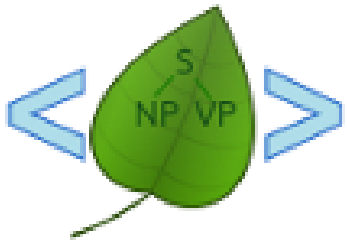
FoLiA: Format for Linguistic Annotation

FoLiA: Format for Linguistic Annotation

Current state of affairs

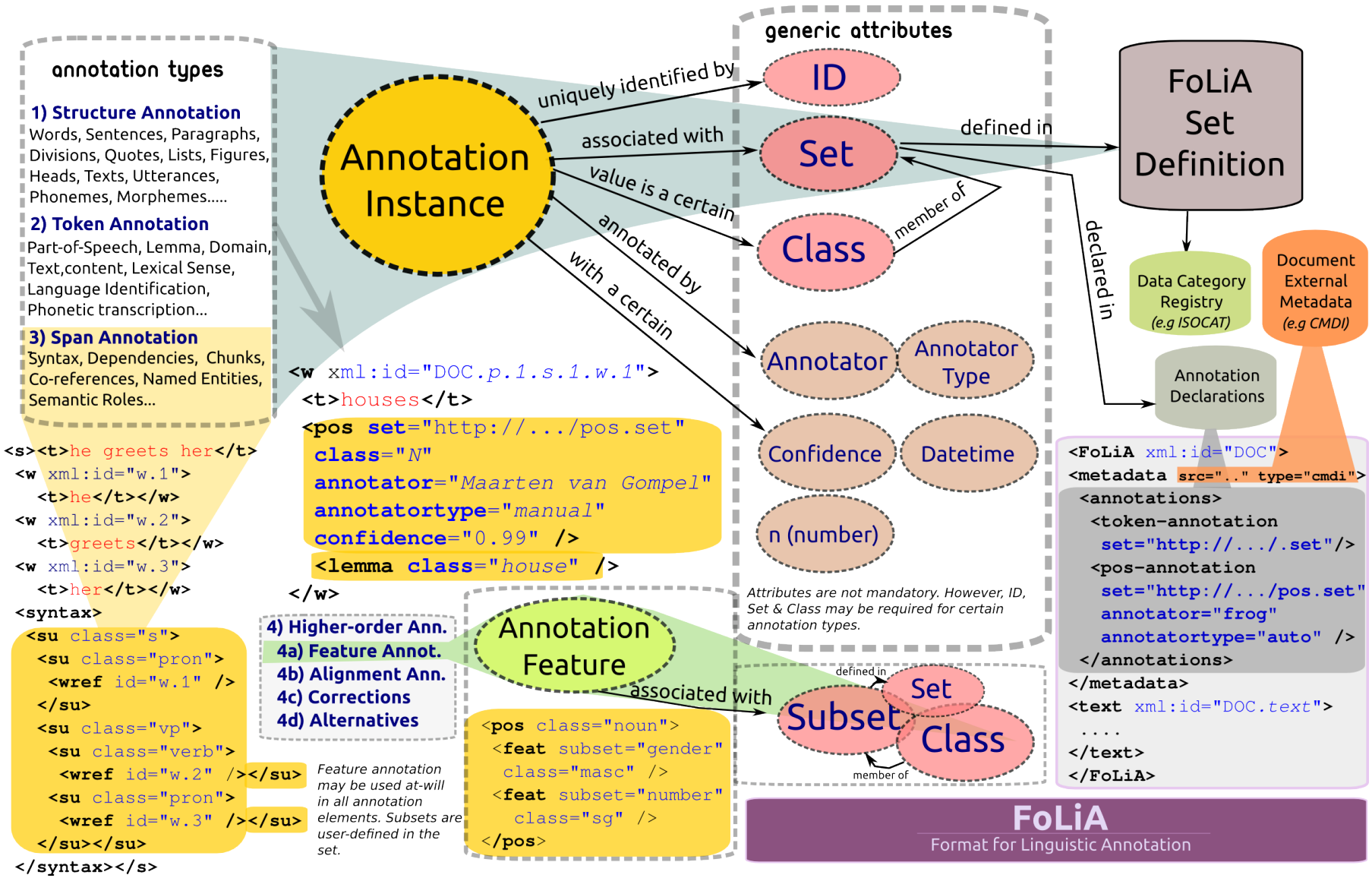


Intended Applications

- as a corpus **storage** format
- as a language resource **exchange** format

FoLiA Characteristics

- **Uniform** setup – Consistent and uniform paradigm unifying various kinds of annotation. Not commited to any label set.
- **Extensible & Flexible** – Easy to extend
- **Expressive** – Verbose expression of annotations, their annotators, timestamps, etc... Moreover, support for *alternative* annotations.
- **Formalised** – Validation on shallow (structural) or deep level. The latter validates the label set and allows for links with for data category registries (e.g. ISOcat).
- **Practical** – Bottom-up development alongside libraries, various applications, different projects.


annotation types
1) Structure Annotation
Words, Sentences, Paragraphs, Divisions, Quotes, Lists, Figures, Heads, Texts, Utterances, Phonemes, Morphemes.....
2) Token Annotation
Part-of-Speech, Lemma, Domain, Text,content, Lexical Sense, Language Identification, Phonetic transcription...
3) Span Annotation
Syntax, Dependencies, Chunks, Co-references, Named Entities, Semantic Roles...
Annotation Instance
uniquely identified by
associated with
value is a certain
annotated by
with a certain
generic attributes
ID
Set
Class
Annotator
Annotator Type
Confidence
Datetime
n (number)
defined in
member of
declared in
FoLiA Set Definition
Data Category Registry (e.g ISOCAT)
Document External Metadata (e.g CMDI)
Annotation Declarations
Attributes are not mandatory. However, ID, Set & Class may be required for certain annotation types.
<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>
<su class="pron"><wref id="w.3" /></su>
</su></su>
</syntax></s>
<w xml:id="DOC.p.1.s.1.w.1">
<t>houses</t>
<pos set="http://.../pos.set" class="N" annotator="Maarten van Gompel" annotatortype="manual" confidence="0.99" />
<lemma class="house" />
</w>
4) Higher-order Ann.
4a) Feature Annot.
4b) Alignment Ann.
4c) Corrections
4d) Alternatives
Annotation Feature
associated with
Subset
Set
Class
defined in
member of
Feature annotation may be used at-will in all annotation elements. Subsets are user-defined in the set.
<pos class="noun">
<feat subset="gender" class="masc" />
<feat subset="number" class="sg" />
</pos>
<FoLiA xml:id="DOC">
<metadata src=".." type="cmdi">
<annotations>
<token-annotation set="http://.../.set"/>
<pos-annotation set="http://.../pos.set" annotator="frog" annotatortype="auto" />
</annotations>
</metadata>
<text xml:id="DOC.text">
....
</text>
</FoLiA>
FoLiA
Format for Linguistic Annotation

Tools for working with FoLiA

- Standard **XML** facilities: XSLT, XPath
- **Python** library: pynlpl.formats.folia
- **C++** library: libfolia *(Ko van der Sloot)*

FoliaTools

Installation: `$ easy_install folia`

- **Converters**: folia2dcoi, dcoi2folia, folia2html, folia2columns, alpino2folia
- **Validator**: foliavalidator
- **Simple search tool**: foliaquery
- **Other**: foliamerge, foliafreqlist

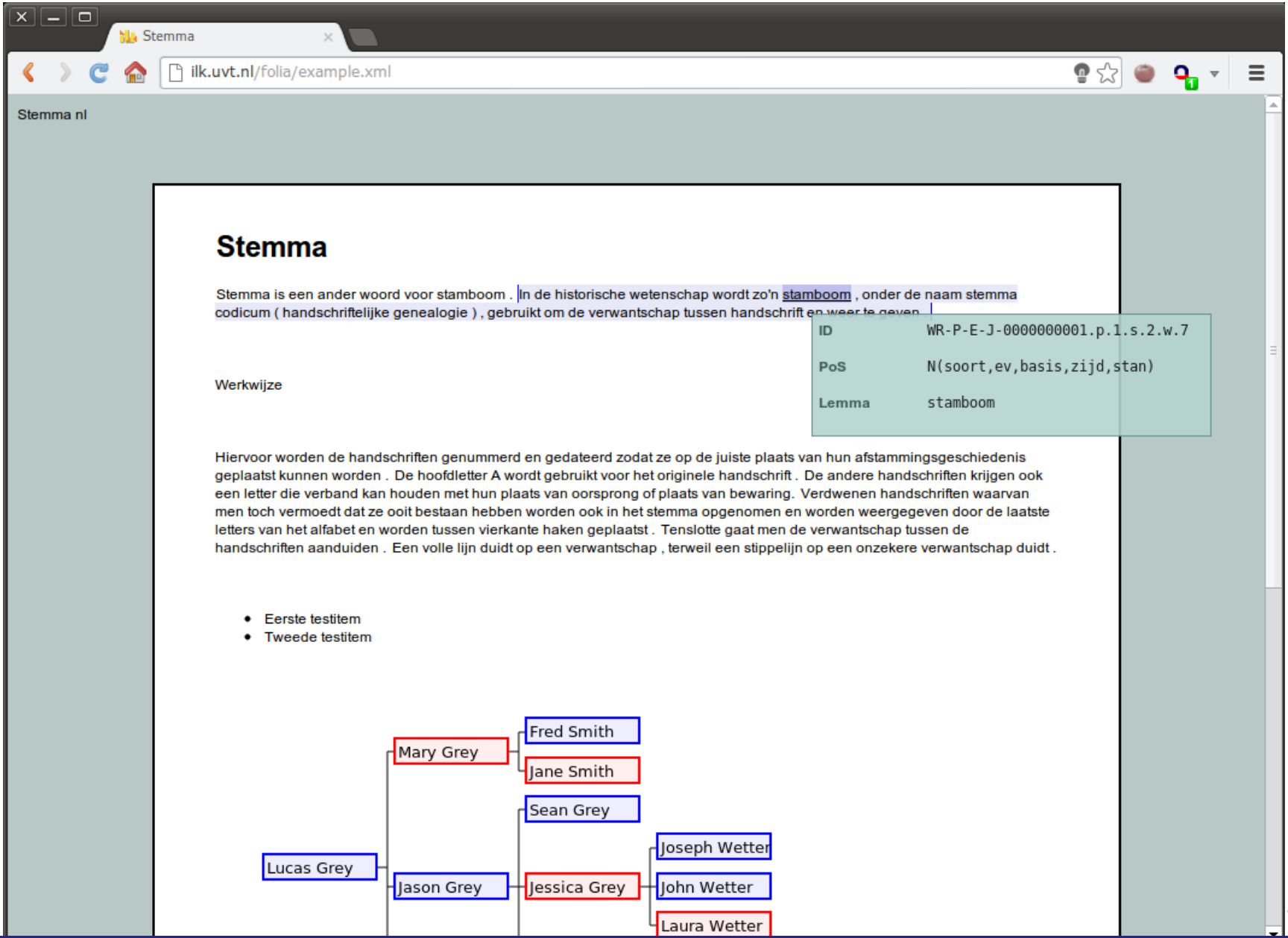
Applications using FoLiA

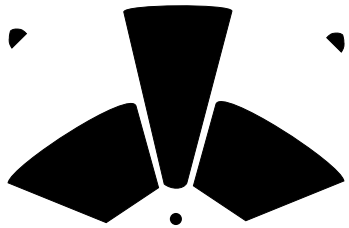
- **Frog** – tagger/lemmatisation/parser suite: FoLiA input & output
- **ucto** – tokeniser: FoLiA input and output
- **Valkuil.net** – Dutch spelling checker
- **Fowlt.net** – English spelling checker
- **Ticcl** – Spelling normalisation

Corpora delivered in FoLiA

- SoNaR *(STEVIN)*, DutchSemCor *(NWO)*, VU-DNC *(CLARIN)*, Basilex *(NWO)*

```python
from pynlpl.formats import folia

#Load FoLiA document
doc = folia.Document(file='/path/to/folia_doc.xml')

#grab a specific sentence from the index
sentence = doc['folia_doc.s.1']

#print words in sentence along with PoS and lemma
for word in sentence.words():
    print word.text()  + ", " + word.pos() + ", " +
        word.lemma()

#add PoS Annotation to a specific word
word = doc['folia_doc.s.1.w.5']
word.append( folia.PosAnnotation, class="N",
  set="http://some/url/cgn.set")

doc.save() #Save edited document
```


Stemma
ilk.uvt.nl/folia/example.xml
Stemma nl
Stemma
Stemma is een ander woord voor stamboom . In de historische wetenschap wordt zo'n stamboom , onder de naam stemma codicum (handschriftelijke genealogie) , gebruikt om de verwantschap tussen handschrift en weer te geven .
ID WR-P-E-J-0000000001.p.1.s.2.w.7
PoS N(soort,ev,basis,zijd,stan)
Lemma stamboom
Werkwijze
Hiervoor worden de handschriften genummerd en gedateerd zodat ze op de juiste plaats van hun afstammingsgeschiedenis geplaatst kunnen worden . De hoofdletter A wordt gebruikt voor het originele handschrift . De andere handschriften krijgen ook een letter die verband kan houden met hun plaats van oorsprong of plaats van bewaring. Verdwenen handschriften waarvan men toch vermoedt dat ze ooit bestaan hebben worden ook in het stemma opgenomen en worden weergegeven door de laatste letters van het alfabet en worden tussen vierkante haken geplaatst . Tenslotte gaat men de verwantschap tussen de handschriften aanduiden . Een volle lijn duidt op een verwantschap , terwijl een stippelijn op een onzekere verwantschap duidt .
Eerste testitem
Tweede testitem
Lucas Grey
Mary Grey
Jason Grey
Fred Smith
Jane Smith
Sean Grey
Jessica Grey
Joseph Wetter
John Wetter
Laura Wetter



New website: `http://proycon.github.com/folia`
Later today: FoLiA demo session including Frog, ucto, Python library.

Questions?