The ELEXIS Interface for Interoperable Lexical Resources

John P. McCrae[1], Carole Tiberius[2], Anas Fahad Khan[3], Ilan Kernerman[4], Thierry Declerck[5,7], Simon Krek[6], Monica Monachini[3] and Sina Ahmadi[1]

[1] Data Science Institute, National University of Ireland Galway
[2] Instituut voor de Nederlandse Taal
[3] CNR- Istituto di Linguistica Computazionale «A. Zampolli»
[4] K Dictionaries
[5] Austrian Centre for Digital Humanities, Austrian Academy of Sciences
[6] Jožef Stefan Institute/University of Ljubljana
[7] DFKI GmbH, Multilinguality and Language Technology Lab

Abstract

ELEXIS is a project that aims to create a European network of lexical resources, and one of the key challenges for this is the development of an interoperable interface for different lexical resources so that further tools may improve the data. This paper describes this interface and in particular describes the five methods of entrance into the infrastructure, through retrodigitization, by conversion to TEI-Lex0, by the TEI-Lex0 format, by the OntoLex format or through the REST interface described in this paper. The interface has the role of allowing dictionaries to be ingested into the ELEXIS system, so that they can be linked to each other, used by NLP tools and made available through tools to Sketch Engine and Lexonomy. Most importantly, these dictionaries will all be linked to each other through the Dictionary Matrix, a collection of linked dictionaries that will be created by the project. There are five principal ways that a dictionary maybe entered into the Matrix Dictionary: either through retrodigitization; by conversion to TEI Lex-0 by means of the forthcoming ELEXIS conversion tool; by directly providing TEI Lex-0 data; by providing data in a compatible format (including OntoLex); or by implementing the REST interface described in this paper.

Keywords: lexicography; linked data; infrastructure; ELEXIS; REST; RDF; TEI; JSON

1. Introduction

ELEXIS is a Horizon 2020 infrastructure project dedicated to lexicography. This new infrastructure will (1) enable efficient access to high quality lexicographic data, and (2) bridge the gap between more advanced and less-resourced scholarly communities working on lexicographic resources. In most European countries, elaborate efforts are put into the development of lexicographic resources describing the language(s) of the community. 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. Consequently, the lexicographic landscape in Europe is rather heterogeneous. Firstly, it is characterized by stand-alone lexicographic resources, which are typically encoded in incompatible data structures due to the isolation of efforts, prohibiting reuse of this valuable data in other fields. Secondly, there is a significant variation in the level of expertise and resources available to lexicographers across Europe. Within ELEXIS, strategies, tools and standards are under development for extracting, structuring and linking lexicographic resources to unlock their full potential for Linked Open Data, NLP and the Semantic Web, as well as in the context of digital humanities. In a virtuous cycle of cross-disciplinary exchange of knowledge and data, a higher level of language description and text processing will be achieved. By harmonizing and integrating lexicographic data into the Linked Open Data cloud, ELEXIS will make this data available to AI and NLP for semantic processing of unstructured data, considerably enhancing applications such as machine translation, machine reading and intelligent digital assistance thanks to the ability to scale to wide coverage in multiple languages. This, in turn, will enable the development of improved tools for the production of structured proto-lexicographic data in an automated process, using machine learning, data mining and information extraction techniques, where the extracted data can be used as a starting point for further processing either in the traditional lexicographic process or through crowdsourcing platforms.

In the context of the ELEXIS project it has been necessary to develop an interface that allows all different kinds of dictionary data to be included in the infrastructure. As such, the ELEXIS interface is a set of common protocols which take the form of a REST API and which allows dictionaries and lexicographic resources to be accessed through a common interface and in a uniform manner. The REST interface will allow users who wish to query a given endpoint to get back the metadata of the different lexicographic resources accessible 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 (at least one of which must be implemented), these comprise the formats for interoperability of the ELEXIS project. The data model ensures that key elements of the dictionary data are referred to in a uniform manner, and as a particular example of this we require that all the part of speech values are mapped to the Universal Dependencies (UD) part of speech tagset (Petrov et al., 2012; Nivre et al., 2016).

In this paper, we describe this interface and its usage as a tool for getting dictionary data into the ELEXIS infrastructure, so that they can be linked to each other, used by NLP tools and made available through tools to Sketch Engine (Kilgarriff et al., 2014) and Lexonomy (Měchura, 2017). Most importantly these dictionaries will all be linked to each other as part of the **Dictionary Matrix**, a collection of linked dictionaries that will be created by the project. There are five principal ways that a dictionary may be entered into the Matrix Dictionary: either through retrodigitization; by conversion to TEI Lex-0 by means of the forthcoming ELEXIS conversion tool; by directly providing TEI Lex-0 data; by providing data in a compatible format (including

OntoLex, Cimiano et al., 2014); or by implementing the REST interface[1] described in this paper.

2. The REST interface

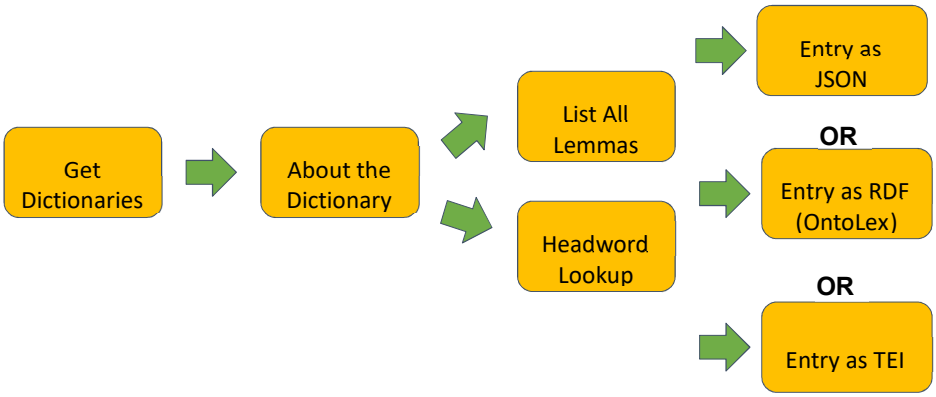

Figure 1: The access protocol for the REST interface

The goal of the REST interface (depicted in Figure 1) is to provide access to the dictionary for the Dictionary Matrix. To this extent it provides a number of basic tools to provide indexing and search over the dictionary interface. As the interface is intended to be implemented with very little effort for the contributors to the ELEXIS network there is a focus on making minimal and simple queries, as such the interface only documents very basic usage. More sophisticated usage can be provided by either custom extensions or by downloading all the data and querying it offline. The first query is to show the set of dictionaries that are available at a particular endpoint, which is done with the following call:

Method Name:	/dictionaries
Parameters:	*None*
Returns:	A list of dictionary IDs
Example Request:	http://www.example.com/dictionaries

Example Response:	{
	"dictionaries": [
	"dict1",
	"dict2" · · · · ·
	]
	}

[1] http://elexis-eu.github.io/elexis-rest/elexis.html

The next call in the interface is normally to retrieve the metadata about this dictionary that is necessary to show the dictionary in the dictionary interface. We require a small number of custom parameters that are especially helpful to the ELEXIS interface, including information about the release level, which is whether the data is public, limited to signed-in academic users or private, as well as information about the genre of the dictionary and languages. For genres, we use the previous categorization at the EU dictionary portal, which is as follows:

- **General dictionaries** are dictionaries that document contemporary vocabulary and are intended for everyday reference by native and fluent speakers.
- **Learners' dictionaries** are intended for people who are learning the language as a second language.
- **Etymological dictionaries** are dictionaries that explain the origins of words.
- **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.
- **Historical dictionaries** are dictionaries that document previous historical states of the language.
- **Spelling dictionaries** are dictionaries which codify the correct spelling and other aspects of the orthography of words.
- **Terminological dictionaries** describe the vocabulary of specialized domains such as biology, mathematics or economics.

For languages, we consider that the dictionary has a single language for its headwords, but that the definitions may be in different languages. As such, a bidirectional, bilingual dictionary is split into two 'dictionaries' based on the direction in which we are querying. In addition, there are over 40 other metadata properties, mostly derived from Dublin Core, which may be included in the metadata, although these have no functional role and are merely reproduced for the user at the dictionary portal.

Method Name:	/about
Parameters:	The dictionary ID
Returns:	An object describing the dictionary
Example Request:	http://www.example.com/about/example-dictionary

Example Response:	{
	"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" }]
	}

The next issue is obtaining individual entries from the dictionary, in which two principle modes are planned: firstly, retrieval of all entries in the dictionary in order and, secondly, search by lemma. Entries in the dictionary are defined by their lemma, their part-of-speech values and the formats that they are available in. For part-of-speech we use the universal dependencies categories as this provides a broad but good categorization of part-of-speech values, and these values have already been documented and tested in a wide range of languages[2]. As such, we believe that these categories are a good general purpose categorization of part-of-speech values. The full list is given below.

adjective	interjection	punctuation
adposition	(common) noun	subordinating conjunction
adverb	numeral	symbol
auxiliary	particle	verb
coordinating conjunction	pronoun	other
determiner	proper noun	

The querying of entries in the order they appear in the dictionary is limited only by the offset and limit that states how many entries into the dictionary to read and how many to return:

[2] See https://universaldependencies.org/u/pos/ for more details.

Method Name:	/list/*dictionary*
Parameters:	A limit and an offset
Returns:	A list of lexical entry descriptions
Example Request:	http://www.example.com/list/example-dictionary?limit=2

Example Response:	[
	{
	"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"]
	}
	]

The lemma lookup requires specifying a lemma, as well as an offset and limit and a flag to say if the query should also look for inflected forms that match this lemma.

Method Name:	/lemma/*dictionary*/*query*
Parameters:	A limit and an offset and flag to state if the entry should be inflected
Returns:	A list of lexical entry descriptions
Example Request:	http://www.example.com/lemma/example-dictionary/works?inflected
Example Response:	*As previous*

The final part of the API is to return the relevant documents in one of the interoperability formats. The interface can be used to access each of the three formats with a URL such as below. It is up to the implementer to decide which of the three (or all three) to implement.

- http://www.example.com/json/dictionary/lemma
- http://www.example.com/ontolex/dictionary/lemma
- http://www.example.com/tei/dictionary/lemma

It should be noted that this interface does not see any modification of the content of the dictionaries, and by participating in the infrastructure content providers allow the ELEXIS infrastructure to provide links and to make public the list of lemmas through the dictionary portal.

2.1 Design considerations

In general, the interface is designed to be lightweight and easy to implement so that many different dictionary providers can contribute their data to the ELEXIS infrastructure. The interface provides only very simple query methods that should be easy to implement with high performance in the database of the third party who is already responsible for ingesting the data into their infrastructure. It also follows that implementations will need to provide their own mapping of their data into one of the formats provided in the next section and in particular find a mechanism for mapping their part-of-speech categories to the universal dependency list. More sophisticated alignment of properties of lexical entries, e.g., domain or region labels, grammatical information, is not covered from this interface as there is little demand and these properties are generally not well-aligned across resources. While the categories presented in universal dependencies are very broad, they are used primarily for indexing and the entries in the formats below can provide very specific part-of-speech categories to be shown to the user.

3. Formats for interoperability

3.1 JSON

The JSON format is provided for the convenience of those who do not have their data already in TEI Lex-0 or OntoLex, and 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. The JSON object returned should have the following fields:

@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" or more specifically "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 Universal Dependency values
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


```
{
 "@context": "https://elexis-eu.github.io/elexis-rest/context.json",
  "@type": "Word",
  "@id": "work-n",
  "canonicalForm": { "writtenRep": "work" },
  "partOfSpeech":
  "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" }]
}}
```


Figure 2: Code example based on http://wordnet-rdf.princeton.edu/lemma/work. NB "commonNoun" is used in the JSON schema for the UD class '(common) noun'.

3.2 OntoLex

The OntoLex-Lemon model was developed by the OntoLex Community Group (Cimiano et al., 2016, see also https://www.w3.org/2016/05/ontolex/ for the Final Community Group Report) based on previous models, in particular the *lemon* model (McCrae et al., 2012; McCrae et al., 2011). This model 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 a model that is inspired by the Princeton WordNet model (Fellbaum, 1998). The OntoLex-Lemon model is based on the Resource Description Framework (Lassila & Swick, 1999), and is divided into five modules, with two more in development

- OntoLex Core: This describes the key elements of the lexicon, e.g., the lexical entry and its forms, the lexical sense and its associated lexical concept and the reference to the 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 the 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.
- Morphology (in development): The morphology module aims to describe the inflectional and agglutinating morphology of rules both in terms of their attested form, but also as a productive phenomenon.

3.2.1 Usage 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 (as specified here: https://www.w3.org/2016/05/ontolex/).

We selected as the original dictionary resource the Algemeen Nederlands Woordenboek (ANW, http://anw.inl.nl/about). The example depicted below shows the

transformation from the ANW entry for the word “wijn” (wine) (see http://anw.inl.nl/article/wijn; Tiberius and Declerck, 2017) into the OntoLex-lemon format, using the Turtle syntax. We focus here on the parameters listed at the beginning of subsection 3.1:

```
: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 .
```

Figure 3: An example of the OntoLex modelling of the ‘wijn’ entry from the AWN dictionary.

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
```

Figure 4: The use of the OntoLex Lexicography module in the interface.

3.3 TEI Lex-0

TEI Lex-0 comprises a subset of the Text Encoding Initiative schema[3] (TEI) developed with the express aim of providing a baseline encoding and target format to better facilitate the interoperability of heterogeneously encoded lexical resources. As such TEI Lex-0 situates itself both within the context of the creation lexical infrastructures such as Ermolaev and Tasovac (2012), as well as in the development of generic TEI-aware tools, including dictionary editing software. Note that although TEI Lex-0 is a subset of TEI it should be not thought of as a replacement of the Dictionary Chapter in the

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

TEI Guidelines[4] and neither is it intended as a format that must be used for editing or managing individual resources – particularly not resources belonging to projects and/or by institutions that already have established workflows based on their own flavours of TEI. Instead it is intended to serve as a format that existing TEI dictionaries can be univocally transformed to in order to be queried, visualized, or mined in a uniform way. At the same time TEI Lex-0 has also been developed with a number of other core use cases in mind, for instance as a best-practice example for didactic purposes, and as a set of best-practice guidelines for new TEI-based projects[5].

Preliminary work for the establishment of TEI Lex-0 started in the Working Group "Retrodigitized 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. According to the Github repository in which the (currently provisional) TEI Lex-0 guidelines are hosted[6], the current status of the schema is, at the time of writing, as a work in progress. However, even though TEI Lex-0 is not currently production-ready, the core elements of the model are said to be in place. It is therefore possible to describe some of the most important features of TEI Lex-0, those that distinguish it from the TEI dictionary chapter. These include the following (a fuller description can be found at the Github repository for TEI LEX-0[7]):

- **The <entry> element:** TEI Lex-0 simplifies and unifies the encoding of dictionary entries by dispensing with the TEI elements <entryFree>, <superEntry>, and <re>. In TEI, the first of these elements is used to encode a single unstructured entry, the second a sequence of entries which are grouped together, and to embed a related lexical entry within another one. Instead in TEI Lex-0 the TEI element <entry> is used (with appropriate adjustments to its content model) in all of these cases as well as for single structured entries (this latter being its usage in the current TEI guidelines), with a recommendation to make use of the type attribute of <entry> to specify the type of entry being encoded.

- **Sense information**: TEI Lex-0 takes a much stricter approach to grouping sense-related information together than the current TEI guidelines. This affects the kinds of elements that can be children of the <entry> element, and in particular <def> which can appear under <sense> and <cit> which can only

[4] https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

[5] To this end TEI Lex-0 aims to stay as aligned as possible with the subset of TEI which comprises the TEI serialization of the updated version of LMF (Lexical Markup Framework) standard (cf. Romary, 2015)

[6] https://github.com/DARIAH-ERIC/lexicalresources/tree/master/Schemas/TEILex0, accessed 6-6-2019

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

appear under <sense> or <dictScrap>.

- The element <hom> is deprecated in TEI Lex-0.

3.3.1 Use of TEI Lex-0 in the interface

Within the context of the ELEXIS project TEI Lex-0 is used both as a target format, to which already existing TEI-encoded dictionaries can be converted, as well as a baseline format into which retrodigitized paper dictionaries and digital native dictionaries in other non-TEI formats will be encoded. This will ensure a sufficient level of homogeneity (both semantic and structural) amongst the resources which have been ingested within the ELEXIS platform (something which it would have been hard to guarantee with TEI), while maintaining compatibility with one of the leading standards for text encoding within the digital humanities, and one which is also becoming increasingly popular for encoding lexical resources.

Below we present some examples of the use of the parameters we have for retrieving lexical information from a resource encoded in TEI Lex-0. 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 ud:norm="VERB">v</pos>
 </gramGrp>
 <sense>
  <cit
   type="translationEquivale
   nt" 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>
```

Figure 5: The entry for the French word 'horrifer' represented in TEI-Lex0

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 the lemma type. 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 normalized (UD) part of speech value for the entry (see introduction). 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.

4. Interoperability in the project architecture

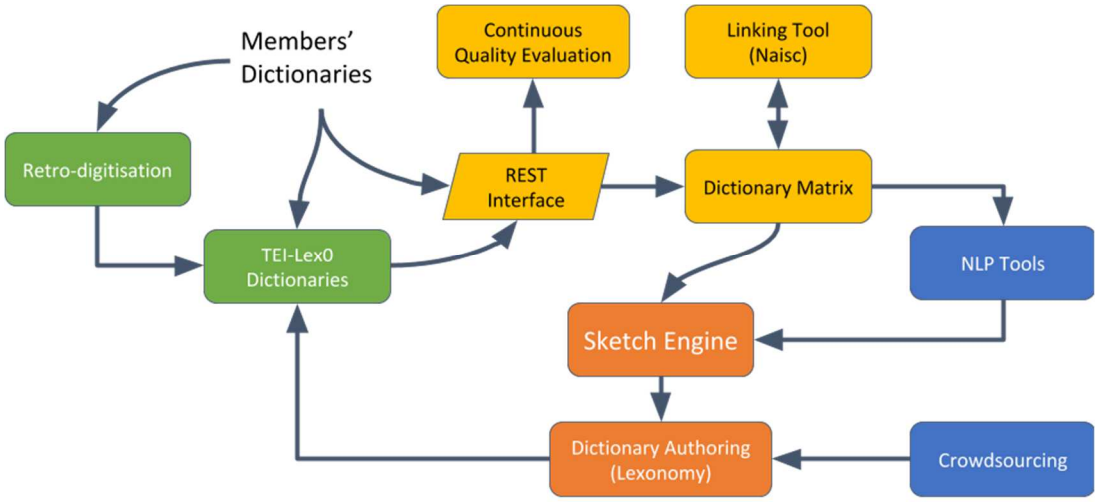

Figure 6: The tools of the ELEXIS infrastructure as an instantiation of the virtuous cycle of eLexicography

The ELEXIS architecture is shown in Figure 6, showing how the REST interface defined above plays an important role in the cycle as the primary interface point. In Figure 7, we show the various ways in which data can enter the infrastructure:

1. From a PDF source or similar OCR is applied and then a semi-automatic tool will be used to identify the structure of the dictionary and output as TEI-Lex0,

2. An existing (non-TEI) XML will be mapped to TEI-Lex0 by identifying the elements that conform to the data model of ELEXIS,

3. TEI-Lex0 documents can be taken directly,

4. Similarly, OntoLex-Lemon can be processed without any modification,

5. Other third-parties may also maintain complete control of their data by implementing the interface above on their own.

Once the data has been provided to the linking infrastructure (yellow in Figure 6), then it will be further processed for NLP applications (blue in Figure 6) and provided to the lexicographic editing interface (orange in Figure 6), which consists of the corpus management tool, Sketch Engine, and the Lexonomy tool for managing and editing lexicographic data, leading to new dictionaries (green in Figure 6).

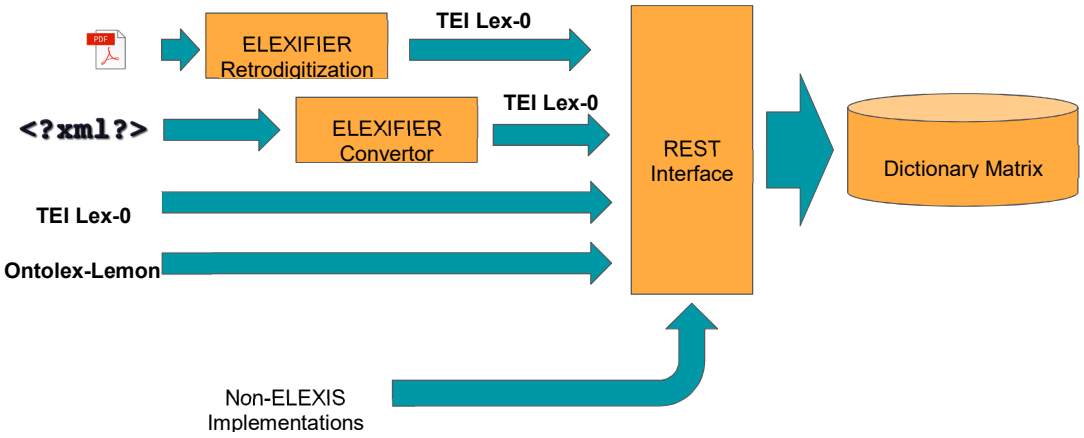

Figure7: Access routes to the ELEXIS architecture depicting the ways data may come into the Dictionary Matrix

4.1 Linking in the ELEXIS infrastructure

There is a plethora of monolingual and multi-lingual resources with a broad range of usage, such as historical dictionaries and terminological resources, available for most European languages. In order to enhance interoperability across resources and languages, ELEXIS provides services for linking resources semi-automatically across languages at various matching levels such as headword, sense and lexeme. Aligned lexical resources, such as Yago (Suchanek et al., 2007), BabelNet (Navigli & Ponzetto, 2012a) and ConceptNet (Speer et al., 2017), have shown to improve word, knowledge and domain coverage and increase multilingualism. In addition, they can improve the performance of NLP tasks such as word sense disambiguation (Navigli & Ponzetto, 2012b), semantic role tagging (Xue & Palmer, 2004) and semantic relations extraction (Swier & Stevenson, 2005).

Lexical data alignment is a challenging task, as lexical information is presented in different structures and dissimilar levels of granularity (Ahmadi et al., 2019). To this end, we are aiming to align lexicographic resources by leveraging ontological properties

and semantic similarity methods. With the current advances in neural networks and resources of significant size available in ELEXIS, we are also interested in applying statistical methods for this task.

4.2 Access to ELEXIS Interface through REST Interface

The retrodigitization tools to be developed in the ELEXIS project will be used for dictionaries that are not already in a digital format. This will apply OCR to the text and then process this text by adding XML markup in the form of TEI-Lex0. 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 the ELEXIS project will be used to convert these resources to TEI-Lex0. If the dictionary is already in TEI-Lex0 or has been converted to TEI-Lex0 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. The implementation in this case is up-to the institute but it must conform to the specification of this document.

4.3 Using legacy and retrodigitized formats (ELEXIFIER)

The ELEXIFIER tool can take dictionaries in two distinct formats as input: (1) XML file with a custom structure/schema and (2) PDF or similar formats originating from word processors (e.g. MS Word). In the custom XML scenario XPath formalisms are used for conversion of the original dictionary to the TEI Lex0-compliant format. In the PDF scenario a more complex process is needed, similar to the one described in Romary and Lopez (2015). In the first step, text and other formatting features (font style, size, colour, etc.) are extracted from the dictionary in PDF form. In the next step, users are asked to manually annotate part of the dictionary in the Lexonomy online dictionary editing tool, according to the ELEXIS data model compatible with TEI-Lex0 standard. In the last step, the annotated text is used as the training material for machine learning algorithms that produce the entire dictionary converted to TEI-Lex0 format. The converted dictionaries can be edited further in the Lexonomy editor.

4.4 Reference implementation for TEI and OntoLex

Figure 8: A screenshot of the reference implementation of the REST interface.

A reference implementation is available for the interface at https://github.com/elexis-eu/dictionary-service, which allows a server to be set up based on either a JSON, OntoLex or TEI document. This interface is implemented in the Rust programming language and as such is available for a wide range of platforms and provides high performance. For JSON files these are directly loaded, however for the TEI and OntoLex it may be necessary to provide some configuration, in particular the mapping of the values used for part-of-speech in the dictionary with the Universal Dependencies categories. It is recommended that those who contribute to the process refer to the existing documentation available from the Universal Dependencies about how to map their categories.

5. Conclusion

eDictionaries are typically in very different stages of digitization, from those where the only digitization is that they have been scanned up to those that have been carefully marked-up with standards such as TEI-Lex0 or 'linked-data native' (Gracia et al., 2017) in OntoLex-Lemon formats. As such there needs to be a highly flexible interface for integrating lexical resources into an ambitious project such as ELEXIS. We have shown a REST interface that will integrate with the retrodigitization and conversion tools in this project to provide multiple ways of entrance into the infrastructure, which ensures that this infrastructure will be open to a wide range of lexicographers.

6. Acknowledgements

All authors are supported by the EU H2020 programme under grant agreements 731015

(ELEXIS - European Lexical Infrastructure). John McCrae is also supported by a research grant from Science Foundation Ireland, co-funded by the European Regional Development Fund, for the Insight Centre under Grant Number SFI/12/RC/2289.

7. References

Ahmadi, S., Arcan, M. & McCrae, J. (2019). Lexical sense alignment using weighted bipartite b-matching. *2nd Conference on Language, Data and Knowledge (LDK 2019)*, p. 5.

Bowers, J., Herold, A. & Romary, L. (2018). TEI-Lex0 Etym–towards terse recommendations for the encoding of etymological information. *JADH 2018*, p. 243.

Cimiano, P., McCrae, J. P. & Buitelaar, P. (2014). Lexicon Model for Ontologies: Community Report.

Ermolaev, N. & Tasovac, T. (2012). Building a lexicographic infrastructure for serbian digital libraries. *Libraries in the Digital Age (LIDA) Proceedings*, 12.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M. & Soria, C. (2006). Lexical markup framework (LMF). In *International Conference on Language Resources and Evaluation-LREC 2006*. p. 5.

Gracia, J., Kernerman, I. & Bosque-Gil, J. (2017). Toward linked data-native dictionaries. In I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubíček & V. Baisa (eds.) *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference, 19-21 September 2017, Leiden, Netherlands*. Brno: Lexical Computing Ltd.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlỳ, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7–36.

McCrae, J., de Cea, G. A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6), pp. 701–709.

McCrae, J. P. & Buitelaar, P. (2018). Linking Datasets Using Semantic Textual Similarity. *Cybernetics and Information Technologies*, 18(1), pp. 109–123.

Měchura, M. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubíček & V. Baisa (eds.) *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference, 19-21 September 2017, Leiden, Netherlands*. Brno: Lexical Computing Ltd.

Navigli, R. & Ponzetto, S. P. (2012a). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, pp. 217–250.

Navigli, R. & Ponzetto, S. P. (2012b). Joining forces pays off: Multilingual joint word sense disambiguation. In *Proceedings of the 2012 joint conference on Empirical*

Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, pp. 1399–1410.

Nivre, J., de Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. & Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA).

Petrov, S., Das, D. & McDonald, R. (2012). A Universal Part-of-Speech Tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Romary, L. & Lopez, P. (2015). GROBID-Information Extraction from Scientific Publications. *ERCIM News*, 100.

Speer, R., Chin, J. & Havasi, C. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI*. pp. 4444–4451.

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M. & Lindström, N. (2014). JSON-LD 1.0.

Suchanek, F. M., Kasneci, G. & Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*. ACM, pp. 697–706.

Swier, R. S. & Stevenson, S. (2005). Exploiting a verb lexicon in automatic semantic role labelling. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 883–890.

Tiberius, C. & Declerck, T. (2017). A lemon Model for the ANW Dictionary. In I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubíček & V. Baisa (eds.) *Proceedings of the eLex 2017 conference*. INT, Trojína and Lexical Computing, Lexical Computing CZ s.r.o., pp. 237–251.

Xue, N. & Palmer, M. (2004). Calibrating features for semantic role labeling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.