellmann, *Linked Data in Linguistics*, Springer, 2012. doi:10.1007/978-3-642-28249-2.

[42] J.P. McCrae, S. Moran, S. Hellmann and M. Brümmer (eds), *Semantic Web 6(4), Special Issue on Multilingual Linked Open Data*, IOS Press, 2015, pp. 313–400. https://content.iospress.com/journals/semantic-web/6/4.

[43] F. Khan, F. Boschetti and F. Frontini, Using lemon to model lexical semantic shift in diachronic lexical resources, in: *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL-2014): Multilingual Knowledge Resources and Natural Language Processing*, 2014, pp. 50–54.

[44] B. Klimek and M. Brümmer, Enhancing lexicography with semantic language databases, *Kernerman Dictionary News* **23** (2015), 5–10.

[45] J. Bosque-Gil, J. Gracia, E. Montiel-Ponsoda and G. Aguado-de-Cea, Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case, in: *GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop Programme*, 2016, p. 65.

[46] F. Abromeit, C. Chiarcos, C. Fäth and M. Ionov, Linking the Tower of Babel: Modelling a Massive Set of Etymological Dictionaries as RDF, in: *Proceedings of the 5th Workshop on Linked Data in Linguistics (LDL-2016): Managing, Building and Using Linked Language Resources*, 2016, p. 11.

[47] J. Gracia, M. Villegas, A. Gómez-Pérez and N. Bel, The apertium bilingual dictionaries on the web of data, *Semantic Web* **9**(2) (2018), 231–240. doi:10.3233/SW-170258.

[48] J. Bosque-Gil, J. Gracia and E. Montiel-Ponsoda, Towards a Module for Lexicography in OntoLex, in: *Proc. of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets at 1st Language Data and Knowledge conference (LDK 2017), Galway, Ireland*, Vol. 1899, CEUR-WS, Galway (Ire-

land), 2017, pp. 74–84. ISSN 1613-0073. http://ceur-ws.org/Vol-1899/OntoLex_2017_paper_5.pdf.

[49] J. Bosque-Gil, D. Lonke, J. Gracia and I. Kernerman, Validating the OntoLex-lemon lexicography module with K Dictionaries' multilingual data, in: *Electronic lexicography in the 21st century. Proceedings of eLex 2019*, 2019, pp. 726–746. doi:10.5281/zenodo.3462317.

[50] B. Klimek, J.P. McCrae, M. Ionov, J.K. Tauber, C. Chiarcos, J. Bosque-Gil and P. Buitelaar, Challenges for the Representations for Morphology in Ontology Lexicons, in: *Electronic lexicography in the 21st century. Proceedings of eLex 2019*, 2019. doi:10.5281/zenodo.3518946. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_33.pdf.

[51] C. Chiarcos, M. Ionov, J. de Does, K. Depuydt, A.F. Khan, S. Stolk, T. Declerck and J.P. McCrae, Modelling Frequency and Attestations for OntoLex-Lemon, in: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC-2020)*, European Language Resources Association (ELRA), 2020, pp. 1–9. doi:10.5281/zenodo.3842633. https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/GLOBALEX2020book.pdf#page=19.

[52] S. Peroni and D. Shotton, FaBiO and CiTO: ontologies for describing bibliographic resources and citations, *Web Semantics: Science, Services and Agents on the World Wide Web* **17** (2012), 33–43. doi:https://doi.org/10.1016/j.websem.2012.08.001.

[53] C. Chiarcos, T. Declerck and M. Ionov, Embeddings for the Lexicon: Modelling and Representation, in: *Proceedings of the 6th Workshop on Semantic Deep Learning (SemDeep-6), held virtually in January 2021, co-located with IJCAI-PRICAI 2020*, Japan, 2021.

[54] S. Stolk, lemon-tree: Representing Topical Thesauri on the Semantic Web, in: *2nd Conference on Language, Data and Knowledge (LDK 2019)*, M. Eskevich, G. de Melo, C. Fäth, J.P. McCrae, P. Buitelaar, C. Chiarcos, B. Klimek and M. Dojchinovski, eds, OpenAccess Series in Informatics (OASIcs), Vol. 70, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2019, pp. 16:1–16:13. ISSN 2190-6807. ISBN 978-3-95977-105-4. doi:10.4230/OASIcs.LDK.2019.16. http://drops.dagstuhl.de/opus/volltexte/2019/10380.

[55] S. Stolk, A Thesaurus of Old English as linguistic linked data: Using OntoLex, SKOS and lemon-tree to bring topical thesauri to the Semantic Web, in: *Proceedings of the eLex 2019 conference*, 2019, pp. 223–247.

[56] A.F. Khan, Towards the Representation of Etymological Data on the Semantic Web, *Information* **9**(12) (2018). doi:10.3390/info9120304. https://www.mdpi.com/2078-2489/9/12/304.

[57] F. Khan, Representing Temporal Information in Lexical Linked Data Resources, in: *Proceedings of the 7h Workshop on Linked Data in Linguistics (LDL-2020)*, European Language Resources Association (ELRA), 2020, pp. 15–22. ISBN 979-10-95546-36-8. https://www.aclweb.org/anthology/2020.ldl-1.3.

[58] A. Burchardt, S. Padó, D. Spohr, A. Frank and U. Heid, Formalising Multi-layer Corpora in OWL/DL – Lexicon Modelling, Querying and Consistency Control, in: *Proc. of the 3rd International Joint Conference on NLP (IJCNLP)*, Hyderabad, India, 2008, pp. 389–396.

[59] K. Verspoor and K. Livingston, Towards Adaptation of Linguistic Annotations to Scholarly Annotation Formalisms on the Semantic Web, in: *Proc. of the 6th Linguistic Annotation Workshop*, Association for Computational Linguistics, Jeju, Republic of Korea, 2012, pp. 75–84.

[60] S. Hellmann, J. Lehmann, S. Auer and M. Brümmer, Integrating NLP using Linked Data, in: *Proc. 12th International Semantic Web Conference, 21-25 October 2013*, Sydney, Australia, 2013, also see http://persistence.uni-leipzig.org/nlp2rdf/. doi:10.1007/978-3-642-41338-4_7.

[61] N. Mazziotta, Building the Syntactic Reference Corpus of Medieval French Using NotaBene RDF Annotation Tool, in: *Proc. of the Fourth Linguistic Annotation Workshop*, Association for Computational Linguistics, 2010, pp. 142–146.

[62] B. Almas, H. Cayless, T. Clérice, Z. Fletcher, V. Jolivet, P. Liuzzo, E. Morlock, J. Robie, M. Romanello, J. Tauber and J. Witt, Distributed Text Services (DTS). First Public Working Draft, Technical Report, Github, 2019, version of May 23, 2019.

[63] S. Cassidy, An RDF realisation of LAF in the DADA annotation server, in: *Proc. of the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation (ISA-5)*, Hong Kong, 2010.

[64] N. Diewald, M. Hanl, E. Margaretha, J. Bingel, M. Kupietz, P. Bański and A. Witt, KorAP Architecture – Diving in the Deep Sea of Corpus Data, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), 2016, pp. 3586–3591.

[65] ISO, ISO 24612:2012. Language Resource Management - Linguistic Annotation Framework, Technical Report, ISO/TC 37/SC 4, Language resource management, 2012. https://www.iso.org/standard/37326.html.

[66] C. Chiarcos, POWLA: Modeling linguistic corpora in OWL/DL, in: *Extended Semantic Web Conference*, Springer, 2012, pp. 225–239. doi:https://doi.org/10.1007/978-3-642-30284-8_22.

[67] N. Ide and L. Romary, International Standard for a Linguistic Annotation Framework, *Natural language engineering* **10**(3–4) (2004), 211–225. doi:10.1017/S135132490400350X.

[68] S. Tittel, H. Bermúdez-Sabel and C. Chiarcos, Using RDFa to Link Text and Dictionary Data for Medieval French, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, European Language Resources Association (ELRA), 2018.

[69] P. Ruiz Fabo, H. Bermúdez Sabel, C. Martínez Cantón and E. González-Blanco, The Diachronic Spanish Sonnet Corpus: TEI and linked open data encoding, data distribution, and metrical findings, *Digital Scholarship in the Humanities* (2020). doi:https://doi.org/10.1093/llc/fqaa035.

[70] A. Gangemi, V. Presutti, D. Reforgiato Recupero, A.G. Nuzzolese, F. Draicchio and M. Mongiovì, Semantic Web Machine Reading with FRED, *Semantic Web* **8**(6) (2017), 873–893. doi:10.3233/SW-160240.

[71] P. Vossen, R. Agerri, I. Aldabe, A. Cybulska, M. van Erp, A. Fokkens, E. Laparra, A.-L. Minard, A.P. Aprosio, G. Rigau et al., Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news,

Knowledge-Based Systems **110** (2016), 60–85. doi:https://doi.org/10.1016/j.knosys.2016.07.013.

[72] N. Ide, J. Pustejovsky, C. Cieri, E. Nyberg, D. DiPersio, C. Shi, K. Suderman, M. Verhagen, D. Wang and J. Wright, The language application grid, in: *International Workshop on Worldwide Language Service Infrastructure*, Springer, 2015, pp. 51–70.

[73] S. Peroni, A. Gangemi and F. Vitali, Dealing with markup semantics, in: *Proceedings of the 7th International Conference on Semantic Systems*, 2011, pp. 111–118. doi:https://doi.org/10.1145/2063518.2063533.

[74] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari and L. Schneider, Sweetening ontologies with DOLCE, in: *International Conference on Knowledge Engineering and Knowledge Management*, Springer, 2002, pp. 166–181. doi:https://doi.org/10.1007/3-540-45810-7_18.

[75] A. Fokkens, A. Soroa, Z. Beloki, N. Ockeloen, G. Rigau, W.R. Van Hage and P. Vossen, NAF and GAF: Linking linguistic annotations, in: *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, 2014, pp. 9–16.

[76] M. Verhagen, K. Suderman, D. Wang, N. Ide, C. Shi, J. Wright and J. Pustejovsky, The LAPPS interchange format, in: *International Workshop on Worldwide Language Service Infrastructure*, Springer, 2015, pp. 33–47. doi:https://doi.org/10.1007/978-3-319-31468-6_3.

[77] N. Ide, K. Suderman, J. Pustejovsky, M. Verhagen and C. Cieri, The language application grid and galaxy, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), 2016, pp. 457–462. doi:https://doi.org/10.1007/978-3-319-31468-6_4. https://aclanthology.org/L16-1073/.

[78] E. Hinrichs, N. Ide, J. Pustejovsky, J. Hajic, M. Hinrichs, M.F. Elahi, K. Suderman, M. Verhagen, K. Rim, P. Stranák et al., Bridging the LAPPS Grid and CLARIN, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018. https://aclanthology.org/L18-1206/.

[79] E. Wilde and M. Duerst, RFC 5147 – URI Fragment Identifiers for the text/plain Media Type, Technical Report, Internet Engineering Task Force (IETF), Network Working Group, 2008.

[80] D. Filip, S. McCance, D. Lewis, C. Lieske, A. Lommel, J. Kosek, F. Sasaki and Y. Savourel, Internationalization Tag Set (ITS) Version 2.0, Technical Report, W3C Recommendation 29 October 2013, 2013.

[81] J. Frey, M. Hofer, D. Obraczka, J. Lehmann and S. Hellmann, DBpedia FlexiFusion the best of Wikipedia> Wikidata> your data, in: *International Semantic Web Conference*, Springer, 2019, pp. 96–112. doi:https://doi.org/10.1007/978-3-030-30796-7_7.

[82] R. Sanderson, P. Ciccarese and B. Young, Web Annotation Data Model, Technical Report, W3C Recommendation, 2017. https://www.w3.org/TR/annotation-model/.

[83] R. Sanderson, P. Ciccarese and B. Young, Web Annotation Vocabulary, Technical Report, W3C Recommendation, 2017. https://www.w3.org/TR/annotation-vocab/.

[84] L. Isaksen, R. Simon, E.T. Barker and P. de Soto Cañamares, Pelagios and the emerging graph of ancient world data, in: *Proceedings of the 2014 ACM conference on Web science*, 2014, pp. 197–201. doi:https://doi.org/10.1145/2615569.2615693.

[85] R. Simon, E. Barker, L. Isaksen and P. de Soto Cañamares, Linked Data Annotation Without the Pointy Brackets: Introducing Recogito 2, *Journal of Map & Geography Libraries* **13**(1) (2017), 111–132. doi:https://doi.org/10.1080/15420353.2017.1307303.

[86] P. Cimiano, C. Chiarcos, J.P. McCrae and J. Gracia, *Linguistic Linked Data in Digital Humanities*, in: *Linguistic Linked Data*, Springer, 2020, pp. 229–262. doi:https://doi.org/10.1007/978-3-030-30225-2_13.

[87] C. Chiarcos and M. Ionov, Ligt: An LLOD-Native Vocabulary for Representing Interlinear Glossed Text as RDF, in: *2nd Conference on Language, Data and Knowledge (LDK 2019)*, M. Eskevich, G. de Melo, C. Fäth, J.P. McCrae, P. Buitelaar, C. Chiarcos, B. Klimek and M. Dojchinovski, eds, OpenAccess Series in Informatics (OASIcs), Vol. 70, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2019, pp. 3:1–3:15. ISSN 2190-6807. ISBN 978-3-95977-105-4. doi:10.4230/OASIcs.LDK.2019.3. http://drops.dagstuhl.de/opus/volltexte/2019/10367.

[88] S. Robinson, G. Aumann and S. Bird, Managing fieldwork data with Toolbox and the Natural Language Toolkit, *Language Documentation & Conservation* **1**(1) (2007), 44–57.

[89] L. Butler and H. Van Volkinburg, Fieldworks Language Explorer (FLEx), *Technology Review* **1**(1) (2007), 1.

[90] M.W. Goodman, J. Crowgey, F. Xia and E.M. Bender, Xigt: extensible interlinear glossed text for natural language processing, *Language Resources and Evaluation* **49**(2) (2015), 455–485. doi:https://doi.org/10.1007/s10579-014-9276-1.

[91] S. Nordhoff, Modelling and Annotating Interlinear Glossed Text from 280 Different Endangered Languages as Linked Data with LIGT, in: *Proceedings of the 14th Linguistic Annotation Workshop*, 2020, pp. 93–104.

[92] C. Chiarcos and C. Fäth, CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way, in: *International Conference on Language, Data and Knowledge*, Springer, Cham, 2017, pp. 74–88. doi:https://doi.org/10.1007/978-3-319-59888-8_6.

[93] M. Marcus, B. Santorini and M.A. Marcinkiewicz, Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics* **19**(2) (1993), 313–330.

[94] F. Mambrini and M. Passarotti, Linked Open Treebanks. Interlinking Syntactically Annotated Corpora in the LiLa Knowledge Base of Linguistic Resources for Latin, in: *Proceedings of TLT, SyntaxFest 2019*, Association for Computational Linguistics, Paris, France, 2019, pp. 74–81. doi:10.5281/zenodo.3474796.

[95] M. Tamper, P. Leskinen, K. Apajalahti and E. Hyvönen, Using biographical texts as linked data for prosopographical research and applications, in: *Euro-Mediterranean Conference*, Springer, 2018, pp. 125–137. doi:https://doi.org/10.1007/978-3-030-01762-0_11.

[96] C. Chiarcos, B. Kosmehl, C. Fäth and M. Sukhareva, Analyzing Middle High German Syntax with RDF and SPARQL, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, European Language Resources Association (ELRA), 2018, pp. 4525–4534.

[97] C. Chiarcos, I. Khait, É. Pagé-Perron, N. Schenk, C. Fäth, J. Steuer, W. Mcgrath, J. Wang et al., Annotating a low-resource language with LLOD technology: Sumerian morphology and syntax, *Information* **9**(11) (2018), 290. doi:https://doi.org/10.3390/info9110290.

[98] C. Chiarcos and C. Fäth, Graph-Based Annotation Engineering: Towards a Gold Corpus for Role and Reference Grammar, in: *2nd Conference on Language, Data and Knowledge (LDK-2019)*, OpenAccess Series in Informatics, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, Germany, 2019, pp. 9:1–9:11. doi:10.4230/OASIcs.LDK.2019.9.

[99] M. Ionov, F. Stein, S. Sehgal and C. Chiarcos, cqp4rdf: Towards a Suite for RDF-Based Corpus Linguistics, in: *European Semantic Web Conference*, Springer, 2020, pp. 115–121. doi:https://doi.org/10.1007/978-3-030-62327-2_20.

[100] C. Chiarcos, K. Donandt, H. Sargsian, M. Ionov and J.W. Schreur, Towards LLOD-based language contact studies. A case study in interoperability, in: *Proceedings of the 6th Workshop on Linked Data in Linguistics (LDL-2018)*, 2018.

[101] C. Chiarcos and L. Glaser, A Tree Extension for CoNLL-RDF, in: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC-2020)*, European Language Resources Association (ELRA), 2020, pp. 7161–7169.

[102] P. Cimiano, C. Chiarcos, J.P. McCrae and J. Gracia, *Modelling Linguistic Annotations*, in: *Linguistic Linked Data*, Springer, 2020, pp. 89–122. doi:https://doi.org/10.1007/978-3-030-30225-2_6.

[103] D. Broeder, M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg and C. Zinn, A Data Category Registry- and Component-based Metadata Framework, in: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA), 2010. https://aclanthology.org/L10-1105/.

[104] D. Broeder, D. van Uytvanck, M. Gavrilidou, T. Trippel and M. Windhouwer, Standardizing a Component Metadata Infrastructure, in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association (ELRA), 2012. https://aclanthology.org/L12-1329/.

[105] M. Windhouwer, E. Indarto and D. Broeder, CMD2RDF: Building a Bridge from CLARIN to Linked Open Data, *Ubiquity Press* (2017), Publisher: Ubiquity Press. doi:10.5334/bbi.8.

[106] D. Broeder, I. Schuurman and M. Windhouwer, Experiences with the ISOcat Data Category Registry, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), 2014. https://aclanthology.org/L14-1171/.

[107] I. Schuurman, M. Windhouwer, O. Ohren and D. Zeman, CLARIN Concept Registry: The New Semantic Registry, in: *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland*, Linköping University Electronic Press, 2016, pp. 62–70. doi:10.1.1.1079.2778.

[108] P. Labropoulou, K. Gkirtzou, M. Gavriilidou, M. Deligiannis, D. Galanis, S. Piperidis, G. Rehm, M. Berger, V. Mapelli, M. Rigault, V. Arranz, K. Choukri, G. Backfried, J.M.G. Peŕez and A. Garcia-Silva, Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid, in: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC-2020)*, European Language Resources Association (ELRA), 2020. doi:10.5281/zenodo.4059210. https://www.aclweb.org/anthology/2020.lrec-1.420/.

[109] M. Gavrilidou, P. Labropoulou, E. Desipri, S. Piperidis, H. Papageorgiou, M. Monachini, F. Frontini, T. Declerck, G. Francopoulo, V. Arranz and V. Mapelli, The META-SHARE Metadata Schema for the Description of Language Resources, in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association (ELRA), 2012. https://aclanthology.org/L12-1593/.

[110] J.P. McCrae, P. Labropoulou, J. Gracia, M. Villegas, V. Rodríguez-Doncel and P. Cimiano, One Ontology to Bind Them All: The META-SHARE OWL Ontology for the Interoperability of Linguistic Datasets on the Web, in: *The Semantic Web: ESWC 2015 Satellite Events*, F. Gandon, C. Guéret, S. Villata, J. Breslin, C. Faron-Zucker and A. Zimmermann, eds, Lecture Notes in Computer Science, Springer International Publishing, 2015, pp. 271–282. ISBN 978-3-319-25639-9. doi:10.1007/978-3-319-25639-9_42. https://link.springer.com/chapter/10.1007/978-3-319-25639-9_42.

[111] S. Piperidis, The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions, in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association (ELRA), 2012. https://aclanthology.org/L12-1647/.

[112] V. Rodriguez-Doncel and P. Labropoulou, Digital Representation of Licenses for Language Resources, in: *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, Association for Computational Linguistics, Beijing, China, 2015, pp. 49–58. doi:10.18653/v1/W15-4206. http://aclweb.org/anthology/W15-4206.

[113] P. Labropoulou, D. Galanis, A. Lempesis, M. Greenwood, P. Knoth, R. Eckart de Castilho, S. Sachtouris, B. Georgantopoulos, S. Martziou, L. Anastasiou, K. Gkirtzou, N. Manola and S. Piperidis, OpenMinTeD: A Platform Facilitating Text Mining of Scholarly Content, in: *WOSP 2018 Workshop Proceedings, Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), 2018, pp. 7–12. http://lrec-conf.org/workshops/lrec2018/W24/pdf/13_W24.pdf.

[114] J.P. McCrae and P. Cimiano, Linghub: a Linked Data based portal supporting the discovery of language resources., *SEMANTiCS (Posters & Demos)* **1481** (2015), 88–91. doi:10.1.1.1083.2922.

[115] J.P. McCrae, P. Cimiano, V. Rodríguez Doncel, D. Vila-Suero, J. Gracia, L. Matteis, R. Navigli, A. Abele, G. Vulcu and P. Buitelaar, Reconciling Heterogeneous Descriptions of Language Resources, in: *Proceedings of the 4th Workshop on Linked Data in Linguistics (LDL-2015): Resources and Applications*, Association for Computational Linguistics, 2015, pp. 39–48. doi:10.18653/v1/W15-4205. https://www.aclweb.org/anthology/W15-4205.

[116] G. Rehm, M. Berger, E. Elsholz, S. Hegele, F. Kintzel, K. Marheinecke, S. Piperidis, M. Deligiannis, D. Galanis, K. Gkirtzou, P. Labropoulou, K. Bontcheva, D. Jones, I. Roberts, J. Hajič, J. Hamrlová, L. Kačena, K. Choukri,

V. Arranz, A. Vasiljevs, O. Anvari, A. Lagzdiņš, J. Meļņika, G. Backfried, E. Dikici, M. Janosik, K. Prinz, C. Prinz, S. Stampler, D. Thomas-Aniola, J.M. Gómez-Pérez, A. Garcia Silva, C. Berrío, U. Germann, S. Renals and O. Klejch, European Language Grid: An Overview, in: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC-2020)*, European Language Resources Association (ELRA), 2020. doi:10.5281/zenodo.4058239. https://www.aclweb.org/anthology/2020.lrec-1.413.

[117] M. Fiorelli, A. Stellato, J.P. Mccrae, P. Cimiano and M.T. Pazienza, LIME: the metadata module for OntoLex, in: *European Semantic Web Conference*, Springer, 2015, pp. 321–336. doi:https://doi.org/10.1007/978-3-319-18818-8_20.

[118] F. Khan and A. Salgado, Modelling Lexicographic Resources using CIDOC-CRM, FRBRoo and Ontolex-Lemon, in: *Proceedings of the International Joint Workshop on Semantic Web and Ontology Design for Cultural Heritage co-located with the Bolzano Summer of Knowledge 2021 (BOSK 2021), Virtual Event / Bozen-Bolzano, Italy, September 20-21, 2021*, A. Bikakis, R. Ferrario, S. Jean, B. Markhoff, A. Mosca and M.N. Asmundo, eds, CEUR Workshop Proceedings, Vol. 2949, CEUR-WS.org, 2021. http://ceur-ws.org/Vol-2949/paper7.pdf.

[119] R. Cyganiak, D. Wood and M. Lanthaler, RDF 1.1 Concepts and Abstract Syntax, Technical Report, W3C Recommendation 25 February 2014, 2014.

[120] A. Phillips and M. Davis, BCP 47 – Tags for Identifying Languages, Technical Report, Internet Engineering Task Force, 2006. http://www.rfc-editor.org/rfc/bcp/bcp47.txt.

[121] F. Gillis-Webber and S. Tittel, The shortcomings of language tags for linked data when modeling lesser-known languages, in: *2nd Conference on Language, Data and Knowledge (LDK 2019)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019. doi:10.4230/OASIcs.LDK.2019.4.

[122] S. Tittel and F. Gillis-Webber, Identification of Languages in Linked Data: A Diachronic-Diatopic Case Study of French, in: *Electronic lexicography in the 21st century. Proceedings of eLex 2019*, 2019, pp. 1–3.

[123] F. Gillis-Webber and S. Tittel, A Framework for Shared Agreement of Language Tags beyond ISO 639, in: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, European Language Resources Association (ELRA), 2020, pp. 3333–3339.

[124] S. Nordhoff, Linked data for linguistic diversity research: Glottolog/langdoc and asjp online, in: *Linked Data in Linguistics*, Springer, 2012, pp. 191–200. doi:https://doi.org/10.1007/978-3-642-28249-2_18.

[125] G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet and C. Soria, Lexical markup framework (LMF), 2006. ISBN 9781118712597. https://aclanthology.org/L06-1348/.

[126] L. Romary, M. Khemakhem, M. George, J. Bowers, F. Khan, M. Pet, S. Lewis, N. Calzolari and P. Banski, LMF Reloaded., in: *Proceedings of the 13th International Conference of the Asian Association for Lexicography (ASIALEX)*, 2019.

[127] V.R. Doncel and E.M. Ponsoda, LYNX: Towards a Legal Knowledge Graph for Multilingual Europe, *Law in Context. A Socio-legal Journal* **37**(1) (2020), 1–4. doi:https://doi.org/10.26826/law-in-context.v37i1.129.

[128] T. Burrows, E. Hyvönen, L. Ransom and H. Wijsman, Mapping Manuscript Migrations: Digging into Data for the History and Provenance of Medieval and Renaissance Manuscripts, *Manuscript Studies: A Journal of the Schoenberg Institute for Manuscript Studies* **3**(1) (2018), 249–252. doi:10.1353/mns.2018.0012.

[129] E. Hyvönen, "Sampo" Model and Semantic Portals for Digital Humanities on the Semantic Web., in: *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020)*, 2020, pp. 373–378.

[130] A. Khan, H. Bohbot, F. Frontini, M. Khemakhem and L. Romary, Historical Dictionaries as Digital Editions and Connected Graphs: the Example of Le Petit Larousse Illustré, in: *Digital Humanities 2019*, 2019.

[131] A. Weingart and E. Giovannetti, A Lexicon for Old Occitan Medico-Botanical Terminology in Lemon., in: *SWASH@ ESWC*, 2016, pp. 25–36.

[132] R. Costa, A. Salgado, A.F. Khan, S. Carvalho, L. Romary, B. Almeida, M. Ramos, M. Khemakhem, R. Silva and T. Tasovac, MORDigital: The Advent of a New Lexicographical Portuguese Project, in: *Electronic lexicography in the 21st century. Proceedings of eLex 2021*, 2021. https://hal.inria.fr/hal-03195362.

[133] V. Propp, *Morphology of the folktale*, Trans., Laurence Scott. 2nd ed., University of Texas Press, 1968.

[134] S. Thompson, *Motif-index of folk-literature: A classification of narrative elements in folktales, ballads, myths, fables, medieval romances, exempla, fabliaux, jest-books, and local legends*, Revised and enlarged edition (1955–1958), Indiana University Press, 1958.

[135] H.-J. Uther, *The Types of International Folktales: A Classification and Bibliography. Based on the system of Antti Aarne and Stith Thompson*, Suomalainen Tiedeakatemia, 2004.

[136] T. Declerck, A. Kostová and L. Schäfer, Towards a Linked Data Access to Folktales classified by Thompsons Motifs and Aarne-Thompson-Uthers Types, in: *Proceedings of Digital Humanities 2017*, ADHO, 2017.

[137] F. Diehr, M. Brodhun, S. Gronemeyer, K. Diederichs, C. Prager, E. Wagner and N. Grube, Modellierung eines digitalen Zeichenkatalogs für die Hieroglyphen des Klassischen Maya, in: *47. Jahrestagung der Gesellschaft für Informatik, Digitale Kulturen, INFORMATIK 2017, Chemnitz, Germany, September 25-29, 2017*, M. Eibl and M. Gaedke, eds, LNI, Vol. P-275, GI, 2017, pp. 1185–1196. doi:10.18420/in2017_120.

[138] G. Zólyomi, B. Tanos and S. Sövegjártó, The Electronic Text Corpus of Sumerian Royal Inscriptions, 2008. http://oracc.museum.upenn.edu/etcsri/.

[139] É. Pagé-Perron, M. Sukhareva, I. Khait and C. Chiarcos, Machine translation and automated analysis of the Sumerian language, in: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 2017, pp. 10–16. doi:10.18653/v1/W17-2202.

[140] M. Hartung, M. Orlikowski and S. Veríssimo, Evaluating the Impact of Bilingual Lexical Resources on Cross-lingual Sentiment Projection in the Pharmaceutical Domain, 2020.

[141] B.-P. Ivanschitz, T.J. Lampoltshammer, V. Mireles, A. Revenko, S. Schlarb and L. Thurnay, A Semantic Catalogue for the Data Market Austria., in: *SEMANTICS Posters&Demos*, 2018.

[142] D. Lonke and J. Bosque Gil, Applying the OntoLex-lemon lexicography module to K Dictionaries' multilingual data, *K Lexical News (KLN)* (2019). https://kln.lexicala.com/kln28/lonke-bosque-gil-ontolex-lemon-lexicog/.

[143] G. Rehm, D. Galanis, P. Labropoulou, S. Piperidis, M. Welß, R. Usbeck, J. Köhler, M. Deligiannis, K. Gkirtzou, J. Fischer, C. Chiarcos, N. Feldhus, J. Moreno-Schneider, F. Kintzel, E. Montiel, V. Rodríguez Doncel, J.P. McCrae, D. Laqua, I.P. Theile, C. Dittmar, K. Bontcheva, I. Roberts, A. Vasiljevs and A. Lagzdiņš, Towards an Interoperable Ecosystem of AI and LT Platforms: A Roadmap for the Implementation of Different Levels of Interoperability, in: *Proceedings of the 1st International Workshop on Language Technology Platforms*, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 96–107. ISBN 979-10-95546-64-1. https://www.aclweb.org/anthology/2020.iwltp-1.15.

[144] Slator, Slator 2021 Data-for-AI Market Report, Technical Report, Slator, 2021.

[145] C. Chiarcos, M. Ionov, M. Rind-Pawlowski, C. Fäth, J.W. Schreur and I. Nevskaya, LLODifying linguistic glosses, in: *Proceedings of Language, Data and Knowledge (LDK-2017)*, Galway, Ireland, 2017. doi:https://doi.org/10.1007/978-3-319-59888-8_7.

[146] H.B.-S. Sabine Tittel and C. Chiarcos, Using RDFa to Link Text and Dictionary Data for Medieval French, in: *Proceedings of the 5th Workshop on Linked Data in Linguistics (LDL-2016): Managing, Building and Using Linked Language Resources*, European Language Resources Association (ELRA), 2018. ISBN 979-10-95546-19-1.

[147] M. Curado Malta, P. Centenera and E. González-Blanco García, POSTDATA – Towards publishing European Poetry as Linked Open Data, *International Conference on Dublin Core and Metadata Applications* **16** (2016), 19–20. doi:10400.22/8564.

[148] M. Curado Malta, P. Centenera and E. Gonzalez-Blanco, Using Reverse Engineering to Define a Domain Model: The Case of the Development of a Metadata Application Profile for European Poetry, in: *Developing Metadata Application Profiles*, IGI Global, 2017, pp. 146–180. doi:10.4018/978-1-5225-2221-8. http://e-spacio.uned.es/fez/view/bibliuned:365-Egonzalez9.

[149] M. Curado Malta, H. Bermúdez-Sabel, A.A. Baptista and E. Gonzalez-Blanco, Validation of a metadata application profile domain model, *International Conference on Dublin Core and Metadata Applications* (2018), 65–75. doi:10.5281/zenodo.1441217.

[150] M. Curado Malta, Modelação de dados poéticos: Uma perspectiva desde os dados abertos e ligados, in: *Humanidades Digitales. Miradas hacia la Edad Media*, D. González and H. Bermudez Sabel, eds, De Gruyter, Berlin, 2019, pp. 24–48. ISBN 978-3-11-058542-1. doi:10.1515/9783110585421-004.

[151] E. González-Blanco, S. Ros Muñoz, M.L. Díez Platas, J. De la Rosa, H. Bermúdez-Sabel, A. Pérez Pozo, L. Ayciriex and B. Sartini, Towards an Ontology for European Poetry, DARIAH Annual Event 2019, Warsaw, Poland, 2019. doi:10.5281/zenodo.3458772. https://zenodo.org/record/3458772#.Xhw_YOhKjIV.

[152] Postdata ERC project, Network of ontologies - POSTDATA, [Online; accessed 2021-01-17]. http://http://postdata.linhd.uned.es/results/.

[153] Postdata ERC project, Postdata-core ontology, [Online; accessed 2021-01-17]. http://http://postdata.linhd.uned.es/results/.

[154] Postdata ERC project, Postdata-prosodic ontology, [Online; accessed 2021-01-17]. http://http://postdata.linhd.uned.es/results/.

[155] Postdata ERC project, Postdata-structural ontology, [Online; 2021-01-17]. http://http://postdata.linhd.uned.es/results.

[156] M.L. Díez Platas, S. Ros Muñoz, E. González-Blanco, P. Ruiz Fabo and E. Álvarez Mellado, Medieval Spanish (12th-15th centuries) Named Entity Recognition and Attribute Annotation System based on contextual information, *JASIST (Journal of the Association for Information Science and Technology)* (2020). doi:https://doi.org/10.1002/asi.24399.

[157] J. De la Rosa, S. Ros Muñoz, E. González-Blanco, Á. Pérez Pozo, L. Hernández and A. Díaz Medina, Bertsification: Language modeling fine-tuning for Spanish scansion, 4th International Conference on Science and Literature (postponed due to COVID-19 crisis), Girona, 2020. doi:https://doi.org/10.1007/s00521-021-06692-2.

[158] J. De La Rosa, S. Ros Muñoz, E. González-Blanco and Á. Pérez Pozo, PoetryLab: An Open Source Toolkit for the Analysis of Spanish Poetry Corpora, Carleton University and the University of Ottawa, Virtual Conference, 2020, DH2020. doi:http://dx.doi.org/10.17613/rsd8-we57. https://hcommons.org/deposits/item/hc:31763/.

[159] J. de la Rosa, S. Ros and E. González-Blanco, Predicting metrical patterns in Spanish poetry with language models, *arXiv preprint arXiv:2011.09567* (2020). doi:10.5281/zenodo.4314596.

[160] S. Krek, I. Kosem, J.P. McCrae, R. Navigli, B.S. Pedersen, C. Tiberius and T. Wissik, European lexicographic infrastructure (elexis), in: *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts*, 2018, pp. 881–892.

[161] P. Bański, J. Bowers and T. Erjavec, TEI-Lex0 guidelines for the encoding of dictionary information on written and spoken forms, in: *Electronic lexicography in the 21st century. Proceedings of eLex 2017*, Lexical Computing CZ s.r.o., 2017.

[162] F. Mambrini and M. Passarotti, Linked Open Treebanks. Interlinking Syntactically Annotated Corpora in the LiLa Knowledge Base of Linguistic Resources for Latin, in: *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, Association for Computational Linguistics, Paris, France, 2019, pp. 74–81. doi:10.18653/v1/W19-7808. https://www.aclweb.org/anthology/W19-7808.

[163] R. Sprugnoli, M. Passarotti, F.M. Cecchini and M. Pellegrini, Overview of the EvaLatin 2020 Evaluation Campaign, in: *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 105–110. ISBN 979-10-95546-53-5. doi:10.5281/zenodo.3819936. https://www.aclweb.org/anthology/2020.lt4hala-1.16.

[164] M. Passarotti, M. Budassi, E. Litta and P. Ruffolo, The Lemlat 3.0 Package for Morphological Analysis of Latin, in: *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, Linköping University Electronic Press, 2017, pp. 24–31. https://aclanthology.org/W17-0506.pdf.

[165] E. Litta, M. Passarotti and F. Mambrini, The Treatment of Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin, in: *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czechia, 2019, pp. 35–43. doi:10.5281/zenodo.3403022. https://www.aclweb.org/anthology/W19-8505.

[166] F. Mambrini and M. Passarotti, Harmonizing Different Lemmatization Strategies for Building a Knowledge Base of Linguistic Resources for Latin, in: *Proceedings of the 13th Linguistic Annotation Workshop*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 71–80. doi:10.5281/zenodo.3349631. https://www.aclweb.org/anthology/W19-4009.

[167] F. Mambrini and M. Passarotti, Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin, in: *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 20–28. ISBN 979-10-95546-46-7. doi:10.5281/zenodo.3862156. https://www.aclweb.org/anthology/2020.globalex-1.3.

[168] G. Franzini, F. Zampedri, M. Passarotti, F. Mambrini and G. Moretti, Græcissâre: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin., in: *Seventh Italian Conference on Computational Linguistics*, J. Monti, F. Dell'Orletta and F. Tamburini, eds, CEUR-WS.org, Bologna, 2020, pp. 1–6. doi:10.5281/zenodo.4319005. http://ceur-ws.org/Vol-2769/paper_06.pdf.

[169] A. Westerski and J.F. Sánchez-Rada, Marl Ontology Specification, V1.1 8 March 2016, 2016. http://www.gsi.dit.upm.es/ontologies/marl/.

[170] G. Franzini, A. Peverelli, P. Ruffolo, M. Passarotti, H. Sanna, E. Signoroni, V. Ventura and F. Zampedri, Nunc Est Aestimandum. Towards an evaluation of the Latin WordNet, in: *Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, R. Bernardi, R. Navigli and G. Semeraro, eds, CEUR-WS.org, Bari, Italy, 2019, pp. 1–8. doi:10.5281/zenodo.3518774.

[171] F.M. Cecchini, R. Sprugnoli, G. Moretti and M. Passarotti, UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works, in: *Seventh Italian Conference on Computational Linguistics*, CEUR-WS. org, 2020, pp. 1–7. doi:10.5281/zenodo.4319001.

[172] C. Fäth, C. Chiarcos, B. Ebbrecht and M. Ionov, Fintan – Flexible, Integrated Transformation and Annotation eNgineering, in: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC-2020)*, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 7212–7221.

[173] T. Declerck, J. McCrae, M. Hartung, J. Gracia, C. Chiarcos, E. Montiel, P. Cimiano, A. Revenko, R. Sauri, D. Lee, S. Racioppa, J. Nasir, M. Orlikowski, M. Lanau-Coronas, C. Fäth, M. Rico, M.F. Elahi, M. Khvalchik, M. Gonzalez and K. Cooney, Recent Developments for the Linguistic Linked Open Data Infrastructure, in: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, European Language Resources Association (ELRA), 2020, pp. 5660–5667.

[174] R. Saurí, L. Mahon, I. Russo and M. Bitinis, Cross-Dictionary Linking at Sense Level with a Double-Layer Classifier, in: *2nd Conference on Language, Data and Knowledge (LDK 2019)*, M. Eskevich, G. de Melo, C. Fäth, J.P. McCrae, P. Buitelaar, C. Chiarcos, B. Klimek and M. Dojchinovski, eds, OpenAccess Series in Informatics (OASIcs), Vol. 70, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2019. ISSN 2190-6807. ISBN 978-3-95977-105-4. doi:10.4230/OASIcs.LDK.2019.20.

[175] B. Lorincz, M. Nutu, A. Stan and G. Mircea, An Evaluation of Postfiltering for Deep Learning Based Speech Synthesis with Limited Data, in: *IEEE 10th International Conference on Intelligent Systems (IS)*, 2020. doi:10.1109/IS48319.2020.9199932.

[176] R. Ion, Teprolin: an Extensible, Online Text Preprocessing Platform for Romanian, in: *Proceedings of the ConsILR-2018*, 2018, pp. 69–76.

[177] A.L. Georgescu, H. Cucu, A. Buzo and C. Burileanu, RSC: A Romanian Read Speech Corpus for Automatic Speech Recognition, in: *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC-2020)*, European Language Resources Association (ELRA), 2020, pp. 6606–6612.

[178] D. Cristea, I. Pistol, S. Boghiu, A. Bibiri, D. Gifu, A. Scutelnicu, M. Onofrei, D. Trandabat and G. Bugeag, CoBiLiRo: a Research Platform for Bimodal Corpora, in: *Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020)*, European Language Resources Association, 2020, pp. 22–27.

[179] D. Gifu, A. Moruz, C. Bolea, A. Bibiri and M. Mitrofan, The Methodology of Building CoRoLa, in: *Revue Roumaine de Linquistique (Romanian Review of Linguistics)/ On design, creation and use of of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo / Conception, création et utilisation du Corpus de Référence du Roumain Contemporain et de ses outils d'analyse. CoRoLa, KorAP, DRuKoLA et EuReCo*, Vol. 64, 2019, pp. 241–253.

[180] A. Li and Z. Yin, Standardization of Speech Corpus, in: *Data Science Journal*, Vol. 6, 2007. doi:https://-doi.org/10.2481/dsj.6.S806.

[181] V. Presutti, E. Blomqvist, E. Daga and A. Gangemi, Pattern-based ontology design, in: *Ontology Engineering in a Networked World*, Springer, 2012, pp. 35–64. doi:https://-doi.org/10.1007/978-3-642-24794-1_3.

Appendix A. The OntoLex-Lemon Model

In order to make the current paper as self-contained as possible, we have decided to include a brief introduction to the OntoLex-Lemon model which constitutes the current appendix. In what follows, we will start by describing the core module of OntoLex-Lemon using an example entry. We will then describe each of its different submodules by developing this example entry. The full guidelines (with additional examples) can be found at: https://www.w3.org/2016/05/ontolex/.

Note that this appendix only covers the very basics and can be skipped by those who already have some familiarity with the OntoLex-Lemon model.

A.1. Introduction and the Core Module

OntoLex-Lemon is (as mentioned in the article itself) an RDF-native model for the modelling and publication of ontology-lexica (that is, language resources that consist both of a lexical and an ontological component) on the Semantic Web. It is an update of the original *lemon* model and, as with this latter model, it aims to enrich or ground linked data ontologies with linguistic information. It has however increasingly come to be used for the modelling and publication of linked data lexica that do not happen to contain any ontological component (and indeed extensions of OntoLex-Lemon such as lexicog 5.1.1 and FrAC 5.1.3 are strongly motivated by lexical use-cases rather than ontological ones).

The model consists of a *core module* and four additional thematic modules, three of which we will describe in the following sections (the metadata module *lime* is described in Section 5.3.3). In the current section, we will look at the core module, presented in Figure 8, with the help of an example. As the figure shows, this module is essentially organised around the class Lexical Entry and its various different properties and relationships; Lexical Entry has the subclasses Word, Multiword Expression and Affix.

We can associate information about the different forms that a Lexical Entry can have (including its lemma or canonical form as well as other morphological variants) via the Form class; the former is linked to the latter via lexicalForm and its subproperties. This class can be associated with written string representations via the representation property and its subproperties.

The semantics of a Lexical Entry can be described using the Lexical Sense and the Lexical Concept classes (a fuller predicate based description of the semantics of an entry can be found in the Syntax and Semantics module which is introduced in Section A.2). The class Lexical Entry is related to Lexical Sense via the sense property (and its inverse isSenseOf). Members of the Lexical Sense class represent:

> the lexical meaning of a lexical entry when *interpreted as referring to the corresponding ontology element*.A link between a lexical entry and an ontology entity via a Lexical Sense object *implies that the lexical entry can be used to refer to the ontology entity in question*[250]. (emphasis ours)

The Lexical Sense object relates to a corresponding ontology entity via the reference property; Lexical Entry individuals can be related directly to these ontology entities via the property denotes[251]. A second class which is used to describe the semantics of an entry is the Lexical Concept class. Members of the latter class represent "a mental abstraction, concept or unit of thought that can be lexicalized by a given collection of senses"[252]; Lexical Concept is a subclass of the SKOS class Concept.

We will demonstrate how some of these classes and properties are used in practise with an example entry from Wiktionary, the open-source online multilingual dictionary. We will take the example of the Urdu word زبان (*zuban*) which means both 'tongue' and 'language'[253]; a screenshot of the entry can be seen in Figure 9.

The entry contains some standard morpho-syntactic and semantic information about the word quite a lot of which can already be encoded using the OntoLex-Lemon core module, along with a select number of properties from the LexInfo vocabulary (described above in Section 4.5).

Figure 10 shows the encoding of زبان as a OntoLex-Lemon Word. Note here the use of LexInfo properties partOfSpeech and gender to specify the part of speech and gender of the word respectively and the use of the *lime* property language to specify the language of the entry. We can also see an example of the use of two sub-properties of the lexicalForm property to link the entry to its different variant forms: with the canonicalForm property linking the entry to its headword form and the otherForm property linking it to its different morphological variants. Furthermore we see the use of the OntoLex-Lemon sense property to link the entry to the two senses mentioned in the original Wiktionary entry, encoded in the example as زبان_sense1 and زبان_sense2 respectively. Both of these senses link to DBpedia entries (in accordance with the original Wiktionary entry) and the OntoLex-Lemon denotes

[250] Definition taken from https://www.w3.org/2016/05/ontolex/#lexical-sense-reference.

[251] The property denotes is equivalent to the property chain sense o reference.

[252] Definition taken from https://www.w3.org/2016/05/ontolex/#lexical-concept.

[253] The entry can be found here https://en.wiktionary.org/wiki/%D8%B2%D8%A8%D8%A7%D9%86#Urdu.

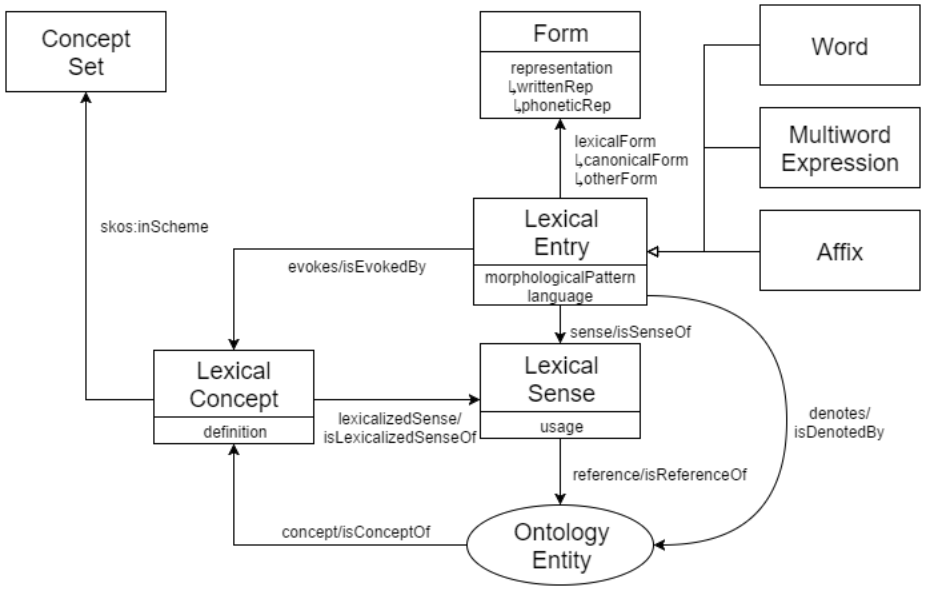

Fig. 8. The OntoLex-Lemon Core.

Urdu [edit]

Etymology [edit]

From Persian زبان (*zobân, zabân*).

Pronunciation [edit]

- IPA(key): /zʊ.bɑːn/, /zə.bɑːn/

Noun [edit]

زبان • (zubān, zabān) *f* (*Hindi spelling* ज़बान)

1. tongue (body part)

   Synonyms: جیبھ (jībh), لسان (lisān)

2. language

   Synonyms: لسان (lisān), بھاشا (bhāśā), بولی (bolī)

Declension [edit]

Declension of زبان		[less ▲]
	singular	plural
direct	زُبان (zubān)	زُبانیں (zubānẽ)
oblique	زُبان (zubān)	زُبانوں (zubānõ)
vocative	زُبان (zubān)	زُبانو (zubāno)

Derived terms [edit]

- مادری زبان (mādrī zabān)

Fig. 9. زبان *(zuban)*

```
:زبان a ontolex:Word;
  lexinfo:gender lexinfo:feminine;
  lexinfo:partOfSpeech lexinfo:noun;
  lime:language "ur"^^xsd:language;
  ontolex:canonicalForm :زبان_lemma;
  ontolex:denotes <http://dbpedia.org/resource/Language>,
                  <http://dbpedia.org/resource/Tongue>;
  ontolex:otherForm :زبان_dir_pl,
                    :زبان_obl_pl,
                    :زبان_obl_sing,
                    :زبان_voc_pl,
                    :زبان_voc_sing;
  ontolex:sense :زبان_sense1, :زبان_sense2 .
```

Fig. 10. The Word زبان.

property links the entry straight to these two DBpedia entries.

In Figure 11 we show two forms of the entry, the direct plural form of the noun, along with a sense, the second sense of the word meaning ‘language’. Note here the use of the OntoLex-Lemon writtenRep property to associate a Form element to the Roman and Arabic script versions of its orthography as well as the use of OntoLex-Lemon property reference to link the sense of the word to its ontological reference.

A.2. Syntax and Semantics

The next module we will look at is the *Syntax and Semantics* module of OntoLex-Lemon, often shortened to *synsem*. This module contains a number of classes and properties for modelling the syntactic behaviour of words and the relationship(s) between this behaviour and the semantic properties of those words. Figure 12 presents these classes and properties in diagrammatic form.

The two classes used to model syntactic behaviour are Syntactic Frame[254] and Syntactic Argument[255]. These can be related to corresponding ontological predicates via the OntoMap class and various other OntoLex-Lemon object properties. To describe these mechanisms here would take us too far beyond our purpose in writing this brief introduction: the interested reader is invited to consult the OntoLex-Lemon guidelines for a fuller description.

A.3. Decomposition

The OntoLex-Lemon *Decomposition* module (*decomp* for short), represented in Figure 13, is intended to indicate "which elements constitute a multiword or compound lexical entry"[256]. It does this via the Component class and the properties subterm, constituent, and correspondsTo.

For instance, to return to our earlier example, take the derived term مادر زبان *madri zuban* listed in the Wiktionary entry for زبان. This expression literally means ‘mother tongue’ and can also be translated as ‘native language’. In our example encoding we class this as a MultiWordExpression that can be decomposed into two subterms using the decomp property subterm. Namely it can be decomposed into the two entries زبان and مادر, as in Figure 14.

As always, further examples and details of the other properties can be found in the OntoLex-Lemon guidelines.

A.4. Variation and Translation

Finally in this appendix we will look at the *Variation and Translation* module (also known as *vartrans*). This module is concerned with representing "relations between lexical entries and lexical senses that are variants of each other." More precisely these entries can

[254]This "represents the syntactic behavior of an open class word in terms of the (syntactic) arguments it requires". See https://www.w3.org/2016/05/ontolex/#syntactic-frames

[255]This "represents a slot that needs to be filled for a certain syntactic frame to be complete." See https://www.w3.org/2016/05/ontolex/#syntactic-frames.

[256]https://www.w3.org/2016/05/ontolex/#decomposition-decomp

```
:زبان_dir_pl a ontolex:Form;
  lexinfo:number lexinfo:plural;
  ontolex:writtenRep "zubānẽ", "زُبانیں"@ur .

:زبان_sense2 a ontolex:LexicalSense;
  ontolex:reference <http://dbpedia.org/resource/Language> .
```

Fig. 11. A form and a sense belonging to زبان.

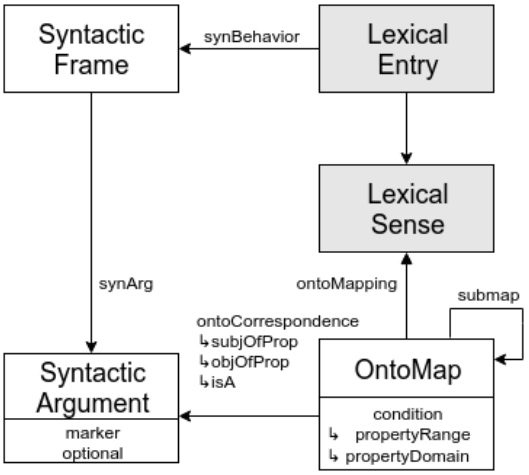

Fig. 12. The OntoLex-Lemon Synsem Module.

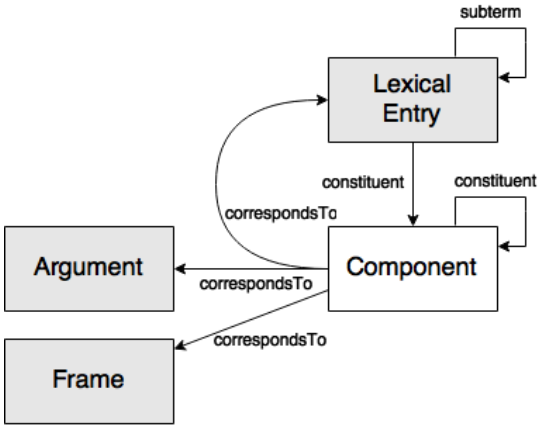

Fig. 13. The OntoLex-Lemon Decomposition Module.

be variants via lexico-semantic relations such as hyponymy and synonymy or because they are translations. The classes and properties of the vartrans module are shown in Figure 16. As the figure shows, the vartrans module reifies lexico-semantic relations between words, defining a class Lexico-semantic Relation with subclasses, Lexical Relation and Sense Relation .

We will illustrate the use of this module with two examples from our running زبان example .In the first case we relate the word زبان to one of its synonyms,

```
:MWE_مادری_زبان a ontolex:MultiWordExpression;
  lime:language "ur"^^xsd:language ;
  decomp:subterm :زبان,
                 :مادری .
```

Fig. 14. Multi-word Expression Example.

بهاشا *bhasa* 'language' – or rather we relate the sense of the word meaning 'language' to a sense of another word which means the same thing. This is represented in Fig 15 using vartrans' classes and properties. We start by specifying this relation as an instance of the class Sense Relation. Next we use the LexInfo class synonym to categorise thie relation (using the vartrans property category) and specify its source (the first sense of *bhasa*) and target (the second sense of our example word) using the vartrans source and target object properties respectively.

In Fig 16 on the other hand we represent a translation relation between the second sense of the word and the first sense of the English word *language*. This is done in much the same way as in the previous example, except that in this case the relation is one of direct equivalence and we use the property category and a relevant term from a terminological vocabulary.

```
:senseRelation a vartrans:SenseRelation;
  vartrans:category lexinfo:synonym;
  vartrans:source :بهاتسا_sense1;
  vartrans:target :زبان_sense2 .
```

Fig. 15. Example of the use of vartrans.

```
:translationRelation a vartrans:Translation;
  vartrans:category <http://purl.org/net/translation-categories#directEquivalent>;
  vartrans:source :زبان_sense2;
  vartrans:target :language_sense_1 .
```

Fig. 16. Example of the use of vartrans.

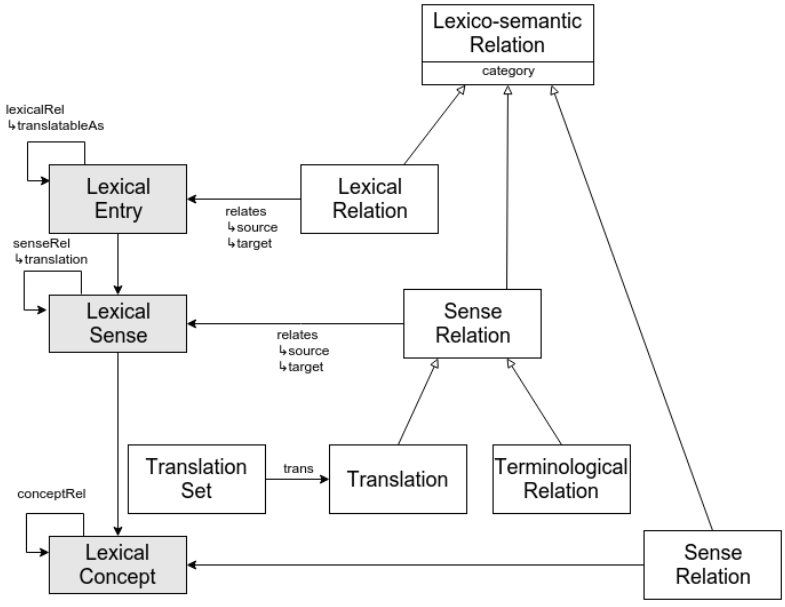

Fig. 17. The OntoLex-Lemon Vartrans Module.