elexis european lexicographic infrastructure

D2.1. Interface for Interoperable Lexical Resources

Author(s): Sina Ahmadi, Mihael Arcan, Thierry Declerck, Ilan Kernerman, Fahad Khan, Simon Krek, John McCrae, Michal Měchura, Monica Monachini, Christophe Roche, Carole Tiberius, Thomas Troelsgård, Ksenia Zaytseva

Date: January 30th 2019

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

H2020-INFRAIA-2016-2017
Grant Agreement No. 731015
ELEXIS - European Lexicographic Infrastructure

D2.1 Interface for Interoperable Resources

Deliverable Number: D2.1
Dissemination Level: Public
Delivery Date: January 29, 2019
Version: Draft
Author(s): Sina Ahmadi, Mihael Arcan, Thierry Declerck, Ilan Kernerman, Fahad Khan, Simon Krek, John McCrae, Michal Měchura, Monica Monachini, Christophe Roche, Carole Tiberius, Thomas Troelsgård, Ksenia Zaytseva

Project Acronym: ELEXIS
Project Full Title: European Lexicographic Infrastructure
Gran Agreement No.: 731015

Deliverable/Document Information

Deliverable No.: D2.1
Deliverable Title: Interface for Interoperable Resources
Author(s): Sina Ahmadi, Mihael Arcan, Thierry Declerck, Ilan Kernerman, Fahad Khan, Simon Krek, John McCrae, Michal Měchura, Monica Monachini, Christophe Roche, Carole Tiberius, Thomas Troelsgård, Ksenia Zaytseva

Dissemination Level: Public

Document History

Version Date	Changes/Approval	Author(s)/Approved by
V0.1 29/01/2019	First Draft of Joint Document	Sina Ahmadi, Mihael Arcan, Thierry Declerck, Ilan Kernerman, Fahad Khan, Simon Krek, John McCrae, Michal Měchura, Monica Monachini, Christophe Roche, Carole Tiberius, Thomas Troelsgård, Ksenia Zaytseva
V0.2 30/01/2018	Corrections incorporated	Fahad Khan, Carole Tiberius

Table of Contents

1 Introduction

The aim of this deliverable is to document the design for the ELEXIS interface for interoperable lexical resources and to describe the steps which are necessary for its implementation; the software that implements this interface will constitute a separate, subsequent, deliverable of the project, namely D 2.2. The interface documented in the current deliverable consists of a set of common protocols which take the form of REST API calls and which will allow dictionaries and lexicographic resources to be accessed through a single interface and in a uniform manner. Section 2 gives a detailed description of this interface. Each API call is described in terms of the objective, parameters and expected responses; an example is also given in each case.

This REST interface implemented by Deliverable 2.2 will allow users who wish to querya given endpoint to get back the metadata of the different lexicographic resources from that endpoint, as well as to query individual dictionaries with the possibility of getting back lexical entries in either JSON-LD, OntoLex or TEI LEX-0[1] (at least one of which must be implemented), these comprise the formats for interoperability of the ELEXIS project. In order to ensure a sufficient level of interoperability between these formats we make sure that all the part of speech values can be mapped to the UD part of speech tagset. In the case of TEI Lex-0 this means using the @norm tag with the TEI pos element to specify a normalised (UD) part of speech value for each entry; in the case of OntoLex (and JSON) we use a set of properties from the Lexinfo vocabulary[2] that can be mapped onto the UD part of speech tagset. Section 3 includes a description (each in a separate subsection) of each of the formats for interoperability of ELEXIS mentioned above.

Deliverable 2.2 will enable those who wish to expose their lexicographic data and who already have their data in one of the three formats to upload that data to a server which will make it available using the REST API interface described in Section 2. Those who do not have their data in the required formats on the other hand will be able to use other tools developed during the project in order to carry out a conversion of their data into one of the formats. This is described in more detail in Section 4 which also includes a description of the role of the REST interface described in Section 2 in the overall ELEXIS architecture.

2 REST Interface Description

2.1 Get Dictionaries

Objective
List all of the dictionaries that are available at the endpoint.

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

[2] https://lexinfo.net/

Parameters
None

Response

dictionaries	*Array of string* An array of identifiers for dictionaries

Table 1: Get Dictionaries Response

Example
http://www.example.com/dictionaries

```
{
  "dictionaries": [
    "dict1",
    "dict2"·····
  ]
}
```

2.2 About the Dictionary

Objective
Get the metadata about the dictionary, including the conditions under which it can be included in the Dictionary Matrix

Parameters

dictionary	*string* Identifier of the dictionary to describe

Table 2: About the Dictionary Parameters

Response

release	*PUBLIC, NONCOMMERCIAL, RESEARCH or PRIVATE* The conditions under which this dictionary (or data from this dictionary) can be included in the ELEXIS Dictionary Matrix
sourceLanguage	*ISO-639 code* The language of the lemmas in the dictionary, as an ISO 639-1,2,3 code. If a (two-letter) ISO 639-1 code exists this should be used in preference
targetLanguage	*Array of ISO-639 codes* The languages of the entries in the dictionary, can be identical to the source language or another language (for example in a bilingual dictionary).
Genre	*Array of strings*[3] The genre of the dictionary. One or more of the following: • **gen** General dictionaries are dictionaries that document contemporary vocabulary and are intended for everyday reference by native and

[3] List of dictionary genres taken from the European Dictionary Portal (http://www.dictionaryportal.eu/).

	fluent speakers. • **lrn** Learners' dictionaries are intended for people who are learning the language as a second language. • **ety** Etymological dictionaries are dictionaries that explain the origins of words. • **spe** Dictionaries on special topics are dictionaries that focus on a specific subset of the vocabulary (such as new words or phrasal verbs) or which focus on a specific dialect or variant of the language. • **his** Historical dictionaries are dictionaries that document previous historical states of the language. • **ort** Spelling dictionaries are dictionaries which codify the correct spelling and other aspects of the orthography of words. • **trm** Terminological dictionaries describe the vocabulary of specialized domains such as biology, mathematics or economics.
license	*URL* The license that can be used to republish this data
title	*string* The title of the resource
creator	*Agent* The creator of the resource
publisher	*Agent* The publisher of this resource

Table 3: About the Dictionary Response

In addition, any number of Dublin Core properties may be included as metadata, they are as follows:

- **abstract** - A summary of the resource.
- **accrualMethod** - The method by which items are added to a collection.
- **accrualPeriodicity** - The frequency with which items are added to a collection.
- **accrualPolicy** - The policy governing the addition of items to a collection.
- **alternative** - An alternative name for the resource.
- **audience** - A class of entity for whom the resource is intended or useful.
- **available** *(date)* - Date that the resource became or will become available.
- **bibliographicCitation** - A bibliographic reference for the resource.
- **conformsTo** - An established standard to which the described resource conforms.
- **contributor** *(array of agents)* - An entity responsible for making contributions to the resource.
- **coverage** - The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
- **created** *(date)* - Date of creation of the resource.
- **date** *(date)* - A point or period of time associated with an event in the lifecycle of the resource.
- **dateAccepted** *(date)* - Date of acceptance of the resource.
- **dateCopyrighted** *(date)* - Date of copyright.
- **dateSubmitted** *(date)* - Date of submission of the resource.
- **description** - An account of the resource.
- **educationLevel** *(date)* - A class of entity, defined in terms of progression through an educational or training context, for which the described resource is intended.

- **extent** - The size or duration of the resource.
- **hasFormat** - A related resource that is substantially the same as the pre-existing described resource, but in another format.
- **hasPart** - A related resource that is included either physically or logically in the described resource.
- **hasVersion** - A related resource that is a version, edition, or adaptation of the described resource.
- **identifier** - An unambiguous reference to the resource within a given context.
- **instructionalMethod** - A process, used to engender knowledge, attitudes and skills, that the described resource is designed to support.
- **isFormatOf** - A related resource that is substantially the same as the described resource, but in another format.
- **isPartOf** - A related resource in which the described resource is physically or logically included.
- **isReferencedBy** - A related resource that references, cites, or otherwise points to the described resource.
- **isReplacedBy** - A related resource that supplants, displaces, or supersedes the described resource.
- **isRequiredBy** - A related resource that requires the described resource to support its function, delivery, or coherence.
- **issued** *(date)* - Date of formal issuance (e.g., publication) of the resource.
- **isVersionOf** - A related resource of which the described resource is a version, edition, or adaptation.
- **mediator** *(array of agents)* - An entity that mediates access to the resource and for whom the resource is intended or useful.
- **modified** *(date)* - Date on which the resource was changed.
- **provenance** - A statement of any changes in ownership and custody of the resource since its creation that are significant for its authenticity, integrity, and interpretation.
- **references** - A related resource that is referenced, cited, or otherwise pointed to by the described resource.
- **relation** - A related resource.
- **replaces** - A related resource that is supplanted, displaced, or superseded by the described resource.
- **requires** - A related resource that is required by the described resource to support its function, delivery, or coherence.
- **rights** -Information about rights held in and over the resource.
- **rightsHolder** *(array of agents)* - A person or organization owning or managing rights over the resource.
- **source** - A related resource from which the described resource is derived.
- **spatial** - Spatial characteristics of the resource.
- **subject** - The topic of the resource.
- **tableOfContents** - A list of subunits of the resource.
- **temporal -** Temporal characteristics of the resource.
- **type** - The nature or genre of the resource.
- **valid** *(date)* - Date of validity of a resource.

Example

http://www.example.com/about/example-dictionary:

```
{
  "release": "PUBLIC",
  "sourceLanguage": "en",
  "targetLanguage": [ "en", "de" ],
```

```
  "genre": [ "gen" ],
  "license": "https://creativecommons.org/licenses/by/4.0/",
  "title": "The Human-Readable Name of this resource",
  "creator": [{
      "name": "Institute of This Resource",
      "email": "contact@institute.com"
  }],
  "publisher": [{
      "name": "Publishing Company"
  }]
}
```

2.3 Get all Lemmas[4]

Objective
Get all of the lemmas contained within this dictionary

Parameters

dictionary	*string* Identifier of the dictionary to list
limit (optional)	*integer >= 1* The maximum number of lemmas to return
offset (optional)	*integer >= 0 (Default: 0)* The offset (index of first lemma) to return

Table 4: Get all Lemmas Parameters

Response

release	*PUBLIC, NONCOMMERCIAL, RESEARCH or PRIVATE* The conditions under which this entry can be included in the ELEXIS Matrix Dictionary
lemma	*string* The lemma of this entry[5]
language	*ISO-639 code* A language code following the ISO 639-1,2,3 standard. If a (two-letter) ISO 639-1 code exists this should be used in preference
id	*string* A unique identifier for the entry
partOfSpeech	*Array of values from "ADJ", "ADP", "ADV", "AUX", "CCONJ", "DET", "INTJ", "NOUN", "NUM", "PART", "PRON", "PROPN", "PUNCT", "SCONJ", "SYM", "VERB"*

[4] Note although in this document we use the term *lemma* we consider this to be the same as 'dictionary headword'

[5] In particular this refers to the string that is marked as a **lemma** in a corresponding TEI Lex-0 document or in the case of OntoLex-Lemon a **written representation** of a **canonical form**.

	or "X" A part of speech tag that the entry has, this must be one of the values from the Universal Dependencies Part-of-Speech Tagset
formats	*Array of values from "tei", "json" or "ontolex"* The formats that this resource is available in. They are as follows • tei: This document is available as TEI Lex-0 from the /tei path • json: This document is available as OntoLex-Lemon in JSON-LD markup from the /json path • ontolex: The document is available as OntoLex-Lemon in Turtle from the /ontolex path

Table 5: Get all Lemmas Response

Example
http://www.example.com/list/example-dictionary?limit=2:

```
[
  {
    "release": "PUBLIC",
    "lemma": "work",
    "language": "en",
    "id": "work-n",
    "partOfSpeech": [ "NOUN" ],
    "formats": [ "tei" ]
  }, {
    "release": "PUBLIC",
    "lemma": "work",
    "language": "en",
    "id": "work-v",
    "partOfSpeech": [ "VERB" ],
    "formats": [ "tei" ]
  }
]
```

2.4 Lemma Lookup

Objective
Given a lemma, find all the entries that are listed under it

Parameters

dictionary	*string* The identifier of the dictionary containing the entries
lemma	*string* The lemma to lookup
language (optional)	*ISO-639 code* A language code following the ISO 639-1,2,3 standard. If a (two-letter) ISO 639-1 code exists this should be used in preference

partOfSpeech (optional)	*One of "ADJ", "ADP", "ADV", "AUX", "CCONJ", "DET", "INTJ", "NOUN", "NUM", "PART", "PRON", "PROPN", "PUNCT", "SCONJ", "SYM", "VERB" or "X"* A part of speech tag that the entry has, this must be one of the values from the Universal Dependencies Part-of-Speech Tagset[6]
limit (optional)	*integer >= 1* The maximum number of entries to return
offset (optional)	*integer >= 0 (Default: 0)* The offset (index of first entry) to return
inflected (optional)	*boolean* If true treat the strings as an inflected form and return all lemmas that may have the string as a form

Table 6: Lemma Lookup Parameters

Response
As for "Get all lemmas"

Example
http://www.example.com/lemma/example-dictionary/work:

```
[
  {
    "release": "PUBLIC",
    "lemma": "work",
    "language": "en",
    "id": "work-n",
    "partOfSpeech": [ "NOUN" ],
    "formats": [ "tei" ]
  }, {
    "release": "PUBLIC",
    "lemma": "work",
    "language": "en",
    "id": "work-v",
    "partOfSpeech": [ "VERB" ],
    "formats": [ "tei" ]
  }
]
```

Example
http://www.example.com/tei/example-dictionary/work-n:

```
<entry xml:id="id">
  <form type="lemma"><orth>work</orth></form>
  <gramGrp>
    <pos norm="NOUN">noun</pos>
  </gramGrp>
  <sense n="1">
    <def>An example TEI Lex-0 Entry</def>
  </sense>
</entry>
```

[6] http://universaldependencies.org/u/pos/

2.5 Entry

Objective
Return the Entry directly as a JSON, OntoLex or TEI document. Services must implement at least one of the actions. Each of these are distinguished by the request path (e.g., /json, /ontolex or /tei) and the services should use standard HTTP responses (e.g., 404) to indicate the unavailability of an entry in a given format, and the HTTP "Content-Type" and "Last-Modified" headers to indicate the status and format of the file.

Parameters

dictionary	*string* Identifier of the dictionary containing the entry
id	*string* The identifier of the entry

Table 7: Entry Parameters

Response
The entry data (see Section 3). The following MIME types should be returned

- JSON: application/json
- OntoLex: text/turtle
- TEI: text/xml

3 Formats for interoperability

3.1 JSON

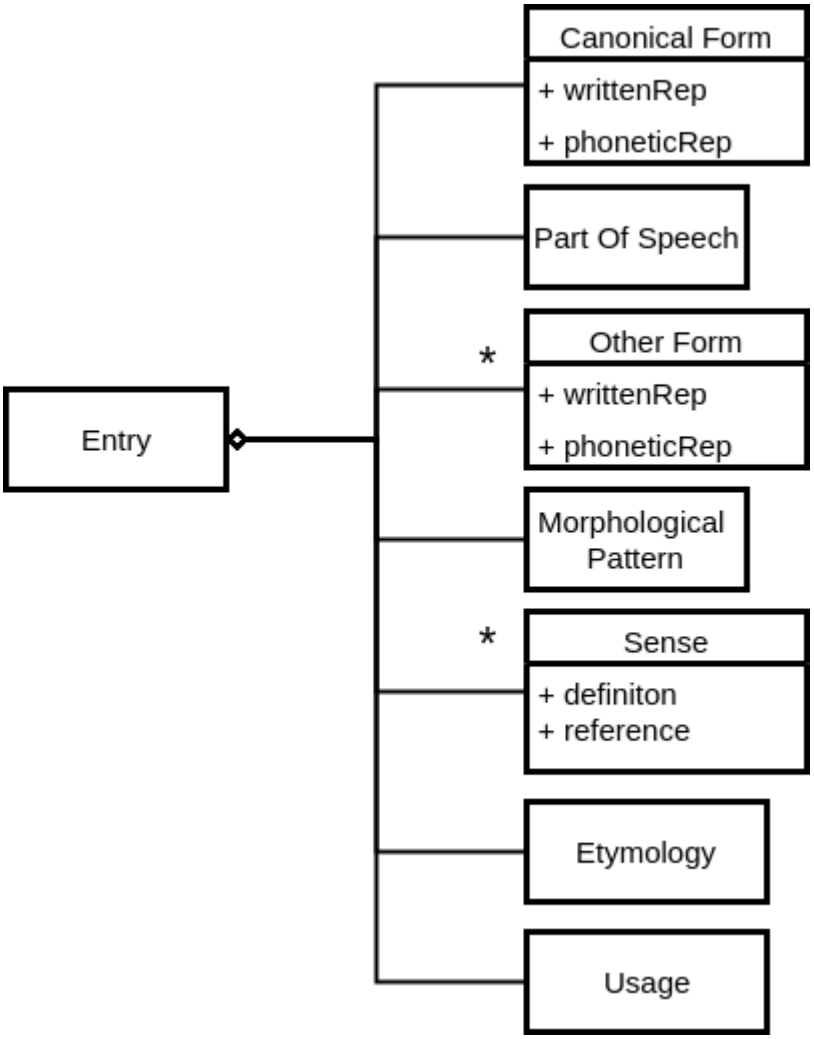

Figure 1: Structure of the JSON model

The JSON format is provided for the convenience of those who do not have their data already in TEI Lex-0 or OntoLex, and who wish to develop an implementation without reference to other standards. This format is a highly reduced version of OntoLex and as such does not capture all the elements that may be present in a dictionary, nor does it preserve the format of the original dictionary. In fact, the JSON document is a version of the OntoLex model using the JSON-LD model. As such the JSON object returned should have the following fields:

Parameters

@context	This should have the fixed value https://elexis-eu.github.io/elexis-rest/context.json
@id	Should be the same as the request ID
@type	One of "LexicalEntry", which can be either "Word", "MultiWordExpression" or "Affix"

canonicalForm	A JSON object with two fields: • **writtenRep:** The lemma goes here • **phoneticRep:** A pronunciation guide (if any)
partOfSpeech	One of the following values, which align with the UD values: • lexinfo:adjective • lexinfo:adposition • lexinfo:adverb • lexinfo:auxiliary • lexinfo:coordinatingConjunction • lexinfo:determiner • lexinfo:interjection • lexinfo:commonNoun (corresponds to UD's noun) • lexinfo:numeral • lexinfo:particle • lexinfo:pronoun • lexinfo:properNoun • lexinfo:punctuation • lexinfo:subordinatingConjunction • lexinfo:symbol • lexinfo:verb • other
otherForm	An array of objects with two fields: • **writtenRep:** The form goes here • **phoneticRep:** A pronunciation guide (if any)
morphologicalPattern	A morphological class if relevant
senses	An array of objects with the following fields: • **definition**: A definition of the sense • **reference:** A URL pointing to an external definition of the entry
etymology	A string giving the etymology of the entry
usage	Notes about the usage of the entry

Table 8: JSON Parameters

Example:

From: http://wordnet-rdf.princeton.edu/lemma/work


```
{
 "@context": "https://elexis-eu.github.io/elexis-rest/context.json",
  "@type": "Word",
  "@id": "work-n",
  "canonicalForm": { "writtenRep": "work" },
  "partOfSpeech": "lexinfo:commonNoun",
  "senses": [{
     "definition": "a product produced or accomplished through the effort
or activity or agency of a person or thing",
      "reference": "http://ili.globalwordnet.org/ili/i61245"
     },{
     "definition": "(physics) a manifestation of energy; the transfer of
energy from one physical system to another expressed as the product of a
```

```
force and the distance through which it moves a body in the
direction of that force;'',
      ''reference'': ''http://ili.globalwordnet.org/ili/i97775''
   }]
}
```

3.2 OntoLex

3.2.1 Overview

The OntoLex-Lemon model was developed by the W3C Ontology-Lexicon Community Group (Cimiano, Philipp, John P. McCrae, and Paul Buitelaar 2016) on the basis on previous models in particular the *lemon* model (McCrae, John, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, et al. 2012) (McCrae, John, Dennis Spohr, and Philipp Cimiano 2011). Ontolex-Lemon provides a general framework for the representation of lexical information relative to ontologies as well as providing for the general modelling of lexical graphs in terms of senses and concepts in an approach that is inspired by the Princeton WordNet model (Fellbaum, 1998). The OntoLex-Lemon model is based on the Resource Description Framework (Lassila, Ora, and Ralph Swick 1999)and is made up of 5 modules with 2 more in development

- **OntoLex Core**: This describes the key elements of the lexicon, e.g., a lexical entry and its forms, a lexical sense and its associated lexical concept and reference to an ontology.
- **Syntax and Semantics:** This module describes how the syntactic frames of an entry can be described and how they can be mapped onto the formal semantics in an ontology.
- **Decomposition:** The decomposition module is concerned with how lexical entries can be decomposed into sub-entries, for example in multi-word expressions.
- **Variation and Translation**: Variation (and specifically translation) represents relations between words and in this model such relations can be across entries, part-of-speech and even whole lexicons. Relations in the model are characterized as purely lexical, purely semantic or lexico-semantic.
- **Linguistic Metadata**: The Linguistic Metadata (LiMe) module allows for general metadata about the lexicon such as the number of entries and senses it contains.
- **Lexicographic (In Development):** This module describes several aspects that are common in print lexicography, including the ordering and grouping of senses, as well as lexico-semantic restrictions, and examples/attestations.
- **Morphology (In Development):** The morphology module aims to describe the inflectional and agglutinative morphology of rules both in terms of their attested form, but also as a generative procedure.

3.2.2 Usage of Ontolex in the interface

In this section we present some examples of the use of the parameters we have for retrieving an entry in the OntoLex-lemon format[7]. We selected as the original dictionary resource *the Algemeen Nederlands Woordenboek*[8] (ANW). The example described below shows a transformation from the ANW entry for the word "wijn"[9] (*wine*) into the OntoLex-lemon format, using the Turtle syntax. We focus here on the parameters listed at the beginning of Section 3.1 of this document:

[7] As specified here: https://www.w3.org/2016/05/ontolex/

[8] http://anw.ivdnt.org/about

[9] See http://anw.ivdnt.org/article/wijn

```
:lex_wijn_182155
  rdf:type ontolex:Word ;
  lexinfo:anw_articleType "\"de\"" ;
  lexinfo:gender lexinfo:masculine ;
  lexinfo:partOfSpeech lexinfo:commonNoun, lexinfo:noun ;
  ontolex:canonicalForm :form_wijn_singular ;
  ontolex:otherForm :form_wijnen_plural ;
  ontolex:sense :sense_wijn1.0, :sense_wijn1.1, :sense_wijn1.2,
                :sense_wijn1.3, :sense_wijn1.4 .
```

In the OntoLex example of a lexical entry above the reader can see the **id** of the entry "lex_wijn_182155" along with the fact that we are dealing with an `ontolex:Word` (a subclass of the OntoLex class `ontolex:LexicalEntry`) as the value of the **type** parameter. We have also the information about the **Part-of-Speech** parameter (`lexinfo:commonNoun`) and (`lexinfo:noun`). The reader should be aware that we do not describe an ambiguity here, as `lexinfo:commonNoun` and `lexinfo:noun` are both instances of the LexInfo class NounPos. In case of ambiguities we would have then 2 entries for the same headword form. The example above also includes a link to the `ontolex:`**`canonicalForm`** and to an `ontolex:`**`otherForm`**, which is the plural form of the word. And finally to a number of `ontolex:`**`sense`** objects. Below we give a single example of such a sense:

```
:sense_wijn1.0
  rdf:type ontolex:LexicalSense ;
  skos:definition "alcoholhoudende drank, verkregen door gisting van het
sap van druiven of van andere vruchten, met een middelmatig alcoholgehalte
van doorgaans ongeveer 12 procent; alcoholhoudende drank van gegist
druivensap"@nl ;
  ontolex:isLexicalizedSenseOf :Concept_325624, :Concept_Stofnaam,
                               :Concept_mass ;
  ontolex:isSenseOf :lex_wijn_182155 ;
  ontolex:reference <https://www.wikidata.org/wiki/Q282> ;
  ontolex:usage lexinfo:massNoun .
.
```

In this example of a sense, we see the implementation of other features that could act as additional parameters. For example, a **definition** is added to the sense description (here in Dutch), and also an external **reference** to an ontology is given (here to a wikidata article). We also give an example of the `ontolex:`**`usage`** element, which here specifies that this sense is valid only in the case we are dealing with the massNoun interpretation of the word, which is possible only when the entry is used in singular. In the next three examples of the OntoLex-lemon code, we see how compound nouns are handled:

```
:lex_wijnfles_182211
  rdf:type ontolex:MultiWordExpression ;
  lexinfo:anw_articleType "\"de\"" ;
  lexinfo:gender lexinfo:feminine, lexinfo:masculine ;
  lexinfo:partOfSpeech lexinfo:commonNoun, lexinfo:noun ;
  rdf:_1 :comp_wijn_1 ;
  rdf:_2 :comp_fles_1 ;
  decomp:constituent :comp_fles_1 ;
  decomp:constituent :comp_wijn_1 ;
  decomp:subterm :lex_wijn_182155, :lex_fles_18089 ;
```

```
  ontolex:sense :sense_wijn1.3 .


:comp_fles_1
  rdf:type decomp:Component ;
  decomp:correspondsTo :lex_fles_18089 .


:comp_wijn_1
  rdf:type decomp:Component ;
  decomp:correspondsTo> :lex_wijn_182155 .
```

In the three snippets of code in the Turtle syntax shown above, the reader can see that the lexical entry is now typed as a "**MultiWordExpression**" (which is also a subclass of `ontolex:`**`LexicalEntry`**). While the gender associated to the entry seems strange (being marked both as a feminine and a masculine name), this reflects a property of the Dutch language as the noun is used in combination with the determiner "de". One can use the Lexinfo value "**commonGender**" instead.

The reader can also see how the decomposition of the word in its two main components is represented in OntoLex-lemon, and how those components refer to corresponding entries. More details on a first version of this conversion of the ANW entries into OntoLex-lemon are given in (Tiberius, Carole, and Thierry Declerck 2017). Here we have shown how the parameters of the format for interoperability of ELEXIS matrix dictionaries are implemented in OntoLex-lemon.

3.2.3 Lexicographic module Example

The OntoLex lexicographic module aims to close the gap between the computational use cases originally envisioned by the OntoLex Community Group and the kind of lexicographic data handled in projects such as ELEXIS. One of the principal differences that has been observed is that OntoLex has a strict and relatively restrictive definition of a lexical entry as having a single lemma and being of a single part-of-speech class. In the Lexicography module this may be handled by super-entries which give a structured and ordered grouping of an entry and its senses, e.g.,

```
:lead-1 a lexicog:SuperEntry ;
  rdf:_1 [ lexicog:describes :lead-n-1 ] ; # As in ''a dog lead''
  rdf:_2 [ lexicog:describes :lead-v-1 ] . # As in ''they lead''


:lead-2 a lexicog:SuperEntry ;
  rdf:_1 [ lexicog:describes :lead-n-2 ] ; # The metal
  rdf:_2 [ lexicog:describes :leaden-a-1 ] . # A derived adjective
```

3.3 TEI Lex-0

3.3.1 Overview

TEI Lex-0 provides a baseline encoding and a target format in order to facilitate the interoperability of heterogeneously encoded lexical resources. This is important both in the context of building lexical infrastructures as such (Ermolaev, N. and Toma Tasovac 2012) 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 by institutions that already have established workflows based on their own flavours of TEI. Instead TEI

Lex-0 should be primarily seen as a format that existing TEI dictionaries can be univocally transformed into in order to be queried, visualised, 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 TEI Lex-0 aims to stay as aligned as possible with the subset of TEI which comprises the TEI serialisation of the updated version of LMF (Lexical Markup Framework) standard, cf. (Romary, 2015). Preliminary work on the establishment of TEI Lex-0 started in the Working Group "Retrodigitised Dictionaries" as part of the COST Action European Network of e-Lexicography (ENeL). Upon the completion of the COST Action in 2017, the work on TEI Lex-0 was taken up by the DARIAH Working Group "Lexical Resources". Currently, the work on TEI Lex-0 is conducted by the DARIAH WG "Lexical Resources" and falls within the ELEXIS project.

3.3.2 Usage in the interface

In this section, we present some examples of the use of the parameters we have for retrieving lexical information from a resource encoded in TEI Lex-0[10] .

The following example is taken from a bilingual dictionary and illustrates the entry for the French verb “horrifier” *(horrify)* in TEI Lex-0.

```
<entry xml:lang="fr" xml:id="horrifier">
 <form type="lemma">
  <orth>horrifier</orth>
 </form>
 <gramGrp>
  <pos norm="VERB">v</pos>
 </gramGrp>
 <sense>
  <cit type="translationEquivalent"
   xml:lang="en">
   <quote>horrify</quote>
  </cit>
  <cit type="example">
   <quote>elle était horrifiée par la dépense</quote>
   <cit type="translation" xml:lang="en">
    <quote>she was horrified at the expense</quote>
   </cit>
  </cit>
 </sense>
</entry>
```

The entry for “horrifier” is enclosed in an <entry> tag, which in the context of TEI Lex-0, is used to encode the basic element of the dictionary microstructure; grouping all the information related to a particular linguistic entity, including further entries related to it (e.g. homographs or compound phrases). The <form> tag on the next line groups all the information on the written and spoken forms of one headword. The above entry is of type lemma. The <gramGrp> (grammatical information group) tag groups morpho-syntactic information about a lexical item. In the context of ELEXIS, a @norm attribute is required to specify a normalised (UD) part of speech value for the entry (see Introduction).

[10] As specified here: https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

Within the <sense> tag, all information relating to one word sense in a dictionary entry is grouped together, for example definitions, examples, and translation equivalents. The example entry for "horrifier" contains a translation in English (**<cit type="translationEquivalent" xml:lang="en">)** and an example (**<cit type="example">)** which also has a translation in English. Note that the translations have a language attribute, identifying the language of the translation.

In the next example, we show the same original ANW entry for the Dutch word "wijn"[11] as in the case of OntoLex-lemon format, which is converted to TEI Lex-0 using the conversion tool developed in WP1, T1.4[12]. The original XML entry uses custom XML schema with elements in bold identified as parameters used in the REST interface and listed at the beginning of section 3.1:

```
<?xml version="1.0" encoding="UTF-8" ?>
<Woordenboek>
<artikel>
        <Lemma>
                <Lemmavorm>wijn</Lemmavorm>
                <Lemmatype>woord</Lemmatype>
        </Lemma>
        <Woordsoort>
                <Type>substantief</Type>
                <Substantief>
                        <Naamtype>soortnaam</Naamtype>
                        <GeslachtBody>
                                <Geslacht>mannelijk</Geslacht>
                        </GeslachtBody>
                        <Lidwoord>
                                <AardLidwoord>de</AardLidwoord>
                        </Lidwoord>
                </Substantief>
        </Woordsoort>
        ...
                <BetekenisEnGebruik>
                <Kernbetekenis pid="325624">
                        <betekenisInfo>
                                <Betekenisnummer>1.0</Betekenisnummer>
                                <Lemma>
                                        <Lemmavorm>wijn</Lemmavorm>
                                        <Lemmatype>woord</Lemmatype>
                                </Lemma>
                                <Woordsoort>
                                        <Type>substantief</Type>
                                        <Substantief>
                                                <Naamtype>soortnaam</Naamtype>
                                                <GeslachtBody>
                                                        <Geslacht>mannelijk</Geslacht>
                                                </GeslachtBody>
                                                <Lidwoord>
                                                        <AardLidwoord>de</AardLidwoord>
                                                </Lidwoord>
                                                <Getal>geen meervoud</Getal>
                                        <Betekenisklasse>stofnaam</Betekenisklasse>
```

Lemma

Part of speech

Sense

[11] See http://anw.inl.nl/article/wijn

[12] http://capybara.ijs.si/janez/elexis/transformDemo.html

```
                        </Substantief>
                    </Woordsoort>
                    ...
<definitieBody>
      <Definitie>alcoholhoudende drank, verkregen door gisting van het sap van druiven of van andere vruchten, met een middelmatig alcoholgehalte van doorgaans ongeveer 12 procent; alcoholhoudende drank van gegist druivensap</Definitie>
</definitieBody>
...
<Voorbeeld pid="328398">
      <Tekst>Vul het vat aan [...] maar zeker niet meer met een suikeroplossing daar de wijn anders weer aan het gisten gaat of te zoet zal worden.</Tekst>
      <BronID>8940</BronID>
            <URL>http://home.hetnet.nl/~grvwijk/index.html</URL>
</Voorbeeld>
```

Definition

Example

The entire entry is converted to TEI Lex-0 with the parameters used in REST interface in the expected TEI Lex-0 elements:

```
<entry>
 <form>
  <orth></orth>
 </form>
 <gramGrp>
  <pos></pos>
 </gramGrp>
 <sense>
  <def></def>
  <cit type="example">
   <quote></quote>
  </cit>
 </sense>
</entry>.
```

Lemma

Part of speech

Sense

Definition

Example

The rest of the the entry is converted to general <seg> and <dictScrap> elements, to be able to retain complete information from the original dictionary, and use parameters from TEI Lex-0 in REST interface. The corresponding parts of the converted “wijn” dictionary entry in TEI LEX-0 format are:

```
 <text>
    <body>
      <entry xml:id="ANW-1" xml:lang="NL">
        <form>
          <orth>wijn</orth>
        </form>
        <dictScrap>
            <seg>woord</seg>
      </dictScrap>
        <gramGrp>
```

Lemma

```
          <pos norm="NOUN">substantief</gram>
        </gramGrp>
        <dictScrap>
            <seg>
…
</dictScrap>
    <dictScrap>
      <seg>
            <seg>6970</seg>
      </seg>
</dictScrap>

     <sense                      xml:id="S1">
            <seg>
                  <seg>1.0</seg>
                  </seg><form><orth>wijn</orth></form><seg><seg>
                        <seg>woord</seg>
                  </seg>
…
<seg>
      <seg>ongeleed</seg>
</seg>
</seg>
<def>
alcoholhoudende drank, verkregen door gisting van het sap van druiven of
van andere vruchten, met een middelmatig alcoholgehalte van doorgaans
ongeveer 12 procent; alcoholhoudende drank van gegist druivensap</def>
<seg><seg>
</seg>
…
<seg>
      <cit type="example">
      <quote>Vul het vat aan [...] maar zeker niet meer met een
      suikeroplossing daar de wijn anders weer aan het gisten gaat of te
      zoet zal worden.</quote>
      </cit>           <seg>8940</seg>
      <seg>http://home.hetnet.nl/~grvwijk/index.html</seg>
</seg>
```

Part of speech

Sense

Definition

Example

4 Role of REST interface in project architecture

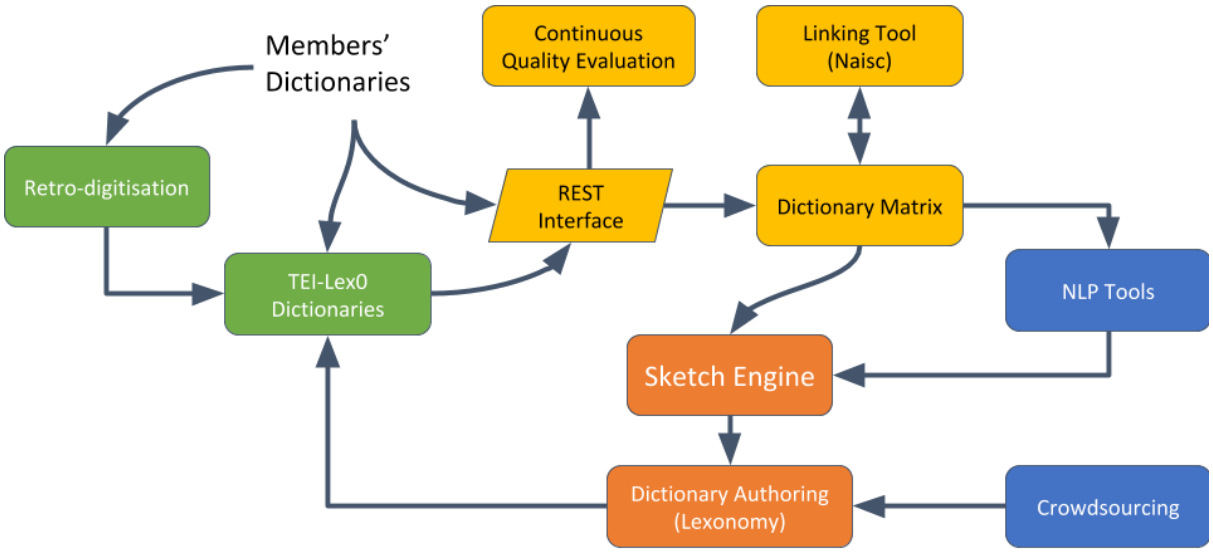

Figure 2: Overview of the ELEXIS architecture

The ELEXIS interface is given in the above figure. This deliverable defines the REST interface, and it will be used by the Dictionary Matrix (D8.1) as well as the validation and benchmarking services (D2.5). As input, it is envisioned that the REST interface can be used in various ways as shown in Figure 3 below:

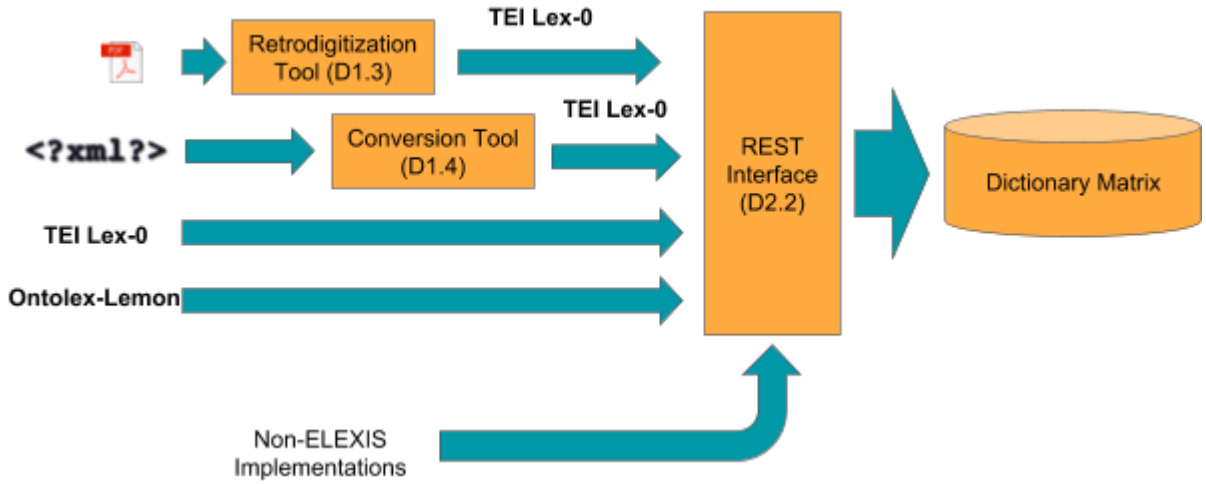

Figure 3: Access to ELEXIS interface through REST interface

- For dictionaries that are not already in a digital format the retrodigitization tools to be developed in D1.3 will be used. This will apply OCR to the text and then process this text by adding XML markup in the form of TEI Lex-0.
- For dictionaries that are already available in a digital form, but not one that is supported directly by the project, the conversion tool developed in D1.4 will be used to convert these resources to TEI Lex-0.

- If the dictionary is already in TEI Lex-0 or has been converted to TEI Lex-0 by one of the two methods described above, then it can be consumed directly by the *interoperable interface* which will be developed in the next year and reported in D2.2.
- If the dictionary is in another format supported by the project, in particular OntoLex-Lemon, then this can also be supported directly in the REST interface
- Finally, it will be possible for other institutes to participate in the interface by implementing the interface described in this document.

5 Conclusion

In this deliverable we have described the design of the REST interface to be used to access lexicographic resources within the ELEXIS project with a description of the different formats in which dictionary data will be made available.

References

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

Ermolaev, N. and Toma Tasovac. (2012). Building a Lexicographic Infrastructure for Serbian Digital Libraries. *Libraries in the Digital Age (LIDA) Proceedings.* Retrieved from http://ozk.unizd.hr/proceedings/index.php/lida/article/view/55

Fellbaum, C. (1998). *WordNet.* John Wiley & Sons, Inc.

Lassila, Ora, and Ralph Swick. (1999). *Resource Description Framework (RDF) Model and Syntax Specification.* W3C. Retrieved from http://www.w3.org/TR/REC-rdf-syntax/

McCrae, John, Dennis Spohr, and Philipp Cimiano. (2011). Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. *Proc. of the 8th Extended Semantic Web Conference*, (pp. 245–49).

McCrae, John, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, et al. (2012). Interchanging Lexical Resources on the Semantic Web. *Language Resources and Evaluation, 46*(6), 701–9.

Romary, L. (2015). TEI and LMF crosswalks. *JLCL*(30).

Tiberius, Carole, and Thierry Declerck. (2017). A lemon Model for the ANW Dictionary. *Proceedings of the eLex 2017 conference*, (pp. 237-251).