2018-01-18

# FoLiA: Format for Linguistic Annotation

Maarten van Gompel & Ko van der Sloot

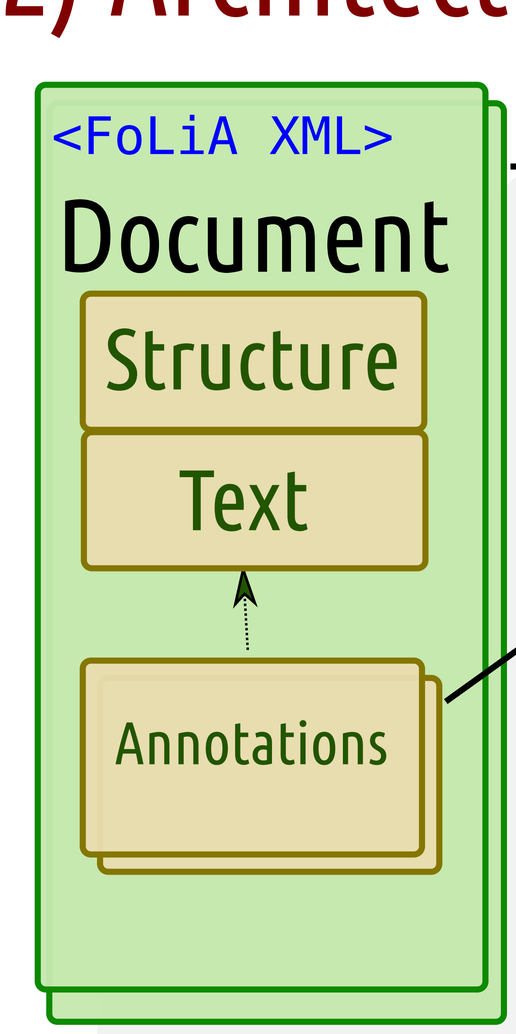
Centre for Language and Speech Technology - Radboud University Nijmegen

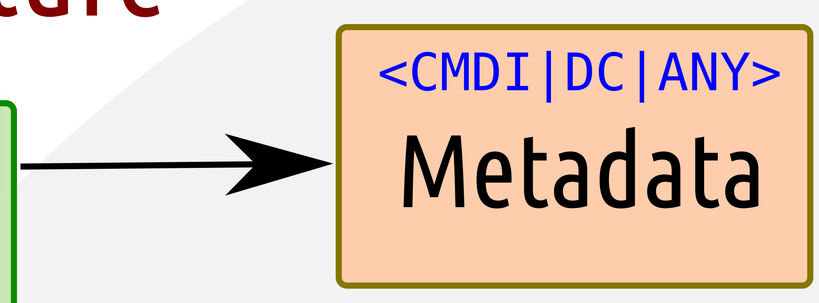
## 1) Features

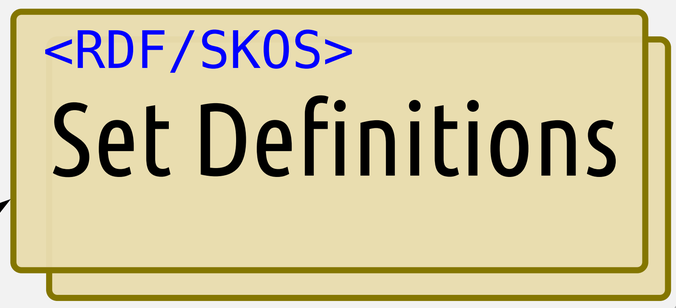
https://proycon.github.io/folia

- Rich, unified, formalised, XML-based document format for the representation of linguistically annotated resources (incl. corpora)
- Facilitates resource **exchange** & **interoperability**
- **Specific** support for many linguistic annotation types
- Language & vocabulary (tagset) agnostic: external **set definitions**
- Focus on **practical** usability: lots of open-source **tools** and libraries
- In active use & development for over 7 years

## 2) Architecture




```
<w xml:id="DOC.p.1.s.1.w.1">
 <t>houses</t>
 <pos set="https://.../pos.set"
  class="N"
  annotator="Maarten van Gompel"
  annotatortype="manual"
  confidence="0.99" />
  <lemma class="house" />
</w>
```

```
<s><t>he greets her</t>
 <w xml:id="w.1">
   <t>he</t></w>
 <w xml:id="w.2">
   <t>greets</t></w>
 <w xml:id="w.3">
   <t>her</t></w>
 <syntax>
  <su class="s">
   <su class="pron">
    <wref id="w.1" />
   </su>
   <su class="vp">
    <su class="verb"
     <wref id="w.2"
    <su class="pron"
     <wref id="w.3"
   </su></su>
 </syntax></s>
```

**Structure annotation**

Document text/speech structure & tokenisation

**Inline annotation & stand-off (span) annotation**

various predefined types of linguistic annotation

**Higher-order annotation**

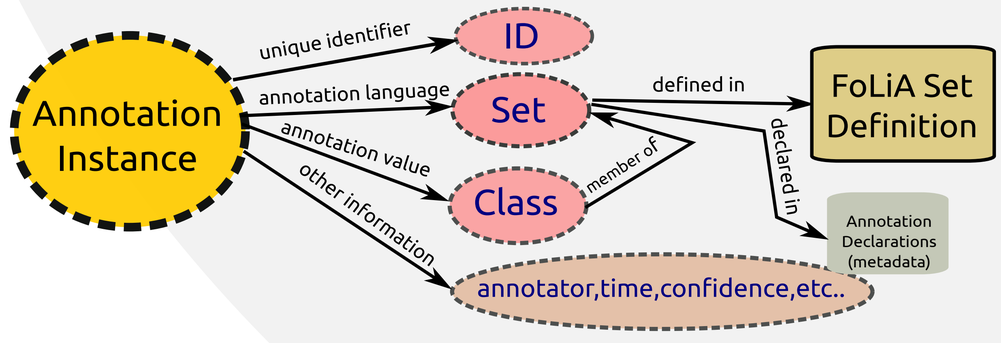
Annotations on annotations

Links to external resources (incl. linked open data)

**Generic Attributes**

ID, set/class and rich annotator & time information;

```
<pos class="noun">
 <feat subset="gender"
  class="masc" />
 <feat subset="number"
  class="sg" />
</pos>
```




LAMA

## 3) Tools

- Validators
- Converters
- Visualisation
- Analysis/statistics
- Libraries (Python, C++)

## 4) Software & Data

- NLP tools (Frog, ucto, TICCL, etc..)
- search tools (MTAS, Black-/Whitelab)
- annotation editors (FLAT)
- corpora (SoNaR, Nederlab, Basilex,etc.)

## 5) Python Library

```
$ pip install pynlpl
```

```
> from pynlpl import folia
> doc = folia.Document(
              file="doc.xml")
> for word in doc.words():
        ....
```

 Funding by CLARIN-NL & CLARIAH