UDK 81’374
81’322
Izvorni znanstveni rad
Rukopis primljen 10. XII. 2022.
Prihvaćen za tisak 28. IV. 2023.
https://doi.org/10.31724/rihjj.49.1.7

Penny Labropoulou
Institute for Language and Speech Processing, Athena RC
Epidavrou & Artemidos, Athens GR-15125
orcid.org/0000-0001-5123-7890
penny@athenarc.gr
David Lindemann
UPV/EHU University of the Basque Country, Faculty of Arts
Unibertsitatearen Ibilbidea, 5 ES-01006 Vitoria-Gasteiz
orcid.org/0000-0002-8261-6882
david.lindemann@ehu.eus
Christiane Klaes
TU Braunschweig, University Library
Universitätsplatz 1, DE-38106 Braunschweig
orcid.org/0000-0003-4870-4392
c.klaes@tu.braunschweig.de
Katerina Gkirtzou
Institute for Language and Speech Processing, Athena RC
Epidavrou & Artemidos, Athens GR-15125
orcid.org/0000-0002-4725-3094
katerina.gkirtzou@athenarc.gr

THE LEXMETA MODEL FOR LEXICAL RESOURCES: THEORETICAL AND IMPLEMENTATION ISSUES

This paper presents LexMeta, a metadata model for the description of lexical resources, such as dictionaries, word lists, glossaries, etc., to be used in language data catalogues mainly targeting the lexicographic and broader humanities communities but also users exploiting such resources in their research and applications. A comparative review of similar models is made in order to show the differences and commonalities with LexMeta. To enhance semantic interoperability and support the exchange of (meta)data across disciplinary and general catalogues, the most influential models for our purposes, namely FRBR (used in library catalogues) and META-SHARE (used for language resources), are selected as a base for the design of LexMeta. We discuss how these models are aligned and extended with new properties as required for the description of lexical resources. The formal representation

of the model following the Linked Data paradigm aims to further enhance the semantic interoperability. The choice to implement it in two formats (as an RDF/OWL and as a Wikibase ontology) facilitates its adoption and hence its enrichment, yet poses challenges as to their synchronisation, which are addressed through automatic workflows. We conclude with ongoing and planned activities for the improvement of the model.

1. Introduction

The rise of data-driven methods in research and the increase of digital and digitised materials used as an object of study and/or ancillary resources applied for its processing has put data in a central place in all scientific disciplines and made its discovery and (re-)use of crucial importance. The FAIR Data Principles (Wilkinson et al. 2016) have defined four properties aiming to optimise the use of data, putting 'specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals': Findable, Accessible, Interoperable and Reusable. Data is available through discipline-specific repositories, aggregated catalogues (e.g., EOSC,[1] Dataset Search[2]) and nowadays Data Spaces (cf. European Strategy for Data,[3] IDSA,[4] Gaia-X,[5] etc.), enhancing their (re-)usability, yet making their discovery a challenging task. **Metadata** (data descriptions), acting as the intermediary between consumers and digital resources, plays an instrumental role in this endeavour. **Semantic interoperability** of the (meta)data and, in consequence, of the metadata vocabularies is key to their re(use).

In this paper, we present **LexMeta**, a metadata model catering for the description of human-readable and computational lexical resources from the perspective of language data catalogues (Lindemann et al. 2022), and discuss the way its design and implementation facilitate interoperability.

We use the term *lexical resource* as a cover term for all types of resources such as term lists, glossaries, dictionaries, morphological lexica, ontologies, etc., organised on the basis of lexical or conceptual units (lexical items, terms, con-

[1] https://eosc-portal.eu/
[2] https://datasetsearch.research.google.com/
[3] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066
[4] https://internationaldataspaces.org/
[5] https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html

cepts, phrases, etc.) with their supplementary information (e.g., grammatical, semantic, statistical information, etc.), regardless of their distribution format and intended use (manuscript or printed paper, digital resource on CD-ROM or accessible through an online interface, for consultation by human users, or structured dataset to be used in natural language processing applications such as spell checking or machine translation).[6]

Section 2 gives an overview of models with a similar scope to LexMeta, shedding light on the different perspectives with which they approach lexical resources. Section 3 presents LexMeta, and the principles upon which it has been built, while Section 4 focuses on its formal implementation in a dual format, explaining the reasons for this, and the implications thereof. We conclude with a summary of ongoing and planned activities for the enhancement of the model.

2. Related work

The landscape of metadata used in catalogues is composed of a wide range of models and vocabularies due to the divergent needs and purposes these catalogues are set to respond to: different objects of description (e.g., datasets, software, publications), target users (e.g., all types of users for general catalogues vs. scholars of specific disciplines, with established descriptive practices and specialised requirements), purposes (e.g., preservation, dissemination, processing), etc.

The most relevant for our work are those used for the description of lexical resources in catalogues catering for digital humanities scholars, as well as the more popular general ones, often used for the exchange of metadata and as foundations for application profiles customised to the needs of specific use cases.

A de-facto standard for general catalogues is **DCAT** (Data Catalog Vocabulary),[7] an RDF vocabulary for data catalogues, in use by European data portals and a lot of disciplinary data catalogues, and adopted in the core ontologies of Data

[6] The definition is taken from the term *lexical/conceptual resource* introduced in the META-SHARE ontology (http://w3id.org/meta-share/meta-share/LexicalConceptualResource), but referred to as *lexical resource* for reasons of brevity.

[7] https://www.w3.org/TR/vocab-dcat-3/

Spaces, such as the IDS Information Model.[8] Although it has no elements specific to lexical resources, we briefly describe it here due to the influence of its design on the description of all types of datasets.[9] Each dataset is described through two core classes: *Dataset* represents 'a collection of data, published or curated by a single agent or identifiable community; the notion of dataset is broad and inclusive, covering data in many forms, including numbers, text, pixels, imagery, sound and other multi-media, and potentially other types', and can, consequently, be used for any lexical resource. The class *Distribution* represents any accessible form of a dataset, such as printed paper, a downloadable file, a dictionary available through a mobile application, etc.

Lexical resources are typically included in library catalogues. The most popular based on Linked Data principles model in these catalogues is **FRBR** (Functional Requirements for Bibliographic Resources), a conceptual model for describing bibliographic metadata (IFLA Study Group for FRBR 2009). A more recent version is included in the FRBR-LRM (Library Reference Model, Riva et al. 2017), and LRMoo, a formal ontology seeking to bring together bibliographic and museum information (International Working Group on LRM, FRBR and CIDOC CRM Harmonisation 2021). FRBR defines as distinct classes the concepts of *Work* (an abstract notion for any creative creation), *Expression* (the realisation of a single work, such as a certain version or edition), and *Manifestation* (the distribution of a single realisation, e.g., on paper, or as a digital dataset).[10]

Digital and digitised lexical resources are also included in catalogues for digital humanities researchers, such as the European Language Grid[11] (Rehm et al. 2021; Piperidis et al. 2022) and META-SHARE[12] (Piperidis 2012) and those maintained in CLARIN national centres[13] (Eskevitch et al. 2020).

The META-SHARE model, initially implemented as an XML schema (Gavrilidou et al. 2012) and later as an RDF/OWL ontology **(MS-OWL** or, hereafter,

[8] https://international-data-spaces-association.github.io/InformationModel/docs/index.html

[9] schema.org, used for the Google Dataset Search engine, is also extremely popular but with no specific elements for lexical resources; with regard to data, it is also based on DCAT (for more details, see https://schema.org/docs/data-and-datasets.html).

[10] The fourth class, *item*, is used for the physical instances, e.g. the copies of a book, and is thus outside our scope.

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

[12] http://www.meta-share.org/

[13] https://www.clarin.eu/content/language-resources

MS) (McCrae et al. 2018; Khan et al. 2022: 27–30) caters for language resources, i.e., data resources (structured or unstructured datasets, lexica, language models, etc.) and technologies used for language processing (taggers, machine translation applications, etc.). It builds around three key concepts: *resource type*, *media type*[14] and *distribution*, which give rise to the core classes of the model. The lexical/conceptual resource covers all types of lexica, dictionaries, glossaries, ontologies, etc. The MS model defines a broad range of mandatory, recommended and optional properties catering for the full lifecycle of the resource, from inception and creation to actual usage; the optionality status aims to support discovery and interoperability both for humans and machines. Following DCAT, the *DatasetDistribution* class (conceived especially for data resources) is used for the description of the distributable forms of a resource.

Properties are assigned to the most relevant class. Thus, a set of properties common across all resources (e.g., descriptive and administrative metadata), are assigned to the *LanguageResource*, while more technical features and classification elements are attached to the appropriate subclasses. For instance, the *LexicalConceptualResource* takes classification properties that encode the subtype (e.g., computational lexicon, ontology, dictionary, etc.), and the contents of the resource (unit of description, types of linguistic and extralinguistic information, etc.). The *DatasetDistribution* class provides information on how and where to access the resource, technical features of the physical files (such as size and format), and licensing terms. Thus, different distributions with the same content can be modelled without redundancies in the content description, since that is attached to one entity of class *LexicalConceptualResource*; several *DatasetDistribution* entities can be linked to the former. This resembles the use of content-describing FRBR *Expression* entities with (different) *Manifestation* entities, such as a dictionary available as a book or on CD-ROM, or even as online publications, or a bilingual computational lexicon in CSV format and in a binary format ready to be consumed by a software program.

CLARIN catalogues do not use a single metadata model. Instead, to better support the varying metadata practices of research communities, CLARIN has initiated the Component MetaData Infrastructure (CMDI, Broeder et al. 2012;

[14] We do not present the *media type* concept, as in the LexMeta current version we have focused on textual resources and multimedia dictionaries that have text as the main feature.

International Organization for Standardization 2020), which provides a framework for the description and reuse of metadata blueprints, in the form of *profiles* created from common building blocks (*components*) that group together semantically similar elements (e.g., elements describing locations, persons, organisations, etc.). Both components and profiles are stored in the same registry[15] and can be re-used and extended. The public registry at the time of writing contains 13 profiles catering for lexical resources, some of which are similar and some targeting specific areas (e.g., historical linguistics). A CMDI-compatible version of the MS model is included among these profiles.

Finally, the **TEI** (Text Encoding Initiative) guidelines for the representation of texts in digital form[16] is one of the standards most widely adopted by the digital humanities community. The TEI specifications for digital dictionaries[17] have a wider scope than ours as they purport to represent the contents, structure and creation process of dictionaries. Of relevance to this work is the TEI header, conceived as roughly corresponding to the title page of an electronic book, including all information related to the encoding of a text. Although the header provides a detailed account of the described resource, including the elements required for a bibliographic record, it is not used for cataloguing purposes. Additional work is contained in the TEI-Lex0 technical specification,[18] part of which (properties from the header) are taken into account.

It should be noted that there are other models relevant for lexical resources which we, however, do not present here due to their limited scope over the type of metadata that concerns us. For instance, the influential Lexical Markup Framework (LMF, International Organization for Standardization 2019), a metamodel for the representation of data in lexical resources, focuses on the representation of their content. With regard to metadata, it includes two classes: *LexicalResource* (container for the whole resource), and *GlobalInformation* (for administrative metadata, mentioning only three attributes, namely a mandatory one for language, and two optional ones for script and character encoding).

[15] https://catalog.clarin.eu/ds/ComponentRegistry/
[16] https://tei-c.org/
[17] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html
[18] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#

3. Presentation of the LexMeta model

3.1 Objectives and approach

LexMeta aims to cater for the description of lexical resources included in catalogues of libraries and language data repositories. It must satisfy the requirements and needs of the respective catalogue users but also have a broader outlook, considering recent developments and initiatives in the metadata and data-related areas.

In order to cover a broad range of target users from different communities and with different user needs, an inventory of potential properties was created, taking into account various applications and descriptive practices of the target communities. For instance, bibliographic metadata (e.g., title, author(s), publication date) required for citing the resource, format and size properties deemed important for discovering tools that can be used for editing or visualising them, licence under which they can be used for developing new products, keywords that can be used for easier identification, relations to publications that describe, or review them and can be consulted for further information, etc. Metadata elements from other catalogues have also been explored.

The formal representation of these properties has also considered related state-of-the-art models and vocabularies to ensure interoperability and support exchange of data with catalogues of the target community as well as of other communities. In addition, extensibility of the model with new properties was one of the desiderata given the evolving data landscape.

Following a survey of the most relevant models, we decided to base the LexMeta conceptual model upon the MS and FRBR models, aligning and extending them with new concepts in accordance with the needs of our application. Elements from other established vocabularies (e.g., DCMI[19] for general properties, and BIBO[20] for bibliographic metadata) are also re-used, when possible, while new elements have been introduced in a uniform model in case of witnessed gaps (cf. next subsection).

[19] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[20] https://www.dublincore.org/specifications/bibo/bibo/

One of the main features of the model is the definition of a minimal subset of properties as mandatory. This ensures that we can accommodate in an optimal way the exchange of metadata with catalogues that may not include all metadata elements, and it tries to minimise the reluctance of human users to add metadata for their resources. At the same time, it supports the creation of a core metadata record with at least the information that is considered sine qua non for the intended uses.

One of the problems encountered when creating a model is the degree of freedom to be allowed for the value range of specific elements. Human editors of metadata prefer free text, yet this creates inconsistencies between the values (e.g., spelling variants or errors, alternative terms). On the other hand, closed vocabularies do not allow for the addition of new values. We thus decided to use controlled vocabularies yet with a continuous curation of the vocabularies to import new terms when needed. For this reason, flexible editing tools that support versioning, ease-of-export of the new versions and direct import into the applications that exploit them are an important asset (cf. Section 4).

3.2 Overview of the model

LexMeta is built around three main classes, which follow the relevant FRBR and MS conceptual distinctions: [21]

- *LCR Series* corresponds to the abstract notion of a lexicographic work;
- *Lexical/Conceptual Resource* (*LCR*) represents the realisation of a single work, such as a certain version or edition;
- *Dataset Distribution* corresponds to the physical form in which a lexical/conceptual resource is realised (e.g., as a printed book or as a digital file).

As depicted in Figure 1, the LexMeta/MS classes are mapped as subclasses (with the relation *rdfs:subclassOf*) to the corresponding FRBR classes. We have opted

[21] Khan and Salgado (2021: 9), in their paper on modelling lexical resources with FRBRoo (previous version of the LRMoo presented in Section 2) and Ontolex, also propose two classes for lexical resources, *Lexicographic Work* and *Lexicographic Expression,* as subclasses of the FRBR classes *Work* and *Expression* respectively, but do not discuss versions or different distributable forms; their focus is on the representation of the inner structure and contents of lexical resources.

for using the *rdfs:subclassOf* and *rdfs:subpropertyOf* when aligning the LexMeta elements to other vocabularies, because an equivalence relation might result in unwanted or inconsistent inference rules for the data described with them.

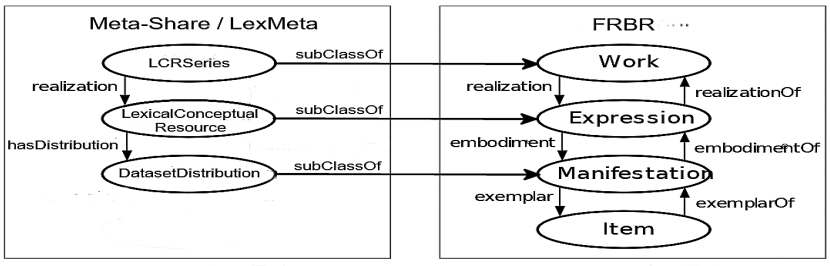

Figure 1. Alignment between LexMeta/MS and FRBR classes

The distinction of the three classes, used in both models, allows us to group and link different publications (e.g., print publications, reprints, and digital versions, or computational lexica distributed in XML or RDF format) with the same content as well as to describe them more consistently by attaching their properties at the appropriate level. *LCR Series* groups the various editions and versions (*LCRs*) of the same abstract *work*. Content-describing metadata are common across *Distributions* of the same *LCR* and are assigned to the *LCR* level. Publication metadata and technical features are attached at the *Distribution* level. Figure 2 depicts the main properties attached at the *LCR* and *Distribution* level.

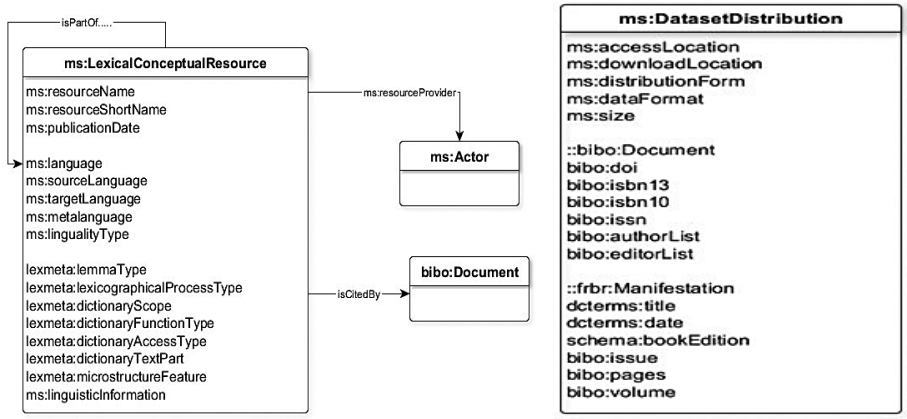

Figure 2: Properties on *LCR* and *DatasetDistribution* (excerpts)

Identification properties (title, identifier) are assigned to all three classes, obviously to be filled in by different values, that help distinguish each instance pertaining to one of these classes; for instance, the publication title and ISBN identifier of a print version is assigned at the *Distribution* level, while a general title is added at the *LCR* level.[22] Similarly, the three classes are linked to each other with appropriate properties (e.g., *hasDistribution* between *LCR* and *Distribution*).

Further properties attached to the *LCR* class represent administrative and provenance metadata (e.g., author, holder of Intellectual Property Rights, creation details, such as information on the digitisation process and used tools, if it was digitised, or the collection and selection process, if it was created from scratch, etc.) that are common across all its *Distributions*, as well as relations to other *LCRs* (e.g., to sources, new editions, etc.). To encode the language(s) of the contents, five distinct properties are included: *source, target* and *pivot language* (for multilingual resources)*, object language* and *metalanguage.*

The *LCR* class also includes a detailed set of classification properties for the structure and type of the LCR. Although these properties are optional, because they are not present in all catalogues, they constitute an important feature for discovery purposes and for assisting users to select the resource that best fits their needs. For instance, users that wish to incorporate the resource in a tagger will look for lexica with morphological information, and for foreign language teaching purposes, learner dictionaries. In the case of resources encoded according to the TEI specifications (TEI Consortium 2020) or the Ontolex model (Cimiano et al. 2016) and the OntoLex-Lemon Lexicography module (Bosque-Gil and Gracia 2019),[23] this information could be automatically extracted from these representations. Furthermore, **MS** provides properties for relating different LCRs to each other (e.g., *is converted version of, replaces, is part of*).

[22] DOIs can be assigned both at *LCR* and *Distribution* level, although as part of the publication metadata it should be assigned at *Distribution* only according to our argumentation. The reason for this is that the *Distribution* class is open to different interpretations (cf. comment on Distribution, https://www.w3.org/TR/vocab-dcat-3/#Class:Distribution) and the resource providers apply the term in different ways. In addition, the metadata schema of DataCite (the main DOI assignment authority) does not distinguish between resource and distribution.

[23] The Ontolex model and the Lexicography module aim to enrich ontologies with linguistic information. The Lime module (LInguistic MEtadata) provides for metadata at the lexicon-ontology interface, but its scope is very narrow compared to our objectives and, therefore, is not discussed here.

For the classification properties and their value spaces we have exploited the corresponding MS properties and the **LexVoc** vocabulary of lexicographic terms.[24] LexVoc is a structured controlled list of terms related to lexicographic and metalexicographic concepts, that has been developed by re-using and extending term lists from various authoritative sources and organising them in semantic domains (Kosem and Lindemann 2021: 761–763); originally, the LexVoc terms were used for the content-describing indexation of meta-lexicographic works (see the description of the LexBib Knowledge Graph in Section 5). As a result, we include seven properties that serve as parameters along which lexicographic works can be classified: for instance, *lemma type*, describing types of headwords included in an LCR (e.g., whether it includes single- or multi-word units, abbreviations, neologisms, etc.), *scope type*, pointing to dictionary typology terms, such as 'learner dictionary', 'dialect dictionary', 'etymological dictionary', etc., and *microstructure feature*, with terms describing microstructural data presentation features and linguistic features of the content.

The *Distribution* class takes properties relevant to publication metadata (e.g., publication date, publisher), as found in library catalogues, as well as properties related to their accessibility, i.e., the mode of access (e.g., 'book publication', 'dictionary app', etc.), the URL where they are available from and the licensing and pricing conditions.

[24] https://lexbib.elex.is/wiki/LexVoc

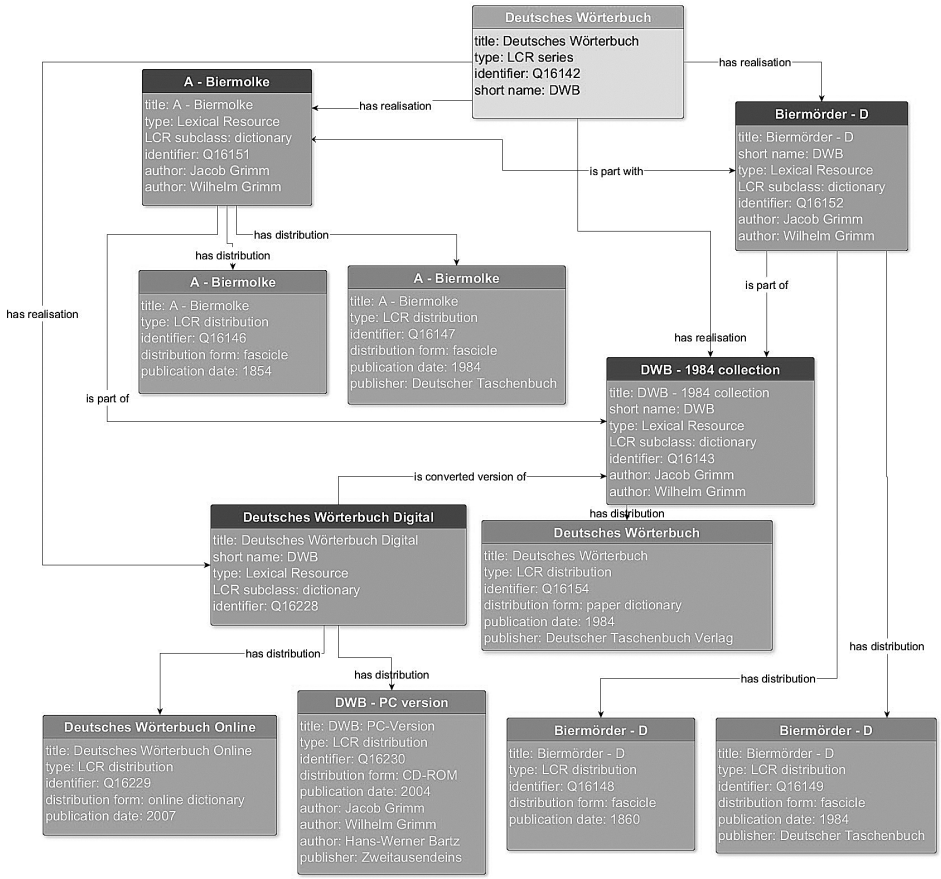

Figure 3: *Deutsches Wörterbuch* encoded with the LexMeta model (simplified)

To ascertain the descriptive efficiency of LexMeta, we have manually created a set of sample entries. One of them is the *Deutsches Wörterbuch*, which illustrates the complexity of relations between entities of the core classes. *Deutsches Wörterbuch* is a lexicographic endeavour started in the mid-1850s by the Grimm brothers, initially released as a set of fascicles (published from 1854 to 1954), with different contributors and content features, and each of them in different forms. After 1984, the fascicles were reprinted, with the same contents as the original ones, and at the same time issued as a complete collection. This 1984 collection was later converted into a digital resource, offered as an offline electronic dictionary, and also made accessible on a web portal.

As Figure 3 shows, this can be represented as an *LCR series* with the fascicles and the collection encoded as *LCRs*, each of which are linked to the relevant distributions. Moreover, the *isPartOf* relation is used to link the fascicles to the collection.

4. Technical implementation

For the formal representation of LexMeta we have adopted the Linked Data paradigm and Semantic Web technologies, since they empower by default semantic interoperability. To accommodate the use cases in which LexMeta is (to be) used, we have decided to implement it in two forms: as an RDF/OWL ontology and as an ontology of Wikibase entities. In both cases, controlled vocabularies are represented as SKOS vocabularies.[25]

The RDF/OWL implementation follows from the fact that this is the most common format for Linked Data. It is also in consistency with the implementation of the MS ontology, on which it is based, and it can, consequently, be used in catalogues that already employ the MS model. The LexMeta ontology is available as a draft version at http://w3id.org/meta-share/lexmeta/.

The Wikibase implementation is intended for use in the framework of the **LexBib** Wikibase Knowledge Graph of Lexicography and Dictionary Research, a research infrastructure targeting the lexicographic community. The LexBib Knowledge Graph (Lindemann et al. 2018; Kosem and Lindemann 2021) is designed to include a catalogue *of* and *about* lexicographic works, along with related entities to them, such as persons, organisations, languages, places, events, and lexicographic terminology. It already contains around 10,000 records for publications, and is currently in the process of being populated with bibliographical data for dictionaries from various catalogues, such as OBELEXdict,[26] Glottolog,[27] Wikidata,[28] and a formerly unpublished comprehensive catalogue of Basque dictionaries. LexBib is implemented as an instance of Wikibase,[29] an

[25] https://www.w3.org/2004/02/skos/
[26] https://www.owid.de/obelex/dict/en
[27] https://glottolog.org/
[28] https://www.wikidata.org/
[29] https://wikiba.se

open source software that collects together applications for creating and sharing structured data as linked data entities and their relationships, following the data model also underlying Wikidata. In the Wikibase implementation, LexMeta classes and properties are represented using URIs from the LexBib Wikibase's own namespace, and following Wikibase naming conventions.

The alignment between the formats is foreseen at both sides. At the LexMeta OWL side, the *schema:sameAs* property links to the LexBib Wikibase entities, while at the LexBib Wikibase, this role is taken up by a dedicated property of type *external identifier.*

This dual implementation has its advantages, as it allows us to outreach communities that are accustomed to different practices and use different tools for the editing and enrichment of the ontologies (i.e., Wikibase on the one side and tools such as Protégé,[30] or VocBench[31] for the RDF/OWL form). Features of Wikibase that distinguish that platform from other software for editing and publishing Linked Data are the possibility of editing every single semantic triple from within a graphical interface, which will display the latest and all previous versions of data describing an entity, advanced user management, and compatibility with Wikidata, which makes federating or transferring data from or to that platform straightforwardly possible.

However, keeping both sides constantly synchronised is a challenging task. The workflows for the update of the two forms from each side to the other have been specified and the one from the Wikibase into the RDF/OWL form has already been implemented, while that for the inverse direction is under construction.[32]

It should be noted that we plan another format, as a CMDI-compliant profile (exploiting the MS profile) in order to facilitate the export of the LexBib records into the CLARIN Virtual Language Observatory (VLO)[33] which aggregates metadata from CLARIN centres and other sources. However, given that the profiles cannot be changed, this will be done when LexMeta is stable.

[30] https://protege.stanford.edu/

[31] http://vocbench.uniroma2.it/

[32] While RDF data dumps from Wikibase, which could be used for this, are also available, we opt here for a script that translates Wikibase properties to their LexMeta OWL equivalents, defining that mapping in the Wikibase itself using a dedicated property. For details, see https://lexbib.elex.is/wiki/LexMeta_OWL.

[33] vlo.clarin.eu/

5. Conclusions and Outlook

We have presented the LexMeta model for lexical resources and the theoretical and technical considerations that were taken into account for its design and implementation.

The current version has been made available to scholars from the lexicographic and Linguistic Linked Data communities, and discussions on its improvements are ongoing. These will continue in the framework of the LD4LT W3C working group,[34] in close collaboration with members of the Nexus Linguarum COST action,[35] two groups that both work in the area of Linguistic Linked Data. In addition, synergies with the TEI-Lex0 working group on the described overlapping interests have been established. Finally, for the improvements of LexMeta we will take into account the most recent developments regarding related resource types, such as knowledge graphs (Dumontier 2022) and word embeddings, as well as recommendations and specifications targeting semantic interoperability in major data sharing initiatives like Data Spaces.

The use of LexMeta in the LexBib catalogue and, consequently, its mapping to the metadata models lying behind the catalogues from which LexBib is populated, as mentioned in Section 4, will further enhance the model.

Acknowledgements

The research presented in this paper has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 731015 (ELEXIS, https://elex.is), the COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology, https://www.cost.eu/), and from the Basque Government (IT1534-22 Research Group).

[34] https://www.w3.org/community/ld4lt/
[35] https://nexuslinguarum.eu/

References

Bosque-Gil, Julia; Gracia, Jorge. 2019. *The OntoLex Lemon Lexicography Module*. W3C Community Group Report. https://www.w3.org/2019/09/lexicog/.

Broeder, Daan; Windhouwer, Menzo; van Uytvanck, Dieter; Goosen, Twan; Trippel, Thorsten. 2012. *Proceedings of the workshop describing language resources with metadata: towards flexibility and interoperability in the documentation of language resources. LREC 2012.* European Language Resources Association. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/10867.

Cimiano, Philipp; McCrae, John P.; Buitelaar, Paul. 2016. *Lexicon Model for Ontologies: Community Report*. W3C. https://www. w3.org/2016/05/ontolex/.

Dumontier, Michel. 2022. Towards a computable standard for Knowledge Graph Metadata.https://www.slideshare.net/micheldumontier/a-metadata-standard-for-knowledge-graphs.

Eskevich, Maria; de Jong, Franciska et al. 2020. CLARIN: Distributed Language Resources and Technology in a European Infrastructure. *Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020, co-located with LREC 2020).* European Language Resources Association. 28–34. https://aclanthology.org/2020.iwltp-1.5.

Fahad, Khan; Salgado, Ana. 2021. Modelling Lexicographic Resources using CIDOC-CRM, FRBRoo and Ontolex-Lemon. *Proceedings of the International Joint Workshop on Semantic Web and Ontology Design for Cultural Heritage*. https://ceur-ws.org/Vol-2949/paper7.pdf.

Gavrilidou, Maria et al. 2012. The META-SHARE Metadata Schema for the Description of Language Resources. P*roceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012).* European Language Resources Association. Istanbul. 1090–1097. http://www.lrec-conf.org/proceedings/lrec2012/pdf/998_Paper.

IFLA Study Group on the Functional Requirements for Bibliographic Records. 2009. *Functional Requirements for Bibliographic Records.* https://repository.ifla.org/bitstream/123456789/811/2/ifla-functional-requirements-for-bibliographic-records-frbr.pdf.

International Organization for Standardization. 2019. *ISO 24613-1:2015: Language resource management – Lexical markup framework (LMF) – Part 1: Core model.* ISO. https://www.iso.org/standard/68516.html.

International Organization for Standardization. 2020. *ISO 24622-1:2015: Language resource management – Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model.* ISO. https://www.iso.org/standard/37336.html.

International Working Group on LRM, FRBR and CIDOC CRM Harmonisation. 2021. *RMoo (formerly FRBRoo): object-oriented definition and mapping from IFLA LRM (version 0.7).* https://www.cidoc-crm.org/frbroo/sites/default/files/LRMoo_V0.7%28draft%202021-06-29%29.pdf.

Khan, Fahad et al. 2022. When Linguistics Meets Web Technologies. Recent Advances in Modelling Linguistic Linked Open Data. *Semantic Web Journal*. *Special issue on Linguistic Linked Data*. https://www.semantic-web-journal.net/system/files/swj2859.pdf.

Kosem, Iztok; Lindemann, David. 2021. New developments in Elexifinder, a discovery portal for lexicographic literature. *Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress.* Eds. Gavriilidou, Zoe; Mitits, Lydia.; Kiosses, Spyros. 759–766. https://euralex2020.gr/wp-content/uploads/2021/09/Pages-from-EURALEX2021_ProceedingsBook-Vol2-p759-766.pdf.

Lindemann, David; Kliche, F.; Heid, Ulrich. 2018. LexBib: a corpus and bibliography of metalexicographcal publications. *Proceedings of EURALEX 2018*. Ljubljana. 699–712.

Lindemann, David; Labropoulou, Penny; Klaes, Christiane. 2022. Introducing LexMeta: A metadata model for lexical resources. *Proceedings of the XX EURALEX Conference.* Mannheim. doi: 10.5281/zenodo.6897062.

McCrae, John Philip et al. 2015. One Ontology to Bind Them All: The META-SHARE OWL Ontology for the Interoperability of Linguistic Datasets on the Web. *The Semantic Web: ESWC 2015 Satellite Events.* Lecture Notes in Computer Science. Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-319-25639-9_42. 271–282.

Piperidis, Stelios. 2012. The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. European Language Resources Association. Istanbul. https://aclanthology.org/L12-1647/.

Piperidis, Stelios et al. 2022. The European Language Grid Platform: Basic Concepts. *European Language Grid: A Language Technology Platform for Multilingual Europe.* Ed. Rehm, Georg. Springer International Publishing. 13–36.

Rehm, Georg et al. 2021. European Language Grid: A Joint Platform for the European Language Technology Community. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021).* Association for Computational Linguistics. Kyiv. 221–230. https://www.aclweb.org/anthology/2021.eacl-demos.26.

Riva, Pat; Le Bœuf, Patrick; Žumer; Maja. 2017. *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information*. IFLA. https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf.

TEI Consortium. 2020. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Zenodo. doi:10.5281/zenodo.3992514.

Wilkinson, M. et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 160018. doi.org/10.1038/sdata.2016.18.

LexMeta model za leksičke resurse: teorija i primjena

Sažetak

Rad opisuje LexMeta, metapodatkovni model za opis leksičkih resursa kao što su rječnici, popisi riječi, glosari i dr., koji će se upotrebljavati u katalozima podataka namijenjenima leksikografskoj i široj humanističkoj zajednici, ali i korisnicima koji upotreblajvaju takve modele u istraživanjima i praktičnoj primjeni. U radu je dan usporedni pregled sličnih modela kako bi se pokazale razlike i sličnosti s LexMetom. Kako bi se poboljšala semantička interoperabilnost i podržala razmjena (meta) podataka između strukovnih i općih kataloga, kao temelj za dizajn LexMeta odabrani su najutjecajniji modeli, naime FRBR koji se upotrebljava u knjižničnim katalozima i META-SHARE koji se upotrebljava za jezične resurse. Rad donosi raspravu o tome kako su ti modeli usklađeni i prošireni novim značajkama potrebnima za opis leksičkih izvora. Formalni prikaz modela koji slijedi paradigmu povezanih podataka ima za cilj dodatno poboljšati semantičku interoperabilnost. Izbor da se implementira u dva formata (kao RDF/OWL i kao ontologija Wikibase) olakšava njegovo usvajanje, a time i obogaćivanje, ali i postavlja izazove koji se tiču sinkronizacije formata, koji se rješavaju automatskim tijekovima rada. Zaključujemo rad s opisom tekućih i planiranih aktivnosti na unapređenju modela.

Keywords: lexicography, metadata model, linked data, lexical resources, Wikibase
Ključne riječi: leksikografija, metapodatkovni model, povezani podatci, leksički izvori, Wikibase