elexis european lexicographic infrastructure

D1.4

ELEXIS CONVERSION TOOLS

Author(s): Andraž Repar, Carole Tiberius, Tina Munda, Simon Krek, Jelena Kallas, Iztok Kosem

Date: April 30, 2021

elexis european lexicographic infrastructure

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

D1.4 ELEXIS conversion tools

Deliverable Number:	D1.4
Dissemination Level:	Public
Delivery Date:	April 30, 2021
Version:	2
Author(s):	Andraž Repar, Carole Tiberius, Tina Munda, Simon Krek, Jelena Kallas, Iztok Kosem

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

Deliverable/Document Information

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

Document History

Version Date	Changes/Approval	Author(s)/Approved by
April 20, 2021	Initial draft	Andraž Repar, Carole Tiberius, Tina Munda
April 26, 2021	Internal revision	Simon Krek, Jelena Kallas
April 29, 2021	External evaluation	Iztok Kosem

Table of Contents

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.

1 About this report

This report describes the software that was developed as part of LEX1 infrastructure of ELEXIS to harmonise the different lexicographic data formats. The work reported on in this report builds on the work described in Deliverable 1.3 *Tools for the automatic segmentation and identification of lexicographic content*, which describes the first version of the ELEXIFIER tool.

After its release, ELEXIFIER proved a useful tool not only for automatically segmenting and identifying lexicographic data, but also for transforming proprietary XML encoded lexicographic data to the standard format required for integration in the ELEXIS infrastructure. Therefore it was decided to invest in further optimising and expanding the functionalities of the ELEXIFIER tool rather than developing yet another set of conversion tools.

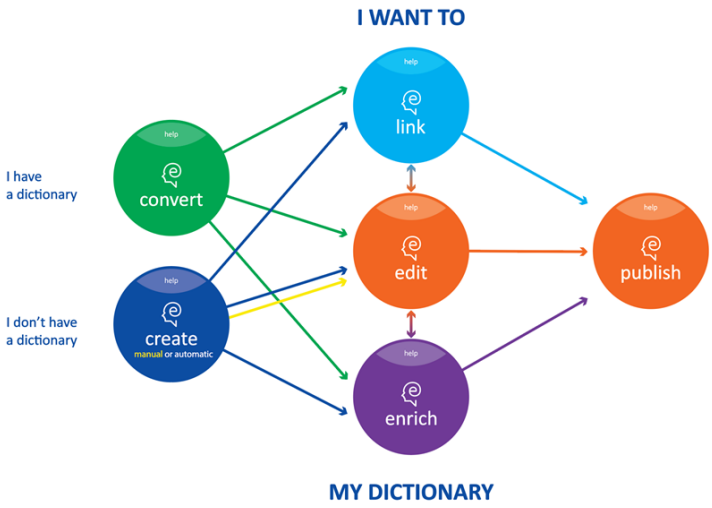

Figure 1: ELEXIFIER is part of the LEX1 infrastructure of ELEXIS and allows users to convert dictionaries into a standard format, and subsequently to link, edit, enrich and publish them

1.1 Changes compared to first report

New elements defined in the Elexis Data Model were added (for details see the list of elements in Section 2.3) and we made several other changes and improvements under the hood (see Section 2.4).

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.

2 ELEXIFIER

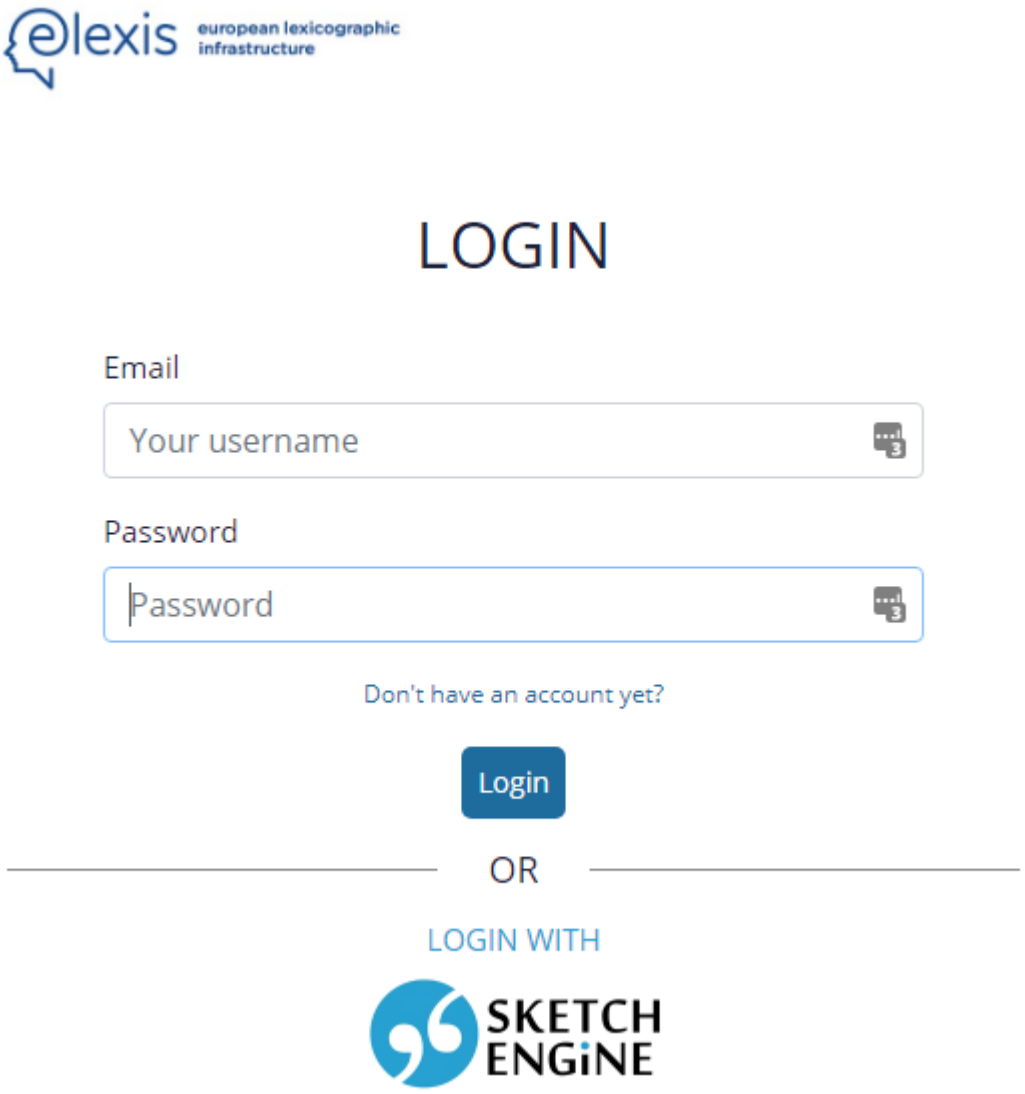

Figure 2: ELEXIFIER login interface.

ELEXIFIER (elexifier.elex.is) is a cloud-based dictionary conversion service for conversion of legacy XML and PDF dictionaries into a shared data format based on the ELEXIS data model[1]. It takes as input an XML or PDF dictionary and produces a TEI Lex-0 compliant XML file in line with the specifications described in the ELEXIS data model.

[1] A first version of the ELEXIS data model is described in Deliverable 1.2 *Best practices for lexicography – intermediate report*.

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.

2.1 Infrastructure

The application consists of two Docker containers:

- front end: https://github.com/elexis-eu/elexifier
- back end: https://github.com/elexis-eu/elexifier-api

The front end is written in Angular. The back end is written in Python Flask and uses a Postgres database. For local installation, see the instructions in the Github repository[2].

2.2 Use

On the login screen, users can create a new account or login with their Sketch Engine credentials (note that users can use their Sketch Engine single sign-on credentials to login to all LEX1 tools). Then they need to select the XML or PDF module and upload a dictionary to get started. For detailed instructions, check the User Guide.

2.3 XML transformation - basic concepts

To transform a proprietary XML dictionary into an ELEXIS data model compliant format, a transformation needs to be defined, which specifies rules for transforming proprietary XML elements into the corresponding ELEXIS data model elements. The script *https://github.com/elexis-eu/elexifier-api/blob/master/app/modules/transformator/dictTransformations3.py* takes as input a JSON object with the following members:

- ***entry*** — describes the selector for entry elements
- ***entry_lang*** — describes the transformer for the language attribute of the entries
- ***sense*** — describes the selector for sense elements
- ***hw*** — describes the transformer for headwords
- ***sec_hw*** — describes the transformer for secondary headwords
- ***pos*** — describes the transformer for part-of-speech tags
- ***hw_tr*** — describes the transformer for translations of headwords
- ***hw_tr_lang*** — describes the transformer for the language of the translations of headwords
- ***ex*** — describes the transformer for examples
- ***ex_tr*** — describes the transformer for translations of examples
- ***ex_tr_lang*** — describes the transformer for the language of the translations of examples

[2] https://github.com/elexis-eu/elexifier

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.

- ***def*** — describes the transformer for definitions
- **gloss** — describes the transformer for sense indicators
- ***usg*** — describes the transformer for labels
- ***note*** — describes the transformer for notes
- ***variant*** — describes the transformer for variant headwords
- ***inflected*** — describes the transformer for inflected forms
- **xr** — describes the transformer for cross-references

2.3.1 Selector descriptions

A selector is a rule that selects 0 or more elements in the input XML tree.

The description of a selector must be a JSON object. This object must contain an attribute named type whose value specifies the type of the selector, plus one or more other attributes whose name and meaning depend on the selector type.

The following types of selectors are currently supported:

Xpath selector: selects the nodes that match a given xpath expression (given in an attribute named expr).

Example:{"type": "xpath", "expr": ".//example/text"}

Union selector: combines the results of several selectors (whose descriptions must be given as a JSON array in an attribute named selectors).

Example:{"type": "union", "selectors": [...]}

Exclude selector: takes two selectors, left and right, and selects all those nodes which were selected by left but not by right.

Example:{"type": "exclude", "left": {...}, "right": {...}}

2.3.2 Transformer descriptions

A transformer is a rule that describes which data from the input document must be transformed into a certain type of element in the output document.

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.

The description of a transformer must be a JSON object. This object must contain an attribute named type whose value specifies the type of the transformer, plus one or more other attributes whose name and meaning depend on the transformer type.

The following types of transformers are currently supported:

(a) Simple transformers

A simple transformer selects a set of elements and extracts an attribute or the inner text from these elements; optionally applies a regular expression to the resulting text and returns the substring matched by a specific group within the regular expression.

The JSON object that describes a simple transformer must contain the following attributes:

type: this must be the string "simple".

selector: a JSON object describing a selector.

attr: the name of an attribute (from the elements selected by the selector) whose value is to be extracted. To extract the inner text of the element, instead of an attribute, use the pseudo-attribute name "{http://elex.is/wp1/teiLex0Mapper/meta}innerText". To extract the inner text of the element and all of its descendants, use "{http://elex.is/wp1/teiLex0Mapper/meta}innerTextRec". To return a constant value instead of extracting the value of an attribute, use the pseudo-attribute name "{http://elex.is/wp1/teiLex0Mapper/meta}constant".

rex: a regular expression that is applied to the value of the attribute attr. If this string does not contain any match for this regular expression, the current element is not transformed (i.e. it is as if it hadn't been selected by the selector at all). If there are several matches, the first one is used. This attribute is optional. If present, it must use the Python regular expression syntax.

rexGroup: this attribute is optional. If present, it must be the name of one of the named groups (?P<name>...) from the regular expression given by the attribute rex. In this case, only the string that matched this named group will be used, rather than the entire value of the attribute attr.

const: this attribute should be present it attr was set to "{http://elex.is/wp1/teiLex0Mapper/meta}constant", and should provide the constant value that you want to return as the result of the transformation.

xlat: this attribute is optional. If present, it should be a hash table that will be used to transform the string obtained from the previous steps (attribute lookup, regex matching). In other words, the

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.

string s will be replaced by xlat[s] if s appears as a key in xlat (otherwise, s will remain unchanged, just as if xlat had not been provided at all).

A simple example:

{ "type": "simple",

"selector": {"type": "xpath", "expr": ".//ExampleCtn//Locale"},

"attr": "lang" }

A more complex example:

{ "type": "simple",

"selector": {"type": "xpath", "expr": ".//sense/seg[1][@type='beleg']"},

"attr": "{http://elex.is/wp1/teiLex0Mapper/meta}innerTextRec"

"rex": "'(?P<insideQuotes>[^'])'",*

"rexGroup: "insideQuotes" }

This transformer selects the first <seg> in each <sense>, builds the inner text and extracts the first substring delimited by single quote marks.

An example of a constant-output transformer (i.e. to assign language codes to XML elements):

{ "type": "simple",

"selector": {"type: "xpath", "expr": ".//type"},

"attr": "{http://elex.is/wp1/teiLex0Mapper/meta}constant",

"const": "en" }

(b) Union transformers

A union transformer takes a set of simple transformers and performs all of their transformations. This might be useful if you need to combine several different transformation rules, e.g. extract attribute @a from instances of the element <b> and also extract attribute @c from instances of the element <d>.

The JSON object that describes a union transformer must contain the following attributes:

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.

type: this must be the string "union".

transformers: an array of JSON objects describing the transformers that are to be combined.

Detailed examples of individual element transformations can be found in the Appendix to this deliverable.

2.4 Changes to the XML module since deliverable D1.3

(a) Support for TEI(-like) input data

We added the option to transform TEI-compliant XML files. During the processing of the partners' dictionaries, we noticed that a number of dictionaries are already in a TEI-compliant or TEI-similar format. Due to several minor specific characteristics of the TEI format, the first version of ELEXIFIER was unable to process these formats. We adapted the transformation code to support TEI to TEI transformations. The advantage of supporting TEI and TEI-like input data is that this also makes it possible to reprocess data, which can be useful, for instance, if the mapping to Universal Dependencies part of speech tags[3] was not provided in the first transformation.

(b) Part of Speech element improvements

In the first version, part of speech information could only be selected from one XML element in the input data. We added the possibility to select multiple elements from the original XML, to take their values and map them to the UD tagset. We also added the option to delete individual POS elements from the mapping list which allows users to map only a subset of the values contained in the selected element.
Furthermore, we changed the way POS elements are transformed and mapped to retain the part of speech tags used in the input data (in the *orig* attribute).

(c) XML:id attribute generation

We changed how the unique XML element IDs (*xml:id*) are generated. Previously, the IDs were assigned with a simple concurrent numbering system (e.g., entry_1, sense_2, entry_3, sense_4 etc.). We have now implemented a more informative system where entry IDs retain information about the dictionary (in the form of an acronym) and sense IDs retain information about the dictionary and the headword they belong to.

[3] https://universaldependencies.org/u/pos/index.html

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.

(d) Various performance improvements under the hood

- A large part of the back end code was refactored for greater efficiency and better performance
- Improvements have been made in terms of transformation speed and full XPath version
- The possibility to select attributes from the parent element was added
- The possibility to bring nested entries to the top level was added
- The functionality to remove redundant dictScrap elements from the result has been improved

2.5 Code table issues

When dealing with XML:id attribute generation, we noticed that the XML specification allows a wider set of characters than some RELAX NG validators. We wanted to generate readable IDs which meant including headwords directly into the IDs - but if the headword in question contained a character not allowed by the RELAX NG validator (in our case jing[4]), then the final XML validation would fail. To avoid this issue, we replace the problematic characters with their numeric code.

2.6 PDF transformation – basic concepts

To transform a PDF dictionary, the user needs to annotate a sample of the PDF file. The PDF is first transformed in flat structure using a pdf2xml conversion script (based on https://github.com/kermitt2/pdf2xml). Then, a chunk of the resulting XML file is sent to Lexonomy[5] for manual annotation. In the next step, the annotations act as training data for the machine learning algorithm. The following features are used by the algorithm: font, font-size, bold, italic, newline and the token content itself.

Machine learning assumes a three-level structure with pages as first level base, entries as second level base and senses as third level base. On the first level, entries are predicted for the second level to work on, which in turn generates third level base - senses. A model is constructed for each level and trained on 75% of the data annotated in Lexonomy. Afterwards, labels for each token (separate word or symbol in the dictionary) of the unlabelled data are predicted for each level. At first,

[4] https://relaxng.org/jclark/jing.html
[5] https://www.lexonomy.eu/

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.

unlabelled data is only available for the first level, but through prediction, second and third level data is generated as well, along with the labels. Labels are then used to wrap tokens into correct containers.

The model used is a recurrent neural network with two inputs for each input token: one-hot encoded token features (such as font, size and so forth) and LSTM-encoded token contents. The two inputs are merged and fed into a bidirectional LSTM, which then outputs a one-hot encoded label. Labels are defined in the annotation and the model can adapt to different labels at different levels, depending on the annotation structure.

Current results show great promise as they often exceed 90% f1 score (varies between levels and datasets) and are achieved within a short training time. However, results are, as always in the field of machine learning, significantly influenced by the quality of the annotation.

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.

3 Evaluation

During development, ELEXIFIER has been continuously tested on data that has been contributed to the infrastructure by ELEXIS partners and observers. To date, 118 different datasets, e.g. general dictionaries, bilingual dictionaries, thesauri, specialised dictionaries (terminology, dialects), and lemma lists have been collected from 32 ELEXIS partner and observer institutions. A sample list of the datasets can be found in the Deliverable 6.3 *Intermediate interoperability report*. The collected data comes in a variety of formats, e.g. XML, TEI, HTML, JSONLD. To put ELEXIFIER to the test, we used the XML and TEI data.

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.

4 Expected impact

As part of the LEX1 infrastructure, ELEXIFIER plays an important role in the ELEXIS project, specifically in terms of feeding data to the Matrix Dictionary. It allows lexicographers with limited computer programming skills to convert their legacy dictionaries into a standardized common format and upload them to the Matrix Dictionary. Within WP5 training materials will be developed to support users.

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.

5 Appendix

Examples of using ELEXIFIER to transform the ELEXIS core elements

In this appendix we will illustrate ELEXIFIER with example data coming from dictionaries from ELEXIS partners. For each of the ELEXIS core elements one or two examples will be given of input data together with the transformed output from ELEXIFIER.

5.1 Entry

entry	part of a lexicographic resource which contains information related to at least one headword

JSI - Slovene Lexical Database (proprietary XML)

Extract from the entry for *mačka* 'cat' in the Slovene Lexical Database.

Input

```
<clanek>
    <glava>
      <oblika>
        <zapis sloleks="LE_S_mačka">mačka</zapis>
       ...
      </oblika>
      ...
      </glava>
   </clanek>
```

Output

```
<entry xmlns:m="http://elex.is/wp1/teiLex0Mapper/meta"
xmlns:a="http://elex.is/wp1/teiLex0Mapper/legacyAttributes"
xmlns="http://www.tei-c.org/ns/1.0" m:e="clanek" xml:lang=""
type="null" xml:id="dict_mačka_1_pos">
 <form type="lemma">
   <orth m:e="zapis" a:sloleks="LE_S_mačka">mačka</orth>
 </form>
...
</entry>
```


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.

5.2 Headword

headword	Organising element of an entry in a lexicographic resource. *In printed dictionaries typically at the top of an entry*

JSI - Slovene Lexical Database (proprietary XML)

Extract from the entry for *mačka* 'cat' in the Slovene Lexical Database.

Input

```
<?xml version="1.0" encoding="UTF-8"?>
  <clanek>
      <glava>
        <oblika>
          <zapis sloleks="LE_S_mačka">mačka</zapis>
         ...
        </oblika>
        ...
        </glava>
      </clanek>
```

Output

```
<entry xmlns:m="http://elex.is/wp1/teiLex0Mapper/meta"
xmlns:a="http://elex.is/wp1/teiLex0Mapper/legacyAttributes"
xmlns="http://www.tei-c.org/ns/1.0" m:e="clanek" xml:lang=""
type="null" xml:id="SLDcat○33○_mačka_1_noun">
 <form type="lemma">
   <orth m:e="zapis" a:sloleks="LE_S_mačka">mačka</orth>
 </form>
```


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.

INT - Algemeen Nederlands Woordenboek (proprietary XML)

Extract from the entry for *kat* 'cat' in the Algemeen Nederlands Woordenboek.

Input

```
<?xml version="1.0" encoding="UTF-8" ?>
<artikel xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
ID="27823"
xsi:noNamespaceSchemaLocation="N:/ANW/Werkstation/Schema/ANWSche
ma-main.xsd"  pid="78332">
  <Lemma>
       <Lemmavorm>kat</Lemmavorm>
       <Lemmatype>woord</Lemmatype>
  </Lemma>
  <Woordsoort>
       <Type>substantief</Type>
       ...
…
</artikel>
```

Output

```
<entry xmlns:m="http://elex.is/wp1/teiLex0Mapper/meta"
xmlns:a="http://elex.is/wp1/teiLex0Mapper/legacyAttributes"
xmlns="http://www.tei-c.org/ns/1.0"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
m:e="artikel" a:ID="27823"
xsi:noNamespaceSchemaLocation="N:/ANW/Werkstation/Schema/ANWSche
ma-main.xsd" a:pid="78332" xml:lang="" type="null" xml:id="ANW○
33○_kat_1_noun">
 <form type="lemma">
   <orth m:e="Lemmavorm">kat</orth>
 </form>
```

5.3 Secondary headword

secondary headword	headword-like lexical item occurring within an entry in a lexicographic resource, for example derived forms, feminine forms, multiword expressions. Often an organising element of a part of an entry

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.

INT - Algemeen Nederlands Woordenboek (proprietary XML)

Extract from the entry for *kat* 'cat' in the Algemeen Nederlands Woordenboek.

Input

```
<?xml version="1.0" encoding="UTF-8" ?>
<artikel xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
ID="27823"
xsi:noNamespaceSchemaLocation="N:/ANW/Werkstation/Schema/ANWSche
ma-main.xsd"  pid="78332">
  <Lemma>
       <Lemmavorm>kat</Lemmavorm>
       <Lemmatype>woord</Lemmatype>

      <Woordfamilie>

      ...
        <Afleidingen>katachtig; katje; kattendom; katterig;
kattig; kattin</Afleidingen>

        ...
      </Woordfamilie>
```

Output

```
<entry xmlns:m="http://elex.is/wp1/teiLex0Mapper/meta"
xmlns:a="http://elex.is/wp1/teiLex0Mapper/legacyAttributes"
xmlns="http://www.tei-c.org/ns/1.0"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
m:e="artikel" a:ID="27823"
xsi:noNamespaceSchemaLocation="N:/ANW/Werkstation/Schema/ANWSche
ma-main.xsd" a:pid="78332" xml:lang="" type="null" xml:id="ANW○
33○_kat_1_noun">
 <form type="lemma">
   <orth m:e="Lemmavorm">kat</orth>
   ...
   <form type="simple">
     <orth m:e="Afleidingen">katachtig; katje; kattendom;
katterig; kattig; kattin</orth>
   </form>
```


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.

5.4 Variant headword

variant headword	lexical item representing one of the alternative forms of the headword, for example spelling variation or regional variation

INT - Woordenboek der Nederlandsche Taal (TEI XML)

Extract from the entry *kat* 'cat' in the Woordenboek der Nederlandsche Taal.

Input

```
<entry source="ind" key="KAT#01#1#M030758" id="M030758"
band="8">
 <GTB-entry id="M030758">
  ...
 </GTB-entry>
 <form type="lemma" n="I" k="1">
   <orth>KAT</orth>
 </form>
 <form type="mdl">kat</form>
<dictScrap><gtb-
woordsoort><gramGrp><pos>znw.</pos>(<gen>v.</gen>,<gen
rend="restricted">m.</gen>)</gramGrp></gtb-woordsoort>
— gewestelijk ook nu nog <form
type="lemmavariant"><orth>KATTE</orth></form> —, znw. Vr.,
...
</dictScrap>
    ...
</entry>
```

Output

```
<entry xmlns:m="http://elex.is/wp1/teiLex0Mapper/meta"
xmlns:a="http://elex.is/wp1/teiLex0Mapper/legacyAttributes"
xmlns="http://www.tei-c.org/ns/1.0" m:e="entry" a:source="ind"
```


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.

```
a:key="KAT#01#1#M030758" a:id="M030758" a:band="8" xml:lang=""
type="null" xml:id="WNTkat○33○_KAT_1_pos">
 <form type="lemma">
   <orth m:e="orth">KAT</orth>
 </form>
 <form type="variant">
   <orth m:e="orth">KATTE</orth>
 </form>
</entry>
```

5.5 Inflected form

inflected form	form of the inflectional paradigm of the headword

INT - Algemeen Nederlands Woordenboek (proprietary XML)

Extract from the entry *zus* 'sister' in the Algemeen Nederlands Woordenboek.

Input

```
...
<Lemma>
      <Lemmavorm>zus</Lemmavorm>
      <Lemmatype>woord</Lemmatype>
</Lemma>
...
<SpellingEnFlexie>
      <SubstantiefVormen>
            <Enkelvoud>
                  <Woordvorm>zus</Woordvorm>
                  <Afbreking>zus</Afbreking>
            </Enkelvoud>
            <Meervoud>
                  <Woordvorm>zussen</Woordvorm>
                  <Afbreking>zus.sen</Afbreking>
            </Meervoud>
      </SubstantiefVormen>
...
```


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.

Output

```
...
<form type="lemma">
   <orth m:e="Lemmavorm">zus</orth>
 </form>
...
<form type="inflected">
   <orth m:e="Woordvorm">zus</orth>
</form>
<form type="inflected">
   <orth m:e="Woordvorm">zussen</orth>
</form>
...
```

5.6 Part of speech

part of speech	any of the word classes to which a lexical item may be assigned, e.g. noun, verb, adjective, etc.

JSI - Slovene Lexical Database (Proprietary XML)

Extract from the entry for *mačka* ‘cat’ in the Slovene Lexical Database.

Input

```
<clanek>
 <glava>
   <oblika>
     <zapis sloleks="LE_S_mačka">mačka</zapis>
     <iztocnica izvor="lbs" status="potr"
tip="neavt">mačka</iztocnica>
   </oblika>
   <korpusi>
     <korpus>
       <frek_lema/>
     </korpus>
```


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.

```
    </korpusi>
    <zaglavje>
      <besedna_vrsta>samostalnik</besedna_vrsta>
    </zaglavje>
```

Output

```
<entry xmlns:m="http://elex.is/wp1/teiLex0Mapper/meta"
xmlns:a="http://elex.is/wp1/teiLex0Mapper/legacyAttributes"
xmlns="http://www.tei-c.org/ns/1.0" m:e="clanek" xml:lang=""
type="null" xml:id="SLDcat〇33〇_mačka_1_noun">
 <form type="lemma">
   <orth m:e="zapis" a:sloleks="LE_S_mačka">mačka</orth>
 </form>
 <gramGrp>
   <gram orig="samostalnik" type="pos">noun</gram>
 </gramGrp>
```

Note that in the output the original part of speech tag is kept as the value of the attribute orig. ELEXIFIER offers the option to map the part of speech tags on the input data to the Universal Dependencies part of speech tags. This mapping is recommended when uploading and transforming lexicographic data with ELEXIFIER.

5.7 Sense

sense	part of an entry which groups together information relating to a meaning of a headword (or secondary headword), for example definitions, examples, and translations

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.

INT - Algemeen Nederlands Woordenboek (proprietary XML)

Extract from the entry for *kat* 'cat' in the Algemeen Nederlands Woordenboek.

Input

```
...
<BetekenisEnGebruik>
   <Kernbetekenis pid="78333">
     <betekenisInfo>
       <Betekenisnummer>1.0</Betekenisnummer>
...
```

Output

```
...
<sense m:e="Kernbetekenis" a:pid="78333" xml:id="ANW○33○
_kat_1_noun_1">
   <gramGrp>
     <gram orig="substantief" type="pos">noun</gram>
   </gramGrp>
   <def m:e="Definitie">klein huisdier dat miauwt en spint en
dat gehouden wordt als gezelschapsdier of voor het vangen van
muizen en ratten; huiskat; poes</def>
   ...
 </sense>
```

5.8 Definition

definition	statement that describes a meaning and permits its differentiation from other meanings within a sense structure of an entry

INT - Algemeen Nederlands Woordenboek (proprietary XML)

Extract from the entry for *06-dealer* (lit. 06-drug dealer) in the Algemeen Nederlands Woordenboek.

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.

Input

```
<artikel ID="71214" pid="190832">
 <Lemma>
   <Lemmavorm>06-dealer</Lemmavorm>
   <Lemmatype>woord</Lemmatype>
 </Lemma>
...
 <BetekenisEnGebruik>
   <Kernbetekenis pid="190834" id="bet1.0">
     <betekenisInfo>
       <Betekenisnummer>1.0</Betekenisnummer>
...
<definitieBody id="el10">
          <Definitie>drugsdealer die via zijn mobiele telefoon
een bestelling voor drugs doorkrijgt en deze vervolgens op
afspraak aflevert</Definitie>
         ...
    </definitieBody>
```

Output

```
<entry xmlns:m="http://elex.is/wp1/teiLex0Mapper/meta"
xmlns:a="http://elex.is/wp1/teiLex0Mapper/legacyAttributes"
xmlns="http://www.tei-c.org/ns/1.0" m:e="artikel" a:ID="71214"
a:pid="190832" xml:lang="" type="null" xml:id="ANW_06-
dealer_1_pos">
 <form type="lemma">
   <orth m:e="Lemmavorm">06-dealer</orth>
 </form>
 <gramGrp>
   <gram orig="substantief" type="pos"/>
 </gramGrp>
...
```


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.

```
 <sense m:e="Kernbetekenis" a:pid="190834" a:id="bet1.0"
xml:id="ANW_06-dealer_1_pos_1">
    ...
    <def m:e="Definitie">drugsdealer die via zijn mobiele
telefoon een bestelling voor drugs doorkrijgt en deze vervolgens
op afspraak aflevert</def>
 </sense>
...
</entry>
```

RAE - Diccionario de la Lengua Española 22a. ed. (TEI XML)

Extract from the entry for *abacería* 'grocery' in the Diccionario de la Lengua Española 22a. ed.

Input

```
<entry key="abacer&#xED;a" pos="sustantivo"
orig="abacer&#xED;a.">
 <orth id="11" type="lemma" norm="abacería" orig="abacería"/>
 <etym>(De<mentioned>abacero</mentioned>)</etym>
 <sense n="1" id="19">
   <gramgrp>
     <pos type="sustantivo"/>
   </gramgrp>
   <usg type="gram" norm="nombre femenino" orig="f."/>
   <def key="abacería" orig="abacería." lPos="nombre femenino"
sPos="f.">Puesto o tienda donde se venden al por menor aceite,
vinagre, legumbres secas, bacalao, etc.</def>
 </sense>
</entry>
```

Output

```
<entry xmlns:m="http://elex.is/wp1/teiLex0Mapper/meta"
xmlns:a="http://elex.is/wp1/teiLex0Mapper/legacyAttributes"
xmlns="http://www.tei-c.org/ns/1.0" m:e="entry" a:key="abacería"
```


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.

```
a:pos="sustantivo" a:orig="abacería." xml:lang="" type="null"
xml:id="R_abacería_1_noun">
 <form type="lemma">
   <orth>abacería</orth>
 </form>
 <sense m:e="sense" a:n="1" a:id="19"
xml:id="R_abacería_1_noun_1">
   <gramGrp>
     <gram orig="sustantivo" type="pos">noun</gram>
   </gramGrp>
   <def m:e="def" a:key="abacería" a:orig="abacería."
a:lPos="nombre femenino" a:sPos="f.">Puesto o tienda donde se
venden al por menor aceite, vinagre, legumbres secas, bacalao,
etc.</def>
 </sense>
</entry>
```

5.9 Sense indicator

sense indicator	short statement that gives an indication of a meaning and permits its differentiation from other meanings within a sense structure of an entry

JSI - Slovene Lexical Database (proprietary XML)

Extract from the entry for *mačka* ‘cat’ in the Slovene Lexical Database.

Input

```
<clanek>
 <glava>
   <oblika>
     <zapis sloleks="LE_S_mačka">mačka</zapis>
     ...
   <pomen>
     <indikator>domača žival; maček</indikator>
     ...
```


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.

Output

```
<entry xmlns:m="http://elex.is/wp1/teiLex0Mapper/meta"
xmlns:a="http://elex.is/wp1/teiLex0Mapper/legacyAttributes"
xmlns="http://www.tei-c.org/ns/1.0" m:e="clanek" xml:lang=""
type="null" xml:id="dict_mačka_1_noun">
 <form type="lemma">
   <orth m:e="zapis" a:sloleks="LE_S_mačka">mačka</orth>
 </form>
  ...
<seg m:e="pomen">
     <gloss m:e="indikator">domača žival; maček</gloss>
...
```

DSL - Den Danske Ordbog (proprietary XML)

Extract from the entry *kat* 'cat' in Den Danske Ordbog.

Input

```
<Artikel Bemaerkninger="synonymer tamkat og huskat ekspanderet"
EntryID="11025614" Historik="a" KommenteretAf="HL"
KommenteretDato="2017-03-21" OprettetAf="KL" OprettetDato="2003-
01-01" PubKlarDato="2013-05-15" Status="publiceret"
ÆndretAf="HL/LTJ" ÆndretDato="2017-03-28">
 <Iddel>
   ...
   <Holem>kat<Homnr>1</Homnr></Holem>
   <Lemklas>sb.</Lemklas>
   ...
   <HolemBeskriver>dyr; bandeord; kortbunke</HolemBeskriver>
 </Iddel>
...
```

Output

```
<entry xmlns:m="http://elex.is/wp1/teiLex0Mapper/meta"
xmlns:a="http://elex.is/wp1/teiLex0Mapper/legacyAttributes"
```


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.

```
xmlns="http://www.tei-c.org/ns/1.0" m:e="Artikel"
a:Bemaerkninger="synonymer tamkat og huskat ekspanderet"
a:EntryID="11025614" a:Historik="a" a:KommenteretAf="HL"
a:KommenteretDato="2017-03-21" a:OprettetAf="KL"
a:OprettetDato="2003-01-01" a:PubKlarDato="2013-05-15"
a:Status="publiceret" a:ÆndretAf="HL/LTJ" a:ÆndretDato="2017-03-
28" xml:lang="" type="null" xml:id="DDOsample○31○_kat_1_noun">
 <form type="lemma">
   <orth m:e="Holem">kat</orth>
 </form>
 <gramGrp>
   <gram orig="sb." type="pos">noun</gram>
 </gramGrp>
    <gloss m:e="HolemBeskriver">dyr; bandeord; kortbunke</gloss>
 ...
```

5.10 Label

label	item from a controlled vocabulary indicating some kind of restriction on the use of the lexical item, for example, time, region, domain, register

INT - Algemeen Nederlands Woordenboek (proprietary XML)

Extract from the entry *fier* 'proud' in the Algemeen Nederlands Woordenboek.

Input

```
<artikel ID="17738" pid="50852">
 <Lemma>
   <Lemmavorm>fier</Lemmavorm>
   <Lemmatype>woord</Lemmatype>
 </Lemma>
...
 <BijzonderhedenGebruik>
```


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.

```
    <Taalvarieteit>(vooral) in België</Taalvarieteit>
    ...
  </BijzonderhedenGebruik>
...
```

Output

```
<entry xmlns:m="http://elex.is/wp1/teiLex0Mapper/meta"
xmlns:a="http://elex.is/wp1/teiLex0Mapper/legacyAttributes"
xmlns="http://www.tei-c.org/ns/1.0" m:e="artikel" a:ID="17738"
a:pid="50852" xml:lang="" type="null" xml:id="ANW_fier_1_pos">
 <form type="lemma">
   <orth m:e="Lemmavorm">fier</orth>
 </form>
 <gramGrp>
   <gram orig="adjectief" type="pos"/>
 </gramGrp>
 <usg m:e="Taalvarieteit" type="hint">(vooral) in België</usg>
...
```

K Dictionaries - Global French-Russian Dictionary (proprietary XML)

Extract from the entry *virus* ‘virus’ in the Global French-Russian Dictionary.

Input

```
...
 <SenseGrp identifier="SE00018356">
   <SidCtn identifier="SI00010414">
     <SenseIndicator>informatique</SenseIndicator>
   </SidCtn>
      <Definition>programme qui endommage un
ordinateur</Definition>
...
```


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.

Output

```
...
 <usg m:e="SenseIndicator" type="hint">informatique</usg>
 <sense m:e="Definition" xml:id="dict_virus_1_noun_2">
   <def>programme qui endommage un ordinateur</def>
 </sense>
...
```

5.11 Example

example	instance of a lexical item's usage in a specific sense

K Dictionaries - Global French-Russian Dictionary (proprietary XML)

Extract from the entry *virus* 'virus' in the Global French-Russian Dictionary.

Input

...

```
<ExampleCtn>
   <Example>le virus de la grippe</Example>
   ...
</ExampleCtn>
```

Output

```
<seg m:e="ExampleCtn">
   <cit type="example">
       <quote m:e="Example">le virus de la grippe</quote>
   </cit>
...
</seg>
```


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.

DSL- Den Danske Ordbog (proprietary XML)

Extract from the entry *glad* 'happy' in Den Danske Ordbog.

Input

```
…
<Dok DokStatus="a">
   <Citat>
         <txt>I Tunesien er atmosfæren en anden, folk er mere
glade, de er ikke stressede</txt>
   </Citat>
  ...
</Dok>
...
```

Output

```
...
<seg m:e="Dok" a:DokStatus="a">
<seg m:e="Citat">
    <cit type="example">
      <quote m:e="txt">I Tunesien er atmosfæren en anden, folk
er mere glade, de er ikke stressede</quote>
    </cit>
</seg>
</seg>
...
```

5.12 Translation

translation	equivalent in another language of any element in an entry

Note that in ELEXIFIER two elements are distinguished. One for the translation of the headword and one for the translation of an example.

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.

K Dictionaries - Global French-Russian Dictionary (proprietary XML)

Extract from the entry *virus* 'virus' in the Global French-Russian Dictionary illustrating the translation of a headword.

Input

```
<DictionaryEntry identifier="FR-011890">
  <HeadwordCtn>
    <Headword>virus</Headword>
    ...
  </HeadwordCtn>
  <SenseBlock>
    <SenseGrp identifier="SE00018355">
      ...
      <Definition>organisme responsable d'une
maladie</Definition>
      <TranslationCluster identifier="TC00034819"
text="organisme  responsable d'une maladie" type="def">
        <Locale lang="ru">
          <TranslationCtn>
            <Translation>ви́рус
              ...
            </Translation>
             ...
          </TranslationCtn>
        </Locale>
      </TranslationCluster>
...
</DictionaryEntry>
```

Output

```
<entry xmlns:m="http://elex.is/wp1/teiLex0Mapper/meta"
xmlns:a="http://elex.is/wp1/teiLex0Mapper/legacyAttributes"
xmlns="http://www.tei-c.org/ns/1.0" m:e="Entry" a:HomNum=""
```


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.

```
    a:hw="virus" a:identifier="EN00010605" a:pos="noun" xml:lang=""
    type="null" xml:id="M_virus_1_noun">
     <form type="lemma">
       <orth>virus</orth>
     </form>
     ...
     <sense m:e="SenseGrp" a:identifier="SE00018355"
    xml:id="M_virus_1_noun_1">
       <def m:e="Definition">organisme responsable d'une
    maladie</def>
       <cit type="translationEquivalent">
         <quote m:e="Translation">ви́рус</quote>
       </cit>
       <cit type="example">
         <quote m:e="Example">le virus de la grippe</quote>
       </cit>
       <cit type="translation">
        <quote m:e="Translation">ви́рус гри́ппа</quote>
       </cit>
 </sense>
```

K Dictionaries - Global French-Russian Dictionary (proprietary XML)

Extract from the entry *virus* 'virus' in the Global French-Russian Dictionary illustrating the translation of an example.

Input

```
    <DictionaryEntry identifier="FR-011890">
      <HeadwordCtn>
        <Headword>virus</Headword>
        ...
      </HeadwordCtn>
      <SenseBlock>
        <SenseGrp identifier="SE00018355">
      ...
           <ExampleCtn>
```


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.

```
            <Example>le virus de la grippe</Example>
            <TranslationCluster identifier="TC00034820" text="le
    virus de la grippe" type="exmp">
              <Locale lang="ru">
                <TranslationCtn>
                  <Translation>ви́рус гри́ппа</Translation>
                </TranslationCtn>
              </Locale>
            </TranslationCluster>
          </ExampleCtn>
      ...
```

Output

```
<entry xmlns:m="http://elex.is/wp1/teiLex0Mapper/meta"
xmlns:a="http://elex.is/wp1/teiLex0Mapper/legacyAttributes"
xmlns="http://www.tei-c.org/ns/1.0" m:e="Entry" a:HomNum=""
a:hw="virus" a:identifier="EN00010605" a:pos="noun" xml:lang=""
type="null" xml:id="M_virus_1_noun">
...
 <sense m:e="SenseGrp" a:identifier="SE00018355"
xml:id="M_virus_1_noun_1">
   ...
   <cit type="example">
     <quote m:e="Example">le virus de la grippe</quote>
   </cit>
   <cit type="translation">
     <quote m:e="Translation">ви́рус гри́ппа</quote>
   </cit>
 </sense>
```


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.

5.13 Cross reference

cross reference	element providing any kind of link or reference to another element within or outside the lexicographic resource

INT - Algemeen Nederlands Woordenboek (proprietary XML)

Extract from the entry *kat* 'cat' in the Algemeen Nederlands Woordenboek.

Input

```
<artikel xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
ID="27823"
xsi:noNamespaceSchemaLocation="N:/ANW/Werkstation/Schema/ANWSche
ma-main.xsd" pid="78332">
 <Lemma>
   <Lemmavorm>kat</Lemmavorm>
   <Lemmatype>woord</Lemmatype>
 </Lemma>
...
       <Woordrelaties>
         <Hyperoniem>
           <link>
             <lemma>zoogdier</lemma>
             <dtype>art</dtype>
             <pid>188698</pid>
           </link>
         </Hyperoniem>
```

Output

```
<entry xmlns:m="http://elex.is/wp1/teiLex0Mapper/meta"
xmlns:a="http://elex.is/wp1/teiLex0Mapper/legacyAttributes"
xmlns="http://www.tei-c.org/ns/1.0"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
m:e="artikel" a:ID="27823"
xsi:noNamespaceSchemaLocation="N:/ANW/Werkstation/Schema/ANWSche
```


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.

```
ma-main.xsd" a:pid="78332" xml:lang="" type="null" xml:id="ANW○
33○_kat_1_noun">
 <form type="lemma">
   <orth m:e="Lemmavorm">kat</orth>
 </form>
...
     <seg m:e="Woordrelaties">
       <seg m:e="Hyperoniem">
         <seg m:e="link">
           <xr m:e="lemma" type="related">
             <ref>zoogdier</ref>
           </xr>
```

5.14 Note

note	free text remark that can accompany any element in a lexicographic resource

INT Algemeen Nederlands Woordenboek (proprietary XML)

Extract from the entry *paasweekeinde* 'Easter weekend' in the Algemeen Nederlands Woordenboek.

Input

```
          ...
        <definitieBody id="el10">
          <Definitie>weekeinde waarin de paasdagen vallen;
weekeinde van Pasen; paasweekend</Definitie>
          <Definitieaanvulling>
            <Opmerking>In tegenstelling tot gewone weekeinden
bevat het paasweekeinde ook de maandag (tweede
paasdag).</Opmerking>
          </Definitieaanvulling>
        ...
        </definitieBody>
```


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.

```
...
```

Output

```
...
   <def>weekeinde waarin de paasdagen vallen; weekeinde van
Pasen; paasweekend</def>
   <note m:e="Opmerking">In tegenstelling tot gewone weekeinden
bevat het paasweekeinde ook de maandag (tweede paasdag).</note>
   ...
```

5.15 References

Data used in the examples in the appendix comes from the following dictionaries:

Acronym of the Institution	Full name of the dictionary	Online version
DSL	Den Danske Ordbog	https://ordnet.dk/ddo
INT	Algemeen Nederlands Woordenboek	http://anw.ivdnt.org/search
INT	Woordenboek der Nederlandsche Taal	https://gtb.ivdnt.org/search/
JSI	Slovene Lexical Database	https://www.clarin.si/repository/xmlui/handle/11356/1030 http://eng.slovenscina.eu/spletni-slovar/leksikalna-baza
RAE	Dictionary of the Spanish language (22nd edition) ('Diccionario de la Lengua Española 22a. ed.')	https://dle.rae.es/
K Dictionaries	K Dictionaries - Global French-Russian Dictionary	

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.