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

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

[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
[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
[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] *CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milan, Italy*
E-mail: marco.passarotti@unicatt.it
[m] *Laboratory of Innovation on Digital Humanities, National Distance Education University UNED, Spain*
E-mail: sros@scc.uned.es
[n] *Department of Computer Science and Engineering, Faculty of Automatic Control and Computers, University Politehnica of Bucharest, Romania*
E-mail: ciprian.truica@upb.ro

Abstract.

1570-0844/0-1900/$35.00 © 0 – IOS Press and the authors. All rights reserved

This article provides an up-to-date and comprehensive survey of models (including vocabularies, taxonomies and ontologies) used for representing linguistic linked data (LLD). It focuses on the latest developments and both builds upon and complement previous works covering similar territory. The article begins with an overview of recent trends which have had an impact on linked data models and vocabularies, such as the growing influence of the FAIR guidelines, the funding of several major projects in which LLD is a key component, and the increasing importance of the relationship of the Digital Humanities with LLD. Next, we give an overview of some of the most well known vocabularies and models in LLD. After this we look at some of the latest developments in community standards and initiatives such as OntoLex-lemon as well as recent work which has been in carried out in corpora and annotation and LLD including a discussion of the LLD metadata vocabularies METASHARE and *lime* and language identifiers. Following this we look at work which has been realised in a number of recent projects and which has a significant impact on LLD vocabularies and models.

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

1. Introduction

The growing popularity of linked data, and especially as linked open data (that is, linked data with an open license) as a means of publishing language resources (lexica, corpora, data categories, etc) has led to the need for a greater focus on models for linguistic linked data (LLD) since these are key to what makes linked data resources so reusable and interoperable. The purpose of this article is to provide an up-to-date and comprehensive survey of models (including vocabularies, taxonomies and ontologies) used for representing linguistic linked data. It will focus on the latest developments and both build upon and complement previous works covering similar territory, avoiding too much repetition and overlap with the latter. In the following Section 2, we give an overview of a number of trends from the last few years which have had, or which are likely to have, a significant impact on the definition and/or use of LLD models. We relate these trends to the rest of the article by highlighting relevant sections of the article (in bold). This overview of trends will help to locate the present work within a wider research context, something that is extremely useful in an area as active as linguistic linked data, as well as assisting readers in navigating the rest of the article. Next, in Section 2.4, we compare the present article with other related work, including an earlier survey of LLD models, in order to help clarify the topics and approach of the present work. Section 3 gives an overview of the most widely used models in LLD. Then in Section 4, we look at recent developments in community standards and initiatives. These include the latest extensions of the OntoLex-Lemon model in Section 4.1, a discussion of relevant work in copora and annotations in Section 4.2, and a section on metadata Section 4.3. Finally there is a section discussing projects, Section 5, and the conclusion, Section 6.

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

The trends we have decided to focus on in this overview are the FAIRification of data in **Section 2.1**, the importance of projects to LLD models in **Section 2.2**, and finally the increasing importance of Digital Humanities use cases in **Section 2.3**.

2.1. FAIR New World

With the growing importance of Open Science initiatives, and especially those promoting the FAIR guidelines (where FAIR stands for Findable, Accessible, Interoperable and Reusable) [1] – and the consequent emphasis on the modelling, creation and publication of language resources as FAIR digital resources – shared models and vocabularies have begun to take on an increasingly prominent role. Although the linguistic linked data community has been active in promoting shared RDF vocabularies and models for years, this new emphasis on FAIR is likely to have a considerable impact in several ways, not least in terms of the necessity for these models to demonstrate a greater coverage, and to be more interoperable one with another. We will look at one series of FAIR related recommendations for models in Section 3 and see how they might be applied to the case of LLD. However in the rest of the 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 well known FAIR principles, the authors clearly state that the criteria proposed by these 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 a "digital research object"
- Its usefulness with respect to tasks to be carried out
- Its usability especially with respect to licensing issues, 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 and research funding bodies, such as the European Commission[1], through several national and international initiatives reflects a wider recognition of the potential of structured and machine actionable data in changing how research is carried out, and especially in helping to support open science practices. The FAIR ideal, in short, is to allow machines as much autonomy as possible in working with data, by the expedient of rendering as much of the semantics of that data explicit (and machine actionable) as possible.

Publishing data using a standardised data model like the Resource Data Framework (RDF) which was specifically intended to facilitate interoperability and interlinking between datasets – along with the other standards proposed in the Semantic Web stack and the technical infrastructure which has been developed in order to support it – obviously goes a long way towards facilitating the publication of datasets as FAIR data. In addition, however, it is also essential that there exist resources that can play the role of domain-specific vocabularies/terminologies/models and data category registries in order to ensure a maximal level of interoperability and re-usability of data produced in a given domain. Such artefacts serve to describe the shared theoretical assumptions held by a community of experts with regard to the semantics of the terms used by that community, and do so in a form that (to varying degrees) is machine readable to computational agents (more so than raw text anyway). The following FAIR principles are especially salient here:

- 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 that the emphasis placed on machine actionability in FAIR resources (that is, with respect to allowing computational agents to take "appropriate action" with respect to a dataset or resource) gives Semantic Web vocabularies/registries a substantial advantage over other (non-Semantic Web native) standards like the Text Encoding Initiative (TEI) guidelines[2] [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, knowledge representation language for describing their semantics in a machine readable way, at least nothing as powerful as the Web Ontology Language (OWL)[3] or the Semantic Web Rule Language (SWRL)[4] (there is no standardised way of describing the meanings of morphemes, lexemes, lemmas, etc. in TEI in a machine actionable way).

The ability to give precise, axiomatic definitions of terms in a formal knowledge representation (KR) language is especially useful in fields like linguistics or literary scholarship, where there can be quite different definitions of the same or similar core concepts with respect to different scholarly traditions or schools of thought. Using a machine readable description in OWL (allied perhaps to a more human readable description given as a string) we can clarify what we mean when we use a concept like 'Sense' or 'Morpheme' in a dataset (even if, given the expressive limitations of OWL, we can't exhaustively describe how we're using those concepts); we can also describe the different kinds of relationship between concepts across languages or different schools of thought.

Secondly, thanks to the use of a shared data model and a powerful native linking mechanism, linguistic linked data datasets can be easily (and in a standard way) integrated with/enriched by (linked data) datasets belonging to other disciplines, for instance geographical and historical datasets or gazetters and authority

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

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

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

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

lists. OWL, and vocabularies, such as PROV-O[5], also allow us to add information pertaining to when something happened, or whether we are describing a hypothesis or not (in which case, also who made it and when). Once again all of these things can be described in a way that makes the semantics of the information (relatively) explicit and machine actionable through the use of pre-existing standards and technologies including the Semantic Web query language *SPARQL Protocol and RDF Query Language* (SPARQL) as well as freely available Semantic Web reasoning engines.

Moreover the pursuit of the FAIR ideal has opened the way to new means of publishing datasets which offer enhanced opportunities for the re-use of such data in an automatic or semi-automatic way. These include for instance *nanopublications*, *cardinal assertions* and *knowlets*.[6] The potential of these new publishing approaches for discovering new facts as well as for comparing concepts and tracking how single concepts change are well described in [5].

The field of language resources offers us a rich array of highly structured kinds of datasets, structured according to a series of widely shared conventions (this is what makes the definition of models and vocabularies for lexica, corpora, etc, so viable in the first place) – 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 as well as to the new data publication approaches previously mentioned. However, the better and more expressive the underlying models are the more effective they will be.

In order to ensure the continued effectiveness of linked data and the Semantic Web in facilitating the creation of FAIR resources, it is vital that pre-existing vocabularies/models/data registries be re-used whenever possible in the modelling of user data; this of course also means ensuring that these models have sufficient coverage and defining extensions when this is not the case, as well as creating training materials suitable for different groups of uers. Part of the intention of this article, together with the foundational work carried out in [6], is to provide an overview of what exists out there in terms of LLD-focused models, to look at the areas which are receiving most attention in order to highlight those which are so far underrepresented. In addition in **Section 3** we look at the most well known LLD models in the light of a recent series of recommendations on the publication of models as FAIR resources.

2.2. The Importance of Projects in LLD

One significant indicator of the success which LLD has had as a means of modelling and publishing data in the last few years is the variety of new projects which have been funded and which have included the publication of linguistic datasets as linked data among their core themes. These include supranational projects, notably European H2020 projects, ERCs and COST actions, as well as projects which have been carried out at both the national and regional levels. This success reflects a more general recognition (both in academia and in industry) of the importance of linked data as a way of making language resources openly available to the research community and to a wider public, as well as demonstrating the continuing maturation of the field. The impact which these recent projects have had in terms of influencing the use as well as the overall perception of LLD across a series of different communities is interesting for the purposes of this article from at least two different points of view.

Firstly, in many cases these projects have driven the definition of new LLD models and/or extensions to existing models, not least by providing numerous new use cases which had not initially been taken into consideration. An analysis of the broader trends which these projects reveal can help us to predict the ways in which models for LLD will evolve in the next few years, reflecting the movement from an initial, formative, period in which the emphasis was more on the definition of 'foundational' vocabularies as building blocks, to one in which the need is more for specialisation. The need for such specialisation is obvious, however the ways in which this specialisation will actually occur are not so obvious given the specific expressive affordances of the RDF data framework. Secondly, these projects also give us clear examples of the use of these vocabularies 'in the wild' so to speak and exemplify their use in a larger number of medium to large scale datasets than was common just a few years

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

[6] *Nanopublications* are defined as the "smallest possible machine readable graph-like structure that represents a meaningful assertion" [5] 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 [5] 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.

ago. A detailed overview of the current situation as regards research projects and LLD, along with extended descriptions of those projects which are – in our opinion – the most significant from the point of view of LLD models and vocabularies is given in **Section 5**.

Note, however, that although the projects which we will discuss in that section have been important, and in some cases have set the agenda, for the development of numerous LLD models and vocabularies, a lot of the actual work on the definition of these resources was carried out – and is being carried out – within community groups and mailing lists such as the W3C OntoLex group, groups whose membership is often open to all (rather than being limited to members of a project or experts nominated by a standards body that is)[7]. This obviously allows for the participation of a wider range of stakeholders as well as the consideration of a greater number of use-cases than otherwise. We include an update on community standards and initiatives in **Section 4**.

2.3. The Relationship of LLD to the Digital Humanities

Several of the projects which we will discuss in this article are either primarily concerned with the Digital Humanities (DH) or at least deal with a number of topics which are more traditionally associated with the former. This is the third major trend which we want to highlight here, since it represents a move away (or rather a branching off) from LLD's beginnings in Computational Linguistics and Natural Language Processing (although these two still perhaps represent the majority of applications of LLD), one that calls for something of a shift in emphases.

This overlap between LLD and DH is particularly clear in the modelling of corpora annotation (**Section 4.2**) and in support for lexicographic use cases (see **Section 4.1.1** and **Section 5.7**). Indeed one obvious example of these shared concerns is the publication of retro-digitised dictionaries as LLD lexica (a major theme of the ELEXIS project, see **Section 5.7**). The latter 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, as well as those aspects which pertain to it as a *physical text* which is being represented in digital form. In the latter case, this includes the representation of (elements of) its *form*, i.e., its structural layout and overall visual appearance[8]. In fact, as we discuss in our description of the OntoLex Lexicography module in **Section 4.1.1**) even the structural division of lexicographic works into textual units such as entries and senses is not always isomorphic to the representation of the lexical content of those units using OntoLex-lemon classes such as LexicalEntry and LexicalSense. We may also wish to model different aspects of the history of the lexicographic work as physical text. For example, in the case of older resources, annotating instances where the content has been superseded by subsequent scholarly work. Or 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[9]. All of which calls for a much richer provision of metadata categories than was previously considered for LLD lexica, both at the level of the whole work as well as at the level of the entry. It also requires the capacity to model salient aspects of the same artefact or resource at different levels of description (somehthing which is indeed offered by the OntoLex Lexicography module, see Section 4.1.1). We discuss metadata challenges in humanities use cases in **Section 4.3.4**.

Retrodigitized dictionaries also reveal the potential for the integration of historical and geographical with lexicographic and linguistic information; this is something which LLD is well placed to do (using appropriately defined classes and properties). Indeed the use made of linked data in DH projects such as Pelagios [8], and Mapping Manuscript Migrations [9] and in the Finnish Sampo datasets [10], among others, strongly testifies to the facility with which heterogeneous humanities data can be modelled in linked data.

An additional series of challenges arises in the consideration of resources for classical and historical languages, or indeed, historical stages of modern languages; and though many of these challenges are not specific to such languages, they do arise most clearly

[7] In fact there is a strong reliance on projects in these community groups since many of their active participants are involved in funded projects. This might suggest a different organisational modality in the future in order to ensure continuity and long term preservation, perhaps in collaboration with an infrastructure like CLARIN or DARIAH.

[8] 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

[9] This was in fact one of the motivations behind the Nénufar project, described in [7].

when we are working with them. So that in the case of lexical resources for historical languages we often come up against the necessity of having to model attestations which sometimes cite reconstructed texts, as well as the desirability of being able to represent different scholarly and philological hypotheses for instance when it comes to modelling etymologies.

The LiLa project [11] (see **Section 5.8** for a more detailed description) provides a good example of the challenges and opportunities of adopting the LLD model to represent linguistic (meta)data of both lexical and textual resources for a classical language (Latin). LiLa aims to engage a wide and diverse scholarly community, which includes linguists, philologists, historians and archaeologists, as well as language learners. In a field like Classics, that relies on a lengthy tradition of scholarship, and is now emphasising the connections between different ancient civilizations, their languages and cultures, LLD offers a powerful and effective solution to the challenges of making information about linguistic and archaeological data from all possible sources (attestation in texts, lexica, thesauri, modern studies, etc.) findable and interoperable.

The relationship between notions such as *word* from the lexical/linguistic and the philological points of view and, more broadly speaking, the relationship between linguistic and philological annotations of text is a topic which is just starting to gain attention within the context of LLD. In particular it is being studied in projects such as the aforementioned LiLa as well as POSTDATA. The latter is another important project which straddles the intersection of LLD and DH and which is discussed in more depth in **Section 5.5**. Other relevant projects described below include The Machine Translation and Automated Analysis of Cuneiform Languages (MTAAC) project (described in **Section 5.6**) and the Text Database and Dictionary of Classic Mayan project (TWKM) (**Section 5.3**)

A full discussion of developments in corpus and annotation from the perspective of LLD is given in **Section 4.2**, including a discussion of the relationship between TEI/XML and RDF in Section 4.2.3. We discuss ongoing work in the modelling of the attestation relation between lexica and textual corpora (along with other kinds of relevant relationships between the two kinds of resource) in **Section 4.1.3**.

2.4. Related Work

The current work is intended both to complement as well as to update a previous general survey on models for representing LLD published by Bosque-Gil et al in 2018 [6]. Although we are now only two years on from the publication of that article, we feel that enough has happened in the intervening time period to justify a new survey article. In addition our focus is also quite different from that of the previous article.

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 however is more focused on the use of LLD vocabularies in practise and on their availability (this is very much how we have approached the survey in **Section 3**). Moreover, the present article includes a detailed discussion of recent work in the use of LLD models and vocabularies in corpora and annotation, **Section 4.2**, as well as an extensive section on metadata, **Section 4.3**, neither of which were given the same detailed level of coverage in [6]. Additionally we would also like to single out the following two initiatives which were not covered in [6] because they had not yet gotten underway:

- The development of new OntoLex-Lemon modules for morphology Section 4.1.2 and frequency, attestations, and corpus Information, described in **Section 4.1.3**
- An important new initiative in aligning LLD vocabularies for corpora and annotation, described in **Section 4.2.7**.

In what follows we will assume that the reader already has some grounding in linked data in general, including familiarity with the Resource Data Framework (RDF), RDF Schema (RDFS) and the Web Ontology Language (OWL), and linguistic linked data in particular. The recently published *Linguistic linked data: representation, generation and applications* [12] will however give the interested reader 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* [13] which aims to describe major developments since 2015, consisting mostly of position papers by researchers from linguistics and the language resource community.

3. LLD Models: An Overview

The current section will give an overview of some of the most well known and widely used models and vocabularies in LLD. An account of some the latest developments with regards to these models can be found in Section 4. We will classify each of the models described in this section according to the scheme given in the linguistic LOD cloud diagram[10] (described in [14]), namely:

- Corpora (and Linguistic Annotations)
- Lexicons and Dictionaries
- Terminologies, Thesauri and Knowledge Bases
- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases

Under each category below we list the most prominent LLD models/vocabularies that were either originally designed to help encode that kind of dataset or have been widely appropriated for that end. For instance, the OntoLex-Lemon model falls under *Lexicons and Dictionaries* since it was initally conceived as a means of enriching ontologies with lexical information in the form of computational lexica and was then widely used for encoding linked data lexica in general, although it can also be used for modelling and publishing other kinds of datasets. This is aside from the category *Linguistic Data Categories* under which we will list linguistic data categories (as opposed to vocabularies for encoding linguistic data categories).

As mentioned above our work is intended, in large part, to be an update of a previous survey paper, [6]. Since that paper gives a detailed description of the models mentioned in this section (with an ample use of illustrative examples) we will not delve into the contents of the models here. Instead we will describe them on the basis of a number of criteria many of which are related to their status as FAIR resources. In particular, we will refer to a recent draft survey on FAIR Semantics [15], the result of a dedicated brainstorming workshop of the FAIRsFAIR project[11]. This report outlined a number of recommendations and best practices for FAIR semantic artefacts where these are defined as "machine -actionable and -readable formalisation[s] of a conceptualisation enabling sharing and reuse by humans and machines" (the term includes: taxonomies, thesauri, ontologies).

In the remaining section we will focus on the following recommendations, selected on the basis of their salience to the set of models and vocabularies under discussion (with justifications for recommendations based on those given in [15]):

- *(P-Rec 2)* Ensure there is a separate URI for the metadata and that they are published separately; this helps in making the resource more findable and supports the extraction of this metadata.
- *(P-Rec 4)* Publish semantic artefacts and their contents in a semantic repository: in order to be able to exploit repository technologies for findability and re-use of semantic artefacts ;
- *(P-Rec 6)* Retrievability through search engines ;
- *(P-Rec 10)* Use a foundational ontology to align semantic artefacts (this enhances re-usability);
- *(P-Rec 13)* Create documented crosswalks and bridges
- *(P-Rec 16)* Ensure clear licensing of semantic artefacts.

The recommendations (P-Rec 2), (P-Rec 4), and (P-Rec 10) have not been followed by any of the models/vocabularies which we look at below, but they would greatly help to make these resources (and the datasets they help to encode) more FAIR. We regard the carrying out of these recommendations as desirable future objectives[12]. On the other hand, as we will see, several of the models mentioned do exist on the Linked Open Vocabulary (LOV)[13] search engine [16] and the DBpedia archivo ontology archive[14]. Note that the LOV site provides a list of criteria for inclusion on their search engine [17][15]. In 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/ we will point this out as it enhances the re-usability of those resources.

In terms of crosswalks, we can mention ongoing work on a TEI-Lex0/OntoLex-Lemon crosswalk described in Section 5.7.

Every one of the models which we will look at is an OWL ontology. We will also list the other vocab-

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

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

[12] The adoption of foundational ontologies, for instance, might help to alleviate some of the problems raised by the proliferation of independently developments as described in [6].

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

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

[15] https://lov.linkeddata.es/Recommendations_Vocabulary_Design.pdf

ularies which they make use of (aside from OWL, RDF, and RDFS that is). These include the well known ontologies/vocabularies: XML Schema Definition[16] (XSD); the Friend of a Friend Ontology[17] (FOAF); the Simple Knowledge Organisation System[18] (SKOS); Dublin Core[19] (DC); Dublin Core Metadata Initiative (DCMI) Metadata Terms[20]; the Data Catalog Vocabulary[21] (DCAT), described also in Section 4.3; the PROV Ontology[22] (PROV-O).

In addition the table also mentions the following vocabularies.

- Activity Streams[23](AS): a vocabulary for activity streams.
- GOLD: an ontology for describing linguistic data, which is described in Section 3.5.
- MARL[24]: a vocabulary for describing and annotating subjective opinions.
- ITSRDF[25]: an ontology used within the Internationalization Tag Set.
- The Creative Commons vocabulary[26] (CC).
- VANN[27]: a vocabulary for annotating vocabulary descriptions
- SKOS-XL[28]: an extension of SKOS with extra support for "describing and linking lexical entities".

Table 1 gives a summary of the models we will discuss. We classify each semantic artifact on the basis of which other models it re-uses, as well as providing license and version information.

3.1. Vocabularies and Models for Corpora and Linguistic Annotations

Linguistic annotation, e.g. for digital editions, corpora, and linking texts with external resources has long been addressed in the context of RDF and linked data. Coexisting with relational databases, XML-based formats (most notably, TEI, see 4.2) or simply text-based formats, RDF-based annotation models have been steadily developing and being used in research and industry. Currently, there are two primary RDF vocabularies widely used for text annotations:**NLP Interchange Format** (NIF), used mostly in language technology and **Web Annotation**, formerly known as *Open Annotation* (abbreviated here as OA), used in digital humanities, life sciences and bioinformatics. Both models have their advantages and shortcomings, and a number of proposals to extend have been voiced. Most importantly, there is a need for synchronization between the two. Both are available on LOV[29] and archivo[30] (the NIF core in the case of NIF[31]). The Web Annotation model, although it is covered by a W3C software and document notice and license, does not express this information in the form of triples in the resource metadata; NIF on the other hand does express licensing information as machine actionable metadata.

More details about both models and their recent developments are described in Section 4.2.

3.2. Lexicons and Dictionaries

The most well known model for the creation and publication of lexica and dictionaries as linked data is the **OntoLex-Lemon model**[32] [19], an output of the W3C ontolex working group which manages its ongoing development and further extension (see Section 4.1). It is based on a previous model, the **LExicon Model for ONtologies** (*lemon*) [18], which was itself strongly influenced by a number of other previous initiatives (including notably LMF).

OntoLex-Lemon, like its predecessor *lemon*, was designed with the intention of enriching ontologies with linguistic information and not of modelling dictionaries and lexicons *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 (including, for instance, retrodigitized dictionaries and wordnets) in general. Resources which have been modelled using OntoLex-Lemon in-

[16] https://www.w3.org/TR/xmlschema-0/
[17] http://xmlns.com/foaf/spec/
[18] https://www.w3.org/2004/02/skos/
[19] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[20] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[21] https://www.w3.org/TR/vocab-dcat-2/)
[22] https://www.w3.org/TR/prov-o/
[23] https://www.w3.org/TR/activitystreams-vocabulary/
[24] http://www.gsi.dit.upm.es/ontologies/marl/
[25] https://www.w3.org/TR/its20/
[26] https://creativecommons.org/ns
[27] https://vocab.org/vann/
[28] https://www.w3.org/TR/skos-reference/skos-xl.html

[29] https://lov.linkeddata.es/dataset/lov/vocabs/nif and https://lov.linkeddata.es/dataset/lov/vocabs/oa
[30] http://archivo.dbpedia.org/info?o=http://www.w3.org/ns/oa
[31] http://archivo.dbpedia.org/info?o=http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core
[32] 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/.

Summary				
Name	**Other Vocabularies/Models Used**	**LLO Category**	**Licenses**	**Versions** (at time of writing 28/01/21)
OntoLex-Lemon	CC, DC, FOAF, SKOS, XSD	Lexicons and Dictionaries	CC0 1.0	Version 1.0, 2016 (but this is closely based on the prior *lemon* model [18])
MMoOn	DC, FOAF, GOLD, LexVo, Ontolex-Lemon, SKOS, XSD	Terminologies, Thesauri and KBs (Morphology)	CC BY 4.0	Version 1.0, 2016
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"
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
META-SHARE	CC, DC, DCAT, FOAF, SKOS, XSD	Linguistic Resource Metadata	CC BY 4.0	Version 2.0 (pre-release)
OLiA	DCT, FOAF, SKOS	Linguistic Data Categories	CC-BY-SA 3.0	Version last updated 27/02/20
Lexinfo	CC, Ontolex, TERMS, VANN	Linguistic Data Categories	CC BY 4.0	Version 3.0, 14/06/2014
LexVo	FOAF, SKOS, SKOSXL, XSD	Typological Databases	CC BY-SA 3.0	Version 2013-02-09

Table 1
Summary of LLD vocabularies

clude: the LLD version of the Princeton Wordnet[33], DBnary (the linked data version of Wiktionary) [20], and the massive multilingual knowledge graph Babelnet [21]. The OntoLex-Lemon model is modular and consists of a core module along with modules for *Syntax and Semantics*[34], *Decomposition*[35], and *Variation and Translation*[36], as well as a dedicated metadata module, *lime*[37] (described in Section 4.3.2).

OntoLex-lemon is not currently available on LOV but its predecessor *lemon* is[38]. Three of its modules are available on archivio, the core, the *lime* metadata module[39] and the Variation and Translation module[40].

All of the OntoLex modules have their licenses (CC0 1.0) described with RDF triples using the CC vocabulary[41] with a URI as an object. Version information is described using owl:versionInfo.

3.3. 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. We will not go into

[33] http://wordnet-rdf.princeton.edu/about
[34] http://www.w3.org/ns/lemon/synsem
[35] http://www.w3.org/ns/lemon/decomp
[36] http://www.w3.org/ns/lemon/vartrans
[37] http://www.w3.org/ns/lemon/lime
[38] https://lov.linkeddata.es/dataset/lov/vocabs/lemon
[39] http://archivo.dbpedia.org/info?o=http://www.w3.org/ns/lemon/lime
[40] http://archivo.dbpedia.org/info?o=http://www.w3.org/ns/lemon/vartrans
[41] Using the cc:license property

any further depth into it here since the vocabulary is applied 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 3.5 – there exists the **MMoOn ontology**[42] for the creation of detailed morphological inventories [22]. MMoon does not currently seem to be available on any semantic repositories/archives/search engines but it does have its own dedicated website[43] which offers a SPARQL endpoint (although this was down at the time of writing). Its license 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 [6]. 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,[44] but on the PHOIBLE website, it is possible only to browse and export selected content into RDF/XML.[45] Since it no longer provides resolvable URIs for its components, PHOIBLE data does not fit within the narrower scope of LLD vocabularies anymore (and we haven't included it in our summary table). It does, however, maintain a non-standard way of linking, as it has been absorbed into the Cross-Linguistic Linked Data infrastructure [23, 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/, see below for additional details.

3.4. Linguistic Resource Metadata

Due to the importance of the topic (and also because it was not discussed in any depth in [6]) we will give a much fuller overview in Section 4.3; here we will only look at accessibility issues here. The two main specialised models which we mention there are the METASHARE ontology[46] and *lime*. The latter is described above. 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.

3.5. Linguistic Data Categories

As of 2010, two major repositories were being widely used 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** [24] which provided human-readable and XML-encoded information about linguistic data categories thatwere relevant for linguistic annotation, the encoding of electronic dictionaries and 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 [25], having been originally developed in the context of the project Endangered Metadata for Endangered Languages Data (E-MELD, 2002-2007)[47]. GOLD stood out with its excellent coverage of low resource languages. In the RELISH project, a curated mirror of GOLD-2010 was incorporated into ISOcat [26]. 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 remains accessible[48], it has not been updated since [27] (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, but no longer as an operational system.

Its successor, the Clarin Concept Registry is briefly discussed below, in Section 4.3. Another one of its successors is the **Lexinfo ontology**[49], the data category register used in OntoLex-Lemon and which reappropriates many of the concepts which were contained in ISOCat for use within the lexical domain (dictionaries, terminologies, lexica).

Currently in its third version, lexinfo can be found both on the LOV search engine [50] and on archivo[51],

[42] https://github.com/MMoOn-Project/MMoOn/blob/master/core.ttl

[43] https://mmoon.org/

[44] https://github.com/clld/phoible/tree/master/phoible/static/data

[45] See, for example, https://phoible.org/inventories/view/161.

[46] http://www.meta-share.org/ontologies/meta-share/meta-share-ontology.owl/documentation/index-en.html

[47] http://emeld.org/

[48] https://linguistlist.org/projects/gold.cfm

[49] https://lexinfo.net/

[50] https://lov.linkeddata.es/dataset/lov/vocabs/lexinfo

[51] http://archivo.dbpedia.org/info?o=http://www.lexinfo.net/ontology/2.0/lexinfo

it appears both times however in its second version. Version 3.0 is under development since late 2019 in a community-guided process via GitHub, and is not registered with either service, yet. Lexinfo's license information is (CC BY 4.0) 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.

For linguistic data categories in linguistic annotation (of corpora and by NLP tools), a separate terminology repository exists with the **Ontologies of Linguistic Annotation** [28, OLiA]. OLiA has been developed 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 a linking with and among 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 [29]. At the moment, OLiA covers annotation schemes for more than 100 languages, for morphosyntax, syntax, discourse and aspects of semantics and morphology.

3.6. Vocabularies for Typological Datasets

Linguistic typology is commonly defined as the field of linguistics that studies and classifies languages based on their structural features [30]. The field of linguistic typology has natural ties with language documentation, and accordingly, considerable work on linguistic typology and linked data has been conducted in the context of the GOLD ontology (see above). We can identify the following relevant datasets.

One of the main contributors and advisors to the scientific study of topology is the Association for Linguistic Typology (ALT)[52]. They provide multiple cross-linguistic diversity and the typological patterns underlying datasets.

One of the most well-known resources that ALT makes available is the World Atlas of Language Structures (WALS)[53] [31, 32] which is a large database of phonological, grammatical, and lexical properties of languages gathered from descriptive materials. This resource can both be used interactively online and can be downloaded. The CLLD[54] (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) [33, 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)[55] incorporates TDS among its repositories.

In terms of vocabularies and models which are relevant for the creation of typological databases we can identify **LexVo**[56] [35]. LexVo bridges the gap between linguistic typology and the LOD community and brings together linguistic resources and the entity relationships provided through the Linked Data Web and the Semantic Web. The project manages to link a large variety of resources on the Web, besides providing global IDs (URIs) for language-related objects. LexVo is available on archivo[57] but is not yet available on LOV.

Finally, another group of datasets relevant for typological research include large-scale collections of lexical data, as provided, for example by PanLex[58] and Starling.[59]

An early RDF edition of PanLex has been described by [36] and was incorporated in the initial version of the Linguistic Linked Open Data cloud diagram. At the time of writing, however, this early RDF version does not seem to be accessible anymore. Instead, CSV and JSON dumps are being provided from the PanLex website. On this basis, [37] describe a fresh OntoLex-Lemon edition of PanLex (and other) data as part of the

[52] https://linguistic-typology.org/

[53] https://wals.info/
[54] https://clld.org/
[55] https://vlo.clarin.eu/
[56] http://lexvo.org/
[57] http://archivo.dbpedia.org/info?o=http://lexvo.org/ontology
[58] http://panlex.org
[59] https://starling.rinet.ru/

ACoLi Dictionary Graph.[60] 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 edition, with the goal to adapt techniques for lexical linking developed in the context of, for example, the on-going series of Shared Tasks on Translation Inference Across Dictionaries.[61] As for modelling requirements of lexical datasets in linguistic typology, these are not fundamentally different from other forms of lexical data, but they adopt OntoLex, resp. its predecessor, *lemon*, see above. They do, however, require greater depth with respect to identifying and distinguishing language varieties, one of the driving forces behind developing Glottolog, see below.

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

The current section comprises an extensive overview of developments in various different LLD community initiatives and standards relating to LLD models and vocabularies. In particular, it focuses on the three areas that we believe have been the most active in the last few years (the first two of the following) or that are starting to gain greater prominence (the third): lexical resources (Section 4.1), annotation and corpora (Section 4.2), and finally metadata (Section 4.3). 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.

The most notable community effort in this context is the Open Linguistics Working Group (OWLG) of Open Knowledge International[62] that introduced the vision of a Linguistic Linked Open Data (LLOD) cloud in 2011 [38], and whose 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 [39], ultimately led to the implementation of LLOD cloud in 2012, celebrated with a special issue of the Semantic Web Journal published in 2015 [40]. The LLOD cloud, now hosted under http://linguistic-lod.org/, was enthusiastically embraced, the linguistics category became a top-level category in the 2014 LOD cloud diagram, and since 2018, it represented the first LOD domain sub-cloud.

At the same time, a number of more specialized initiatives emerged, as mentioned below, for which the Open Linguistics Working Group acted and continues to act as an umbrella that facilitates information exchange among them and between them and the broader circles of linguists interested in linked data technologies and knowledge engineers interested in language. Currently, 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[63] and continued information exchange via mailing list[64]

Over the years, however, the focus of discussion moved 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,[65] originally working on ontology lexicalization, the group extended their activities after the publication of the OntoLex vocabulary (May 2016) and now represents the main locus to discuss the modelling of lexical resources with web standards and in LL(O)D.
- the W3C Community Group Linked Data for Language Technology,[66] with a focus on language resource metadata and linguistic annotation with W3C standards

Most recently, these activities converge in also in funded networks, especially, the Cost Action Nexus Linguarum.

Also, Linked Data plays a certain role in the context of older standardization initiatives, e.g., the TEI Consortium[67], or the ISO TC37/SC4 committee[68].

[60] Data available under https://github.com/acoli-repo/acoli-dicts.
[61] https://tiad2020.unizar.es
[62] https://linguistics.okfn.org/
[63] http://linguistic-lod.org/
[64] 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/.
[65] https://www.w3.org/community/ontolex/
[66] https://www.w3.org/community/ld4lt
[67] https://tei-c.org/
[68] https://www.iso.org/committee/297592.html

We take the standards and initiatives proposed by these communities as our basis of the topics in this section, but we will also look at significant developments respecting these standards and initiatives outside and independent of these groups (see Section 4.1.4) in the interests of completeness and to understand current trends.

Note that our intention has been to go for completeness in our description of these developments. At the same time, however, as has been mentioned we have tried not to include too much material that was already available in the sources cited in Section 2.4.

4.1. Lexical Resources: OntoLex-Lemon and its Extensions

In this section we will describe some of the recent work that has been carried out on the OntoLex-Lemon model, both within the ambit of the W3C Ontolex group as well as outside of it. With regards to the former we will discuss three of the latest extensions to the model (one of which has been published and two which are still currently under development) in Sections 4.1.1, 4.1.2, and 4.1.3. In Section 4.1.4 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 [6] (for an in-depth discussion of such developments prior to 2018 please refer to the latter paper). Note, moreover, that the use of OntoLex-Lemon in a number of different projects is described in Section 5.

*4.1.1. The OntoLex 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. The core module has proven to be effective in capturing the most typical kinds of lexical information contained in dictionaries and lexical resources (e.g., [41–45]. However, there are certain situations in which the model falls short, most notably in the representation of particular elements in dictionaries and other lexicographic datasets [46]. This is not surprising since *lemon* was initially conceived as a model to expand ontologies with lexical information but not to model lexical resources per se.

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 Lexicography Module (*lexicog*)[69]. This module which was the result of collaborative work with contributions by lexicographers, computer scientists, dictionary industry practitioners, and other stakeholders 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 [47] and can be considered in a stable status right now.

4.1.2. OntoLex Morphology Module

Since November 2018, the W3C Ontolex community group has been developing another extension of the core model that would allow for better representation of morphological data in lexical resources.

Morphology plays an important role in many languages, and its description has played an important role in the work of lexicographers. The extent of its presence in concrete resources can vary, ranging from 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.[70]

The core Ontolex-lemon model provides means to encode basic morphological information: for lexical entries, morphosyntactic categories such as part of speech can be provided, whereas basic inflection information (i.e. morphological relationship between a lexical entry and forms) can be modelled by creating any additional inflected forms with corresponding morphosyntactic features (e.g. case, number, etc.)

This, however, covers only a small portion of the potential morphological data present in the resources. Neither derivation (i.e. morphological relationship between lexical entries) nor additional inflectional information (e.g. declension type for Latin nouns) could be properly modelled with the core model. The new module has been proposed to address these limitations.

The scope of the module is threefold:

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

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

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

Figure 1 presents the diagram for the module.

The central class of the module, which is 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 have 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:constistOf which point to instances of morph:Morph.[71] These instances can have morphosyntactic 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 ;
             constistsOf :stem_drive_v, :suff_s .

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

The module[72] has not yet been released and is still very much under development. At the time of writing, a consensus was reached on the first two parts of the module, and their overview has been published [48]. 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.

[71] 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.

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

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

In parallel with 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 and other language resources. Most notably, this includes the representation of attestations (such as illustrative examples in a dictionary). These were originally discussed within LexiCog 4.1.1, but this discussion quickly outgrew the confines of computational lexicography alone. Furthermore, it was observed that OntoLex 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 community group felt the need to specifically address the requirements of modern human 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 dictated the order and of development for various parts of the module.

So far, two papers have been published on parts of the module following its development, and these can thus considered to be relatively stable. This includes 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 [49].

The central element introduced in this module is frac:Observable: 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.

The module provides means to model only absolute frequency, because "relative frequencies can be derived if absolute frequencies and totals are known" [49, p. 2]. To represent frequency, a property frequency with an instance of CorpusFrequency as an object

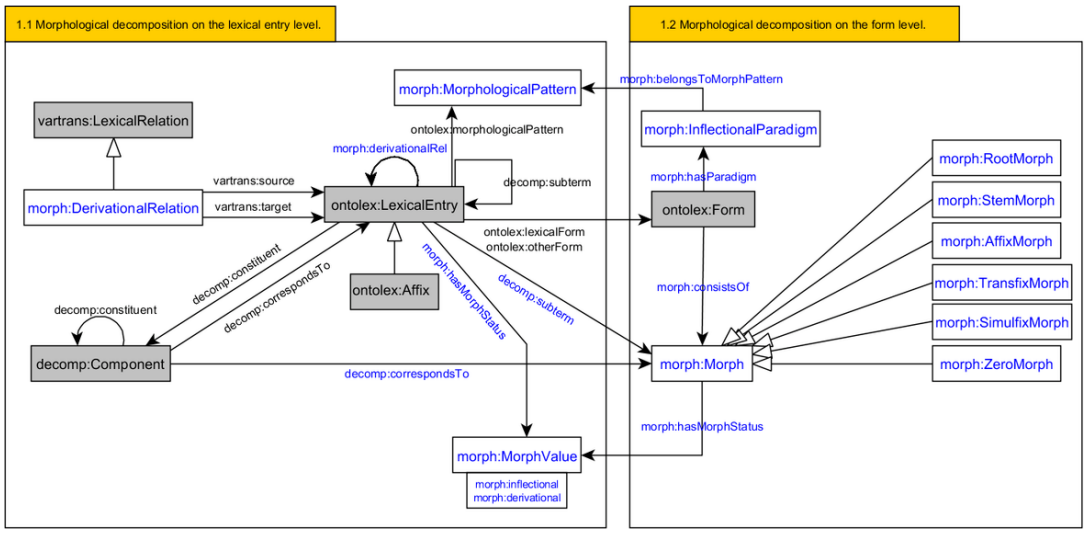

Fig. 1. Preliminary diagram for the Morphology Module

should be provided. This instance must have properties corpus and rdf:value:[73]

```
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 one corpus.

For attestations, i.e. corpus evidence in FrAC is "a special form of citation that provide evidence for the existence of a certain lexical phenomena; they can elucidate meaning or illustrate various linguistic features". As with frequency, there is a class Attestation, an instance of which should be an object of a property attestation. It should have two properties: attestationGloss – text of the attestation and locus – 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 [50] for a citation object and NIF or WebAnnotation (see 4.2) to define a locus.

Another, more recent paper focused on representing embeddings in lexical resources is [51]. 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 ($f : X \to Y$)[74]. 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:FixedSizeVector;
   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 2 presents a diagram of the latest version of the module.

At the moment 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

[73] Examples in this section are taken from [49].

[74] An injective structure-preserving map.

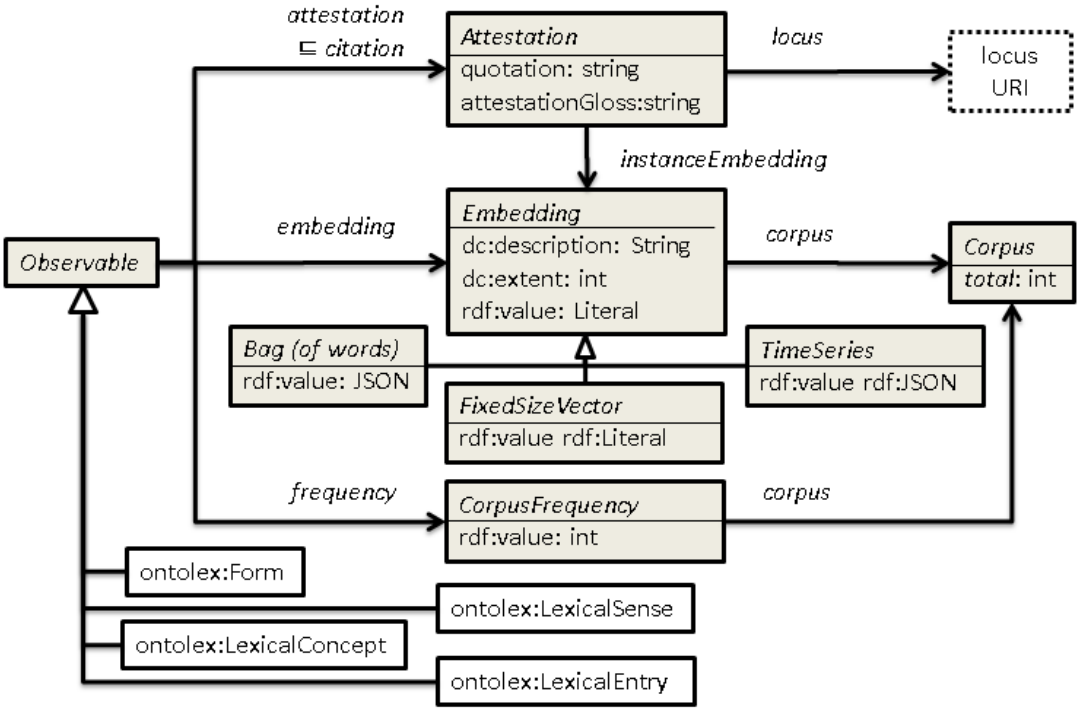

Fig. 2. Preliminary diagram for the FrAC Module

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 pose serious interest for 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.

4.1.4. Selected individual contributions

OntoLex extensions pursued by individual research groups are manifold, and while not being discussed as candidates for future OntoLex modules in the W3C Communitiy Group OntoLex yet, they may represent a nucleus and a cumulation point for future directions.

Selected recent extensions include *lemon-tree* [52], 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[75]. The lemon-tree model has already been used to publish the Thesaurus of Old English [53] and reveals the flexibility of the OntoLex-lemon/LLD approach in allowing for the modelling of different kinds of lexical resources and more specialised kind of linguistic information. As does *lemonEty* [54] another 'unofficial', so to speak, extension of the OntoLex-Lemon model proposed as a means of encoding etymological information: both the kinds of etymological information contained in lexica and dictionaries as well in other kinds of resources. The *lemonEty* model does this by exploiting the graph-based structure of RDF data and by rendering the status of etymologies as prospective hypotheses explicit.

In both of these cases, the RDF data model along with the various different standards and technologies which make up the Semantic Web stack as a whole, and in conjuction with specialist models and vocabularies, enable us to structure such information in salient and 'meaningful' ways that help to enhance the machine actionability of strongly heterogeneous linguis-

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

tic data. This is something which the author attempts to demonstrate in [55], by way of a proposal to extend OntoLex-Lemon with temporal information in a ontologically well motivated way (while being careful to remain within the expressive limitations of RDF in order to exploit standard technologies for that framework including reasoning tools for OWL), and allowing the integration of lexical data with information relating to textual attestations and other historical information.

4.2. Annotation and Corpora

Linguistic annotation by NLP tools and within corpora has long been addressed in the context of RDF and linked data, with different proposals grounded in traditions from natural language processing [56], web technologies [57], knowledge extraction [58], but also from linguistics [59], philology [60], and the development of corpus management systems [61, 62].

A practical introduction to the various different vocabularies used for linguistic annotation existing today – for different purposes, by different user communities and with different capabilities – is given over the course of several chapters in [12]. In brief, the most widely used RDF vocabularies in this area comprise the NLP Interchange Format (NIF, in language technology) and Web Annotation (OA, in bioinformatics and digital humanities), as well as customizations of these. We describe NIF in Section 4.2.1 and Web Annotation in Section 4.2.2.

TEI/XML, a standard which is widely used in the digital humanities and computational philology only comes with partial support for RDF and does not represent a publication format for Linked Data. In Section 4.2.3, however, we discuss the relationship between TEI and RDF vis-a-vis the encoding of corpora and in particular the use of RDFa in TEI.

Notable pre-RDF vocabularies include those developed by ISO TC37/SC4, in particular, the Linguistic Annotation Framework (LAF) that represents "universal" data structures shared by the various, domain- and application specific ISO standards [63]. LAF defines essential data structures that exceed the expressivity of NIF and Web Annotation and is thus significant for their future development. RDF serializations of the LAF do exist as well, but do not seem to be widely used. The topic of LAF , LAF-based RDF vocabularies and its relationship with NIF and WebANnotation is discussed in Section 4.2.4.

Platform specific solutions to modelling annotations and linked data are discussed in Section 4.2.5. In section 4.2.6 Ligt an RDF-native vocabulary for representing interlinear glossed text as linked data and CONLL-RDF, a data model based on the well known CoNLL formats.

Finally, the prospects for convergency between the solutions discussed in the whole of Section 4.2 are described in Section 4.2.7.

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).

4.2.1. NLP Interchange Format

The NLP Interchange Format (NIF),[76] developed at AKSW Leipzig, was designed to facilitate the integration of NLP tools in knowledge extraction pipelines, as part of building a Semantic Web tool chain and a technology stack for language technology on the web [58]. 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 extend, syntactic and semantic parsing. In addition to providing a technology 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 the obligatory use of String URIs as fragment identifiers for anything annotatable by NIF. Every element that can be annotated in NIF has to be a string.[77] NIF does support different fragment identifier schemes, e.g., the offset-based scheme defined by RFC 5147. [64] 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.

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

[77] 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.

Consider the sentence "Stay, they said."[78] The Stanford PCFG parser[79] analyzes *Stay* as a verb phrase contained in (and only constituent of) a sentence. In NIF, both would be conflated. Likewise, zero elements 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 as are frequently found in linguistically annotated corpora. Nevertheless, NIF has been used as a publication format for corpora with entity annotations.[80]

NIF 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 [65, 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 [66]. 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 below.

4.2.2. Web Annotation

Documents in the web come in various forms, and often, it is not possible to embed metadata and annotations directly into them, e.g., because the annotator is not the owner of the document and distributing a local copy may be restricted. Standoff formalisms support the physical separation of annotated material and annotations. The Open Annotation community and their Web Annotation Data Model provide a RDF-based approach for standoff annotation of web documents, with JSON-LD as its designated serialization. The Web Annotation Data Model provides a flexible means to represent standoff annotations relative to any kind of document on the web. It is being applied to linguistic annotations, primarily in the biomedical domain, although prototypical adaptions in other domains have been described as well, e.g., for NLP [57] or Digital Humanities [8].

The Web Annotation data model and vocabulary have been published as W3C recommendations in 2017 [67, 68]. The aim of Web Annotation is to be applicable across different media formats, the most common use case being *"attaching a piece of text to a single web resource"* [67]. In order to achieve this, 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 thr 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 [58].

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 [57], 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 digital humanities, e.g., for the annotation of geographical entities with tools such as Recogito [69], especially since support for creating standoff annotations for static TEI/XML documents has been added (around March 2018 [70, p.247]).

[78] From Stephen Dunn (2009), 'Don't Do That', poem published in the New Yorker, June 8, 2009.

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

[80] 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.

4.2.3. TEI/XML

Together with the Association for Computational Linguistics (ACL, founded in 1962 as the Association for Machine Translation and Computational Linguistics), the European Association for Digital Humanities (EADH, founded in 1973 as the Association for Literary and Linguistic Computing) and the US-American Association of Computers in the Humanities (ACH, founded in 1978), have been working towards establishing general guidelines for the electronic representation of linguistic and literary resources since the mid-1970s. Following similar events in 1977 in San Diego and 1980 in Pisa [71], a 1987 workshop organised by the ACH and held at Vassar College [72] led to the formulation of the "Poughkeepsie Principles" and foundation of the Text Encoding Initiative (TEI). Already in 1986, the SGML Markup Language [73, Standard Generalized Markup Language; ISO 8879:1986] had been standardised to facilitate sharing machine-readable documents, albeit with no special emphasis on (or even concern for) linguistically annotated data. Building on the Poughkeepsie Principles, the TEI Guidelines defined a broad range of SGML (resp., since 2002, XML) tags and accompanying DTDs for encoding language data according to the needs in the humanities. The TEI subsequently contributed greatly to the popularization of XML in the language resource community. In its current edition, P5, TEI/XML continues to be the dominating paradigm for the digital edition of textual data in Digital Humanities.

As an XML-based standard, the TEI takes a necessary focus on standardizing the *form* of language resources for the humanities. Semantic Web formalisms allow to complement this 'syntactic approach' with a formal and standardised way to assess the *meaning* of attributes, markup elements and textual elements in text.

As we discussed in Section 2.3, Linked Open Data is increasingly being recognised as an essential component among the technological approaches that define the area of Digital Humanities, and it continues to grow in importance in the participating communities. As an example, there is a TEI Special Interest Group on Ontologies already since 2004.[81]

The traditional perspective of the TEI on RDF has been that there may be value in being able to link from a digital edition (or another TEI/XML document) to a knowledge graph, e.g., for managing prosopographical, bibliographical or geographical information. Linking between (elements of) electronic editions created with the TEI was addressed by means of specialised XML attributes with narrowly defined semantics. Accordingly, electronic editions in TEI/XML do normally not qualify as Linked Data, even if they use and provide resolvable URIs (TEI pointers). This may not considered to be drastic for electronic editions of historical manuscripts for which it is conceivable to complement them with information drawn from the LLOD cloud, but less that they represent sources of Linked Data. The situation is, however, 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 did, however, begin to change in the last years and long-standing efforts to develop technological bridges between both TEI and LOD are beginning to yield concrete results. As an example, DitMaO, an editor for the conjoint development of lexical and ontological resources in the context of a project on medico-botanical terminology of the Old Occitan language, was originally designed to operate on the basis of lemon, the predecessor of OntoLex-Lemon, as a native LOD tool [74]. More recently, however, it is being integrated into the TEI technology stack under its new name, Lex0 [75]. At the same time, different tools for the conversion of lexical resources in different TEI dialects to OntoLex-Lemon have been presented in the last years.[82]

Beyond lexical resources, the annotation *of* rather than *within* TEI documents 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. A main result of these efforts is the development of the Recogito editor mentioned above, and its extension to TEI/XML. In this way, the annotation is not part of the TEI document, but stored as standoff annotation in a JSON-LD format, and thus, 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 does no longer change. Otherwise, the efforts for synchronization will by far

[81] https://tei-c.org/activities/sig/ontologies/

[82] Among others, this includes a converter for TEI Dict/FreeDict dialect, https://github.com/acoli-repo/acoli-dicts/tree/master/stable/freedict, [37]. For ELEXIS related developments TEI/RDF related developments, see Section 5.7.

outweigh any benefit that the use of W3C standards for encoding the annotation brings.

Therefore, there is a need for encoding RDF triples directly inline in a TEI document. Recently, it has been demonstrated that this can be done in a W3C- and XML-compliant way by incorporating RDFa attributes into TEI [76, 77]. As a result and after more than a decade of discussions, the TEI started in May 2020 to work on a customization that allows the use of RDFa in TEI documents.[83]

4.2.4. LAF and LAF-based vocabularies

The Linguistic Annotation Framework [78, LAF] provides generic data structures for representing *any* kind of linguistic annotation. 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. These roughly correspond to selectors (or target URIs) of Web Annotation.
- markables (nodes) elements that constitute and define the scope of the annotation by reference to anchors and regions. These roughly correspond to annotation elements in the sense of Web Annotation.
- values (labels) elements that represent the content of a particular annotation. These correspond roughly to the body objects of Web Annotation.
- relations (edges) links (directed relations) that hold between two nodes and can be annotated in the same was as markables. In Web Annotation, relations as data structures are not foreseen.

A tentative interpretation of LAF concepts in terms of Web Annotation data structures has been given above. It is important to note that Web Annotation lacks any formal counterpart of edges or relations as defined by LAF. An RDF vocabulary that extends Web Annotation with LAF data categories has been sketched by [57], but apparently, never been applied in practice.

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, it conflates the roles of regions and markables is 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 the 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. We are aware of the following LAF-based RDF vocabularies (in chronological order):

- Cassidy (2010) utilised an RDF graph to express linguistic data structures over a corpus backend natively based on an RDBMS [61]. This included an RDF vocabulary for nodes, labels and edges.
- Chiarcos (2012) introduced POWLA [79], an OWL2/DL serialization of PAULA, a standoff-XML format that implemented the LAF as originally described by [80]. POWLA complements LAF core data structures with formal axioms and a slightly more refined data structures that support, for example, the effective navigation in tree annotations. On current applications of POWLA see the CoNLL-RDF Tree Extension below.
- Verspoor et al. (2012) described a prototypical extension of Web Annotation with an RDF interpretation of the LAF [57]. As mentioned above, this does never have to reached a level of maturity for wider application.
- LAPPS Interchange Format (LIF, see below) [81] is historically and conceptionally an instance of the LAF, but has been designed for specific NLP tasks and concrete workflows, see below under platform-specific vocabularies.

4.2.5. Platform-specific solutions

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 [82], NewsReader [83] or the LAPPS Grid [84].

FRED provides output based on NIF or EARMARK [85], with annotations partially grounded in DOLCE [86], but enriched with lexicalised ad hoc properties for aspects of annotation covered by these. 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).

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

The NewsReader Annotation Format (or NLP Annotation Format) NAF, is an XML-standoff format for which a NIF-inspired RDF export has been described [87], and LIF, the LAPPS Interchange Format [81], a JSON-LD format used for NLP workflows by the LAPPS Grid Galaxy Workflow Engine [88]. 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 [89].

Both LIF and NAF-RDF are, however, not generic formats for linguistic annotations but rather, they provide (relatively rich) inventories of vocabulary items for specific NLP tasks.[84] Neither seem to have been used as a format for data publication, and we are not aware of their use independently from the software they have originally been created for or are being created by.

4.2.6. Domain-specific solutions

Aside from software- or platform-specific formats, a number of vocabularies has been developed that address specific problems or user communities. Here, we list two examples, the Ligt vocabulary that addresses the gap that community standards such as NIF and Web Annotation have with respect to morphology and the annotation of morphologically rich langugages, and CoNLL-RDF, a vocabulary and an associated library that aims to mirror the structure and contents of popular NLP formats in RDF and to provide a round-tripping between these formats and RDF graphs.

Ligt Ligt is a native RDF vocabulary for representing interlinear glossed text as Linked Data [90]. Interlinear glossed text (IGT) is a notation for texts and linguistic examples that provide readers with a way to understand linguistic phenomena. This notation is used in education and various language sciences such as language documentation, linguistic typology, philological studies.

IGT data can consist of different layers, such as translations and transliterations, and usually has layers with morpheme-level alignment, which is not supported by any established vocabularies for representing linguistic corpora or automated annotations. There are several formats specifically designed for storing and exchanging interlinear glossed text, varying in their design and concepts, but these formats are not being reused across different tools, which limits reusability of the annotated data. To overcome this and to improve data interoperability, Ligt brings a tool-agnostic representation for interlinear glossed text in RDF. Beyond structural interoperability, this also allows using shared vocabularies and terminology repositories published in the (Linguistic) Linked Open Data cloud.

Ligt is developed as a generalization over data structures employed by established tools for creating IGT annotations, most notably Toolbox [91], FLEx [92] and Xigt [93]. It is to be noted that these tools are currently incompatible with each other and information between them can only exchanged if manual corrections are applied. Ligt has been designed with the goal to facilitate a pivot format that faithfully captures their respective linguistic information in a uniform way for subsequent processing.

After its publication, Ligt has been applied by third party users to model and annotate interlinear glossed text from 280 endangered languages and their publication as Linked Open Data [94].

Although designed for very specific requirements within linguistics, we consider Ligt an important contribution as it provides data structures that are relevant for low-resource and morphologically rich languages but that, so far, have been neglected by earlier RDF vocabularies for linguistic annotation on the web, in particular, NIF and Web Annotation. It would be possible to encode Ligt information with a generic LAF-based vocabulary such as POWLA, but not with the more established community standards.

CoNLL-RDF In language technology and language sciences, tab-separated values (TSV) represent a frequently used formalism to represent linguistically annotated natural language, often addressed as "CoNLL formats". A large number of such formats do exist, but although they share a number of common features, they are not interoperable, as different pieces of information are encoded differently in these dialects. In the field of NLP, such formats have become defacto standards for the most frequently used types of annotations, as they have been popularised in longstanding series of shared tasks in the last two decades. In corpus linguistics and computational lexicography, the use of TSV formats as the basis for the most popular tools in this context, the Corpus Workbench [95] and SketchEngine [96], had a similar effect.

[84] Historically, LIF is grounded in LAF concepts and has been developed by the same group of people, however, 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.

CoNLL-RDF [97] provides a data model and a programming library that aim to facilitate processing and transforming such data in a serialization-independent way: 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), sentence after sentence is converted to an RDF graph in accordance to the label information provided by the user.

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 part-of-speech tag. The PARSE column contains a full parse in accordance with the Penn Treebank [98].

The CoNLL-RDF library reads such data as a continuous stream, every sequence of rows enclosed in empty lines will be processed as a block, assigned a URI and the type nif:Sentence, every row will be assigned a URI and the type nif:Word, and the annotation of every column will be stored as value of a property in the conll: namespace that is generated from the column label.[85] 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 data, a CoNLL-RDF edition of the Universal Dependencies corpora[86] is available as such 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 so far.

Indeed, previous applications of CoNLL-RDF include linking between corpora and dictionaries [99] and knowledge graphs [100]. Beyond this, syntactic parsing of historical languages [101, 102], the consolidation of syntactic and semantic annotations [103], corpus querying [104], and language contact studies [105] have been conducted on the basis of CoNLL-RDF.

CoNLL-RDF tree extension The example above illustrates one of several possible ways to encode syntactic parses in TSV formats. CoNLL-RDF originally stored such information as a plain string, so that phrase structure has to be be subsequently decoded. It has be demonstrated that this can be done by SPARQL Updates, but this is relatively slow. Chiarcos and Glaser [106] thus provide an extension of CoNLL-RDF with native support for tree structures.

Focusing on the modelling part here, they extend NIF/CoNLL-RDF data structures with POWLA [107], a basic vocabulary that encodes directed labelled multigraphs in accordance with LAF.

As a result, the phrase structure of the example above can now be decoded:

```
: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 RDFS-inferred from the use of these properties.

4.2.7. Towards a Convergency

The sheer number of vocabularies mentioned above already points to 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 may be mutually incompatible. So far, no clear consensus has emerged, albeit NIF and Web Annotation appear to enjoy relatively high popularity in their respective user communities, they are not compatible with each other nor do they support linguistic annotation to the same or even a sufficient extent, thus motivating the continuous development of novel, more specialised vocabularies.

On the other hand, synergies between Web Annotation and NIF have been explored relatively early on

[85] The columns HEAD (for dependency annotation) and PREDARGS (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.

[86] https://universaldependencies.org/

[58], and Cimiano et al. [108, p.89-122] describe how they can be used in combination with each other, more specialised vocabularies such as CoNLL-RDF, and more general vocabularies such as POWLA to model data in a way

- that is applicable to any kind of primary data, including non-textual data (via Web Annotation selectors),
- that can express reference to primary data in a compact fashion (via NIF String URIs),
- that permits round-tripping between RDF graphs and conventional formats (via CoNLL-RDF and the CoNLL-RDF library), and
- that 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 among, 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 is intractable, and thus, the W3C Community Group Linked Data for Language Technology (LD4LT) is currently engaged in a process to develop a harmonization of these vocabularies. While this has been worked on since about mid-2018, regular discussions via LD4LT began in early 2020 only. Concrete results so far are a survey over requirements for any vocabulary for linguistic annotation on the web and the degree to which NIF, Web Annotation and other vocabularies support these at the moment.[87] So far, 51 requirements have been identified, clustered in 6 groups:

1. LLOD compliancy (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, dialog annotation).

So far, this is still work in progress, but if indeed, these challenges can be resolved at some point in the future, and a coherent vocabulary for linguistic annotations emerges, we expect a similar rise in popularity for the adoption of the Linked Data paradigm for encoding linguistic annotations as we have seen in the last years for lexical resources. Also, this was largely driven by the existence of a coherent and generic vocabulary, and indeed, the drift in application that the OntoLex model has recently faced (originally, designed as a model for ontology lexicalization, it became popular for lexical resources independent from any ontology and is now on the way to develop towards a general standard for lexical data on the web) 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 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 capability to integrate, use and re-use pieces of information from different sources. In linguistic annotation, a still largely unsolved problem 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 [56, 59, 61, 79], but only recently, systems that support RDF as an output format have emerged [62]. While it is clear that standoff is a solution for 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,[88] albeit support for RDF

[87] 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.

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

serializations (RDFa) or standoff (Web Annotation in JSON-LD) seems to grow.

4.3. Metadata

The rise of data-driven approaches that use Machine Learning, and in particular recent breakthroughs in the field of Deep Learning field, have given data a central place in all scientific and technological areas. Cross-disciplinary research has also boosted the sharing of data arising across different communities. Thus, a huge volume of data has become available through various repositories, but also via aggregating catalogues, such as the European Open Science Cloud[89] and the Google dataset search service[90]. Metadata play an instrumental role in the discovery, interoperability and hence (re-)use of digital objects, given the fact that they act as the intermediary between consumers (humans and machines) and digital objects. For this reason, FAIR principles [1] include specific recommendations for metadata (see also section1).

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. We thus 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 expression and 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, which obviously extends the search space for data to catalogues other than the purely linguistic ones. Currently, the three models that dominate the description of datasets are DCAT[91], schema.org[92] and DataCite[93]. DCAT profiles are used in various open data catalogues, such as the EU Open Data portal[94], 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 with a small set of core properties which have been selected for the accurate and consistent identification of a resource for citation and retrieval purposes. There are various initiatives for the collection of crosswalks of community-specific metadata models with these models[95], as well as recommendations for extensions for specific data types (e.g., CodeMeta[96] and Bioschemas[97] 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 for them in this paper.

Among models for the description of linguistic resources in general (and not just LLD resources), the Component Metadata Infrastructure (CMDI) profiles [109, 110], 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 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[98] 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 the collection and individual text levels. Both models, however, are XSD-based[99], and therefore not discussed further in this section.

We should also mention the CLARIN Concept Registry (CCR)[100], which is a collection of linguistic concepts [112, 113]. It is the successor to the ISOcat data category registry (described in Section 3.5) and is currently maintained by CLARIN. The CCR is imple-

[89] https://www.eosc-portal.eu
[90] https://toolbox.google.com/datasetsearch
[91] https://www.w3.org/TR/vocab-dcat-2/
[92] https://schema.org/
[93] https://schema.datacite.org/
[94] https://data.europa.eu/euodp/en/data/
[95] See for instance, https://rd-alliance.github.io/Research-Metadata-Schemas-WG/
[96] https://codemeta.github.io/
[97] https://bioschemas.org/
[98] https://www.clarin.eu
[99] The conversion of CMDI metadata records offered in CLARIN into RDF [111] should not be confused with the construction of an RDF model for CMDI profiles
[100] https://concepts.clarin.eu/ccr/browser/

mented in SKOS and includes a concept scheme for metadata, but is a structured list without ontological relations, either internally or externally to other vocabularies. It mainly serves as the semantic interoperability layer of CLARIN; this is achieved through linking metadata fields included in CMDI profiles to concepts from the CCR.

4.3.1. Language Resource Metadata: The META-SHARE ontology

The META-SHARE[101] (or MS-OWL in short) model [114], implemented 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 [115]. The first version of MS-OWL was (semi-)automatically created from the META-SHARE XSD schema [115, 116] (which was designed to support the META-SHARE infrastructure [117]) and discussed in the framework of the LD4LT group. The second version, which is described here, has evolved from it, 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 [114].

MS-OWL has been constructed according to three key concepts: *resource type*, *media type* and *distribution*, which 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) 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 together 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 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 corpus of audio recordings and their transcripts), language resources are represented as *consisting* of at least one media part: 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 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[102].

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

[102] 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.)

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 licences commonly used for language resources, reusing existing vocabularies and extending the Open Digital Rights Language[103] core model [118].

To better illustrate the structure of the MS-OWL, figure 3 depicts a subset of the mandatory and recommended properties for the description of a corpus.

One of the additions made between the two versions of the MS ontology is the development of another vocabulary, again implemented as an OWL ontology, OMTD-SHARE[104] [119]. OMTD-SHARE can be considered as complementary to MS-OWL. It covers *functions* (tasks performed by software components), *annotation types* (types of information extracted or annotated by such software), *methods* (classification 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 [105], which focused on Text and Data Mining resources, and 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 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. The 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, the first version of MS-OWL is used in LingHub[106], a data portal aggregating metadata records for language resources hosted in various repositories and catalogues[120, 121], while the second version, the one described here, is used in the European Language Grid[107], which is a platform for language resources with a focus on industry-relevant Language Technology in Europe[122]. Among the immediate plans, crosswalks with DCAT and schema.org are a priority, to ensure wider uptake and interoperability with (meta)data from other communities.

4.3.2. Linguistic Metadata for Lexical Resources: lime

Another metadata model that has relevance here is OntoLex-Lemon's own dedicated metadata module that, in keeping with the overall citric theme, is called *lime* or the *LInguistic MEtadata* module [123].

In line with the *lemon* conceptual model, *lime* distinguishes three main metadata entities: the lexicon (i.e. a collection of lexical entries), the reference dataset

[103] https://www.w3.org/ns/odrl/2/
[104] http://w3id.org/meta-share/omtd-share/
[105] https://www.openminted.eu
[106] http://linghub.org/
[107] https://live.european-language-grid.eu/

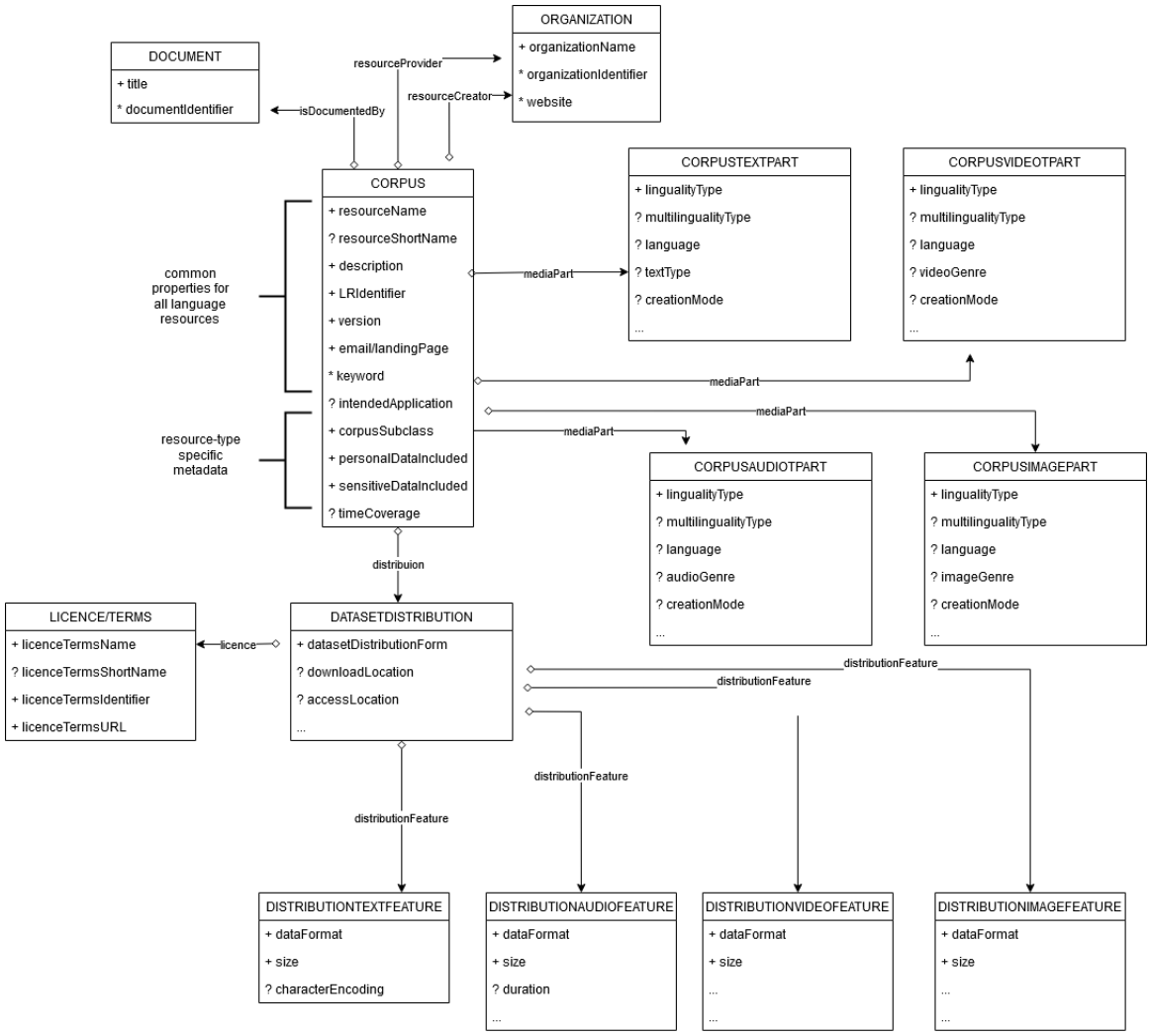

Fig. 3. Simplified subset of the MS-OWL for corpora

(e.g. an ontology that describes the semantics of a domain), and the concept set (an optional set of lexical concepts), and aims to provide quantitative and qualitative information for these entities and their relations (links between them).

The lime:Lexicon class, defined as a subclass of void:Dataset[108]), represents a set of lexical entries, (linked to lime:Lexicon via the property lime:entry) in a certain language (specified with the property lime:language). The property lime:linguisticCatalog) specifies the linguistic model, i.e. the catalogue of linguistic categories used for the annotation of the lexical entries.

The lime:LexicalizationSet class (again a subclass of void:Dataset) represents a collection of *lexicalizations*, which are defined as pairs of a lexical entry and an associated entry in the reference dataset (ontology). The metadata properties on the lime:LexicalizationSet allow us to describe, among other things[109] 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 many ontology elements have been lexicalised on average.

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

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

In addition, *lime* defines the class *Lexical Linkset* (lime:LexicalLinkSet, 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 (ontology). For this class, *lime* defines properties describing, for example, the number of links between the two resources in question. Last, the *Conceptualization Set* (lime:ConceptualizationSet) is analogous to the lime:LexicalizationSet but caters for the links between the the lexicon and the concept set.

4.3.3. Language Identification

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

The handling of language identifiers requires special consideration, as two alternative ways of encoding language information are currently being employed, a URI-based mechanism that builds on terminology repositories, and the use of language tags as a way of typing RDF literals. The difference is that an RDF language tag can be attached to any literal to indicate its language and is internally treated similar to a data type. In particular, language information provided in this way does not entail an additional RDF statement over the literal, so this is allows a compact, readable and efficient identiciation with minimal overhead on data modelling.

On the other hand, most RDF vocabularies provide means to mark the language of a resource with explicit RDF statements, e.g., with properties such as dc:language (for language URIs or strings representations) or lime:language (for string representations). We elaborate on the differences in practice below.

The RDF specification [124] already contains recommendations for language identification by means of language tags that can be attached to RDF strings, e.g., the flag @en in the Turtle literal "report"@en to indicate that the word in question is English.

RDF language codes are defined by BCP47 and the IANA registry and grounded on using standardised language codes as defined in the ISO 639 standard, and this is formulated as part of the RDF specification . Machine-readable identifiers for language varieties are necessary as the same language may be referred to by different names. For example, the Manding language *Bamanakan* (bm) is also known as *Bambara*.

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,[110] provides an extensive set of two letter codes for major languages that date back to the beginning of modern-age 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 was realised that two-letter codes are not sufficient to reflect the linguistic diversity of the world in its past and present, currently estimated to comprise more than 6,000 language varieties. In response to this insight, 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 in human-readable form and as RDF.[111] However, it has to be recognised that the primary use case of ISO 639-2 was an application in libraries, i.e., languages with extensive literature, whereas the demands of linguistics and lexicography, especially historical linguistics and language documentation, exceed far beyond this and also 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).[112]

For applications in linguistics, SIL International acts as maintainer of ISO 639-3, another, and more extensive set of three-letter codes. Differently from ISO 639-1/2 codes, which are meant to be stable and develop at a slow pace, if at all,[113] ISO 639-3 codes are actively maintained by the research community and a continuous process of monitoring, approving (or rejecting) updates, additions and deprecation requests is in place. At the moment, ISO 639-3 codes

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

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

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

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

are published by means of human-readable code tables only,[114] 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 [125] 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 represents the basis only for language tags as specified by BCP47 [126, Best Common Practices 47, also referred to as IETF language tags or RFC 4646] that are incorporated into the RDF specification: BCP47 defines how ISO 639 language tags can be bundled with information about the geographical use, script and other information as follows:

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

These tags are composed from the following elements:

- **Language**: The language as an ISO 639-1 tag if available or otherwise an ISO 639-3 tag, e.g., `en` for English and `ang` for Old English.
- **Script** (optional): The ISO 15924 4-letter code for script, e.g., `Latn` for Latin.
- **Region** (optional): The ISO 3166 2-letter (or UN M.49 3-number) region code, e.g., `DE` (or `276`) for Germany or `US` (or `840`) for the USA
- **Variant**: Zero or more registered variants.[115] Of particular interest to linguists are the tags `fonipa` and `fonxsamp` used to mark phonetic representations in the International Phonetic Alphabet (IPA) or X-SAMPA (ASCII rendering of IPA), respectively.
- **Extension**: Zero or more extensions in custom schemes
- **Private use** (optional): Used for internal notes about identification within a single application.

The W3C provides means for validating BCP 47 language tags and part of the specification is also that language tags should be registered at the Internet Assigned Numbers Authority. The IANA language subtag registry[116] currently provides registered language tags in XML, HTML and plain text. As of 2020, discussions about providing a machine-readable view in RDF and by means of resolvable URIs have set in and are expected to bear fruits 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.[117]

However, the very notion of language tags has been criticised as being too unflexible and basically inapplicable to address the needs of linguistics, e.g., recently by [127, 128], and alternatives are being explored [129].

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

Another provider of language identifiers to be considered here is Glottolog [130],[118] a repository of identifiers for language varieties as with a specific focus on (bot not restricted to) low-resource languages and for applications in linguistic typology and language documentation. Language identification is essential here, as often, language names are used to refer to quite different varieties. For instance, *Saxon* has been the self-designation of Old English (Anglo-Saxon, ISO 639-3 `ang`), but also for a number of historical and modern varieties of Low German (Old Saxon, `osx`; Low Saxon, `nds`). Further, it continues to denote different dialects of High German (Upper Saxon, `sxu`; Transylvanian Saxon [no ISO language code]).

This example does not only illustrate how language codes help to differentiate language varieties, but also that language codes may not be sufficient for distinguishing certain variants (e.g., for Transylvanian Saxon). To complement ISO 639 language codes and provide a more fine-grained vocabulary to distinguish between language variants, Glottolog maintains an independent set of language variety identifiers accessible in human- and machine-readable (RDF) form via resolvable URIs, along with additional metadata, asso-

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

[115] The current list of registered variants is provided under https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry (accessed 10-07-2019).

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

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

[118] https://glottolog.org/

ciated bibliography and a view on their phylogenetic structure.

An important design decision of Glottolog is to avoid the notion of "language", as it comes with unintended political connotations.[119] Instead, Glottolog uses the more neutral term 'languoid', defined as a language variety about (or in) which written literature does exist. Accordingly, language families, protolanguages, national languages, historical varieties, dialects and sociolects can receive a unified treatment. A Glottolog ID combines a 4-letter alphabetic core with a 4-letter numerical code, e.g., `stan1293` for (Standard) English. These IDs come as a native URI: http://glottolog.org/resource/languoid/id/stan1293, which resolves via content negotiation to an HTML visualization or to RDF data, which then provides further links to ISO 639, lexvo, etc.

In addition to providing mere identifiers, Glottolog also features relations, e.g. phylogenetic relations, between languoids are provided in a machine-readable way. For example, English is a subconcept of (skos:broader) 'Macro-English' (macr1271, which groups together Modern English with a number of English Pidgins), etc., and it has further subconcepts (skos:narrower) such as Indian English (`indi1255`), New Zealand English (`newz1240`), etc. Glottolog is designed to be descriptively adequate, but as being extensible rather than exhaustive. Suggestions about novel or incorrect langoids can be reported via the website and will be addressed by the maintainers. Thus, even where a distinction may be missing, it may be introduced upon request, and if properly justified by the accompanying scientific literature (i.e., bibliographical references), it will be accepted.

As a critical remark, we have to note that Glottolog is biased towards endangered modern languages and thus rather sketchy in its historical dimension. Yet, Glottolog is now widely used beyond the language documentation community, e.g., in Wikipedia, and we expect that with intensified use beyond the academic world, Glottolog codes for historical language varieties may become available – and can be suggested for insertion already now.

4.3.4. Future Metadata Challenges: Humanities Use Cases

We have described *lime* in some depth as well as given an extended description of META-SHARE in this section, in the hope that this article can stand as a self-contained survey of the current situation with regards to dedicated metadata vocabularies for LLD (with special emphasis on the case of lexical resources). Such a survey is useful, in particular, because of the ever more pronounced emphasis on the need for good quality metadata in language resources.

It is important for the kinds of humanities-oriented use cases which we have discussed above, including the publication of retrodigitised dictionaries as linked data lexica and the modelling of historical and scholarly lexical resources as LLD, that there is appropriate metadata provision at both the level of the lexicon as well as individual entries. For instance, in the case of retrodigitised LLD dictionaries, metadata for the resources in question should contain information on who compiled the original 'paper' version of the dictionary and when, what edition of the dictionary the resource is based on, who digitised it, what tools were used, etc. Similarly more scholarly lexical resources may require detailed bibliographic information which identifies where certain hypotheses were made and by whom. However this kind of metadata provision is still at a very incipient stage.

This does not, of course, necessarily entail the creation of new ad-hoc metadata vocabularies for LLD since there already exist (non-specialised) models, such as the Semantic Publishing and Referencing suite of ontologies for bibliographic information[120], which can be used in conjunction with META-SHARE and *lime* and others in creating metadata solutions, and potentially application profiles, for the cases which we have mentioned. It does however require that whatever solutions are proposed are then made accessible to a wider community of users. Fortunately, two of the initiatives/projects which we describe in this article, ELEXIS (Section 5.7) and NexusLinguarum (Section 5.10), and which are currently at the forefront of tackling such humanities-oriented use cases also include a strong emphasis on dissemination and on outreach to users (who are not already proficient at linked data) along with the provision of tools and teaching materials. This kind of emphasis would seem to augur well for the wide diffusion and accessibility of future metadata proposals produced under the aegis of these or related projects.

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

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

5. Projects

As we mentioned in the introduction, the funding of an ever rising number of projects at the European, national, and regional levels in which LLD plays a key role is evidence of the success of the latter as a means of modelling and publishing language resources. It also gives us an important picture of the use which is being made of LLD models and vocabularies across different disciplines as well as indicating where future challenges may lie. In this section, therefore, we will give an overview of some of the more prominent of these projects, with a special focus on what they have meant, and what they will mean, for the development and use of LLD models.

5.1. An Overview

As part of the NexusLinguarum COST action (see Section 5.10) several of the authors of the current article decided to carry out a survey of existing and completed 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 searches on CORDIS[121] and the OpenAIRE explorer site[122], as well as through a study of the literature and by requesting input from the other participants of the COST action. In preparation of the survey, we set up a Wikipedia page on OntoLex,[123] extended another Wikipedia page on Linguistic Linked Open Data[124] and encouraged partners from our respective networks to contribute and extend those pages, especially with respect to applications of OntoLex and LLOD in general. Information retrieved in this way has been used to complement our survey on relevant projects.

As a disclaimer, it is to be noted, that while a survey based on publications is a principled process and was guided by following major conferences (in particular, the Language Resource and Evaluation Conference (LREC) series and associated workshops), 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) and selected seminal publications and collections [131, 132], we had a natural selection bias in the project overview towards projects that publish at these venues. As noted, we did a systematic survey over EU-funded activities via CORDIS and OpenAIRE, inevitably however their coverage of national and non-European projects is severely limited, and can only partially be compensated by information retrieved via the active consultation of our respective networks, in particular, NexusLinguarum. At the same time, not all European projects that address Linked Data as applied to language resources have been reported below. We have excluded, for example the Horizon 2020 project "Lynx. Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe" (2017-2021) [133]. That is because, although it does address and build on linguistic linked open data in its multilinguality aspect, its main contribution in terms of data modelling is the area of machine-readable licensing, a topic that is much broader than the area covered by our survey. Similarly, the project "Quality Translation by Deep Language Engineering Approaches" (QTLeap, 2013-2016) did, to some extent, address aspects of Linked Data, but was primarily focusing on Natural Language Processing.

Based on this exploratory work we were able to make a number of observations. The most important probably is that this effort has not been dependent on a single, large-scale project, but was always conducted by a larger community, within which several parallel, large-scale and smaller funded projects have been pursued. But what came through quite strongly, both from the research carried out as part of the survey and the personal experience of the authors, is the importance of international projects for supporting and sustaining LLD models and vocabularies.

This can be demonstrated by the development of OntoLex-Lemon. It ultimately goes back to the Lexical Markup Framework (LMF) [3], a conceptual model for representing machine-readable dictionaries currently developed from a long-standing series of projects that covered lexical resources in NLP and related use cases, most notably the projects Expert Advisory Group on Language Engineering Standards (EAGLES, 1993-1995), and International Standards for Language Engineering (ISLE, 2000-2002), and subsequently further developed within ISO TC37. LMF is a conceptual model, defined in the Uniform Markup Language, with serializations in multiple formats, but with an XML

[121] https://cordis.europa.eu/projects
[122] https://explore.openaire.eu/
[123] https://en.wikipedia.org/wiki/OntoLex
[124] https://en.wikipedia.org/wiki/Linguistic_Linked_Open_Data

serialization defined as part of the standard. In 2007, a prototypical RDF/OWL serialization of LMF has been developed by Gil Francopoulo.[125]

On the basis of LMF, the project Multilingual Ontologies for Networked Knowledge (MONNET, 2010-2013)[126] developed the original *lemon* model as a conceptual model for the lexicalization of ontologies. In 2011, MONNET project members initiated the formation of W3C Community Group Ontology-Lexica, and in this community group, OntoLex-Lemon was developed as a revision of the original lemon model for the specific application of ontology lexicalization. OntoLex-Lemon was further developed in the subsequent **LIDER** project (2013-2015).[127] LIDER contributed to the formation of numerous W3C community groups as a means to provide a long-term perspective for 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 represents the basis for most modern uses of OntoLex-Lemon, and its development towards a general-purpose community standard for publishing lexical resources on web.

Other relevant European projects include the FP7 project **Eurosentiment**[128] which leveraged the *lemon* model to model language resources for sentiment analysis as well as **FREME**[129] (which leveraged the use of datasets in NIF and *lemon*) and **SemaGrow**[130] (which along with the LIDER project helped to support the development of the *lime* metadata module). Due to their impact on LLD models and vocabularies we will dedicate specific sections to the European H2020 projects **ELEXIS** (Section 5.7), **Prêt-à-LLOD** (Section 5.9), and the ERC projects **LiLa** (Section 5.8) and **POSTDATA** (Section 5.5) below.

Aside from the French national project **Nénufar** (mentioned above), we can list the German project **Linked Open Dictionaries** (described in Section 5.4), and the transatlantic project **Machine Translation and Automated Analysis of Cuneiform Languages** (Section 5.6); the Italian national project (Progetti di Rilevante Ieresse Nazionale or PRIN) **Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models** (currently ongoing) which aims to publish a linked data lexicon of the ancient Italic languages[131] using the OntoLex-Lemon model and its extensions; as well as the Italo-German project **DiTMAO**, funded by the DFG (Deutsche Forschungsgemeinschaft), (completed) which produced a lexicon of Old Occitano medical terminology and which also proposed an extension of *lemon* to deal with the specifics of the use-case[132] [134].

In figure 4, we provide an overview over selected projects and vocabularies that they contributed to or developed in the form of a matrix. We distinguish three qualities of contributions: A project is said to have

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

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, or to have

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

Note that this survey provides a partial view only. In particular, contributions by community groups (Open Linguistics WOrking Group, OntoLex, Linked Data for Language Technology, etc.) are not covered, whereas in the reality of Linguistic Linked Open Data, these communities and the networks are actually the basis for subsequent research projects, as they foster collaboration and provide a more sustainable framework for dissemination, maintenance and extension of vocabularies – and, sometimes, are set up with the explicit goal to continue project work after the original funding ceased (e.g., the OntoLex Community Group was originally an extension of the MONNET project). Moreover, these communities are essential in vocabulary development as they exceed beyond the narrow focus of a research project by including contributions of external partners and creating synergies between independent projects. The interested reader may notice, for example, that very few of the projects in Fig. 4 address

[125] Linked under http://www.lexicalmarkupframework.org/, not otherwise published.
[126] https://cordis.europa.eu/project/id/248458
[127] http://lider-project.eu/lider-project.eu/index.html
[128] https://cordis.europa.eu/project/id/296277
[129] https://cordis.europa.eu/project/id/644771
[130] https://cordis.europa.eu/project/id/318497
[131] https://www.prin-italia-antica.unifi.it/
[132] https://www.uni-goettingen.de/en/ditmao/487498.html

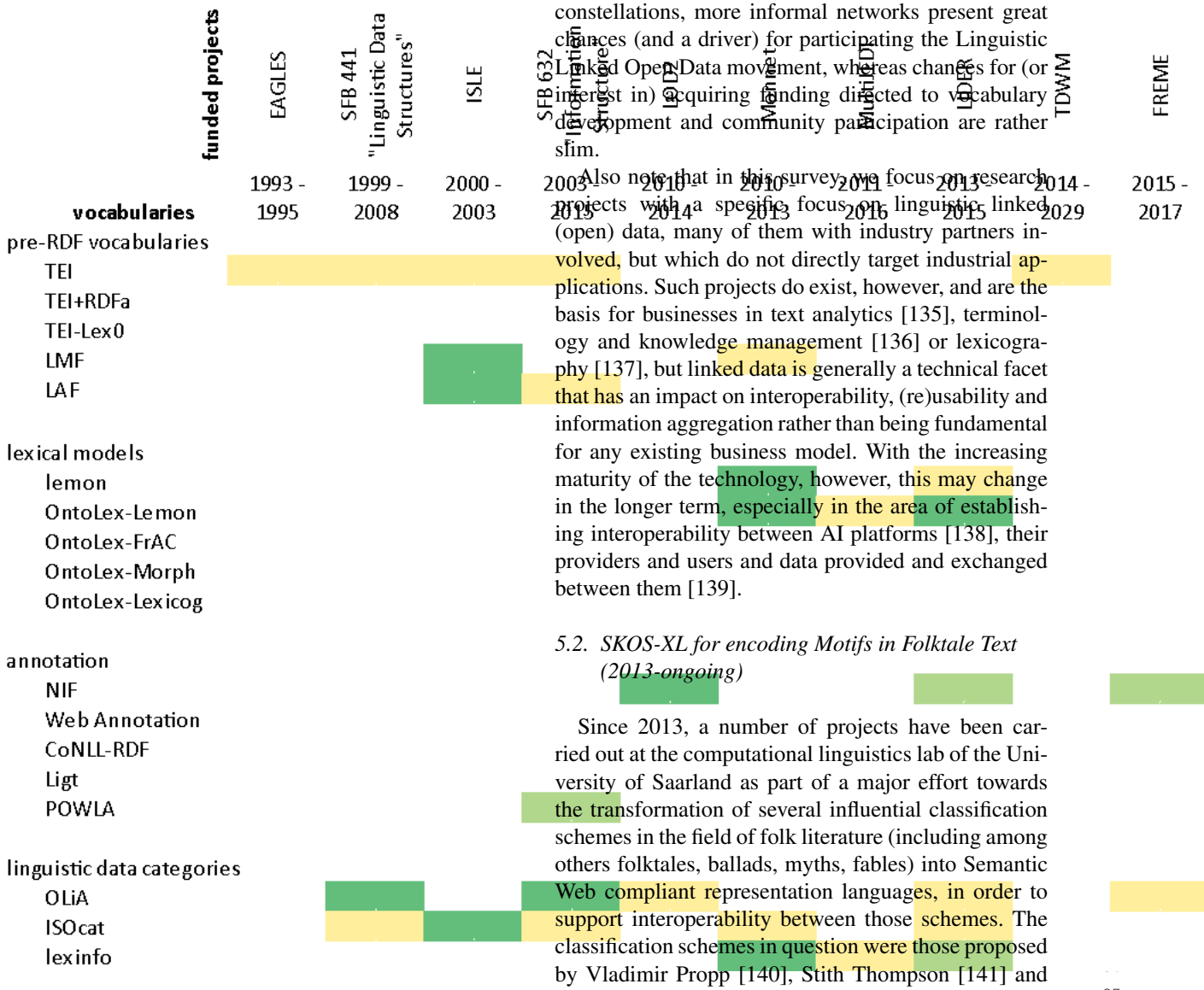

Fig. 4. Usage of and contribution to major LLOD vocabularies by selected research projects

the area of linguistic typology. This is not to say that such projects do not exist, but that the interaction between linguistic typology and language technology operates primarily on informal contacts on mailing lists and via workshops, less in terms of large-scale infrastructure projects, and that, thus, standard development is rarely a priority in typological projects.[133] For such constellations, more informal networks present great chances (and a driver) for participating the Linguistic Linked Open Data movement, whereas chances for (or interest in) acquiring funding directed to vocabulary development and community participation are rather slim.

Also note that in this survey, we focus on research projects with a specific focus on linguistic linked (open) data, many of them with industry partners involved, but which do not directly target industrial applications. Such projects do exist, however, and are the basis for businesses in text analytics [135], terminology and knowledge management [136] or lexicography [137], but linked data is generally a technical facet that has an impact on interoperability, (re)usability and information aggregation rather than being fundamental for any existing business model. With the increasing maturity of the technology, however, this may change in the longer term, especially in the area of establishing interoperability between AI platforms [138], their providers and users and data provided and exchanged between them [139].

5.2. *SKOS-XL for encoding Motifs in Folktale Text (2013-ongoing)*

Since 2013, a number of projects have been carried out at the computational linguistics lab of the University of Saarland as part of a major effort towards the transformation of several influential classification schemes in the field of folk literature (including among others folktales, ballads, myths, fables) into Semantic Web compliant representation languages, in order to support interoperability between those schemes. The classification schemes in question were those proposed by Vladimir Propp [140], Stith Thompson [141] and Anti Aarne, Stith Thompson and Hans-J. Uther [142]. [143] describes how this representation of classification schemes in one machine-readable format, using RDF, RDF(s) and OWL, allowed for their intertwining.

[133] 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.

The work was not limited to classes (and properties), but also to the terms used in the original classification schemes. Those were transformed into (multilingual) SKOS-XL labels. Beyond the encoding of the original terms in SKOS-XL labels, work was dedicated to use of such labels for encoding folktale text sequences, extracted from a manually annotated multilingual folktale corpus, and which are detected as representing motifs that are listed in [141]. Using SKOS-XL for this was very useful, as this opens up the possibility of annotating such motifs as they occur in different (versions of) tales, also in a multilingual context.

5.3. Text Database and Dictionary of Classic Mayan (2014-2029)

The project "Text Database and Dictionary of Classic Mayan" (TWKM, University of Bonn, Germany, 2014-2029) is a long-term project whose goal is to develop a corpus-based dictionary of Mayan hieroglyphic writing. One challenge, specific for ancient languages, is that multiple interpretations of characters and texts need to be represented and re-assessed with the inclusion of new data during dictionary development. This is a common problem due to damaged sources but in case of Mayan language also due to it not being fully deciphered.

The ambitious goal of the project is to develop a near-exhaustive corpus of Classic Mayan which would allow for verifying different reading hypotheses and aid in finishing the decipherment of Maya writing. This poses specific challenges:

- There exist several sign catalogues which might cluster different signs into meanings differently. In order to reflect the state of the art, the catalogue must be linked with other catalogues developed in the fields.
- Each sign in the catalogue should not only link to one or more possible readings, but also provide provenance and associated metadata for each of the reading hypotheses.

To satisfy both requirements, the sign catalogue is formalised in SKOS/RDF. Using top-level properties and concepts from the CIDOC-CRM vocabulary[134] and GOLD, the project develops a 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. [144] provide a detailed description.

This project 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 the project *Sharing Ancient Wisdoms (SAWS, 2010-2013)*[135], 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.

5.4. LiODi (2015-2022)

The *Linked Open Dictionaries* project (LiODi, 2015-2022)[136] aims at developing LLOD-enabled methodologies and infrastructures to facilitate language research for low-resource languages, validating the developments mostly on the languages of the Caucasus.

Within the project, a set of loosely connected tools are being developed, aimed to facilitate language contact studies over lexical and corpus data. One of the primary developing goals is the environment for detecting semantically and phonologically similar words between different languages. Taken together, both components facilitate the detection of possible cognates. Other tools include interfaces for converting, validating, and exploring linguistic data to aid in linguistic research within and outside of the project.

Development and linguistic research are both integral components of the project and tools and pipelines are tested on the data generated and used in the project [105, 145].

From a modelling perspective, the most important contributions of LiODi are that its members have developed and are developing LLD vocabularies for a wide-range of applications in the language sciences, with a particular emphasis on the requirements of low-resource languages, esp. morphologi-

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

[135] http://www.ancientwisdoms.ac.uk/

[136] https://acoli-repo.github.io/liodi/

cally rich languages for which adequate technical support and machine-readable, interoperable data formats currently does not exist. This includes individual, task-specific vocabularies such as Ligt and CoNLL-RDF (see 4.2.6), but also an extension of OntoLex for diachronic relations (cognate and loan relations) [44]. In addition to that, the LiODi project (along with Prêt-à-LLOD, see 5.9) is the main contributor to the ACoLi Dictionary Graph [37].[137] To the best of our knowledge, the ACoLi Dictionary Graph currently represents the most extensive collection of machine-readable bilingual open source dictionaries available, with currently more than 3000 substantial data sets for more than 430 ISO 639-3 languages (including full OntoLex editions of PanLex, Apertium, FreeDict, MUSE, Wikidata, the Open Multilingual WordNets, the Intercontinental Dictionary Series, XDXF and StarDict – the latter only to the extent that the copyright could be clarified and an open license was confirmed).[138]

More significant than lexical resources and novel vocabularies are, however, the contributions of LiODi to the development of community standards for the area. This does includes, among other aspects, significant contributions to the emerging OntoLex Morphology module, initiating and moderating the development of the OntoLex FrAC module and the LD4LT initiative on harmonizing vocabularies for linguistic annotation on the web. Furthermore, LiODi has a strong dedication to disseminating, advocating and education on linked data approaches for linguistics. 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, LDL 2020) and collaborated with international partners in the publication of the first text book on the topic [131] and a number of edited volumes (aside from five volumes of proceedings of the aforementioned events, this includes a collection on linked data for collaborative, data-intense research in the language sciences [146]).

Out of the context of conjoined activities at summer schools/datathons, the project supports numerous external partners in expertise with data modelling and language resource management. Aside from close ties that LiODi thus has with most projects listed here, a notable collaboration with the POSTDATA project and the Academy of Sciences in Heidelberg, Germany, led to the first practical applications of RDFa within TEI editions in the Digital Humanities [77, 147], and ultimately to the development of an official TEI+RDFa customization (see above).

[137] https://github.com/acoli-repo/acoli-dicts

[138] https://panlex.org/, https://www.apertium.org/, https://freedict.org, https://github.com/facebookresearch/MUSE, https://www.wikidata.org/, http://compling.hss.ntu.edu.sg/omw/, https://sourceforge.net/projects/xdxf/, http://stardict.sourceforge.net/

5.5. POSTDATA (2016-2021)

The *Poetry Standardization and Linked Open Data* project [139], aims to bridge the digital gap between traditional cultural assets and the growing sophistication of data modelling and publication practises in the field of DH. The project is built upon two main pillars: linked open data and PoetryLAb, a set of dedicated Natural Language Processing (NLP) tools. It focuses on poetry analysis, applying Semantic Web standards and technologies for enhanced interoperability to numerous different poetry assets. As a result, it will help to share important knowledge about the domain of poetry and publish literary works in the linked open data cloud.

Throughout the project, Postdata have been developing a poetry ontology. This ontology is based on the analysis and comparison of different data structures and metadata coming from eighteen projects and databases devoted to poetry in different languages at the European level, [148–151]. The Postdata ontology is *an encapsulated ontology model*, where domain knowledge has been implemented in 3 layers: *Postdata-core*, *Postdata metrical and literary analysis* and *Postdata-transmission*. This layered ontology, which is based on the re-use of relevant ontologies targeting the project's domain of interest, covers different levels of description from the abstract concept of the poetry work to its bibliographic representation [152–156]. The model is intended to support tasks associated with the analysis of poetry such as close reading, distant reading or critical analysis. The ontologies are modelled in OWL language and will be exposed via SPARQL endpoints.

The Postdata metrical layer is devoted to representing knowledge related to the poetical structure and prosody of a poem and contains salient linguistic, phonetic and metrical concepts. From the metrical point of view, a poem is formed by *stanzas* that contain *lines* that can be understood as a list of *words*. The concept, *word*, is present in OntoLex-Lemon and in NIF.

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

In both cases, the definition of the concept is not sufficient to capture all the knowledge needed for the analysis and description of a word in the metrical context. In the latter case the concept *word* is associated with linguistic information such as *lemma*; phonetic features as a *syllable, foot, feet type onset or coda*; as well as other types of metrical information. However, the intention is to link the Word concept in the Postdata metrical ontology with the Ontolex-Lemon concept Word through the property wordsense, allowing us to capture the range of meanings of the concept. Moreover, the Postdata Word class will also be linked to the NIF Word class since because of the shared relationship to NLP operations.

The second pillar of Postdata consists of a set of tools called PoetryLab, which encompass the several different levels of poetry scholarship, from the most formal analyses relating to scansion, to more cognitive ones which concern metaphor understanding as well as others related to knowledge and subjective perception involving IA 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 including an Name-Entity Recognition system (NER) for medieval place names and organizations, [157] 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, [158–160]. The goal is to use the results of these tools in order to build a knowledge graph in RDF, that is compliant with Postdada ontology.

5.6. MTAAC (2017-2020)

The Machine Translation and Automated Analysis of Cuneiform Languages is a Digging into Data international funded project which associated specialists of cuneiform languages and computational linguists in the development of cutting edge tools for the annotation and distribution of linguistic data of the cuneiform corpus. The objectives of the project were to:[161]

- Formulate, test and evaluate methodologies for the automated analysis and machine translation of transliterated (i.e., transcribed sign-by-sign) cuneiform documents, and provide state-of-the-art technology to specialists in the field of Assyriology;
- Make available the translation of a specific and representative set of cuneiform documents to scholars in related disciplines and to a networked public (namely the Ur III corpus);
- Provide new data for the study of the language, culture, history, economy and politics of the ancient Near East by harvesting the linguistic byproducts of the translation and information extraction processes;
- Formalize these new data utilizing Linked Open Data (LOD, including Linguistic LOD) vocabularies, and foster the practices of standardization, open data and LOD as integral to projects in digital humanities and computational philology.

At the end of 2020, the project had successfully attained its objectives, offering a range of tools, the integration of some of them in a web platform and new data in the form of linguistic annotations and translations, all which are available under open licenses[140] and soon also accessible through the new web platform of the Cuneiform Digital Library Initiative (CDLI https://cdli.ucla.edu) in many forms, including (L)LOD.

Although the overall project's objective was to open the way to developing tools and producing richer linguistic data for all cuneiform languages, MTAAC has focused on a specific corpus of cuneiform: a group of unannotated Sumerian texts issued from the bureaucratic apparatus of the Ur III period (21th century BC)[141]. Another corpus composed of royal inscriptions in the Sumerian language[162], annotated with morphology, was also employed.

Textual information found in the main databases storing Sumerian texts are formatted using variants of the ATF (ASCII Transliteration Format) format. While ORACC[142] stores linguistic annotations (lexical and morphological information only) interlinearly, the Canonical format, released earlier, was deployed paired with a XML annotation format which fell into disuse. In context of this project, the robust and flexible method of storing multi-layer annotations, the CoNLL format, was chosen. A derivative internal format, called CDLI-CoNLL is employed to store the data

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

[141] These texts were extracted from CDLI

[142] Oracc: The Open Richly Annotated Cuneiform Corpus https://oracc.museum.upenn.edu

locally[143] but it can be exported in CoNLL-U format, as well at Brat Standalone format, for better compatibility. The project also employs CoNLL-RDF, in particular so we can integrate with LLOD.

Starting with the original C-ATF transliterations and associated metadata, linguistic annotations are added and stored using thee CDLI-CoNLL format. The data can be further represented as CoNLL-RDF, state in which it can easily be queried, transformed and linked. Capitalizing on this potential, we linked the annotations, lexical information, and metadata.

The ETCSRI morphological annotations[144] were mapped to Unimorph[145] using Turtle-RDF[146], rendeing Sumerian material accessible for cross-lunguistic inquieries. SPARQL is leveraged through CoNLL-RDF for syntactic annotation which are mapped to Universal Dependencies for POS and dependencies lables. Lexical data is linked to guide word entries in the ePSD, employing lemon/ontolex compliant index. The metadata concerning the analysis of the medium of the text and other meta classifications of the texts are mapped to the CIDOC-CRM, following the British Museum's appoach.

Overall, MTAAC prepared 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[105]: 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.

5.7. ELEXIS (2018-2022)

Following on from the European Network for e-Lexicography Cost Action[147], the ELEXIS project is currently undertaking the construction of a European infrastructure for electronic lexicography [163]. LLD will play a key role in the ELEXIS infrastructure, as means of connecting dictionaries and other lexicographic resources both within and across language boundaries. Indeed, the idea of ELEXIS is to eventually construct an network of interlinked electronic lexica and other lexicographic and language resources in several different languages: what is known as a *Matrix Dictionary*. Another important aspect of the project concerns the conversion of legacy lexicographic resources into structured data, and potentially, linked data in order to feed into this Matrix Dictionary.

The main models being used in the project are OntoLex-Lemon and the TEI-Lex0 model [164]. Here it will perhaps be useful to give a brief description of the latter. TEI-Lex0 is a custmomization of the TEI schema[148] that is adapted to the encoding of lexical resources. In particular 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-Lex0 has deprecated elements such as `superEntry` or `entryFree`). This makes the possibility of a crosswalk from (at least a subset of) TEI-Lex0 to OntoLex-Lemon more feasible than, say, a crosswalk from say, a minimal customisation of TEI based on the TEI dictionary guidelines. TEI-Lex0 is being developed by a special working group and (pre-Covid) organised regular in-person training schools, with support from ELEXIS too.

Both OntoLex-Lemon and TEI-Lex0 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-Lex0 and a joint workshop was held between these projects at eLex 2019. Work is also underway on a crosswalk between TEI-Lex0 and OntoLex-Lemon. The latest version of a proposed TEI-Lex0 to OntoLex converter can be found at https://github.com/elexis-eu/tei2ontolex.

The project is also promoting the standardisation of OntoLex-Lemon and TEI-Lex0 through the OASIS working group on Lexicographic Infrastructure Data

[143] This was an essential step to support the preservation of domain specific annotation which are richer than their counterparts found in inguisitc all-encompassing models.

[144] vhttp://oracc.museum.upenn.edu/etcsri/parsing/index.html

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

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

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

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

Model and API (LEXIDMA)[149], which will lead to a new unifying standard for lexicographic data that will be serialised in both OntoLex-Lemon and TEI-Lex0.

ELEXIS aims to provide support for the creation and editing of dictionary resources using both models. 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 as a means of creating and publishing dictionaries and lexica or for converting retrodigitised legacy and historical dictionaries into computational resources. It should be noted that the availability of manuals and targeted teaching materials plays an important factor in 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 exposure to linked data or conceptual modelling in the past. This is usually because 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. 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 can be bridged by such specialist manuals and materials. This is also one of the motivations behind the strong emphasis on training in Nexus Linguaram (see Section 5.10).

Both the production of training materials and the push to promote OntoLex-Lemon as a serialisation format for a standard targeted specifically to lexicographic use cases seems to promise much in terms of the future use of linked data in the context of electronic lexicography. It is inevitable that the experiences of lexicographers and linguists in using OntoLex-Lemon (and its lexicographic extension, see Section 4.1.1) 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. However the fact that it was chosen for ELEXIS shows that OntoLex-Lemon is now regarded as an established model for the creation of lexical resources. It also shows that it is flexible and expressive enough to be regarded as, at least, an acceptable starting point for working towards developing discipline-wide practices for encoding a category of resources as heterogeneous and often as complex in their informational structure as lexica and dictionaries are, as structured, linkable datasets.

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

5.8. LiLa (2018-2023)

The *LiLa: Linking Latin* ERC project[150] aims to connect language resources developed for the study of Latin, bringing the worlds of textual corpora, digital libraries, lexica and tools for Natural Language Processing together. To this end, LiLa makes use of the LOD paradigm and of a set of ontologies from the LLD 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 that involve multiple layers of linguistic annotations and knowledge, such as: what subjects are constructed with verbs formed with a certain prefix [165]? What WordNet synsets do they belong to?

As Latin is characterised by a very rich morphology (where, for instance, a single verb could yield more than 100 forms, excluding the nominal inflection of participles), LiLa focuses on lemmatization as the key task that allows a meaningful and functional connection. 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)[151] as a preliminary step for more advanced forms of analysis.[152]

LLD standards such as OntoLex-Lemon (see Section 4.1) provide an adequate framework to model the relations between the different classes of resources via lemmatization, while also offering a robust solution to model most lexica.

The central component in LiLa's framework, the gateway between the different projects, is the Knowledge Base of canonical forms that are used to lemmatize texts (called 'lemma bank'). This collection was

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

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

[152] 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 [166]. 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.

created starting from the lexical database of the morphological analyzer Lemlat [167], and currently includes a set of about 190,000 forms that are potentially used as lemmas in corpora or lexica.[153]

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 [168]. For a section of ca 36,000 lemmas, the lemma bank includes also derivational information, listing the morphemes (i.e. the prefixes, affixes and lexical bases) that can be identified in each lemma [169].

The fact that OntoLex-Lemon' forms are allowed to have multiple written representations is a particularly helpful feature for a language attested across ca. 25 centuries and a wide spectrum of genres, which is characterised by a substantial amount of spelling variation. Harmonizing different lemmatization solutions adopted by corpora and NLP tools, however, requires practitioners to deal with other kinds of variation as well [170]. 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"), projects may vary considerably in the adopted strategies. For this reasons, the LiLa ontology introduces one sub-class of the Lemma and two object properties that connect forms to forms. The property lemma variant connects two lemmas that can be alternatively used to lemmatize forms of the same words. Hypolemma is a sub-class of the 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, which was modelled using the OntoLex-Lemon extension *lemonEty* [54], respectively on inherited Indo-european lexicon[154] [171] and loans from Greek[155] [172]. The polarity lexicon *LatinAffectus* connects a polarity value (expressed using the Marl ontology) to a general sense for 1,998 entries[156] [173]. Finally, 1,421 verbs from the *Latin WordNet* have been manually revised and published as LOD[157] [174].

In addition to lexica, two annotated corpora are currently linked to the LiLa lemma bank. The *Index Thomisticus* Treebank[158] provides morpho-syntactic annotation for 375,000 tokens from the Latin works of Thomas Aquinas (13th century CE), while the *Dante Search* corpus[159] includes the lemmatised 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 [175].[160] 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[161], which takes an instance of the Lemma class as its range, since no existing vocabulary provided a suitable way to express this relation.

5.9. Prêt-à-LLOD (2019-2022)

The Prêt-à-LLOD project's goal is to make linguistic linked open data 'ready-to-use' and part of this mission is to support the development of new vocabularies for linguistic linked data. In particular, Prêt-à-LLOD has provided significant support to the development of the OntoLex-Lemon module, including the development of a model for lexicography, morphology and corpus information (all of which are discussed in Section 4.1).

Additionally, 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 govern-

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

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

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

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

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

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

[159] http://lila-erc.eu/data/corpora/DanteSearch/id/corpus

[160] https://universaldependencies.org/guidelines.html

[161] https://lila-erc.eu/lodview/ontologies/lila/

ment services, pharmaceutical industry, and finance, details of which are described in [176] 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 [177].

5.10. NexusLinguarum (2019-2023)

The *European network for Web-centred linguistic data science* (NexusLinguarum)[162] is a COST Action project that involves researchers from 42 countries. The network started in October 2019 and will continue its activities for four years. The Action promotes synergies across Europe between linguists, computer scientists, terminologists, language professionals, and other stakeholders from both industry and society, in order to investigate 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.

On 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 4). 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[163]), and training schools (e.g., EuroLAN 2021[164]), where linguistic linked data will take 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 some prototypical cases. In a first phase, the definition of use cases covers Humanities and Social Sciences, Linguistics (Media and Social Media, and Language Acquisition), Life Sciences, and Technology (Cybersecurity and FinTech). NexusLinguarum also places a strong emphasis on lesser resourced languages. As an example of the kinds of complex, heterogeneous resources which have been proposed by consortium members as candidates for 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[165] is working towards adding the Romanian language to the multilingual Linguistic Linked Open Data cloud[166]. There are four different ReTeRom components. These are CoBiLiRo, SINTERO[167], TEPRO-

[162] https://nexuslinguarum.eu/

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

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

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

[166] 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/)

[167] 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 [178]

LIN[168] and TADARAV[169]. We will focus on CoBiLiRo,

CoBiLiRo (*Bimodal Corpus for Romanian Language*), coordinated by the "Alexandru Ioan Cuza" University from Iași (UAIC), is working with a large collection of parallel speech/text data [181]. This collection is annotated on different levels on both the acoustic and the linguistic components [182], which facilitates searching, editing and statistical analysis operations over it. 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 by various language resource builders; (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 [183] and based on the idea of alignment between speech and the text components, taking into consideration several annotation conventions proposed in 2007 by Li and Zhi-gang [184]. At present, the header of this format includes the following metadata: *source of the object stored*; *speaker's gender*; *speaker's identity (if she/he agreed to this)*; *vocal type (spontaneous or inreading)*; *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 the preceeding example (one of many within the project, falling within several different disciplines or technical domain) demonstrates NexusLinguaram's potential as a testing ground for many of the new vocabularies and modules mentioned above (as well as for the potential of linked data as a paradigm for the modelling and publication of language data) through the analysis of complex multifaceted use cases involving several different types of language resources.

6. Conclusions

In this article we have attempted to give a comprehensive description of the current state of affairs with respect to the use, definition and availability of LLD models and vocabularies. As we hope that the article has demonstrated this is an very active, very dynamic area of research, with numerous projects and initiatives underway (or due to commence in the short term) which promise to bring further changes and improvements in expressivity and coverage (i.e., in addition to the changes discussed here). For this reason we have tried to place our description of recent advances in the field within a discussion of more general, ongoing trends. There are of course numerous future challenges to be confronted, as well as areas of immense opportunity, many of which have been detailed in this survey. But this is natural in a field as relatively youthful and as full of potential as linguistic linked data.

Acknowledgments

This article is based upon work from COST Action NexusLinguarum (CA18209), supported by COST (European Cooperation in Science and Technology). The authors thank Milan Dojchinovski for several very helpful suggestions.

[168] TEPROLIN (*Technologies for Processing Natural Language - Text*) which is coordinated by the Research Institute for Artificial Intelligence "Mircea Drăgănescu" (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 – subsyntactic (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 [179]

[169] 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 [180] 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.

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, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet and C. Soria, Lexical markup framework (LMF), 2006.

[4] E. de la Clergerie and L. Clément, MAF: a morphosyntactic annotation framework, *Actes de LTC* (2005), 90–94.

[5] 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.

[6] 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.

[7] 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: *The sixth biennial conference on electronic lexicography, eLex 2019*, 2019.

[8] 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.

[9] 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.

[10] E. Hyvönen, "Sampo" Model and Semantic Portals for Digital Humanities on the Semantic Web., in: *DHN*, 2020, pp. 373–378.

[11] 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.

[12] P. Cimiano, C. Chiarcos, J.P. McCrae and J. Gracia, *Linguistic Linked Data: Representation, Generation and Applications*, Springer International Publishing, Cham, 2020.

[13] A. Pareja-Lora, M. Blume, B.C. Lust and C. Chiarcos (eds), *Development of linguistic linked open data resources for collaborative data-intensive research in the language sciences*, MIT Press, Cambridge, 2019. ISBN 978-0-262-53625-7.

[14] 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.

[15] Y. Le Franc, J. Parland-von Essen, L. Bonino, H. Lehväslaiho, G. Coen and C. Staiger, D2.2 FAIR Semantics: First recommendations (2020). doi:10.5281/ZENODO.3707985. https://zenodo.org/record/3707985.

[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.

[17] P.-Y. Vandenbussche and B. Vatant, Metadata recommendations for linked open data vocabularies, *Version* **1** (2011), 2011–12.

[18] 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.

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

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

[21] 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 Evaluation (LREC'14)*, N.C.C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014. ISBN 978-2-9517408-8-4.

[22] B. Klimek, N. Arndt, S. Krause and T. Arndt, Creating Linked Data morphological language resources with MMoOn (2016).

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

[24] 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)*, 2008.

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

[26] H. Aristar-Dry, S. Drude, M. Windhouwer, J. Gippert and I. Nevskaya, Rendering endangered lexicons interoperable through standards harmonization: the relish project, in: *LREC 2012: 8th International Conference on Language Resources and Evaluation*, European Language Resources Association (ELRA), 2012, pp. 766–770.

[27] 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.

[28] C. Chiarcos and M. Sukhareva, Olia – ontologies of linguistic annotation, *Semantic Web* **6**(4) (2015), 379–386.

[29] C. Chiarcos, C. Fäth and F. Abromeit, Annotation Interoperability for the Post-ISOCat Era, in: *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 5668–5677.

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

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

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

[33] 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.

[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.

[35] G. De Melo, Lexvo.org: Language-related information for the linguistic linked data cloud, *Semantic Web* **6**(4) (2015), 393–400.

[36] P. Westphal, C. Stadler and J. Pool, Countering language attrition with PanLex and the Web of Data, *Semantic Web* **6**(4) (2015), 347–353.

[37] C. Chiarcos, C. Fäth and M. Ionov, The ACoLi dictionary graph, in: *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 3281–3290.

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

[39] C. Chiarcos, S. Nordhoff and S. Hellmann, *Linked Data in Linguistics*, Springer, 2012.

[40] J.P. McCrae, S. Moran, S. Hellmann and M. Brümmer (eds), Semantic Web, Vol. 6, 2015, pp. 313–400. https://content.iospress.com/journals/semantic-web/6/4.

[41] 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.

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

[43] 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.

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

[45] 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.

[46] 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 (Ireland), 2017, pp. 74–84. ISSN 1613-0073. http://ceur-ws.org/Vol-1899/OntoLex{_}2017{_}paper{_}5.pdf.

[47] 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 the eLex 2019 conference*, 2019, pp. 726–746.

[48] 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: *Proceedings of Sixth Biennial Conference on Electronic Lexicography, eLex 2019*, 2019. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_33.pdf.

[49] 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 Globalex Workshop on Linked Lexicography (@LREC 2020)*, 2020, pp. 1–9. https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/GLOBALEX2020book.pdf#page=19.

[50] 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.

[51] 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.

[52] S. Stolk, lemon-tree: Representing Topical Thesauri on the Semantic Web, in: *2nd Conference on Language, Data and Knowledge (LDK 2019)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[53] 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.

[54] A.F. Khan, Towards the Representation of Etymological Data on the Semantic Web, *Information* **9**(12) (2018), 304.

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

[56] 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.

[57] 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.

[58] 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/.

[59] 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.

[60] 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.

[61] 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.

[62] 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)*, 2016, pp. 3586–3591.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

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

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

[69] 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.

[70] 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.

[71] N.M. Ide and C.M. Sperberg-McQueen, The TEI: History, Goals, and Future, in: *Text Encoding Initiative: Background and Context*, N. Ide and J. Véronis, eds, Springer Netherlands, Dordrecht, 1995, pp. 5–15. ISBN 978-94-011-0325-1. doi:10.1007/978-94-011-0325-1_2.

[72] L. Burnard, Report of Workshop on Text Encoding Guidelines, *Literary and Linguistic Computing* **3**(2) (1988), 131–133. https://academic.oup.com/dsh/article-abstract/3/2/131/1020443.

[73] C.F. Goldfarb, Information Processing: Text and Office Systems: Standard Generalized Markup Language (SGML), Technical Report, International Organization for Standardization (ISO), Geneva, Switzerland, 1986.

[74] A. Bellandi, E. Giovannetti and A. Weingart, Multilingual and multiword phenomena in a lemon old occitan medico-botanical lexicon, *Information* **9**(3) (2018), 52.

[75] A. Bellandi and E. Giovannetti, Involving Lexicographers in the LLOD Cloud with LexO, an Easy-to-use Editor of Lemon Lexical Resources, in: *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, 2020, pp. 70–74.

[76] 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)*, J.P. McCrae, C. Chiarcos, T. Declerck, J. Gracia and B. Klimek, eds, European Language Resources Association (ELRA), Paris, France, 2018. ISBN 979-10-95546-19-1.

[77] 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).

[78] ISO, Language Resource Management - Linguistic Annotation Framework (LAF), Standard, International Organization for Standardization, Geneva, 2012, Project leader: Nancy Ide.

[79] C. Chiarcos, POWLA: Modeling Linguistic Corpora in OWL/DL, in: *The Semantic Web: Research and Applications*, E. Simperl, P. Cimiano, A. Polleres, O. Corcho and V. Presutti, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 225–239. ISBN 978-3-642-30284-8.

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

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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)*, 2016, pp. 457–462.

[89] 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.

[90] 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.

[91] 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.

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

[93] 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.

[94] 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.

[95] S. Evert and A. Hardie, Twenty-first Century Corpus Workbench: Updating a Query Architecture for the New Millennium, in: *Proceedings of the Corpus Linguistics 2011 Conference*, Birmingham, UK, 2011, pp. 1–21.

[96] A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý and V. Suchomel, The Sketch Engine: Ten Years On, *Lexicography* **1**(1) (2014), 7–36.

[97] C. Chiarcos and C. Fäth, CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way, in: *Language, Data, and Knowledge*, J. Gracia, F. Bond, J.P. McCrae, P. Buitelaar, C. Chiarcos and S. Hellmann, eds, Springer, Cham, Switzerland, 2017, pp. 74–88. ISBN 978-3-319-59888-8.

[98] 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.

[99] 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)*, 2019, pp. 74–81.

[100] 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.

[101] 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)*, Miyazaki, Japan, 2018, pp. 4525–4534.

[102] 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.

[103] 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)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[104] 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.

[105] 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.

[106] 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)*, ELRA, Marseille, France, 2020, pp. 7161–7169.

[107] C. Chiarcos, POWLA: Modeling Linguistic Corpora in OWL/DL, in: *The Semantic Web: Research and Applications*, Vol. 7295, D. Hutchison, T. Kanade, J. Kittler, J.M. Kleinberg, F. Mattern, J.C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M.Y. Vardi, G. Weikum, E. Simperl, P. Cimiano, A. Polleres, O. Corcho and V. Presutti, eds, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 225–239, Series Title: Lecture Notes in Computer Science. ISBN 978-3-642-30283-1 978-3-642-30284-8.

[108] P. Cimiano, C. Chiarcos, J.P. McCrae and J. Gracia, *Modelling Linguistic Annotations*, in: *Linguistic Linked Data*, Springer, 2020, pp. 89–122.

[109] 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), Malta, 2010, pp. 43–47. http://www.lrec-conf.org/proceedings/lrec2010/pdf/163_Paper.pdf.

[110] D. Broeder, D. van Uytvanck, M. Gavrilidou, T. Trippel and M. Windhouwer, Standardizing a Component Metadata Infrastructure, in: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N.C.C. Chair), K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Associ-

ation (ELRA), Istanbul, Turkey, 2012. ISBN 978-2-9517408-7-7.

[111] 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.

[112] 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)*, N.C.C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014. ISBN 978-2-9517408-8-4.

[113] 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.

[114] 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 12th Language Resources and Evaluation Conference (LREC 2020)*, N. Calzolari, F. Bćhet, P. Blache, C. Cieri, K. Choukri, T. Declerck, H. Isahara, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 3421–3430.

[115] 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. http://www.lrec-conf.org/proceedings/lrec2012/pdf/998_Paper.pdf.

[116] 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. https://link.springer.com/chapter/10.1007/978-3-319-25639-9_42.

[117] S. Piperidis, The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions, in: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N.C.C. Chair), K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, eds, European Language Resources Association (ELRA), Istanbul, Turkey, 2012. ISBN 978-2-9517408-7-7.

[118] 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.

[119] 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), Miyazaki, Japan, 2018, pp. 7–12. ISBN 979-10-95546-20-7. http://lrec-conf.org/workshops/lrec2018/W24/pdf/13_W24.pdf.

[120] 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.

[121] 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: Resources and Applications*, Association for Computational Linguistics, Beijing, China, 2015, pp. 39–48. doi:10.18653/v1/W15-4205. https://www.aclweb.org/anthology/W15-4205.

[122] 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 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 2020, pp. 3366–3380. ISBN 979-10-95546-34-4. https://www.aclweb.org/anthology/2020.lrec-1.413.

[123] 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.

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

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

[126] 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.

[127] 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.

[128] 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 the eLex 2019 conference*, 2019, pp. 1–3.

[129] F. Gillis-Webber and S. Tittel, A Framework for Shared Agreement of Language Tags beyond ISO 639, in: *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 3333–3339.
[130] S. Nordhoff, Linked data for linguistic diversity research: Glottolog/langdoc and asjp online, in: *Linked Data in Linguistics*, Springer, 2012, pp. 191–200.
[131] P. Cimiano, C. Chiarcos, J.P. McCrae and J. Gracia, *Linguistic Linked Data*, Springer, 2020.
[132] 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.
[133] 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.
[134] A. Weingart and E. Giovannetti, A Lexicon for Old Occitan Medico-Botanical Terminology in Lemon., in: *SWASH@ ESWC*, 2016, pp. 25–36.
[135] M. Hartung, M. Orlikowski and S. Veríssimo, Evaluating the Impact of Bilingual Lexical Resources on Cross-lingual Sentiment Projection in the Pharmaceutical Domain, 2020.
[136] 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.
[137] 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/.
[138] 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. Vasiļjevs 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, Marseille, France, 2020, pp. 96–107. ISBN 979-10-95546-64-1. https://www.aclweb.org/anthology/2020.iwltp-1.15.
[139] Slator, Slator 2021 Data-for-AI Market Report, Technical Report, 2021.
[140] V. Propp, *Morphology of the folktale*, Trans., Laurence Scott. 2nd ed., University of Texas Press, 1968.
[141] 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.
[142] 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.
[143] T. Declerck, A. Kostová and L. Schäfer, Towards a Linked Data Access to Folktales classified by Thompson's Motifs and Aarne-Thompson-Uther's Types, in: *Proceedings of Digital Humanities 2017*, ADHO, 2017.
[144] 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.
[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.
[146] A. Pareja-Lora, B. Lust, M. Blume and C. Chiarcos, *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*, The MIT Press, 2019.
[147] H.B.-S. Sabine Tittel and C. Chiarcos, Using RDFa to Link Text and Dictionary Data for Medieval French, in: *Proc. of the 6th Workshop on Linked Data in Linguistics (LDL-2018): Towards Linguistic Data Science*, European Language Resources Association (ELRA), Paris, France, 2018. ISBN 979-10-95546-19-1.
[148] 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.
[149] 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.
[150] 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.
[151] 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. https://doi.org/10.1515/9783110585421-004.
[152] 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.
[153] P.E. project, Network of ontologies - POSTDATA, Postdata ERC project, [Online; accessed 2021-01-17]. http://http://postdata.linhd.uned.es/results/.
[154] P.E. project, Postdata-core ontology, Postdata ERC project, [Online; accessed 2021-01-17]. http://http://postdata.linhd.uned.es/results/.
[155] P.E. project, Postdata-prosodic ontology, Postdata ERC project, [Online; accessed 2021-01-17]. http://http://postdata.linhd.uned.es/results/.
[156] P.E. project, Postdata-structural ontology, Postdata ERC project, [Online; 2021-01-17]. http://http://postdata.linhd.uned.es/results.
[157] 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.

[158] 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.

[159] 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/.

[160] 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).

[161] 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.

[162] G. Zólyomi, B. Tanos and S. Sövegjártó, The Electronic Text Corpus of Sumerian Royal Inscriptions, 2008.

[163] 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.

[164] 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 Conference*, 2017.

[165] 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.

[166] 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. https://www.aclweb.org/anthology/2020.lt4hala-1.16.

[167] 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.

[168] C. Chiarcos and M. Sukhareva, Olia – ontologies of linguistic annotation, *Semantic Web* **6**(4) (2015), 379–386, Publisher: IOS Press.

[169] 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. https://www.aclweb.org/anthology/W19-8505.

[170] 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:10.18653/v1/W19-4009. https://www.aclweb.org/anthology/W19-4009.

[171] 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, Marseille, France, 2020, pp. 20–28. ISBN 979-10-95546-46-7. https://www.aclweb.org/anthology/2020.globalex-1.3.

[172] 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. http://ceur-ws.org/Vol-2769/paper_06.pdf.

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

[174] 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.

[175] 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.

[176] 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)*, N. Calzolari, F. Béchet, P. Blache, C. Cieri, K. Choukri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk and S. Piperidis, eds, ELRA, 2020, pp. 5660–5667, ELRA.

[177] 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)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

[178] 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.

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

[180] A.L. Georgescu, H. Cucu, A. Buzo and C. Burileanu, RSC: A Romanian Read Speech Corpus for Automatic Speech Recognition, in: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 2020, pp. 6606–6612.

[181] 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.

[182] 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.

[183] C.M. Sperberg-McQueen and L. Burnard, Original editors, revised and expanded under the supervision of the Technical Council of the TEI Consortium. TEI P5: Guidelines for Electronic Text Encoding and Interchange, 2018.

[184] A. Li and Z. Yin, Standardization of Speech Corpus, in: *Data Science Journal*, Vol. 6, 2007.