

D1.2

Best practices for lexicography – intermediate report

Authors: Carole Tiberius, Rute Costa,
Tomaž Erjavec, Simon Krek, John
McCrae, Christophe Roche, Toma
Tasovac

Date: 31 January 2020

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

D1.2 Best practices for lexicography – intermediate
report

Deliverable Number: D1.2

Dissemination Level: public

Delivery Date: 31.01.2020

Version: 1.0

Author(s): Carole Tiberius, Rute
Costa, Tomaž Erjavec,
Simon Krek, John
McCrae, Christophe
Roche, Toma Tasovac

Project Acronym: ELEXIS

Project Full Title: European Lexicographic Infrastructure

Grant Agreement No.: 731015

Deliverable/Document Information

Deliverable Number: D1.2

Deliverable Title: Best practices for lexicography – intermediate report

Authors: Carole Tiberius, Rute Costa, Tomaž Erjavec, Simon Krek, John McCrae,
Christophe Roche, Toma Tasovac

Dissemination Level: public

Document History

Version	Date	Changes/Approval	Author(s)/Approved by
V0.1	19.11.2019	First draft	Carole Tiberius, Rute Costa, Christophe Roche
V0.2	16.12.2019	First draft and review	Tomaž Erjavec, Simon Krek, John McCrae, Toma Tasovac
V0.3	15.01.2020	Incorporation of feedback	Carole Tiberius, Rute Costa, Tomaž Erjavec, Simon Krek, John McCrae, Christophe Roche, Toma Tasovac
V0.4	20.01.2020	Review	Karlheinz Moerth, Tanja Wissik
V0.5	23.01.2020	Incorporation of feedback	Carole Tiberius, Rute Costa, Tomaž Erjavec, Simon Krek, John McCrae, Christophe Roche, Toma Tasovac
V0.6	27.01.2020	Review	Bob Boelhouwer
V1.0	31.01.2020	Final version	Carole Tiberius, Rute Costa, Tomaž Erjavec, Simon Krek, John McCrae, Christophe Roche, Toma Tasovac

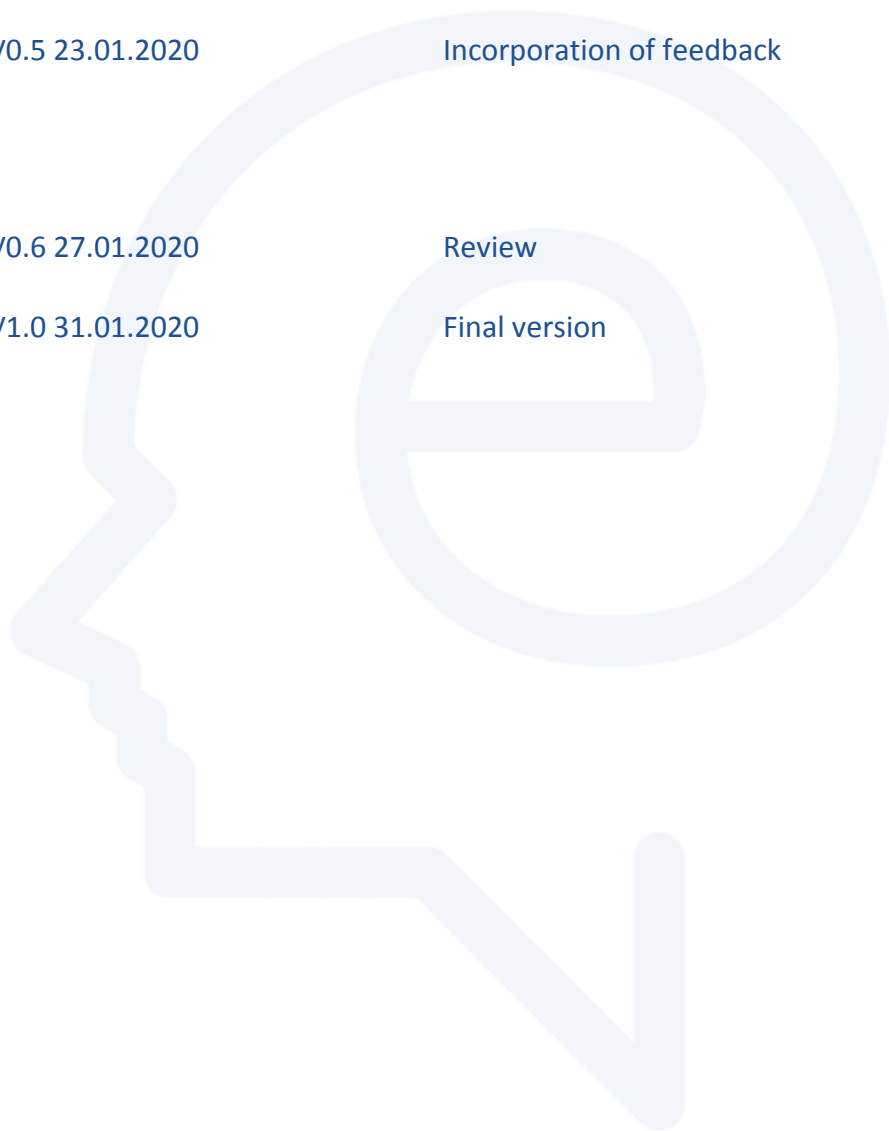


Table of Contents

1	Introduction	3
2	Existing Standards	4
2.1	TEI and Dictionaries	4
2.1.1	An overview of TEI	4
2.1.2	TEI ODDs.....	4
2.1.3	The TEI ecosystem.....	5
2.1.4	The TEI Dictionary module	5
2.2	ISO Standards.....	6
2.2.1	International Organisation for Standardisation.....	6
2.2.2	ISO 1951: Entries in dictionaries	8
2.2.3	ISO 24613-x: Language resource management -LMF - x	9
2.2.4	Other ISO Standards.....	15
2.3	Standard Format Marker and Multi-Dictionary Format Codes.....	18
3	ELEXIS interoperability formats	20
3.1	TEI Lex-0	20
3.2	Ontolex-Lemon	22
4	Data formats and standards within ELEXIS	25
5	References	28
6	Annex: Sample dictionary entries	29
6.1	TEI encoded dictionaries.....	29
6.1.1	Dictionary of Old Dutch (IVDNT)	29
6.1.2	A machine-readable Dictionary of Dagaare (OEAW)	31
6.1.3	Mittelhochdeutsches Handwörterbuch von Matthias Lexer (TCDH).....	32



6.2	ISO encoded dictionaries	34
6.2.1	ISO 1951: Dictionary of Karelian (Institute for the Languages of Finland)	34
6.2.2	LMF: OMBI Arabic-Dutch (IVDNT).....	35
6.3	TEI Lex-0	37
6.3.1	Dicionário da Academia das Ciências de Lisboa.....	37
6.4	Ontolex-Lemon	40
6.4.1	Global English resource (KD).....	40



1 Introduction

This deliverable is part of task 1.3. The aim of this task is to facilitate the creation of lexicographic resources in European institutions, by creating robust documentation, guidelines and collections of best practices in order to promote clearly defined workflows for producing, describing and annotating lexicographic resources (both synchronic and diachronic) in accordance with international standards and interoperability formats.

This deliverable constitutes an intermediate report and focuses on data formats and standards used in lexicography. It forms the basis for defining guidelines and best practices for producing, describing and annotating lexicographic resources which will be presented in deliverable D1.5 at the end of the project. The deliverable is structured as follows. First we will give an overview of existing standards, i.e. TEI, ISO Standards and Standard Format Marker codes. Second, we will introduce the reader to more recent developments and we will describe the ELEXIS interoperability formats, TEI Lex-0 and Ontolex-Lemon. Third, we will discuss the ongoing work on data formats and standards within ELEXIS. This work is particularly relevant for the integration of data from different lexicographic data providers into the ELEXIS infrastructure (see also D1.3).



2 Existing Standards

In this section, we give an overview of existing standards that are used in lexicography, i.e. TEI, ISO Standards and the use of Standard Format Marker codes in Field Linguistics.

2.1 TEI and Dictionaries

This section briefly introduces the Text Encoding Initiative Guidelines, which provide also a module for encoding dictionaries. The TEI is especially important in the context of ELEXIS, as a) TEI is the most common encoding format for the existing dictionaries according to the ELEXIS survey on User Needs (Kallas et al. 2019), and b) the TEI Lex-0 recommendation, discussed in Section 3.1, is a parametrisation of TEI and also extensively uses the TEI infrastructure, in particular the ODD schema and documentation language.

2.1.1 An overview of TEI

The [Text Encoding Initiative Guidelines](#) are an ambitious attempt, which started in 1987, to propose an encoding scheme that would apply to texts in any language, of any date, and of any text type, without restriction on form or content. While the target audience is primarily the scholarly research community, they are also useful for librarians, publishers and others distributing or creating electronic texts. The design goals of the Guidelines were that they should provide a standard format for data interchange; provide guidance for the encoding of texts in this format; support the encoding of all kinds of features of all kinds of texts studied by researchers; and be application independent. It should be stressed that TEI is a descriptive, rather than a prescriptive recommendation, i.e. it tries to allow a loss-less encoding of any document, rather than attempting to enforce that each document is encoded in an exactly specified manner. The advantage of this approach is coverage, the disadvantage that it is possible to encode the same document in various ways: while the TEI is a good standard for interchange, it is less for interoperability.

While the TEI could be taken as a text metamodel, it is very firmly entrenched in XML, and, essentially, defines several hundred elements and their attributes that make useful textual distinctions, and documents their semantics (i.e. intended use) in the Guidelines.

2.1.2 TEI ODDs

The TEI is based on XML, which uses XML schemas (either DTDs, inherited from the time of SGML, or W3C XML schemas, or ISO RelaxNG schemas) to validate a particular document type. However, given its generality, the TEI is too large and meant for too many different types of text and analysis for it to be sensible to have a single XML schema for the whole of the TEI. Furthermore, the TEI from its very beginning took on the idea originating from the so called literate programming, that the schema should also contain its documentation, in other words, One Document Does it all, or ODD.

Therefore, the TEI Guidelines are structured as an ODD, i.e. contain both the element and attribute definitions as well as their documentation, while the complete TEI is divided into a number of modules, with the option to use only chosen modules, elements and attributes in a particular parametrisation of



the TEI. Furthermore, the TEI also allows changing or adding to the TEI elements or attributes to cater for situations not covered by the TEI.

Once the TEI has been parameterised, it is then possible to generate from a particular ODD both standard XML schemas (DTDs, W3C or RelaxNG) as well as the accompanying documentation (in TEI, HTML or PDF) with the use of the TEI XSLT stylesheets, as further explained in the next section. With this relatively straightforward way of parametrising the Guidelines (including the documentation) it becomes simple to develop very specific schemas, as is the case of the TEI serialisation of the TEI Lex-0 proposal.

2.1.3 The TEI ecosystem

While TEI is the longest continuously running attempt at developing general guidelines for text encoding, it is also much more than just a set of formal specifications and accompanying documentation. First, it is organised as a consortium with a well-defined governing structure, where the TEI Council is charged with continuous development of the TEI Guidelines, mostly as a response to issues raised by the large and varied community that uses them. This continuous development is one of its major strengths, as bugs and inconsistencies get resolved, often after a detailed debate, while at the same time it keeps up with technical developments. Second, the TEI maintains a very active mailing list, tei-l, where questions, even by novices, are answered quickly and exhaustively. Third, the Guidelines are available under the open CC BY licence directly from [GitHub](https://github.com), unlike e.g. the standards by ISO, as well as on the tei-c.org website, where, e.g. each defined element has its URL with its description, links to the prose of the Guidelines, context, examples, and definition.

Very important is also the openly available TEI tool-chest that is developed and maintained by the TEI Consortium. Its most important part are the TEI Stylesheets, i.e. a collection of XSLT scripts that support the transformation of many different text formats (e.g. docx, html, markdown) into TEI and the conversion of TEI into these formats. This gives a simple way of both up-converting legacy data into TEI, and preparing TEI documents for reading. Similarly to the Guidelines, these stylesheets can also be parametrised and changed, to make them better suitable for particular projects.

The stylesheets, when invoked in a dedicated ODD mode, also support, as mentioned in the previous section, transforming a TEI ODD parameterisation into a) the XML schema in any of the three XML schema languages and b) into the accompanying documentation, which covers the elements used in the parameterisation. Such a project specific ODD can then be maintained with Git and published on the Web.

2.1.4 The TEI Dictionary module

The TEI Guidelines, from the very first edition, also contain [a module for encoding dictionaries](#). The elements defined in this module are primarily meant for encoding human-oriented dictionaries or glossaries, but can also be useful in the encoding of computational lexicons.



The TEI defines all the basic building blocks for encoding dictionaries as well as computational lexicons, it supports the typographical, editorial and lexical views on dictionaries, and structures a lexical entry into meaningful high-level chunks, such as <form> and <sense>.

The immensely varied structures of dictionaries has stretched the TEI descriptive goals to their utmost, with TEI even offering an <entryFree> element, where “anything goes anywhere”. The standard dictionary <entry> element makes limited attempts at mandating unifying practices, although selected encodings of structures in several real (printed) dictionaries are used as examples, e.g. from [OALD](#):

```
<entry>
  <form>
    <orth>competitor</orth>
    <hyph>com|peti|tor</hyph>
    <pron>k@n"petit@ (r)</pron>
  </form>
  <gramGrp>
    <pos>n</pos>
  </gramGrp>
  <def>person who competes.</def>
</entry>
```

Much information can also be encoded in attributes with open vocabularies. Therefore the TEI schema for dictionaries is far from being prescriptive, and it does not really offer a good encoding for interchange and interoperability of dictionaries.

The TEI thus offers a very good basis for further defining a generally applicable but much more constrained model for encoding dictionaries. Taking into account also the TEI ODD language for schema (re)definition and simple but effective methods of development, publication and maintenance, it provides an infrastructure to develop a truly useful model for dictionary interchange and interoperability, exactly as is done in the TEI Lex-0 model as discussed in Section 3.1.

2.2 ISO Standards

This section is about ISO Standards for Lexicography that are useful for the ELEXIS project. It concerns the following standards specified by the ISO/TC37/SC2 and the ISO/TC37/SC4: ISO/TC37/SC2: ISO 1951; ISO/TC37/SC4: ISO 24613-1, -2, -3, -4, -5. After a short introduction on the International Organisation for Standardisation, we will describe the individual standards.

2.2.1 International Organisation for Standardisation

2.2.1.1 ISO

ISO (the International Organisation for Standardisation) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees (TC). Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organisations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO



collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardisation.

International Standards are drafted in accordance with the rules given in the [ISO/IEC Directives, Part 2](#). The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75% of the member bodies casting a vote. [Foreword section of the ISO Standards].

In the framework of the ELEXIS project, we focus on the work from Subcommittee 2 (SC) and Subcommittee 4 of the Technical Committee (TC) 37.

2.2.1.2 TC 37

The TC 37, titled "[Language and terminology](#)", serves the language, content and knowledge industries as well as users of terminology and language technology products. It is increasingly of immediate relevance for developers and service providers of software and other forms of content.

The language industry is of increasing importance in the world-wide economy, supporting international trade and communications. It also plays a major role in linking social networks, such as government to citizen, and developed to developing markets. Language has a special and growing interest in the computing industry, which sees mastering language in computing environments as the greatest challenge for developing "next generation" computing technologies.

All of these areas where the language industry figures predominantly have at least one thing in common: the computer is the primary means of information exchange. Language has to work effectively in a computerised environment otherwise there will be major breakdowns in communication and immeasurable losses both economic and social. In order for language to be understood and processed by computers and computerised applications of all sorts, it must be "structured."

ISO/TC37 plays a key role in enabling language for computing environments. It creates standards for structuring language resources. "Structurable" language resources include terminologies, lexical resources (dictionaries, lexicons, etc.), language-based commercial data (names, properties, catalogues, etc. of products and services), signs and symbols, codes and formulae, corpora (text, speech, audio), taxonomies and ontologies. [[Business Plan of ISO/TC 37](#)].

The ISO/TC37 is structured into 5 Subcommittees (SC), each of them in charge of Standards in relation with their topic:

- SC1: Principles and methods
- SC2: Terminology workflow and language coding
- SC3: Management of terminology resources
- SC4: Language resource management
- SC5: Translation, interpreting and related technology

As mentioned above, ELEXIS is mainly concerned with the standards of the SC2 and the SC4.



2.2.1.3 SC2

Subcommittee 2 (SC2) is titled “Terminology workflow and language coding”, whose chairperson is Prof. Rute Costa from the University NOVA of Lisbon. The standards and guidelines produced by SC2 cover the application of the principles and methods of terminology work, with a focus on terminography and lexicography, reference coding, cultural diversity management, assessment and quality management, translation and interpretation processes and certification schemes.

Several standards are under the responsibility of the ISO/TC 37/SC 2. The ISO 1951: 2007 “Presentation/representation of entries in dictionaries — Requirements, recommendations and information” is one of the ISO Standards useful for the ELEXIS project and is described in section 2.2.2.

2.2.1.4 SC4

Subcommittee 4 (SC4) is titled “Language resource management”. The research areas of SC4 include computational linguistics, computerised lexicography, and language engineering. Text and speech corpora, lexicons, ontologies and terminologies are typical instances of language resources to be used for language and knowledge engineering. In both monolingual and multilingual environments, language resources play a crucial role in preparing, processing and managing the information and knowledge needed by computers as well as humans.

2.2.2 ISO 1951: Entries in dictionaries

ISO 1951:2007 is the standard dedicated to “Presentation/representation of entries in dictionaries — Requirements, recommendations and information”. The scope of ISO 1951 is to deal with monolingual and multilingual, general and specialised dictionaries. It specifies a formal generic structure independent of the publishing media and it proposes means of presenting entries in print and electronic dictionaries.

ISO 1951 considers dictionary entries as comments about topics, which are lexical units. An entry has a main topic (the headword). Other topics (e.g. variants, translations) are said to be “related topics”. Topics and comments are data elements. Each data element has a content model. Data elements are grouped into compositional elements in order to produce an unambiguous and fully computable entry. Open lists of data elements and compositional elements are provided herein, and are extendable by the user for specific purposes.

Below are examples of lexical units and comments which should be used in a standardised dictionary entry:

Name	Generic identifier	Explanation
abbreviated form	AbbreviatedForm	Lexical unit formed by omitting words or letters from a longer form [...]. [Adapted from ISO 1087-1:2000, definition 3.4.9]
derivation	Derivation	A change in the form of a lexical unit, usually modification in the base/root or affixation which signals a change in part-of-speech information.
part of speech	PartOfSpeech	A category assigned to a lexical unit based on its grammatical and semantic properties. [Adapted from ISO 12620:1999, A.2.2.1]
subjectfield	SubjectField	An area of human knowledge. [Adapted from ISO 12620:1999, A.4]

Example of the XML encoding of a sample entry for the English headword ‘administrator’:

```

<DictionaryEntry identifier ='pocketdict-en-fr-administrator'>
  <HeadwordCtn>
    <Headword>administrator</Headword>
  </HeadwordCtn>
  <SenseGroup>
    <TranslationCtn>
      <Translation>administrateur (<Suffix>-trice<GrammaticalGender
        value='feminine'/></Suffix><GrammaticalGender value='masculine'/>
      </Translation>
    </TranslationCtn>
  </SenseGroup>
</DictionaryEntry>

```

2.2.3 ISO 24613-x: Language resource management -LMF - x

The ISO 24613 (LMF - Lexical Markup Framework) multi-part standard is based upon the definition of an implementation-independent meta-model combining a core model and additional models that onomasiological lexical content may take.



LMF also provides guidelines for various implementation contexts, and where appropriate describes LMF compliant serialisations for various application contexts.

2.2.3.1 ISO/DIS 24613-1:2018(E)

The “Lexical Markup Framework (LMF) — Part 1: Core Model” is a metamodel for representing data in monolingual and multilingual lexical databases used with computer applications. LMF provides mechanisms that allow the development and integration of a variety of electronic lexical resource types.

It also provides definitions of terms useful for the ELEXIS project. Let us give some examples:

data category

elementary descriptor used in a linguistic description or annotation scheme

form

instance of a word, multi-word expression, root, stem, or morpheme

lexical entry

container for managing one form or several forms and possibly one or several meanings in order to describe a lexeme

part of speech (lexical category, word class)

category assigned to a lexeme based on its grammatical properties

LMF relies on key standards among which:

Unicode for character encodings

ISO 12620 Data Category Registry (DCR) providing a set of data category specifications

Unified Modeling Language (UML) of which a subset is used for description of the specification.

The LMF core package is a metamodel that provides a flexible basis for building LMF models and extensions. LMF models are represented by UML classes, associations among the classes, and a set of data categories that function as UML attribute-value pairs. The data categories are used to adorn the UML diagrams that provide a high level view of the model. The LMF core package is described by the following figure:

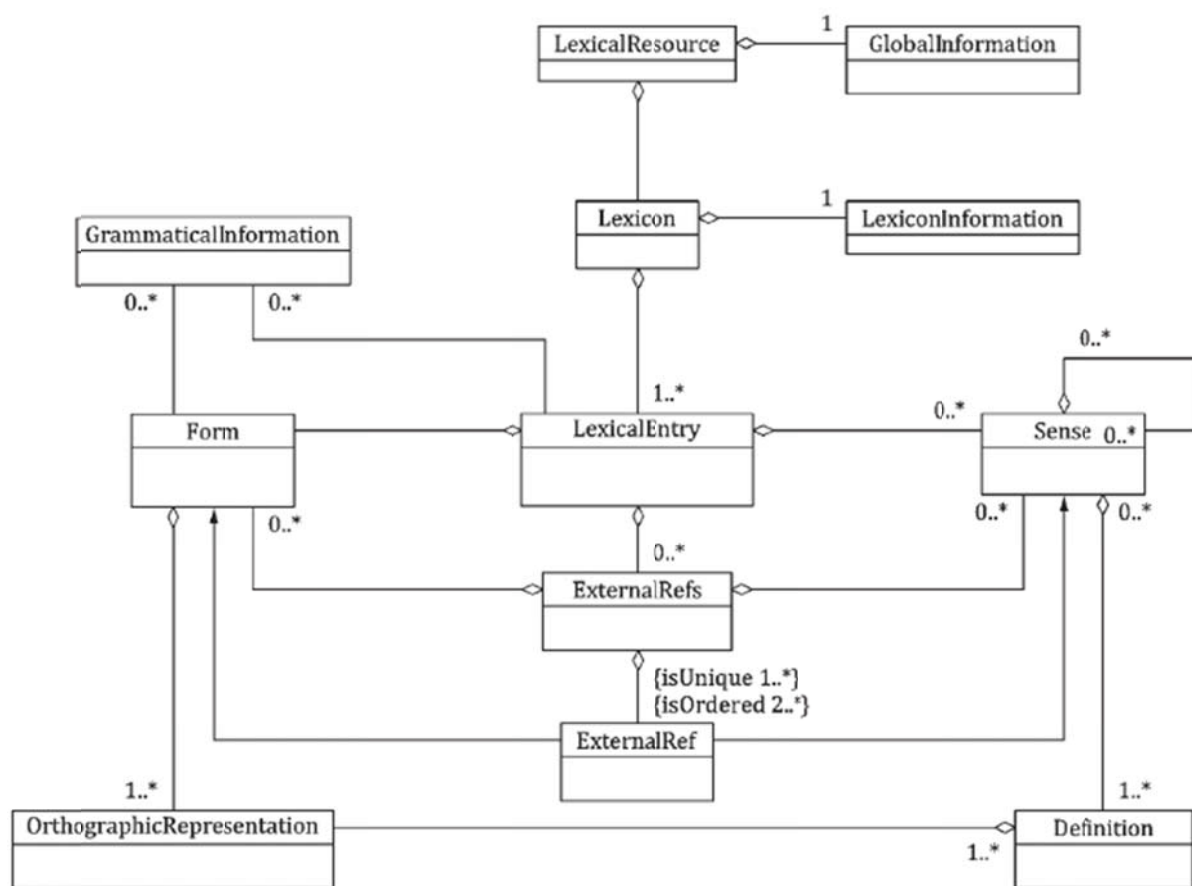


Figure 1: LMF core package (ISO/DIS 24613-1:2018(E), p. 5)

2.2.3.2 ISO/FDIS 24613-2:2019

The “Language resource management — Lexical markup framework (LMF) — Part 2: Machine readable dictionary (MRD) model” Standard extends the LMF Part 1, Core model, through the use of the processes and mechanisms described in LMF Part 1. The objective is to enable flexible design methods to support the development of machine readable dictionaries for different purposes while enabling cross comparisons of different designs and a basis for developing assessments of standards conformance. The scope of supported design goals ranges from simple to complex human-oriented MRDs, both monolingual and bilingual; lexicons that support conceptual-lexical systems through links with ontological resources; rigorously constrained lexicons for supporting machine processes; and lexicons that provide an extensional description of the morphology of lexical entries.

The Machine Readable Dictionaries (MRD) model is organised as presented in the following figure:

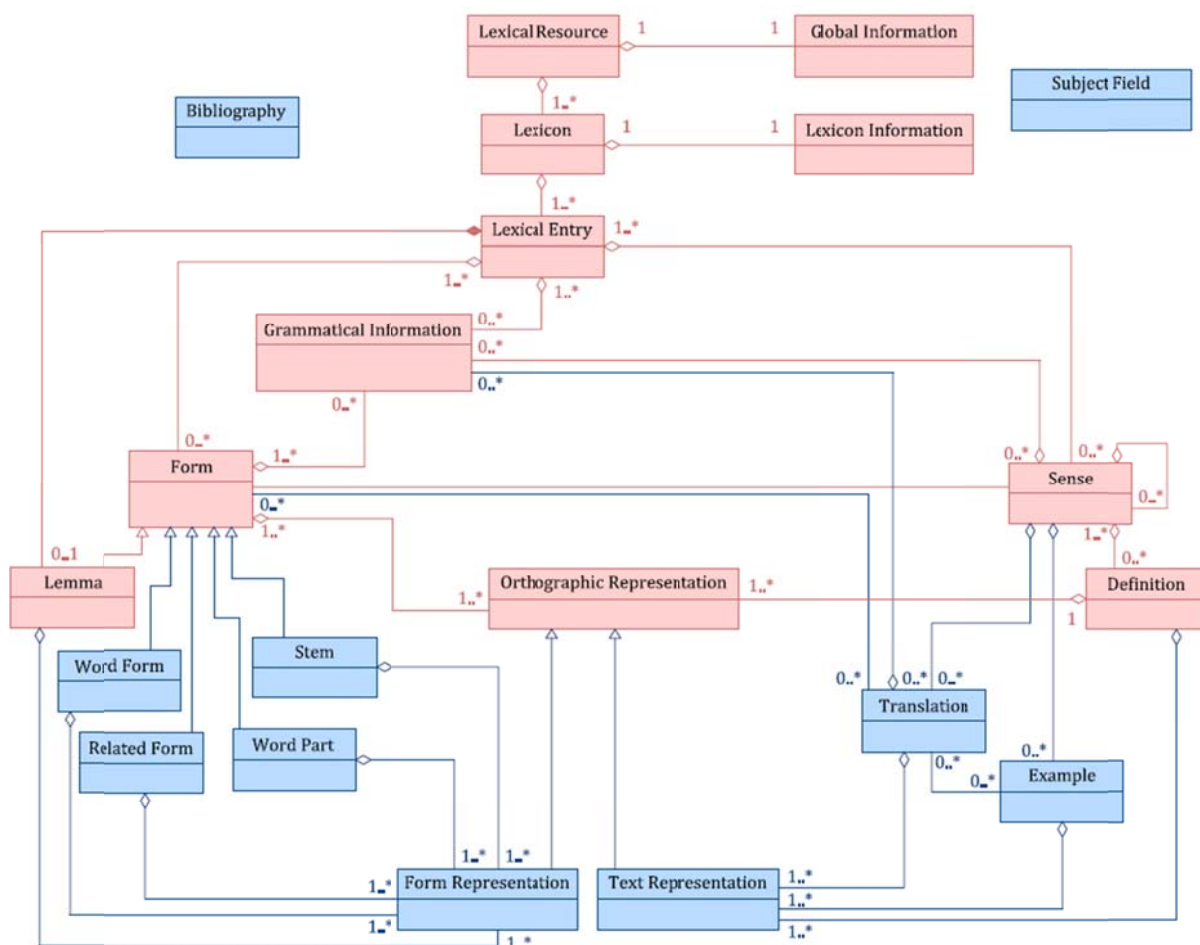


Figure 2: MRD class model (ISO/FDIS 24613-2:2019(E), p. 2)

The MRD model is represented by UML classes such as:

WordForm class

WordForm is a Form subclass containing a word form, such as an inflected form, that a lexeme can take when used in a sentence or a phrase. The WordForm class is in a zero-to-many aggregate association with the LexicalEntry class (inheriting the Form multiplicity). The WordForm class can manage simple lexemes, compounds, multi-word expressions, and sub-lexemes such as affixes and roots.

Lemma class

Lemma is a Form subclass representing a lexeme or sub-lexeme used to designate the LexicalEntry (part of the Form-Sense paradigm). The Lemma class is in a zero-to-one aggregate association with the LexicalEntry class that overrides the multiplicity inherited from the Form class (see ISO 24613-1 for a more complete description of the Lemma).

Stem class

Stem is a Form subclass containing a stem or root. The Stem class can be typed as a specific type of stem or root (e.g. type="arabicRoot"). The Stem class is in a zero-to-one aggregate association with the LexicalEntry class (overriding the multiplicity inherited from the Form class).

Translation class

In a bilingual MRD, the Translation class represents the translation equivalent of the word form managed by the Lemma or WordForm class. The Translation class is in a zero-to-many aggregate association with the Sense class, which allows the lexicon developer to omit the Translation class from a monolingual dictionary.

2.2.3.3 ISO/DIS 24613-3:2020(E)

The goal of the ISO/DIS 24613-3 “Language resource management – Lexical markup framework (LMF) – Part 3: Etymological extension” Standard is to support the development of detailed descriptions of the various etymological phenomena and/or diachronic links between lexical entries in born-digital and/or retro-digitised lexicons. It provides both a meta-model for such an extension as well as the relevant data categories. The etymology extension of LMF relies on several classes. For example, the Etymologisable class provides a means of referring to the set of linguistic elements that can have etymologies as described in the following figure:

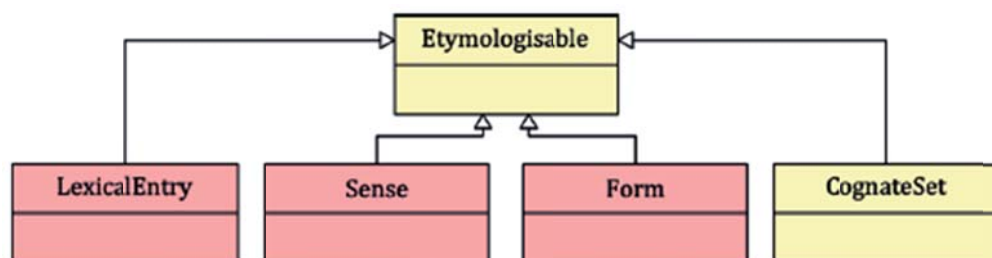


Figure 3: The Etymologisable class and its subclasses (ISO/DIS 24613-3:2020(E), p 2)

The figure below is an example of a diachronic etymological process with phonological change.

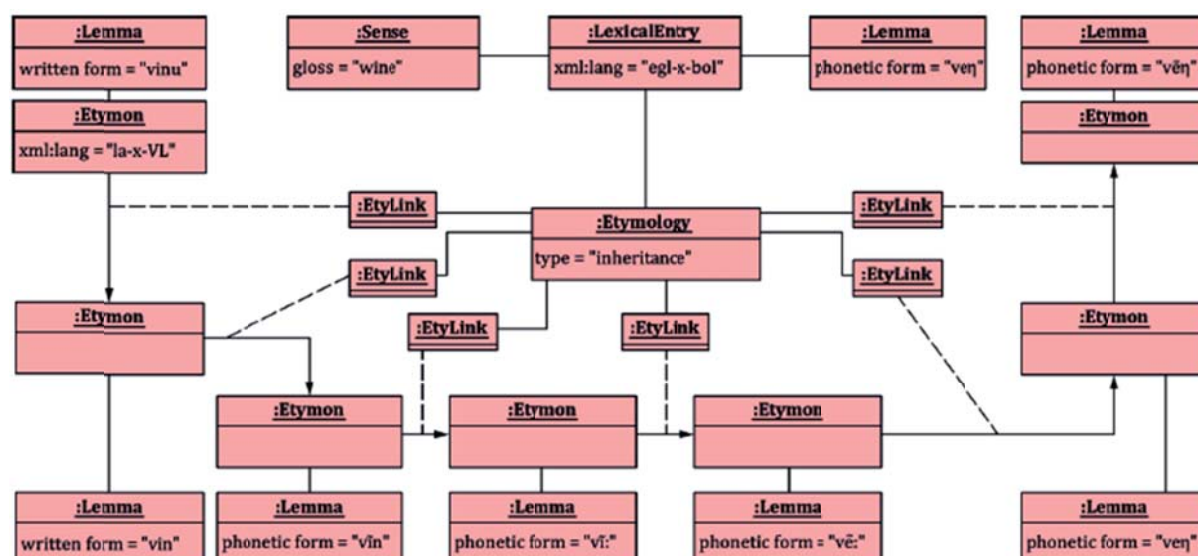


Figure 4: Diagram of multi-stage inheritance and phonological change in Bolognese (ISO/DIS 24613-3:2020(E), p. 6)

2.2.3.4 ISO/DIS 24613-4:2020 (E)

The “Language resource management — Lexical markup framework (LMF) — Part 4: TEI serialisation” describes the serialisation of the LMF standard defined as an XML model compliant with the TEI guidelines. This serialisation covers both the classes of the LMF core model and classes provided by the following additional parts of ISO 24613: machine readable dictionaries, etymology, etc. It will be refined when the new LMF configuration will be better advanced. The following example in French illustrates the encoding of a simple dictionary entry:

```

<entry>
  <form type="lemma">
    <orth>langouste</orth>
    <pron>lågust</pron>
    <gramGrp>
      <pos>n.</pos>
      <gen>f.</gen>
    </gramGrp>
  </form>
  <sense n="1">
    <def>Grand crustacé marin (Décapodes macroures) aux pattes antérieures dépourvues de pinces, aux antennes longues et fortes, et dont la chair est très appréciée.</def>
  </sense>
  <sense n="2">
    <usg type="register">Fig. et fam. (vulg.)</usg>
    <def>Femme, maîtresse.</def>
  </sense>
  <etym>XIIIe; languste, v. 1120, «sauterelle»; encore dans Corneille (Hymnes, 7); anc. provençal langosta, altér. du lat. class. locusta «sauterelle».</etym>
</entry>

```

2.2.3.5 ISO NP 24613-5:2018(E)

ISO/NP 24613-5:2018: “Language resource management — Lexical markup framework (LMF) — Part 5: Lexical base exchange (LBX) serialization”

LMF Part 5, Language base exchange (LBX) serialisation is a W3C XML serialisation for machine readable dictionaries (MRD) that describes the basic hierarchy of information of an MRD, including information on the form, sense, and metadata. The MRD part is supplemented by various resources that are part of the definition of LMF described in Part 1, Core Model and Part 2, Machine readable dictionaries (MRD).

The LBX serialisation supports the instantiations of LMF described in LMF Part 2, including electronic lexical resources, such as electronic monolingual, bilingual and multilingual lexical databases, as well as extensional morphologies.

2.2.4 Other ISO Standards

Other ISO standards related to ELEXIS include the following:

ISO 24611:2012 - Language resource management — Morpho-syntactic annotation framework (MAF), confirmed in 2018. This standard provides a framework for the representation of annotations of word-forms in texts; such annotations concern tokens, their relationship with lexical units, and their morpho-syntactic properties. It describes a metamodel for morpho-syntactic annotation that relates to a reference to the data categories contained in the ISOCat data category registry (DCR, as defined in ISO 12620). It also describes an XML serialisation for morpho-syntactic annotations, with equivalences to the guidelines of the TEI (text encoding initiative). This standard provides a set of definitions related to ELEXIS. Let us quote:

inflected form: form that a word can take when used in a sentence or a phrase

lexical entry: container for managing a set of word-forms and possibly one or more meanings to describe a lexeme

lexicon: resource comprising a collection of lexical entries for a language

part of speech: category assigned to a word based on its grammatical and semantic properties

The purpose of **ISO 639** is to establish internationally recognised codes (either 2, 3, or 4 letters long) for the representation of languages or language families. ISO 639 is a set of standards about the representation of names for languages and language groups. Let us quote the first three of them.

ISO 639-1:2002 - Codes for the representation of names of languages - Part 1: Alpha-2 code, confirmed in 2019, provides a code consisting of language code elements comprising two-letter language identifiers for the representation of names of languages. The language identifiers according to this part of ISO 639 were devised originally for use in terminology, lexicography and linguistics, but may be adopted for any application requiring the expression of language in two-letter coded form, especially in computerised systems.



ISO 639-2:1998 - Codes for the representation of names of languages — Part 2: Alpha-3 code: Because the ISO 639-1 standard uses only two-letter codes for languages, it is not able to accommodate a sufficient number of languages. ISO 639-2:1998, confirmed in 2016, provides two sets of three-letter alphabetic codes for the representation of names of languages, one for terminology applications and the other for bibliographic applications. This part of ISO 639 also includes guidelines for the creation of language codes and their use in some applications.

ISO 639-3:2007 - Codes for the representation of names of languages — Part 3: Alpha-3 code for comprehensive coverage of languages, was confirmed in 2016. Whereas ISO 639-1 and ISO 639-2 are intended to focus on the major languages of the world that are most frequently represented in the total body of the world's literature, ISO 639-3 extends the ISO 639-2 alpha-3 codes in order to cover all known natural languages. The language identifiers were devised for use in a wide range of applications, especially in computer systems, where there is potential need to support a large number of the languages that are known to have ever existed.

ISO 12620: 2019: Management of terminology resources — Data category specifications

ISO 12620 provides guidelines and requirements governing data category specifications for language resources, i.e. class of data items that are closely related from a formal or semantic point of view, such as /part of speech/, /subject field/, /definition/. ISO 12620 specifies mechanisms for creating, documenting, harmonising and maintaining data category specifications in a data category repository. It also describes the structure and content of data category specifications.

The two following standards do not focus specifically on lexicography or lexical resources, but can still be considered relevant in the context of ELEXIS:

ISO 30042:2019: Management of terminology resources — TermBase eXchange (TBX)

TBX is a framework for representing structured terminological data. It specifies useful notions for lexicography such as data category (e.g. /part of speech/, /subject field/, /definition/). Nevertheless, following the ISO principles on Terminology, TBX is concept-oriented and not word-oriented: the "terminological entry" is a "concept entry" defined as "part of a terminological data collection which contains the terminological data related to one concept". TBX is more dedicated to terminography than to lexicography.

ISO/IEC 19505-1:2012: Information technology - Object Management Group Unified Modeling Language (OMG UML), Infrastructure. The [Unified Modeling Language \(UML\)](#) is a general-purpose modeling language with a semantic specification, a graphical notation, an interchange format, and a repository query interface. It is designed for use in object-oriented software applications, including those based on technologies recommended by the Object Management Group (OMG).

ISO/IEC 19505-2:2012: Information technology — Object Management Group Unified Modeling Language (OMG UML) — Part 2: Superstructure, confirmed in 2017, defines the [Unified Modeling Language \(UML\), revision 2](#). The objective of UML is to provide system architects, software engineers, and software developers with tools for analysis, design, and implementation of software-based systems



as well as for modeling business and similar processes. UML is used for description of the specification of concepts in ISO Standards such as the LMF core package (ISO 24613-1) and the LMF Machine Readable Dictionaries Model (MRD) model (ISO 24613-2).

ISO 24156-1:2014: Graphic notations for concept modelling in terminology work and its relationship with UML — Part 1: Guidelines for using UML notation in terminology work. This standard gives guidelines for using a subset of UML symbols independent of their normal UML meaning, to represent concepts in concept models that result from concept analysis. It describes how UML symbols can be used for that. It does not describe the principles and methods of terminology work, which is covered in ISO 704.



2.3 Standard Format Marker and Multi-Dictionary Format Codes

In Field Linguistics, the use of standard format marker (SFM) codes has become a de facto standard for encoding lexicon structures. SFM codes are backslash codes marking the individual fields in a dictionary entry. For instance, \lx marks the lexeme field and \se the sense field in an SFM entry. Only the beginning of a field is marked, and so the format can contain ambiguities in fields that span more than one line e.g. senses, examples and etymologies. Simple tools that read SFM files make no assumptions about the data and are unable to enforce data integrity.

Multi-Dictionary Formatter (MDF) is a restricted set of these codes. This set was originally defined as part of the MDF software program which was developed at the end of the past century by [SIL International](#), a faith-based nonprofit organisation which is especially active in the area of field linguistics. SIL observed that formatting and printing of a dictionary formed a continual source of frustration for many linguists and anthropologists who compile dictionaries. Getting the information from this format to a printed document could be so frustrating to the ordinary computer user that it would not get done at all—or at least not until one could get the help of a computer whiz. MDF was designed to bridge this gap. The only requirement to use the MDF program, was that the data had to be marked up with field codes recognised by MDF. Using this system of field markers, MDF could then automatically format lexical data as a traditional print double-column formatted dictionary.

The MDF tool has now become obsolete and it has been replaced by Toolbox and more recently by the FieldWorks Language Explorer tool. Both tools can import SFM/MDF files. Toolbox opens MDF files directly but does not enforce any integrity on the data, whereas in FieldWorks, the data has to comply with the FieldWorks conceptual model. In this model, fields are grouped into objects or classes that represent entries, senses, example sentences, etc. as is illustrated in the following diagram:



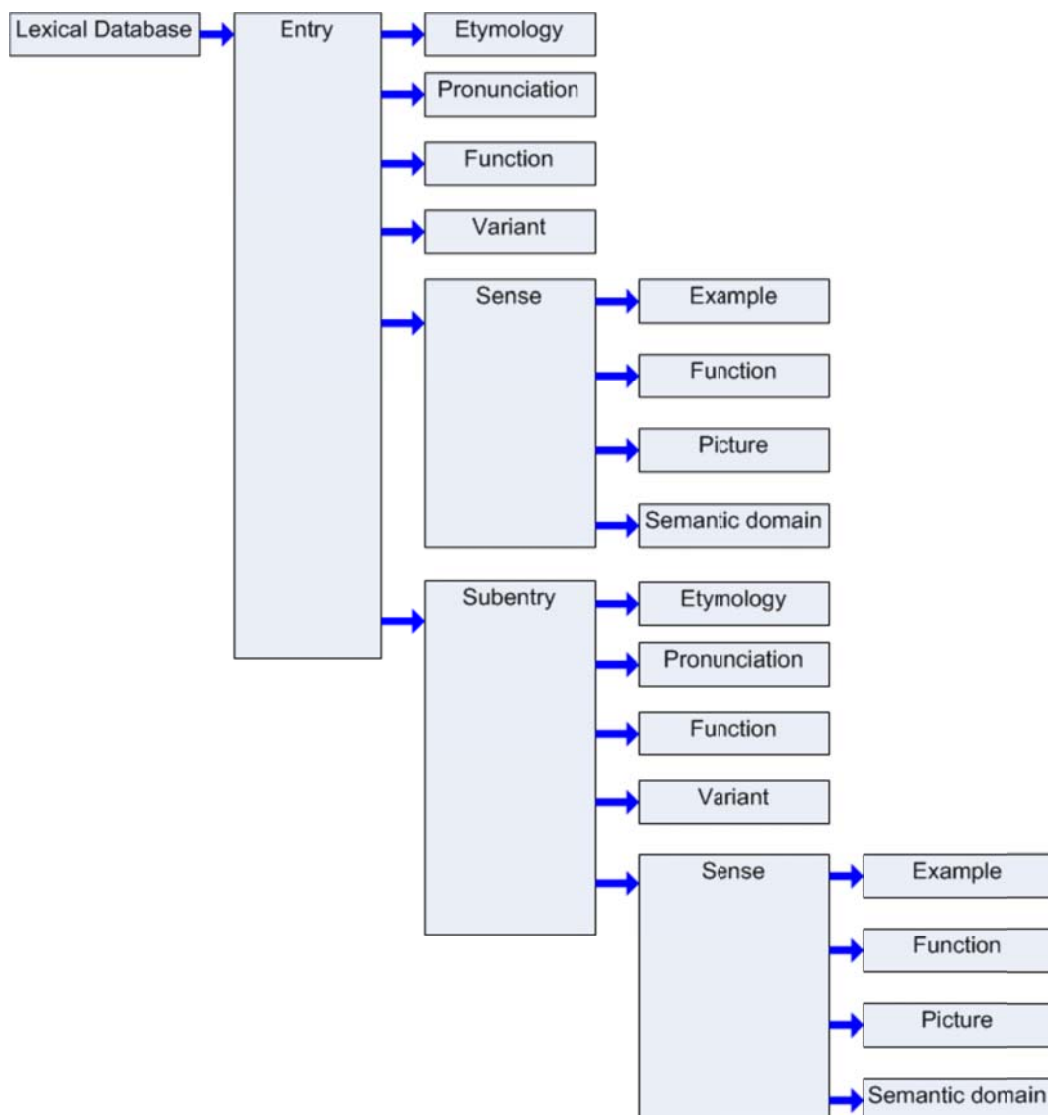


Figure 5: FieldWorks conceptual model (Zook 2009: 3)

Fieldwork's import wizard allows the user to map each SFM/MDF marker to the corresponding field and writing system. Importing SFM/MDF data into FieldWorks is not always trivial on account of the possible ambiguity inherent in the input data. FieldWorks attempts to make reasonable assumptions while importing and describes those assumptions to the user in detail during the import process. Pre-processing the SFM/MDF file is preferable and enables an import without assumptions. FieldWorks data is stored in an XML file structured as a series of key, value pairs. There is no simple conversion from that XML format to any other, however FieldWorks can export the data in a variety of formats, including XHTML, SFM and XML.

3 ELEXIS interoperability formats

For both XML and RDF, there is currently work done on standardisation. There is the initiative of TEI Lex-0 with a special focus on retro-digitised dictionaries. In addition, in the Linked Open Data community, the Ontolex-Lexicon Community group is working on a module for dictionaries, the lexicog module. Both TEI Lex-0 and Ontolex-Lexicon are envisaged as standards for best practices in lexicography, and are supported within ELEXIS. In this section, we provide a short summary of both.

3.1 TEI Lex-0

TEI Lex-0 aims at establishing a baseline encoding and a target format to facilitate the interoperability of heterogeneously encoded lexical resources. This is important both in the context of building lexical infrastructures as such and in the context of developing generic TEI-aware tools such as dictionary viewers and profilers.

TEI Lex-0 should not be thought of as a replacement of the Dictionary Chapter in the *TEI Guidelines* or as the format that must be used for editing or managing individual resources, especially in those projects and/or institutions that already have established workflows based on their own flavours of TEI. TEI Lex-0 should be primarily seen as a format which implements a set of constraints on top of those provided by the TEI Guidelines so that existing TEI dictionaries, once univocally transformed, can be queried, visualized, or mined in a uniform way. At the same time, however, there is no reason why TEI Lex-0 could not or should not be used as a best-practice example in educational settings or as a set of best-practice guidelines for new TEI-based projects, especially considering the fact that the specification for TEI Lex-0 aims to stay as aligned as possible with the TEI subset developed in conjunction with the revision of the ISO LMF (Lexical Markup Framework) standard (section 2.2.3).

TEI Lex-0 is hosted by the DARIAH WG "Lexical Resources" in a [GitHub repository](#). The TEI Lex-0 format is actively and openly discussed using the GitHub [ticketing system](#).

The TEI Lex-0 repository consists of:

- [TEI Lex-0 specification](#) which is defined in an ODD file ("One Document Does it All"), a single XML resource which contains explanatory prose, examples of usage and formal declarations for components of the TEI Abstract Model (elements and attributes, modules, as well as classes and macros).
- the [RelaxNG schema](#) generated from the ODD file which can be used to validate the conformance of dictionary files with TEI Lex-0.
- a human-readable [HTML version](#) of the TEI Lex-0 specification.

TEI Lex-0 imposes different types of constraints vis-a-vis TEI:

- reducing the number of available elements (for instance, TEI Lex-0 uses only <entry>, whereas TEI has several elements for the basic unit of the dictionary microstructure: <entry>, <entryFree>, <superEntry>, <re> (related entry) and <hom> (homonym).
- making certain attribute values required (for instance, xml:lang and xml:id on <entry>)
- reducing the number of possible attribute values on certain elements (such as <usg>)



- enforcing additional syntactic constraints (for instance, <def> can only appear inside a <sense>) or, when necessary, allowing new syntactic constructs (for instance, nested entries)

```

<entry xml:lang="pt" xml:id="caixa-de-óculos">
  <form type="lemma">
    <orth>caixa-de-óculos</orth>
    <pron>kajʃed'ɔkuluf</pron>
  </form>
  <gramGrp>
    <gram type="lexicalConstruction" value="polylexical"/>
    <gram type="pos" norm="NOUN">s.</gram>
    <gram type="gen">m.</gram>
    <lbl>e</lbl>
    <gram type="gen">f.</gram>
  </gramGrp>
  <sense xml:id="caixa-de-óculos_1" n="1">
    <usg type="attitude">Deprec.</usg>
    <usg type="socioCultural">Fam.</usg>
    <usg type="attitude">Joc.</usg>
    <def>Pessoa que usa óculos.</def>
  </sense>
  <form type="inflected">
    <gramGrp><gram type="number">Pl.</gram></gramGrp>
    <orth>caixas-de-óculos</orth>
  </form>
</entry>

```

Figure 6: TEI Lex-0 encoded entry from DACL: *Dicionário da Academia das Ciências de Lisboa, 2020*. Ana Salgado, (coord.). Lisboa: Academia das Ciências de Lisboa.

In addition to being used as an ELEXIS interoperability format, TEI Lex-0 has been used in a number of training events, both in the context of DARIAH and ELEXIS.



3.2 OntoLex-Lemon

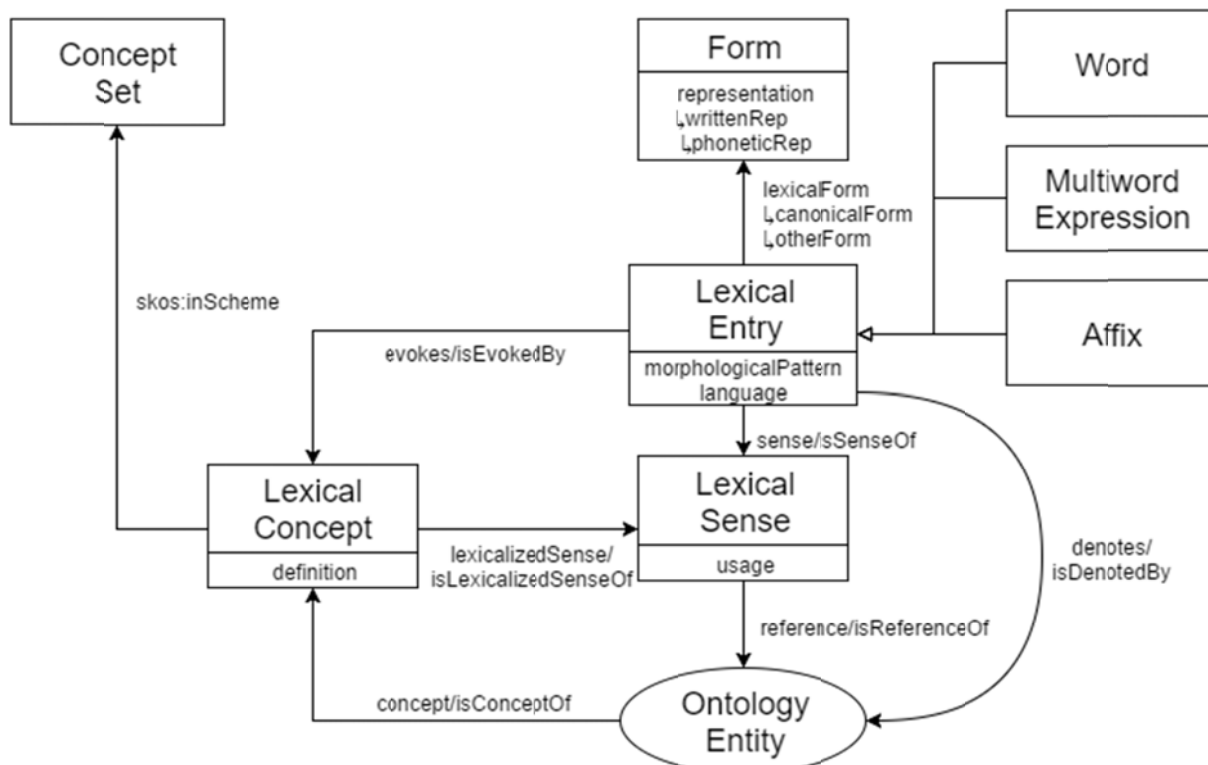


Figure 6: The OntoLex Core Model

The OntoLex-Lemon model has been developed by the OntoLex Community Group of the World Wide Web Consortium to act as a model for the representation of lexical information in ontologies. It has since developed into the de facto standard for representing lexical information as RDF. The group took as its starting point the lemon (Lexicon Model for Ontologies) model and developed the following modules for representing lexical information:

- Core module: This module was directly inspired by LMF in defining lexical entries as the core element of the lexicon. The lexical entry is then composed of a number of forms and semantically of a number of lexical senses. The meaning of these lexical senses can be explained either by a formal ontology or a language-independent lexical concept.
- Syntax and Semantics module: This module describes the interaction between the lexicon and ontology, by describing how ontological predicates may be expressed in natural language.
- Decomposition module: This module describes how a multi-word lexical entry can be decomposed into its forms.
- Variation and Translation module: This module provides vocabulary for describing relations between entries and in particular translation.
- Metadata module (LIME): Metadata is a key issue in describing the vocabulary of the model and many predicates for describing a lexicon are given here

In addition to the 5 modules described in the [initial standard](#), further modules have been developed to enhance the usage. In particular, it was observed that many users of OntoLex-Lemon were not using an

ontology to describe the semantics of their lexical entries, and so modules were introduced to provide better representation of general lexicographic resources.

- [Lexicography Module](#): This module was developed to better represent traditional dictionaries. One of the major issues encountered was that the lexical entry defined in the core had strict requirements that made it suitable for natural language processing applications. In particular, it required that each lexical entry had a single lemma, part-of-speech, morphology and etymology. As this is frequently not the case in traditional dictionaries and more general Entry was introduced that can group lexical entries or lexical senses together. In addition, two further innovations were included in the module: namely, the ability to restrict meanings to certain forms (e.g., meanings that only apply to plural forms of nouns) and to provide examples of the usage of a term.

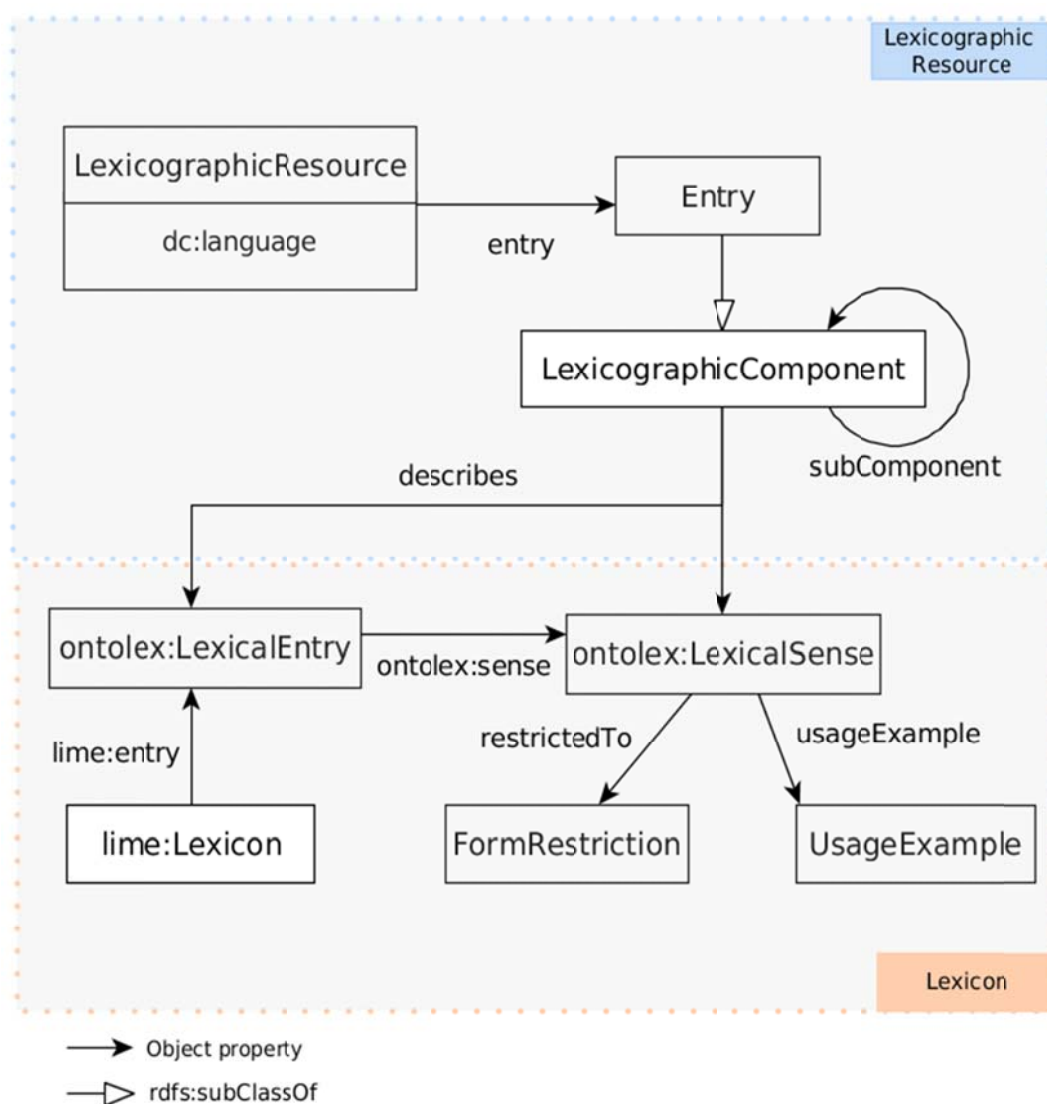


Figure 7: The Lexicography Module for OntoLex

- Morphology Module: This module is focused on describing the morphological decomposition of forms into *morphs*, it provides both a declarative mode for detailing the decomposition of a specific form, as well as a generative mode, which allows forms to be generated according to paradigms.
- Frequency, Attestation and Corpus Information (FrAC) Module: This module describes how a lexicon can be connected to a corpus and as such general linguistic information such as the frequency of an entry and its attested occurrences in a text can be documented. This is achieved in combination with the Web Annotation Data Model.



4 Data formats and standards within ELEXIS

In each and every European country, elaborate efforts are put into the development of lexicographic resources describing the language(s) of the community. These lexicographic resources are often the result of long-term projects in which literally thousands of person years were and continue to be dedicated to their compilation in national and regional projects.

Although confronted with similar problems relating to technologies for producing and making these resources available, cooperation on a larger European scale has long been limited. In addition, standardisation efforts have not been particularly successful within the field of lexicography before the digital age, an observation which was confirmed by the results from the WP1 survey on User Needs (Kallas et al. 2019). More specifically, the results from the survey for the lexicographic partner institutions show that:

- most lexicographic projects use XML or databases, but some projects are still working with non-structured data and text format.
- custom XML and TEI are the most commonly used XML formats.
- most institutions do not use existing standard vocabularies for encoding their lexicographic data. Two institutions pointed out TEI as the standard vocabulary they use for their projects, and one institution mentioned IsoCat, GOLD, TEI (most likely for different projects).

While the survey only covered the 10 ELEXIS lexicographic partner institutions, we think it is safe to conclude that the lexicographic landscape in Europe is still rather heterogeneous. It is characterised by stand-alone lexicographic resources, typically encoded in incompatible formats due to the isolation of efforts, which prohibits reuse of this valuable data in other fields, such as natural language processing, linked open data and the Semantic Web, as well as in the context of digital humanities.

The data from the ELEXIS partner institutions comes in the following formats, i.e. custom-XML for contemporary dictionaries, various versions of TEI mainly for retrodigitised dictionaries, relational databases (Oracle or MySQL) and more recent also API access is offered¹.

In order to ensure semantic interoperability between these diverse dictionary structures, ELEXIS will establish a common model. Such a model is necessary to a) streamline the integration of lexicographic data into the ELEXIS infrastructure, b) to allow reliable linking of the data in the dictionary matrix, and c) to form a basic template for the creation of new lexicographic resources, such that they can automatically benefit from the tools and services provided by the ELEXIS infrastructure.

¹ For instance, the API from the Estonian Language Institute can be found at <https://github.com/tripledev/ekilex/wiki/Ekilex-API> and the one from KDictionaries at <https://www.lexicala.com/>.



The aim of the project is not to develop a fully-fledged data model. Neither does the project aim to replace existing models. The main aim is to ensure semantic interoperability between lexicographic resources predominantly using their own custom format. Our first intermediate goal during this grant period was to establish a common reference model where the main concepts are unambiguously defined (e.g. translation in resource X refers to the same object as a translation in resource Y) so that the process of mapping them can be done consistently and without too much anguish.

To start with a detailed analysis of sample data (provided by the lexicographic partners and observing institutions) has been carried out identifying the “core” requirements and their current encoding in the different data sets. The following core elements have been identified:

- **entry**
- **headword and secondary headword**
- **part of speech**
- **language (source language and target language)**
- **sense**
- **sense structure**
- **definition**
- **translation**
- **example**
- explanation (gloss and/or sense indicator)
- cross reference
- lexical relation (synonym, antonym, hypernym, hyponym)
- form paradigm
- inflected form
- label
- multi word expression
- phrase
- collocation

The elements highlighted in bold are currently supported in the ELEXIFIER tool (see D1.3) and were prioritised. The next steps are to refine and finalise the definitions for these core elements and to express the ELEXIS data model in a formalism like UML. This way the serialisations to the two ELEXIS interoperability formats, i.e. Ontolex-Lemon and TEI Lex-0 can be realised.

Within the project, a number of meetings have already been organised supporting further development of Ontolex-Lemon and TEI Lex-0 in the context of ELEXIS. In July 2018 and in January 2019 a TEI Lex-0 meeting was organised by BCDH. In November 2018 an Ontolex-Lemon meeting was organised by NUIG in Leiden. In October 2019, a joint TEI Lex-0 and Ontolex-Lemon meeting was held in conjunction with the eLex conference in Sintra bringing the two communities together.

In addition, ELEXIS will recommend the use of standard vocabularies (e.g. for POS tags the Universal Dependencies tagset), and special attention will be given to metadata. For the latter, we will work in close collaboration with existing infrastructures such as CLARIN and DARIAH.



In the first 12 months of the project, a standardisation body was established in ELEXIS, currently with 16 members from 9 ELEXIS partners. The first face-to-face meeting of this Standards Committee was in Vienna at the observer event in February 2019. During the first 24 months the Standards Committee had several meetings mainly defining the role of the committee and discussing the KPI of producing a new standard (in OASIS) defined in the Grant Agreement under Objective 2.

In OASIS standardisation organisation, Lexicographic Infrastructure Data Model and API (LEXIDMA) Technical Committee was established at the end of 2019, with a [GitHub repository](#) designed for use in development of TC chartered work products and test suites. The committee's work can be followed at the [TC public web page](#). The Chair of the committee is Tomaž Erjavec (Jožef Stefan Institute, Slovenia) and the Secretary is David Filip (Trinity College Dublin, Ireland). From ELEXIS consortium, the following participants are also members of the committee as of January 2020: Simon Krek (JSI), Iztok Kosem (JSI), Miloš Jakubiček (LC), Ilan Kernerman (KD) and John McCrae (NUIG). Participation in the LEXIDMA TC is open to all interested parties. The main purpose is defined as:

“The LEXIDMA TC's purpose is to create an open standards based framework for internationally interoperable lexicographic work. The TC will develop a simple, modular, and easy to adopt data model that will be attractive for all lexicographic industry actors across companies and academia as well as geographic locations. Adoption of that model will facilitate exchange of lexicographic and linguistic corpus data globally and also enable effective exchange with adjacent industries such as language services, terminology management, or technical writing.

The TC will describe and define standard serialization independent interchange objects based predominantly on state of the art in the lexicographic industry. Defining specific serializations, transaction models, standard interfaces, and web services based on the defined objects and object models is also in scope as far as it facilitates the high level purpose set out here. It aims to develop this lexicographic infrastructure as part of a broader ecosystem of standards employed in Natural Language Processing (NLP), language services, and Semantic Web.”

The resulting ELEXIS data model which will be (among other uses in ELEXIS) exploited for the purpose of creating the ELEXIS “dictionary matrix”: a universal repository of linked senses, meaning descriptions, etymological data, collocations, phraseology, translation equivalents, examples of usage and all other types of lexical information found in all types of existing lexicographic resources, monolingual, multilingual, modern, historical etc., available through a RESTful web service developed as part of LEX1 infrastructure. ELEXIS dictionary matrix will be also available as part of Linguistic Linked Open Data cloud (LLOD), and it will serve as the source for providing links to (particular headwords, senses etc. in) dictionaries available online, through the [European Dictionary Portal](#), and included in the matrix.



5 References

- ISO 639-1:2002** *Codes for the representation of names of languages - Part 1: Alpha-2 code*
- ISO 639-2:1998** *Codes for the representation of names of languages — Part 2: Alpha-3 code*
- ISO 639-3:2007** *Codes for the representation of names of languages — Part 3: Alpha-3 code*
- ISO 1951:2007** *Presentation/representation of entries in dictionaries – Requirements, recommendations and information.*
- ISO 24611:2012** *Language resource management — Morpho-syntactic annotation framework (MAF)*
- ISO/IEC 19505-1:2012** *Information technology - Object Management Group Unified Modeling Language (OMG UML), Infrastructure*
- ISO/IEC 19505-2:2012** *Information technology - Object Management Group Unified Modeling Language (OMG UML), Superstructure*
- ISO 24156-1:2014** *Graphic notations for concept modelling in terminology work and its relationship with UML — Part 1: Guidelines for using UML notation in terminology work*
- ISO/CD 24613-1:2018(E)** *Language resource management — Lexical markup framework (LMF) — Part 1: Core model.*
- ISO/CD 24613-2:2019(E)** *Language resource management — Lexical markup framework (LMF) — Part 2: Machine Readable Dictionary (MRD) model.*
- ISO/WD 24613-3:2020(E)** *Language resource management — Lexical Markup Framework (LMF) — Part 3: Etymological Extension*
- ISO/WD 24613-4:2020** *Language resource management — Lexical Markup Framework (LMF) — Part 4: TEI serialisation*
- ISO NP 24613-5:2018** *Language resource management — Lexical markup framework (LMF) — Part 5: Lexical base exchange (LBX) serialization*
- ISO 12620: 2019** *Management of terminology resources — Data category specifications*
- ISO 30042:2019:** *Management of terminology resources — TermBase eXchange (TBX)*
- Kallas, J., Koeva, S., Kosem, I., Langemets, M. & Tiberius, C. (2019). ELEXIS deliverable 1.1 Lexicographic Practices in Europe: A Survey of User Needs. Available at: https://elex.is/wp-content/uploads/2019/02/ELEXIS_D1_1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf
- Zook, K. (2009) *Technical Notes on SFM Database Import.* Available at: <http://online.fliphtml5.com/bfud/cwvf/>



6 Annex: Sample dictionary entries

In this annex, we give a few examples of dictionary entries which are encoded in the standards described in this deliverable.

6.1 TEI encoded dictionaries

6.1.1 Dictionary of Old Dutch (IVDNT)

The Dictionary of Old Dutch (ONW) is a scientific dictionary that describes Dutch from around 500 to 1200. The ONW contains 8954 keywords and around 30,000 quotes. The dictionary was created between 1998 and 2008. See: <http://gtb.ivdnt.org/search/>

Below, the entry for *katta* ‘cat’ in the online version of the Dictionary of Old Dutch.

KATTA

Woordsoort: znw., v.
 Modern lemma: Kat, kat

Oudste attestatie: 1165

Frequentie: totaal: 6, appellatieven: 1, toponiemen: 5

Etymologie: *Zie voor de etymologie, EWN II, 651. Cognaten: Oudfries katte.*

Morfologie: *ongeleed.*

Flexie: in Latijnse context cath (1)

als deel van een toponiem cat- (3), cate- (1), catu- (TW leest cat-) (1)

Overige historische woordenboeken: VMNW: *catte* (znw.v.), MNW: *catte* (znw.v.), WNT: *kat* (I) (znw.v.)

↔ 1. Kat, wilde kat. *In het Oudnederlands alleen als toponymisch element en als toenaam overgeleverd, vgl. Debrabandere 2003: 231.*

+ Als eerste deel van een toponiem

Literatuur:

Debrabandere 2003 231

EWN II 651

Koppelingen:
 Vorig artikel: *katēl*
 Volgend artikel: *kegila*
GTB Woordenboeken:
 VMNW: *catte* (znw.v.),
 MNW: *catte* (znw.v.),
 WNT: *kat* (I) (znw.v.)

The TEI source encoding for this entry is as follows:

```
<entry xml:id="ID2684" type="main">
  <interpGrp type="GTB-entry">
    ...
  </interpGrp>
  <form type="lemma">
    <orth extent="full">katta</orth>
  </form>
  <form type="mdl">Kat, kat</form>
  <dictScrap>
    ...
  </dictScrap>
  <sense>
    <interpGrp type="metadata">
      ...
    </interpGrp>
  </sense>
</entry>
```



```

</interpGrp>
<def>Kat, wilde kat. <note place="unspecified" type="def">In het Oudnederlands alleen als
toponymisch element en als toenaam overgeleverd, vgl. Debrabandere 2003: 231.</note></def>
<cit>
  <q>de allodiis walteri <hi rend="font-weight:bold">cath</hi>.</q>
  <note place="unspecified" type="translation">M.b.t. het erfgoed van Wouter Kat.</note>
  <bibl>
    <title><idno n="OnwBr0031"/>Leys 1958: 154</title>
    <placeName>
      <settlement>&#xA0;[z.p.]</settlement>
    </placeName>
    <date type="dateRange" atLeast="1165" atMost="1165"/>
  </bibl>
</cit>
<sense>
  <interpGrp type="metadata">
    ...
  </interpGrp>
  <def>Als eerste deel van een toponiem</def>
  <re xml:id="ID2684.re.4" type="toponym">
    <form type="lemma">
      <orth extent="full">kattafurda ?</orth>
    </form>
    <form type="lemma">
      <orth extent="full">kattawurth ?</orth>
    </form>
  <sense>
    <interpGrp type="metadata">
      ...
    </interpGrp>
    <def><placeName>*Katvoorde, *Katwierde</placeName> onbekende plaats, mog. bij
    Saaksum, Baflo, prov. Groningen</def>
    <note place="unspecified">Het eerste element is onzeker. Het tweede element is
    eigenlijk ofri. <hi rend="font-style:normal">uur&#xF0;</hi>, onl. <hi rend="font-
    style:normal">wurth</hi> 'bewoonde hoogte', maar dit is geherinterpreteerd als <hi
    rend="font-style:normal">furda</hi>.</note>
    ...
  </sense>
</re>
</sense>
</sense>
<interpGrp type="listBibl-metadata">
  ...
</interpGrp>
</entry>

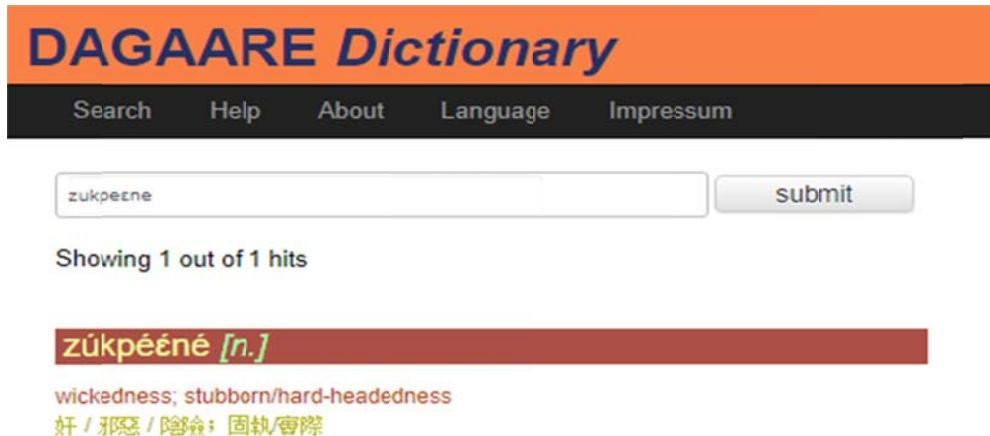
```



6.1.2 A machine-readable Dictionary of Dagaare (OEAW)

The Online Dagaare - English Lexicon was compiled by Adams Bodomo and published in this form in 2015. It comprises more than 1250 entries. See: <https://dagaare.acdh.oeaw.ac.at/>

Below the entry for *zúkpééné* ('wickedness'):



The screenshot shows the 'DAGAARE Dictionary' website. At the top, there is a navigation bar with links for 'Search', 'Help', 'About', 'Language', and 'Impressum'. Below the navigation bar is a search input field containing the text 'zukpeene' and a 'submit' button. Underneath the search field, it says 'Showing 1 out of 1 hits'. The main entry for 'zúkpééné [n.]' is displayed in a dark red box. Below the entry title, the English translation 'wickedness; stubborn/hard-headedness' is shown, followed by the Chinese translation '奸 / 邪惡 / 陰險; 固執/實際'.

The TEI source encoding for this entry is as follows:

```
<entry xml:id="sid_01247">
  <form type="lemma">
    <orth>zukpeene</orth>
    <orth type="diacritisized">zúkpééné</orth>
  </form>
  <gramGrp>
    <gram type="pos">n.</gram>
  </gramGrp>
  <sense>
    <cit type="translation" xml:lang="zh-yue-Latn">
      <quote>gaan1 / ce4 ngok3, jam1 him2; gu3 zap1 / sat6 zai3</quote>
    </cit>
    <cit type="translation" xml:lang="zh-yue">
      <quote>奸 / 邪惡 / 陰險 ; 固執/實際</quote>
    </cit>
    <cit type="translation" xml:lang="eng">
      <quote>wickedness; stubborn/hard-headedness</quote>
    </cit>
    <cit type="translation" xml:lang="deu">
      <quote></quote>
    </cit>
  </sense>
</entry>
```

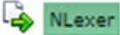



6.1.3 Mittelhochdeutsches Handwörterbuch von Matthias Lexer (TCDH)

Only a few years after completing the Middle High German dictionary (BMZ), which G. F. Benecke, W. Müller and F. Zarncke had created as a family dictionary of words, the need arose for easier access to the Middle High German keywords. This is how the Middle High German dictionary by Matthias Lexer was created as an alphabetical index to the BMZ. At the same time, Lexer made an effort to create a convenient handheld dictionary, and also added words and evidence from newly published sources. In addition to the BMZ, the Lexer particularly includes texts from the late Middle Ages and extends the range of text types primarily to include chronic, legal and religious literature. In this way, the Lexer records approximately 34,000 new keywords. However, the hand dictionary remains very closely related to the BMZ as an index; Articles on such keywords that are also used in the BMZ must always be read together with the corresponding BMZ articles.

See: <http://www.woerterbuchnetz.de/cgi-bin/WBNetz/setupStartSeite.tcl>

Below the entry for *blas* ('pale') as it appears in the Wörterbuchnetz:

  **blas** *adj.* (BMZ I. 200^a) *kahl.* ich liez mich roufen, daz ich blas (: was) wurd an mînem houbet *Ls.* 1. 298,64; *bildl. schwach, gering, nichtig* NEIDH. 48,18 *u. anm.; oft bei JER. auch in heutiger bedeut. blass, bleich* s. PFEIFF. s. 132. — *der grundbegriff ist wol »scheinen, leuchten«: altn. blasa nhd. erscheinen, ags. blase nhd. fackel. vgl. WEIG. 1,158. DWB. 1,73;*

The corresponding TEI encoding looks like this:

```
<entry xml:id="LB02975" type="fam" n="100296.41">
  <form type="headword">
    <form type="lemma">
      <ref target="NB01667 FB01827">blas</ref>
    </form>
  </form>
  <gramGrp>
    <gram type="adj">adj.</gram>
    <ref type="BMZ" n="1.200.a" target="BB01297">(I. 200<hi rend="sup">a</hi></ref>
  </gramGrp>
  <sense>
    <def>kahl.</def>
    <cit type="example"><q>ich liez mich roufen, daz ich blas (: was) wurd an mînem
      houbet</q></cit>
    <title n="QL0037" type="sigle"><bibl><author>Ls.</author></bibl> <ref>1.
      298,64</ref></title>;
  </sense>
  <sense>
    <def>bildl. schwach, gering, nichtig</def>
```



```

<title n="QN0005" type="sigle"><bibl><author>Neidh.</author></bibl> <ref>48,18 <hi
rend="italic">u. anm.</hi></ref></title>;
</sense>
<sense>
  <hi rend="italic">oft bei</hi>
  <title n="QJ0007" type="sigle">
    <bibl>
      <author>Jer.</author>
    </bibl>
  </title>
  <hi rend="italic">auch in heutiger bedeut.</hi>
  <def>blass, bleich</def>
  <hi rend="italic">s.</hi>
  <title type="sigle">
    <bibl>
      <author>Pfeiff.</author>
      <hi rend="italic">s.</hi>
    </bibl>
    <ref>132.</ref>
  </title>
</sense>
<etym>— <hi rend="italic">der grundbegriff ist wol „scheinen, leuchten“:</hi>
  <hi rend="italic">altn.</hi> <lang rend="altn">blasa
    <lang rend="nhd" n="trans"><hi rend="italic">erscheinen</hi></lang></lang>,
  <hi rend="italic">ags.</hi> <lang rend="ags">blase
    <lang rend="nhd" n="trans"><hi rend="italic">fackel</hi></lang></lang>. <hi
rend="italic">vgl.</hi>
  <title n="QW0039" type="sigle"><bibl><author>Weig.</author></bibl>
  <ref>1,158.</ref></title>
  <title n="QD0048" type="sigle"><bibl><author>Dwb.</author></bibl>
  <ref>1,73</ref></title>; </etym>
</entry>

```



6.2 ISO encoded dictionaries

6.2.1 ISO 1951: Dictionary of Karelian ([Institute for the Languages of Finland](#))

The *Dictionary of Karelian* is a dialect dictionary of Karelian, which is a Finnic language. The commentaries are in Finnish. The dictionary describes the vocabulary of the two main dialects of Karelian: Karelian Proper and Olonets Karelian (Livvi-Karelian). The dictionary, comprising six volumes with a total of 3,800 pages and almost 83,000 entries, has been published both in print and online. See: https://www.kotus.fi/en/dictionaries/dictionary_of_karelian.

Below the entry for *kuusi* ('spruce') in the online version of the dictionary.

The corresponding XMLEX ISO encoding looks like this:

```
<!DOCTYPE Dictionary
  SYSTEM "http://kaino.kotus.fi/dtd/xmllex/XmLex_V00_kotus.dtd">
<Dictionary xmlns:xlink="http://www.w3.org/TR/xlink" xmlns:kotus="http://www.kotus.fi/"
  sourceLanguage="karjala" targetLanguage="suomi">
  <DictionaryEntry sortKey="27311" identifier="kuusi01" homographNumber="I">
    <HeadwordCtn>
      <Headword>kuusi</Headword>
      <SearchForm>kuusi</SearchForm>
      <PartOfSpeechCtn>
        <PartOfSpeech display="no" freeValue="s." value="noun"/>
      </PartOfSpeechCtn>
      <GrammaticalNote display="yes">s.</GrammaticalNote>
      <Definition>kuusi (Picea excelsa).</Definition>
      <ExampleBlock>
        <ExampleCtn>
          <Example>
            <Fragment>kuusi</Fragment>. </Example>
            <FreeTopic type="levikki">
              <GeographicalUsage freeType="pitäjä" class="pitäjä">Oulanka</GeographicalUsage>
            </FreeTopic>
          </ExampleCtn>
          <ExampleCtn>
            <Example>
              <Fragment>kuuši</Fragment>. </Example>
              <FreeTopic type="levikki">
                <GeographicalUsage freeType="pitäjä" class="pitäjä">Kiestinki</GeographicalUsage>
                <GeographicalUsage freeType="pitäjä" class="pitäjä">Pistoj</GeographicalUsage>
                <GeographicalUsage freeType="pitäjä" class="pitäjä">Uhtua</GeographicalUsage>
                <GeographicalUsage freeType="pitäjä" class="pitäjä">Vuokkin</GeographicalUsage>
              </FreeTopic>
            </ExampleCtn>
            <ExampleCtn>
              <Example>
                <Fragment>kuuži</Fragment>. </Example>
                <FreeTopic type="levikki">
                  <GeographicalUsage freeType="pitäjä" class="pitäjä">Jyskyj</GeographicalUsage>
                  <GeographicalUsage freeType="pitäjä" class="pitäjä">Tunkua</GeographicalUsage>
                </FreeTopic>
              </ExampleCtn>
            </ExampleBlock>
  </DictionaryEntry>
</Dictionary>
```




```

    <GeographicalUsage freeType="pitäjä" class="pitäjä">Repola</GeographicalUsage>
    <GeographicalUsage freeType="pitäjä" class="pitäjä">Mäntys</GeographicalUsage>
    <GeographicalUsage freeType="pitäjä" class="pitäjä">Poraj</GeographicalUsage>
    <GeographicalUsage freeType="pitäjä" class="pitäjä">Tver</GeographicalUsage>
  </FreeTopic>
</ExampleCtn>
<ExampleCtn>
  <Example>
    <Fragment>kuužešta luajitah astieda</Fragment>. </Example>
  <FreeTopic type="levikki">
    <GeographicalUsage freeType="pitäjä" class="pitäjä">Rukaj</GeographicalUsage>
  </FreeTopic>
</ExampleCtn>
</ExampleBlock>
</HeadwordCtn>
</DictionaryEntry>
</Dictionary>

```

6.2.2 LMF: OMBI Arabic-Dutch (IVDNT)

OMBI-Arabic-Dutch and OMBI-Dutch-Arabic are bilingual lexical resources which were originally compiled within the framework of the project “Woordenboek Nederlands-Arabisch, Arabisch-Nederlands, Nijmegen” in the period of 1998 till 2002 at the Radboud University of Nijmegen. This project was part of a large government initiative in the Netherlands and Flanders in the 1990s aimed at improving and stimulating the production of bilingual dictionaries and lexical databases with Dutch as source or target language. The printed dictionaries for Arabic and Dutch (Hoogland et al 2003) were published in 2003 by Bulaaq, Amsterdam. To ensure future interchangeability and interoperability of these bilingual lexical resources, the original format was converted to XML-LMF ([Maks et al. 2008](#)).

The LMF encoding of the entry for ائتلافِي ‘coalition’ is given below:

```

<LexicalEntry LE-id="6" LE-homonymnr="1" sy-pos="adj">
  <LE-admin osrcfuid="119415"/>
  <Form-A>
    <LemmatisedForm-A writtenForm="ائتلافِي"/>
    <Morpho-syntax-A mor-type="" mor-comparisonType="" mor-declinability="" flec-flectionalType="" sy-adverbialUsage="" sy-position=""/>
  </Form-A>
  <Sense-A S-id="13" S-seqnr="1" sem-resume="pol">
    <S-admin o-srcluid="119416" o-srcfuid="119415"/>
    <Semantics semtype="" sem-shift="">
      <Definition>
        <sem-def></sem-def>
        <sem-defSource></sem-defSource>
      </Definition>
    </Semantics>
    <Syntax-A>
      <Sy-complementation>
      </Sy-complementation>
    </Syntax-A>
  </Sense-A>
</LexicalEntry>

```



```

</Syntax-A>
<Pragmatics prag-connotation="" prag-geography="" prag-subjectField="" prag-style="formal" prag-origin=""
  prag-socGroup="" prag-chronology=""/>
<Translations-Sense>
  <Description Descr-id="17" description="coalitie-">
    <Descr-admin o-did="638094"/>
  </Description>
</Translations-Sense>
<Examples>
  <Example Ex-id="18" Ex-seqnr="1" canonicalForm="ائتلافية حكومة" textualform="">
    <Ex-admin o-srcexid="119498" o-srcluid="119416" o-srcfuid="119415"/>
    <Syntax-Ex sy-category="" sy-type="free"/>
    <Semantics-Ex exDefinition=""/>
    <Pragmatics prag-connotation="" prag-geography="" prag-subjectField="" prag-style="" prag-origin=""
      prag-socGroup="" prag-chronology=""/>
    <Translations-Ex>
      <Translation-Ex TrEx-id="5" TrEx-seqnr="1" TrEx-equivalent="coalitieregering" TrEx-pos="noun"
        TrEx-equivalency="complete equivalent">
        <TrEx-admin o-transid="119499" o-tarfuid="119415" o-tarluid="119416" o-tarexid="119498" tr-
          resume="regering door coalitiepartijen" tr-form="coalitieregering" tr-pos="noun"/>
      </Translation-Ex>
      <Translation-Ex TrEx-id="6" TrEx-seqnr="2" TrEx-equivalent="coalitiekabinet" TrEx-pos="noun"
        TrEx-equivalency="unmarkAN">
        <TrEx-admin o-transid="205341" o-tarfuid="51815" o-tarluid="51816" o-tarexid="119498" tr-
          resume="kabinet van meerdere partijen" tr-form="coalitiekabinet" tr-pos="noun"/>
      </Translation-Ex>
      <Translation-Ex TrEx-id="7" TrEx-seqnr="3" TrEx-equivalent="regeringscoalitie" TrEx-pos="noun"
        TrEx-equivalency="unmarkAN">
        <TrEx-admin o-transid="322572" o-tarfuid="51815" o-tarluid="51816" o-tarexid="119498" tr-
          resume="coalitie in een regering" tr-form="regeringscoalitie" tr-pos="noun"/>
      </Translation-Ex>
    </Translations-Ex>
  </Example>
</Examples>
</Sense-A>
</LexicalEntry>

```



6.3 TEI Lex-0

6.3.1 Dicionário da Academia das Ciências de Lisboa

The Academia das Ciências de Lisboa has digitised the DLPC: Dicionário da Língua Portuguesa Contemporânea which was published in 2001. The digital version is known as DACL: Dicionário da Academia das Ciências de Lisboa (Ana Salgado, (coord.) 2020).

The dictionary is currently being revised both in terms of lexicographic content and the data model (TEI to TEI Lex-0). The Academy plans to keep both formats —TEI in the backend, and TEI Lex-0 for data interchange and interoperability. Below the entry for *antepassado* ('ancestor') from the print dictionary.

antepassado¹, a [ẽtĩpẽsádu, -ẽ]. *adj.* (De *ante-* + *passado*).
 Que pertence ou viveu numa época anterior. ≈ ANTECESSOR, PREDECESSOR. ≠ DESCENDENTE, SUCESSOR.

antepassado² [ẽtĩpẽsádu]. *s. m.* (De *ante-* + *passado*).
 1. Pessoa que é ascendente de outra ou outras. ≈ ASCENDENTE. ≠ DESCENDENTE. *Certos povos crêem-se descendentes de um antepassado comum. «o vaqueiro, pai do vaqueiro, o avô e outros antepassados mais antigos haviam-se acostumado a percorrer veredas, afastando o mato com as mãos.»* (G. RAMOS, *Vidas Secas*, p. 36). 2. *pl.* Pessoas anteriormente ao momento actual. ≈ ANTECESSORES. ≠ VINDOUROS. *Herdámos estes costumes dos nossos antepassados. Culto dos antepassados.*

Figure taken from DLPC: *Dicionário da Língua Portuguesa Contemporânea*. 2001. João Malaca Casteleiro (coord.), 2 vols. Lisboa: Academia das Ciências de Lisboa & Editorial Verbo.

The corresponding XML encoding for the first entry of *antepassado* looks like this:



```

<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Dicionário da Academia das Ciências de Lisboa</title>
      </titleStmt>
      <publicationStmt>
        <publisher>Academia das Ciências de Lisboa</publisher>
      </publicationStmt>
      <sourceDesc>
        <bibl>
          <title>Dicionário da Língua Portuguesa Contemporânea</title>
          <extent>2 volumes</extent>
          <extent>3809 pp.</extent>
          <author>Academia das Ciências</author>
          <publisher>Editorial Verbo</publisher>
          <date>2001</date>
        </bibl>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <entry type="derivativeWord" xml:lang="pt" xml:id="antepassado.1" n="1">
        <form type="lemma">
          <orth>antepassado</orth>
        </form>
        <form type="inflected">
          <orth>antepassado</orth>
          <pron>ẽtipẽs 'adu</pron>
          <gramGrp>
            <gram type="gen">m.</gram>
          </gramGrp>
        </form>
        <form type="inflected">
          <orth>antepassada</orth>
          <gramGrp>
            <gram type="gen">f.</gram>
          </gramGrp>
          <pron>ẽtipẽs 'ade</pron>
        </form>
        <gramGrp>
          <gram type="pos" norm="ADJ">adj.</gram>
        </gramGrp>
        <etym type="grammaticalization">
          <seg type="desc">De</seg>
          <cit type="etymon">
            <form>
              <orth extent="pref">ante-</orth>
            </form>
          </cit>
        </etym>
      </entry>
    </body>
  </text>
</TEI>

```



```

    <cit type="etymon">
      <form>
        <orth>passado</orth>
      </form>
    </cit>
  </etym>
  <sense xml:id="antepassado_1">
    <def>Que pertence ou viveu numa época anterior.</def>
    <xr type="synonymy">
      <ref type="sense">antecessor</ref>
    </xr>
    <xr type="synonymy">
      <ref type="sense">sucessor</ref>
    </xr>
    <xr type="antonymy">
      <ref type="sense">descendente</ref>
    </xr>
    <xr type="antonymy">
      <ref type="sense">sucessor</ref>
    </xr>
  </sense>
</entry>
</body>
</text>
</TEI>

```

Encoding taken from: DACL: *Dicionário da Academia das Ciências de Lisboa*, 2020. Ana Salgado, (coord.). Lisboa: Academia das Ciências de Lisboa.



6.4 Ontolex-Lemon

6.4.1 Global English resource (KD)

This lexicographic set consists of an English lexicographic core with translation equivalents in other languages, and is part of the Global series. The English core has over 17,000 entries including 27,000 senses and 30,000 examples of usage, and can be used on its own as a monolingual dictionary or serve as a base for developing bilingual and multilingual dictionaries, for learners or translation purposes. The first version was compiled from 2007 to 2010, the entries continue to be updated and new bilingual versions are added.

An extract of the ontolex-lemon encoding for the entry *cat* in JSON is given below:

```
{
  "@context": "https://api.lexicala.com/contexts/entry_context.json",
  "@id": "kd-lex:EN/cat-n",
  "@type": "ontolex:LexicalEntry",
  "entryId": "EN00001439",
  "dictionaryEntryId": "EN_DE00001695",
  "lexicon": {
    "@id": "kd-lex:EN",
    "@type": "lime:Lexicon",
    "language": "en"
  },
  "dictionary": {
    "@id": "kd-dictionary:EN",
    "@type": "kd:Dictionary"
  },
  "version": 1,
  "pos": "lexinfo:noun",
  "forms": [
    {
      "@id": "kd-lex:EN/cat-n-form",
      "@type": "Ontolex:Form",
      "text": {
        "en": "cat"
      },
      "pronunciation": {
        "en-fonipa": "kæt"
      }
    }
  ],
  "sense": [
    {
      "@id": "kd-lex:EN/cat-n-EN_SE00002867-sense",
      "@type": "ontolex:LexicalSense",
      "reference": {
        "@id": "kd-base:EN_SE00002867-concept",
        "@type": "skos:Concept",
        "definition": {
```



```

    "en": "an animal often kept as a pet"
  }
},
"usage": {
  "en": " "
},
"example": [
  {
    "@id": "kd-lex:EN/cat-n-EN_SE00002867-sense-TC00006643-ex",
    "@type": "kd:UsageExample",
    "value": {
      "en": "We have two cats."
    }
  },
  "relation": {
    "@id": "TC00006643-trans-ex-cl",
    "@type": "kd:TranslationExampleCluster",
    "relates": [
      {
        "@type": "kd:UsageExample",
        "value": {
          "br": "Temos dois gatos."
        }
      },
      {
        "@type": "kd:UsageExample",
        "value": {
          "da": "Nous avons deux chats."
        }
      },
      {
        "@type": "kd:UsageExample",
        "value": {
          "dk": "Vi har to katte."
        }
      },
      {
        "@type": "kd:UsageExample",
        "value": {
          "fr": "Nous avons deux chats."
        }
      },
      {
        "@type": "kd:UsageExample",
        "value": {
          "ja": "私たちは猫を2匹飼っている。"
        }
      },
      {
        "@type": "kd:UsageExample",
        "value": {

```




```

      "text": {
        "dk": "kat"
      }
    }
  },
  ...
],
"translation": [
  {
    "@id": "kd-trans:EN-ID/cat-n-EN_SE00002868-sense--cat-n-EN_SE00002868-sense-
    TC00006644-trans",
    "@type": "ontolex:Translation",
    "target": {
      "@id": "kd-lex:ID/-cat-n-EN_SE00002868-sense",
      "@type": "ontolex:LexicalSense",
      "reference": {
        "@id": "kd-base:EN_SE00002868-concept",
        "@type": "skos:Concept"
      },
      "sense_entry": {
        "@id": "kd-lex:ID/",
        "@type": "ontolex:LexicalEntry",
        "form": {
          "@id": "kd-lex:ID/-form",
          "@type": "ontolex:Form"
        }
      }
    }
  },
  {
    "@id": "kd-trans:EN-DK/cat-n-EN_SE00002868-sense-vildkat-cat-n-EN_SE00002868-sense-
    TC00006644-trans",
    "@type": "ontolex:Translation",
    "target": {
      "@id": "kd-lex:DK/vildkat-cat-n-EN_SE00002868-sense",
      "@type": "ontolex:LexicalSense",
      "reference": {
        "@id": "kd-base:EN_SE00002868-concept",
        "@type": "skos:Concept"
      },
      "sense_entry": {
        "@id": "kd-lex:DK/vildkat",
        "@type": "ontolex:LexicalEntry",
        "form": {
          "@id": "kd-lex:DK/vildkat-form",
          "@type": "ontolex:Form",
          "text": {
            "dk": "vildkat"
          }
        }
      }
    }
  }
]

```



```
    }  
  }  
},  
...  
{  
  "@id": "kd-lex:EN/cat-n-EN_SE00002869-sense",  
  "@type": "ontolex:LexicalSense",  
  "reference": {  
    "@id": "kd-base:EN_SE00002869-concept",  
    "@type": "skos:Concept"  
  }  
}  
]  
}
```

