Clariah-WP3

Integrating Diachronous Conceptual Lexicons
through Linked Open Data

Conversion Scripts Documentation

Author:
Isa Maks (isa.maks@vu.nl)

Version:
1.0

July 13, 2016

Contents

1 Introduction

This document describes java scripts that convert xml input to RDF triples (trig) output, for the Clariah-WP3 project *Integrating Diachronous Conceptual Lexicons through Linked Open Data* The project aims at collecting, publishing as LOD different existing lexicons, that were previously in their proprietary formats and - in some cases - hard to discover. In a second phase, it aims at linking the underlying ontologies of the different resources. The first step consists of the conversion of the data to one common data model that follows the LEMON - Lexicon Model for Ontologies standard (Lemon) in combination with the Lexicon Markup Framework (http://www.lexicalmarkupframework.org). This document makes reference to 2 scripts that convert different sets of data to the common LOD model. As input formats differ per resource, the scripts probably need to be adjusted to accommodate other resources.

2 conversion script 1: "flat"

This scripts converts entries that are related to an classification or ontology by being related to one or more related classes. The classes are not hierarchically organised. It can be used to convert hisco data (https://datasets.socialhistory.org/dataset.xhtm and data with similar content.

2.1 script

- name of the script: xml2lod_flat.jar , java code
- java -jar xml2lodflat.jar [filename]
- the first argument [filename] is the first part of the inputfile [name].xml that should be in the same directory as the jar file. The name should not be too long and meaningful as it is also used as a prefix in the output data.
- the output file containing the generated RDF-triples

[0]http://www.w3.org/2016/05/ontolex/

2.2 inputfile

This is an example of an entry inputfile. The original data set needs to be converted to this format.

```
<entries>
  <entry id="9946">
        <lemma>vlasboer</lemma>
        <pos>noun</pos>
        <classes>
                <class type="broader">Agriculture, animal husbandry and forestry w
                <class type="broader">Beroep</class>
                <class type="broader">61110</class>
        </classes>
        <usages>
                <usageGeoTime>
                        <geographicAreas>
                                <geographicArea>Netherlands</geographicArea>
                        </geographicAreas>
                        <usagePeriod start="1800" end="2000" />
                </usageGeoTime>
        </usages>
        <otherForms>
                <otherForm>vlas boer</otherForm>
                <otherForm>vlasboeren</otherForm>
        </otherForms>
        <provenance>https://datasets.socialhistory.org/dataset.xhtml?persistentId=
  </entry>
</entries>
```

- entry id: unique identifier (copied from source)
- lemma:canonical form
- pos: part of speech
- classes: contains a set of classes the word (in this meaning) is related to. *type* specifies the relationship (e.g. broader, related)
- usages: combinations of reference to time periods and locations in which this word (in this meaning is used). It refers to both both *lemma*

and *otherForm*. A *usageGeoTime* instance may have zero or more *geographicalArea* and zero or one *usagePeriod*. An entry may have zero or more *usageGeoTime* instances.

- provenance: the original source of the data
- otherForms: spelling and form variants

2.3 outputfile

The output files contains rdf triples in trig format. The example is based on the preceding xml input.

```
hisco:entry-9946 a ontolex:LexicalEntry .
hisco:entry-9946 ontolex:canonicalForm "vlasboer" .
hisco:entry-9946 rdf:label "vlasboer"@nl .
hisco:entry-9946 lexinfo:partOfSpeech lexinfo:noun .
hisco:entry-9946 ontolex:otherform "vlas boer"  .
hisco:entry-9946 ontolex:otherform "vlasboeren"  .
hisco:entry-9946 ontolex:sense hisco:sense-9946 .
hisco:sense-9946 a ontolex:LexicalSense .
hisco:sense-9946 ontolex:reference hisco:concept-9946 .
hisco:sense-9946 ontolex:usage hisco:hisco:usage-99460 .
hisco:hisco:usage-99460 a ontolex:LexicalUsage .
hisco:hisco:usage-99460 time:hasBeginning time:Instant-1800 .
time:Instant-1800 time:year "1800"^^xsd:byte .
hisco:hisco:usage-99460 time:hasEnd time:Instant-2000 .
time:Instant-2000 time:year "2000"^^xsd:byte .
hisco:usage-99460 lemon-cltl:geographicalArea "Netherlands" .

hisco:concept-9946 a skos:concept .
hisco:concept-9946 ontolex:isReferenceOf hisco:sense-9946 .
hisco:concept-9946 skos:broader hisco:class-Agriculture,_animal_husbandry_and_fore
hisco:concept-9946 skos:broader hisco:class-Beroep .
hisco:concept-9946 skos:broader hisco:class-61110 .
```

3 conversion script 2: wordnet

This script converts standard wordNet-LMF (XML) to RDF triples (trig).

In addition to the previous script an element "lexical concept" is added that corresponds to the synsets. For more details http://www.w3.org/2016/05/ontolex/ (section 3.6)

3.1 script

- name of the script: odwnrbn2lod.jar , java code
- java -jar odwnrbn2lod.jar [filename]
- the first argument [filename] is the first part of the inputfile [name].xml that should be in the same directory as the jar file. The name should not be too long and meaningful as it is also used as a prefix in the output data.
- the output file containing the RDF triples

3.2 inputfile

The inputfile is in the format WordNet-LMF; ODWN (open dutch wordnet) can be found here https://github.com/MartenPostma/OpenDutchWordnet