



ELEXIS Data Model

Carole Tiberius and Simon Krek
Vienna, 18 February 2019



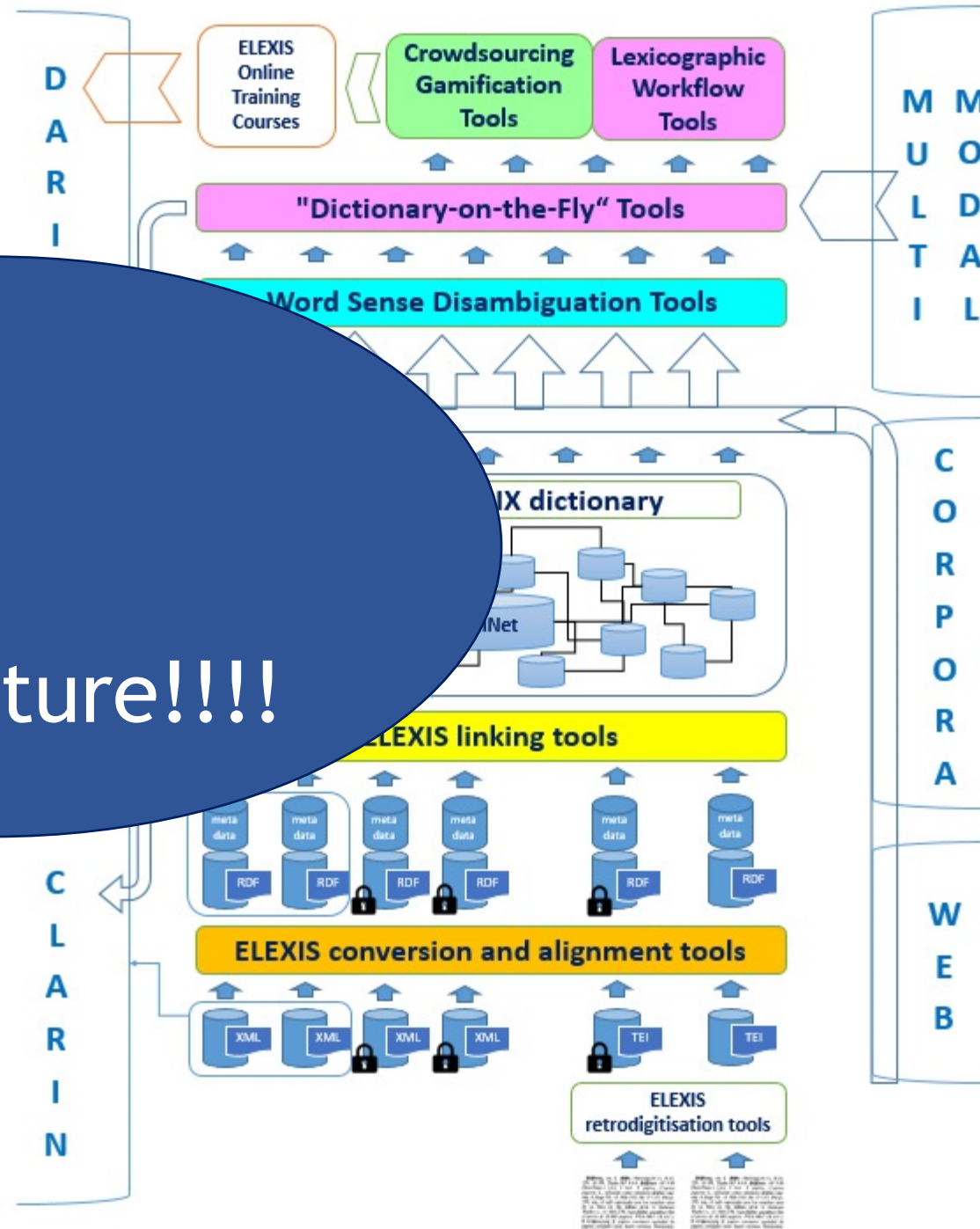
Why do we need a data model?



(multi)lingual linking

- “To ensure that there is integration at even the most basic level we will define a **minimal common data model** capturing the basic concepts of a lexicographic resource such as entries (single-word, multi-word), senses, syntactic frames, etymologies etc. and linguistic relationships such as synonymy/antonymy, translation, domain/region/register classification, relatedness, etc. that will be compatible with existing models used in the community, including TEI, LMF and OntoLex-Lemon.”

No Data
Empty
Infrastructure!!!!



ELEXIS data sources

Lexicographic data will be provided by

- Consortium partners
- Observing institutions with resources included in the European Dictionary Portal (<http://www.dictionarportal.eu/en/>) resulting from the COST ENeL action
- Other open access resources containing lexicographic data (CLARIN, DARIAH, etc.).



ELEXIS input data: consortium partners

partner	dictionary or lexicographic resource	language
JSI	Slovene Lexical Database	Slovene
INT	Dictionary of Contemporary Dutch (ANW)	Dutch
INT	Dictionary of the Dutch Language (WNT)	Dutch
INT	Dictionary of Old Dutch (ONW)	Old Dutch
INT	Dictionary of Early Middle Dutch (VMNW)	Early Middle Dutch
INT	Dictionary of Middle Dutch (MNW)	Middle Dutch
OEAW	Dictionary of Bavarian Dialects of Austria	Austrian
OEAW	Dagaare-Cantonese-English Dictionary	Dagaare, Cantonese, English
OEAW	Hausa-English Dictionary	Hausa, English
OEAW	Database of Bavarian Dialects of Austria	Austrian Variants
OEAW	Russian Dialect Dictionary	Russian
OEAW	Tunico	Tunisian

ELEXIS input data: consortium partners

partner	dictionary or lexicographic resource	language
BCHD	Karadžić, Serbian Dictionary (1818, 1852)	Serbian
BCHD	Miklošič, Lexicon Palaeoslovenico-Graeco-Latinum (1862–1865)	Old Church Slavic
BCHD	Daničić, Dictionary of Serbian Literary Antiquity (1863-4)	Serbian (medieval)
BCHD	Bojanić & Trivunac, Dictionary of Dubrovnik Dialect	Serbian (dialect)
BCHD	Elezović, Dictionary of Kosovo-Metohija Dialect	Serbian (dialect)
BCHD	Zlatanović, Dictionary of Southern Serbian Dialects	Serbian (dialect)
BCHD	Žugić, Dictionary of Jablanica Region	Serbian (dialect)
RILMTA	Hungarian Concise Dictionary	Hungarian
IBL	Dictionary of synonyms	Bulgarian
IBL	Dictionary of antonyms	Bulgarian
IBL	Dictionary of new words	Bulgarian
IBL	Dictionary of Bulgarian	Bulgarian

ELEXIS input data: consortium partners

partner	dictionary or lexicographic resource	language
KD	K English Multilingual Dictionary	English multilingual
KD	Global French Multilingual + L2-French	French multilingual
KD	Random House Webster's College Dictionary	English
DSL	The Danish Dictionary	Danish
DSL	Dictionary of the Danish Language	Danish
DSL	Moths Dictionary	Danish
DSL	Old Danish Dictionary	Old Danish
DSL	Danish Thesaurus	Danish
UCHP	Dictionary of Danish Insular Dialects	Danish
UCHP	Dictionary of Old Norse Prose	Old Norse

ELEXIS input data: consortium partners

partner	dictionary or lexicographic resource	language
TCDH	The German Dictionary by Jacob and Wilhelm Grimm (first edition)	German
TCDH	Rhenish Dictionary	German (dialect)
TCDH	Palatinate Dictionary	German (dialect)
TCDH	Dictionary of the German-Lorraine Dialects	German (dialect)
TCDH	Dictionary of the Alsatian Dialects	German (dialect)
TCDH	Grammatical-Critical Dictionary of the High-German Idiom (Adelung, 2nd)	German
TCDH	Middle High German Dictionary (Benecke, Müller, Zarncke)	Middle High German
TCDH	Middle High German Dictionary (Lexer)	Middle High German
EKI	The Dictionary of Standard Estonian ÕS 2013	Estonian
EKI	The Explanatory Dictionary of the Estonian Language	Estonian
EKI	The Dictionary of Foreign Words	Estonian
EKI	The Estonian Etymological Dictionary	Estonian
EKI	The Basic Estonian Dictionary	Estonian
EKI	The Estonian-Russian Dictionary	Estonian-Russian
EKI	The Russian-Estonian Dictionary	Russian-Estonian
RAE	Diccionario de la lengua española, 22nd Ed. (2001)	Spanish



Disparate lexicographic sources



ELEXIS input data

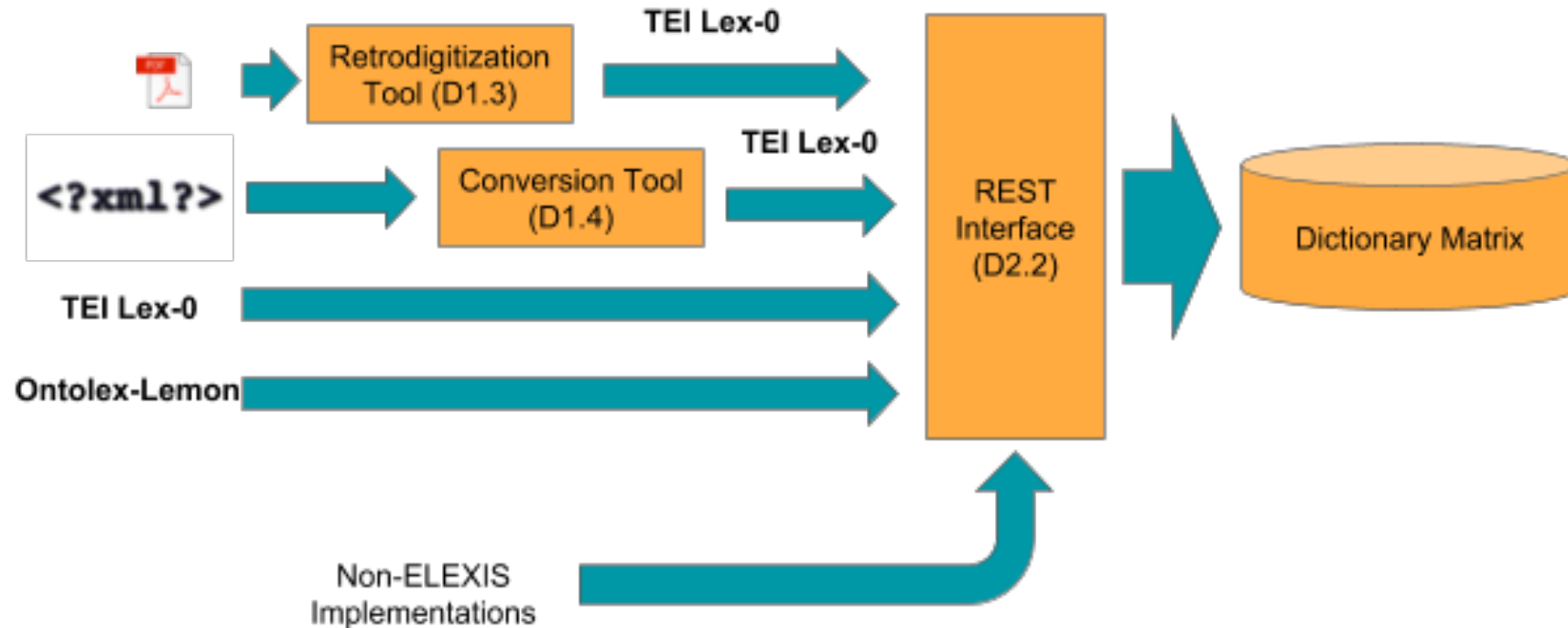
Data formats used by Consortium partners:

- **XML**, mainly contemporary dictionaries (JSI, INT, DSL, KD)
- **TEI**, mainly retrodigitised dictionaries (INT, OEAW, BCHD, TCDH)
- **HTML** (RAE, KD)
- **JSONLD** (KD)
- **Relational database** (Oracle (UCHP) or MySQL (IBL, UCHP))
- **API** (EKI)



How can we integrate this data?

Common protocols for input data



Access to ELEXIS interface through REST interface

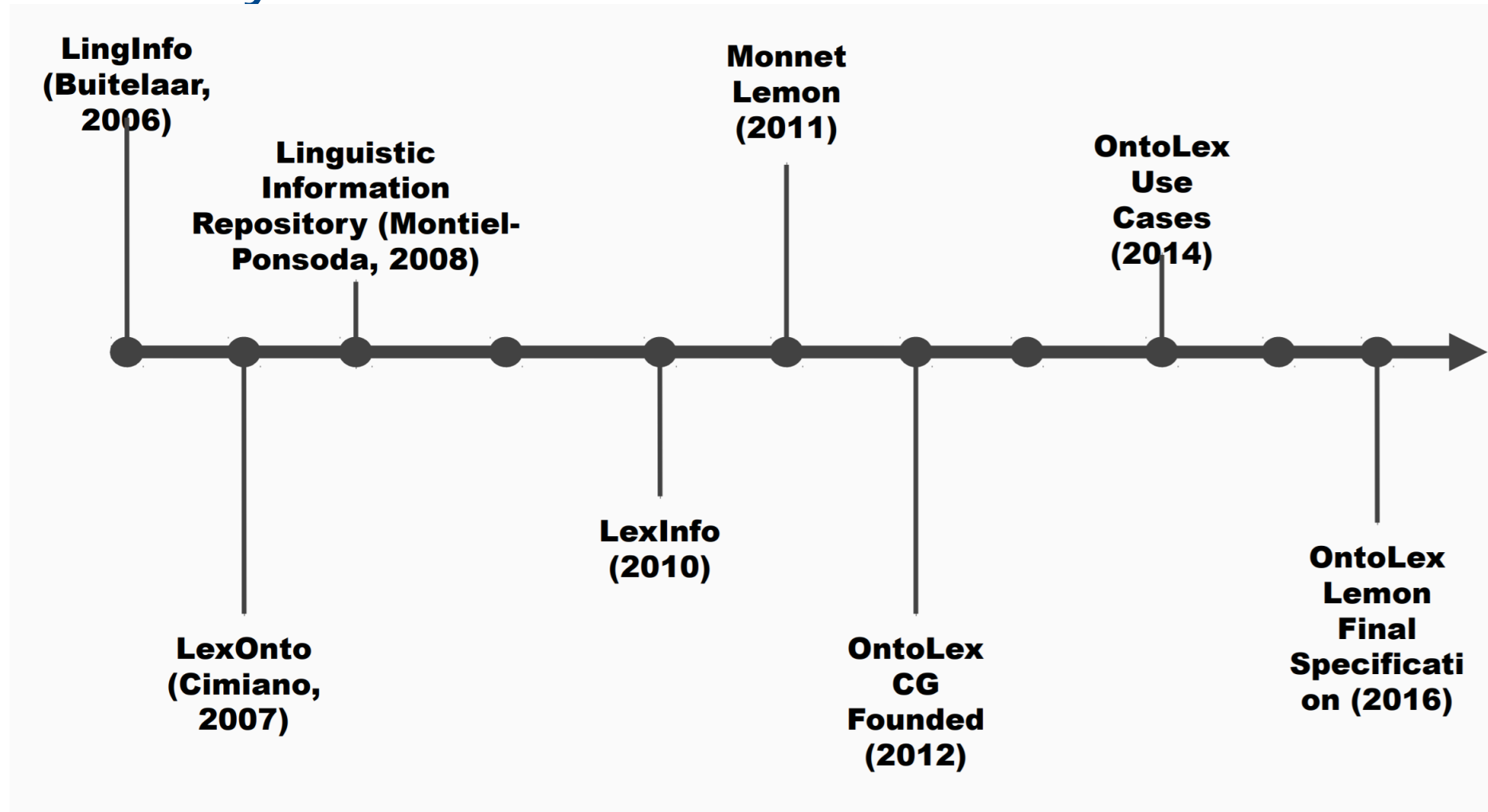
Ontolex-Lemon

- developed by the W3C Ontology-Lexicon Community Group
- 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
- based on the Resource Description Framework

Ontolex-Lemon Modules

- Ontolex Core
- Syntax and Semantics
- Decomposition
- Variation and Translation
- Linguistic Metadata
- Lexicography (in development)
- Morphology (in development)

History of the model



TEI Lex-0

Institutional background

- COST Action European Network of e-Lexicography (ENeL)
 - Working Group 2 “Retrodigitised Dictionaries” (Vera Hildenbrandt and Toma Tasovac)
- DARIAH working group “Lexical Resources” (Laurent Romary and Toma Tasovac)
- ELEXIS



TEI Lex-0

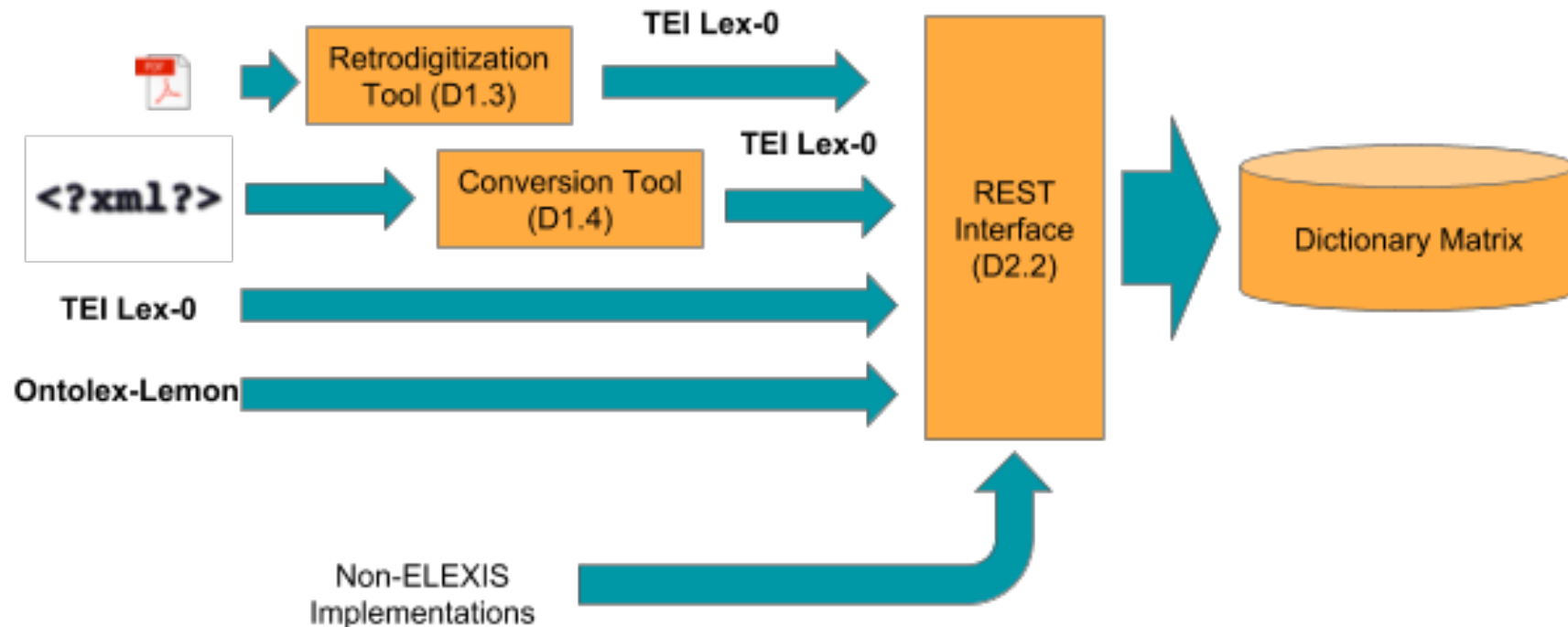
- TEI Lex-0 is a subset of the TEI schema (Text Encoding Initiative) serving as a baseline encoding and a target format to facilitate the interoperability of heterogeneously encoded lexical resources.
- This is important both in the context of building lexical infrastructures as such and in the context of developing generic TEI-aware tools such as dictionary viewers and profilers.

What TEI Lex-0 is and what it isn't

- 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 flavors of TEI.
- TEI Lex-0 should be primarily seen as a format that existing TEI dictionaries can be univocally **transformed to** in order to be queried, visualised, or mined in a uniform way.
- **AND LINKED!**

How can we integrate this data?

Common protocols for input data



Access to ELEXIS interface through REST interface

What do we have? Syntactic interoperability

relying on specified data formats, communication protocols, and the like to ensure communication and data exchange. The systems involved can process the exchanged information, but there is no guarantee that the interpretation is the same. (Ide and Pustejovsky 2010)



What do we need? Semantic interoperability

Semantic interoperability exists when two systems have the ability to automatically interpret exchanged information meaningfully and accurately in order to produce useful results via reference to a common information exchange reference model.

How do we achieve this?

- Common Vocabulary (WP1)
- Common Metadata (WP2)
- (Common Licensing schema (WP6))

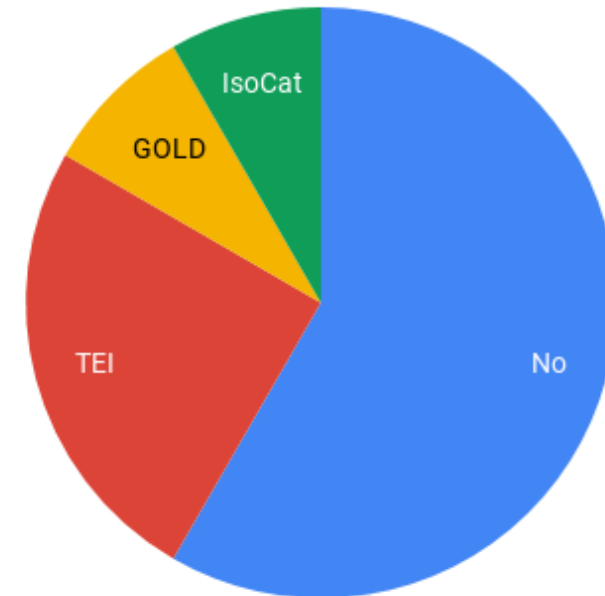


Common Vocabulary

- Survey results (for institutions):

Most of the responses (7) show that the lexicographic projects do not use existing standard vocabularies for encoding lexicographic data. Two institutions pointed out TEI as the standard vocabulary they used for their projects and, one institution uses IsoCat, GOLD, TEI (most likely for different projects).

Use of Standard Vocabulary



Common vocabulary

A reference model where the main concepts are unambiguously defined:

- lemma,
- part of speech
- sense
- multi word expression
- ...



We are not reinventing the wheel, but rely on existing standards, i.e. ISO standards (including LMF), etc. as well as the ELEXIS interoperability formats, i.e. Ontolex and TEI Lex-0.

Common Core Vocabulary

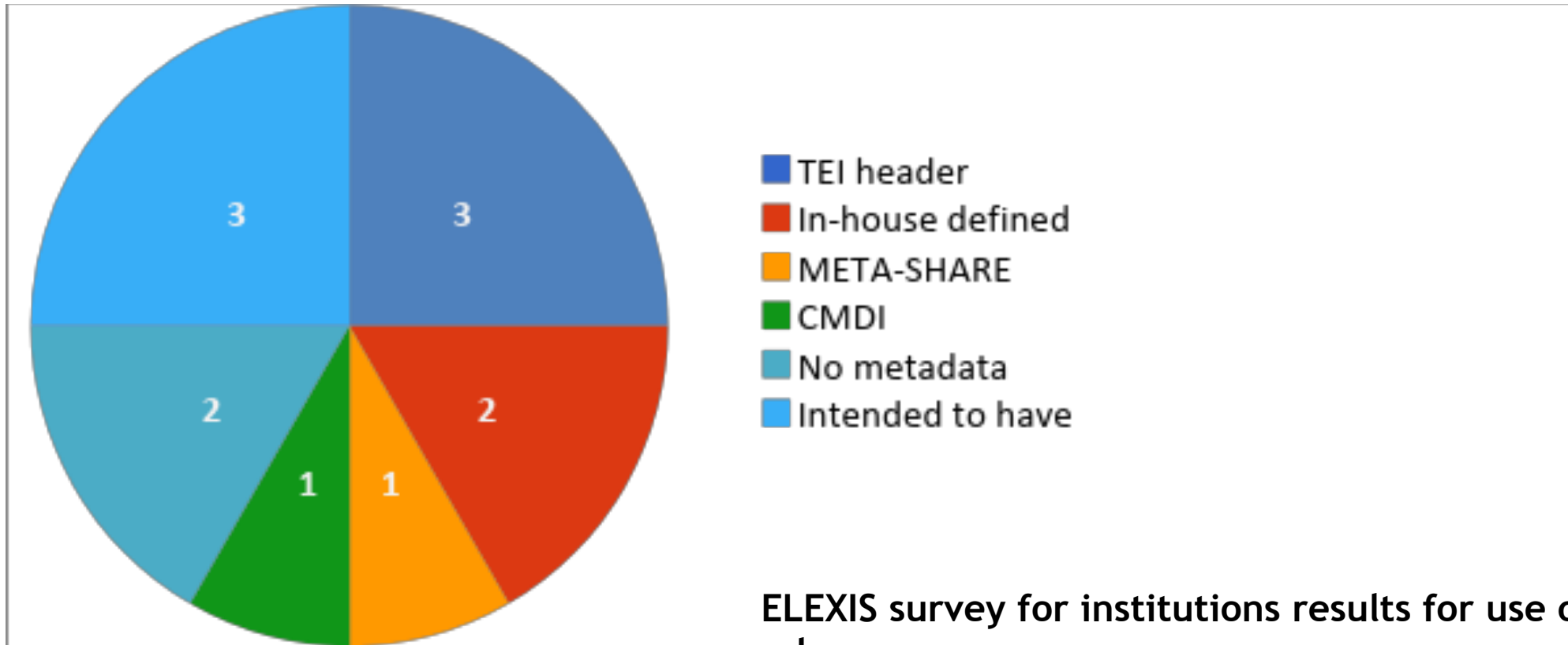
Element list (11/2018)	TEI Lex-0	Ontolex-Lemon	LMF	ISO1951:2007	ELEXIS
Lemmas (incl variant forms of lemmas)			3.4 lemma, lemmatized form, canonical form = conventional word form chosen to represent a lexeme ##### 3.2 word form = instance of a word, multi-word expression, root, stem, or morpheme ##### 3.5 lexeme = abstract unit generally associated with a set of word forms (3.2) sharing a common meaning	3.7 lemma: base word = lexical unit (3.8) according to lexicographical conventions to represent the different forms of an inflectional paradigm; ##### lexical unit(3.8) = unit of language, belonging to the lexicon of a given language and which is described or mentioned in a dictionary.	
Part of speech	attribute of <gram> element: pos (part of speech) any of the word classes to which a word may be assigned in a given language, based on form, meaning, or a combination of features, e.g. noun, verb, adjective, etc.	We can specify the part of speech of a word as follows using the lexinfo vocabulary: <rdfs:comment> Term used to describe how a particular word is used in a sentence. </rdfs:comment> <rdfs:comment> A category assigned to a word based on its grammatical and semantic properties. </rdfs:comment>	3.11 part of speech, lexical category, word class = category assigned to a lexeme (3.5) based on its grammatical properties ##### 2008 version 3.37 part of speech: lexical category: word class = category assigned to a lexeme based on its grammatical properties NOTE Typical parts of speech for European languages include: noun, verb, adjective, adverb, preposition, etc.	part of speech = A category assigned to a lexical unit based on its grammatical and semantic properties. [Adpated from ISO 12620:1999, A 2.2.1;	



Current state of affairs

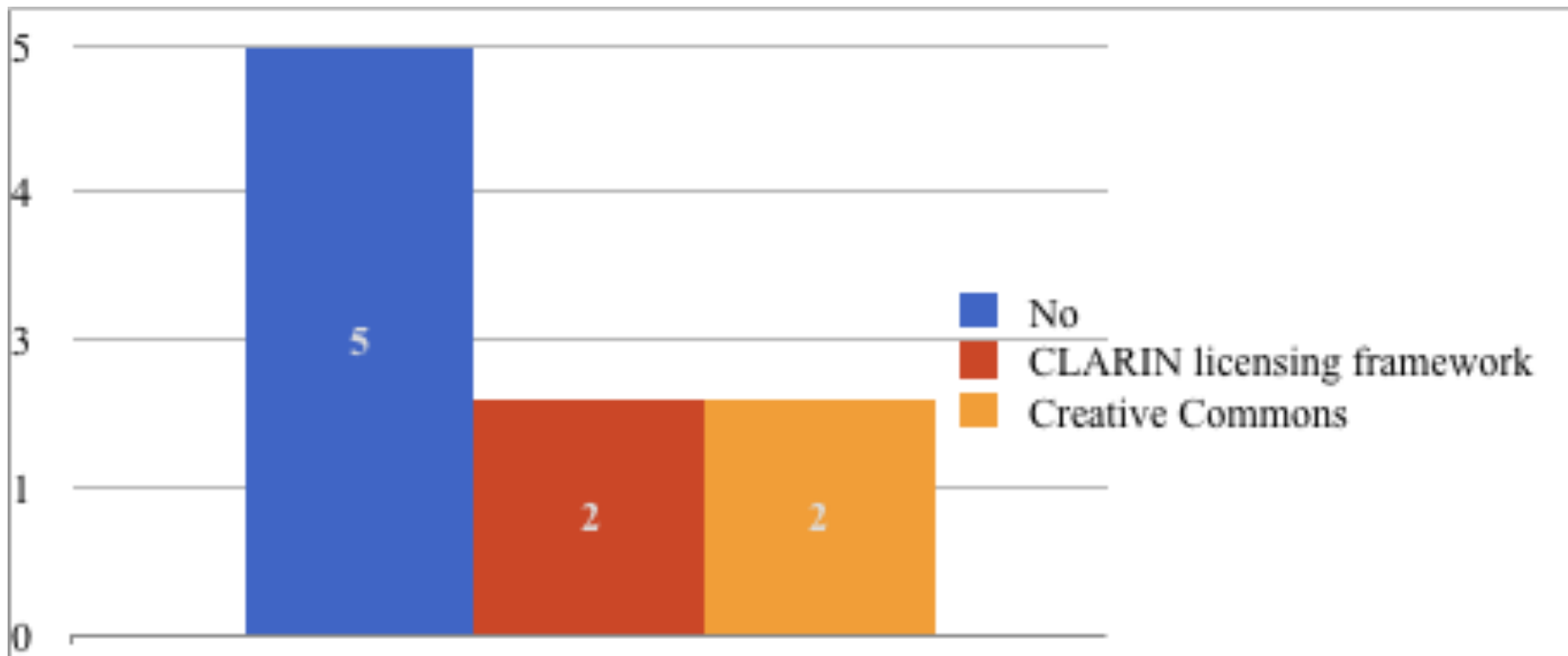
Information category	Examples of possible containers
lemma	lemmavorm, iztonica, Holem, @class='f'...
part of speech	gramGrp, pos, woordsoort, szofaj
glosses (senses) (anything that describes the sense, e.g. definition, sense indicator)	definition, sense, minidefinitie, translation
examples, attestations	zglesd, voorbeeld, Citat, cit type=.., concordance lines
phrases, collocations, multiword expressions

Common metadata



ELEXIS survey for institutions results for use of metadata schema

Standard licensing schema




ELEXIS survey for institutions results for use of licensing schema

ELEXIS data model

- a **minimal common data model** capturing the main concepts of lexicographic resources
- takes into account international standards.
- can be expressed by both interoperability formats (Ontolex-Lemon and TEI Lex-0)

The main goal is to ensure **semantic interoperability** between lexicographic resources to enable **integration** in the ELEXIS infrastructure.

Use case

- Lexicographic Project X: “We would like to contribute data to ELEXIS. We have a monolingual dictionary of contemporary Dutch with quite a complex microstructure.”
-  : “Wonderful. To be able to integrate your data in the ELEXIS infrastructure, you need to map the labels used for the elements in your data onto the ELEXIS data model, like this”:

Mapping

```
<?xml version="1.0" encoding="UTF-8" ?>
<Woordenboek>
<artikel>
  <Lemma>
    <Lemmavorm>wijn</Lemmavorm>
    <Lemmatype>woord</Lemmatype>
  </Lemma>
  <Woordsoort>
    <Type>substantief</Type>
  ...
  <BetekenisEnGebruik>
    <Kernbetekenis pid="325624">
      <betekenisInfo>
        <Betekenisnummer>1.0</Betekenisnummer>
      </betekenisInfo>
    </Kernbetekenis>
  </BetekenisEnGebruik>
  <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>
```

lemma

part of speech

sense

example

Etc.

In TEI Lex-0

```
<text>↓
  <body>↓
    <entry xml:id="ANW-1" xml:lang="NL">↓
      <form>↓
        <orth>wijn</orth>↓
      </form>↓
      <gramGrp>↓
        <pos ud:norm="NOUN">substantief</gram>↓
      </gramGrp>↓
      <sense xml:id="S1">↓
        <seg>1.0</seg>↓
        <cit type="example"><quote>Vul het vat aan [...] maar zeker ↓
        |hiet meer met een suikeroplossing daar de wijn anders ↓
        weer aan het gisten gaat of te zoet zal worden.</quote></cit>↓
        ...↓
      </sense>
    </entry>
  </body>
</text>
```

Development plans (2019)

- Common vocabulary for the main concepts
- Implementation of main concepts in TEI Lex-0
- Implementation of main concepts in Ontolex-Lemon



Thank you for your attention

