

CLARIAH Interoperability: FoLiA



Introduction

Introduction

FoLiA is a rich XML-based format for linguistic annotation

Intended Applications

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

Aim

- Provide one solution for a wide variety of linguistic annotation types, using a generic paradigm
- Provide one canonical and validatable serialisation (XML-based)
- Provide a **practical** solution that can be readily adopted
 - FoLiA is developed bottom-up alongside tools and libraries and extended as needed



Characteristics (1/3)

FoLiA Paradigm

Consistent and **generic** framework accross various linguistic annotation types.

- **Seperation of label semantics** from the format itself. FoLiA does not commit to any label set or language. Label sets are defined in separate *set definitions*. Annotations carry *classes* that are defined by these sets.
- **Specific** – Various specific annotation types are defined.
- **Expressive** – Verbose expression of annotations: globally unique identifiers, their annotators, timestamps, confidence values, etc... Support for *alternative* annotations, corrections of any annotation.
- **Formalised** – Validation on shallow (structural) or deep level. The latter validates the label set and allows for links with for data category registries.



Characteristics (2/3)

FoLiA Format

- **Document-based** – One document, one XML file
- Retains **structural** information (like TEI, but not as expressive), **integrated** approach.
- **Localised** – Annotations are stored in close proximity to the textual items they describe. This facilitates streaming parsers.



Characteristics (3/3)

Annotation types

Annotation types can be grouped as follows:

- **Structural Annotation** – Annotation of document structure: divisions, paragraphs, sentences, words/tokens, lists, tables, etc..
- **Token Annotation** – Annotation pertaining to a token (or other structural element): part-of-speech, lemmas, lexical semantic senses
- **Span Annotation** – Annotation spanning multiple tokens: (named) entities, dependency relations, semantic roles, coreference relations, syntactic parsers
- **Higher-order annotations** – Annotations on annotations: corrections, alternatives, comments, descriptions, references to internal or external sources (alignments), annotation of arbitrary substrings



Context: Metadata

Metadata

Documents can ...

- ...refer to external metadata files (any format)
- ...embed other XML-based metadata schemes (e.g. CMDI, Dublin Core)
- ...use FoLiA's native metadata scheme (a simple key value store)



Context: Set Definitions

Set definitions

- Documents declare what annotation types they make use of and what sets are used.
- ...declarations refer to external set definition files
- A set definition defines what classes are part of a particular label set.
- ... classes may be linked to data category registries/ontologies for formal semantic closure.
- Annotations refer to classes
- Set definitions are currently expressed in a simple XML format
- Considering moving these to RDF



Context: Converters (1/2)

Converters: To FoLiA from ...

- Alpino XML
- ALTO XML (OCR output)
- DCOI (legacy)
- HTML (via INL's *OpenConvert*, limited)
- HOCR (HTML OCR output from *Tesseract*)
- NAF (work-in-progress)
- plain text (via tokeniser *ucto* or *OpenConvert* (INL))
- Political Mashup XML
- ReStructuredText
- TEI (via INL's *OpenConvert*, limited)
- MS Word (via INL's *OpenConvert*, limited)



Context: Converters (2/2)

Converters: From FoLiA to ...

- DCOI
- HTML (for rich visualization)
- NAF (work-in-progress, by Antske)
- plain text
- plain text enriched with token annotations (limited)
- simple columned format (one token per line, limited)
- ReStructuredText
- TEI (via INL's *OpenConvert*, limited)



Context: Software Libraries

Software Libraries

- for Python: part of *PyNLPI*
- for C++: *libfolia* (by Ko van der Sloot, RU)



Context: Utilities

Utility collections

- foliatools (python based)
- foliautils (c++ based)

Utilities (CLI)

- **foliavalidator** or **folialint** - FoLiA validators
- foliacorrect (tool to deal with corrections in FoLiA)
- foliacount (compute simple statistics)
- folia-idf (IDF statistics for a directory of FoLiA documents)
- folia-langcat (language detection)
- foliamerge (merge multiple folia documents)
- foliaquery (query FoLiA documents, using FQL or CQL)
- foliacorrect (tool to deal with corrections in FoLiA)
- various of the earlier mentioned converters



Context: FoLiA-aware Software (1/3)

FoLiA-aware NLP Software

- **ucto** - Tokeniser, multilingual, can read and produce FoLiA
- **Frog** - NLP Suite for dutch: tokenisation (via ucto), PoS-tagging, lemmatisation, morphological analysis, shallow parsing, named entity recognition, dependency parsing. Can read and output FoLiA.
- **Gecco** - Generic Environment for Context-aware Correction of Orthography: spelling correction software, backend for **Valkuil.net**. Outputs FoLiA.
- **TICCL** - Text-induced Corpus Cleanup: Text normalisation and post-OCR correction system. Outputs FoLiA.
- **Colibri Core** - Pattern counting and extraction, limited FoLiA input support



Context: FoLiA-aware Software (2/3)

Search software

- **BlackLab** - Corpus retrieval engine with (limited) indexing support for FoLiA (INL)
- **WhiteLab** - Web application for the search and exploration of large corpora with linguistic annotations (Matje van de Camp, TaalMonsters)
- **PaQu** - Parse and search in syntactically annotated corpora (Groningen University)



Context: Folia-aware Software (3/3)

FLAT (FoLiA Linguistic Annotation Tool)

- A web-based collaborative annotation tool
- Eventually aims to support annotation for all annotation types supported in FoLiA.
- Used for multiple projects; recently adopted by PARSEME for annotation of MWEs as well
- backend: FoLiA Document Server (*foliadocserve*)
- communication: **FoLiA Query Language (FQL)**
- Public installation: <http://flat.science.ru.nl>



Context: Folia-aware Software (3/3)

Correction set:
<https://raw.githubusercontent.com/namedentitycorrection...>

Legend - Named Entity
(hide)

- Location
- Person
- Miscellaneous
- Organisation
- Product
- Location - Nature elements
- Location - Building

OPERATION MARKET GARDEN

18th Sept 1944 1100 hours **Saltby Aerodrome**

Emplaned on **US Dakota aircraft** for **Arnhem**. Some Gliders observed ditched in **English Channel**. Anti-aircraft 1400 hours Dropped on DZ (Ginkel Heath) Light anti-aircraft and machine gun fire as we dropped. DZ was be twisted leg on landing so I took over as **No 1 Rendezvous** was on edge of **DZ** which we both made.

1500 hours Moved along railway line in direction of **Arnhem** then through **Wolfheze**. We were held up by enemy houses to try to take out enemy. **Nick** and I were on top of a garden wall and about to drop when enemy machine great speed. We found it impossible to proceed via the rear of the houses as the **Germans** had it well covered. short time but no enemy seen. Orders were given to withdraw back along **railway line**. We bedded down in the **Bren Gun**. Enemy were firing tracer bullets trying to attract our fire so they could find our positions. **I lay pylon above my head**. I was relieved after about 2 hours. And was glad to rejoin my comrades. Attempted to mortar bomb.

19 Sept

In morning ad could identify Believe there separated from with **Sgt Sheph** from our front

No sooner had was. **I** received think now for was spotted and realise now ho

was brought in and placed beside me. **I** remember seeing him in no different terms about being dumped off t to speak as I did under normal conditions. **I** had every respect for the **RSM**. A morphine injection quietened m

They tried to evacuate us on a **Bren gun carrier** but when we went up the track a warning was shouted that a retreated, the 88 shell just missed. We were later evacuated by jeep to the **Schoonoord Hotel** in **Oosterbeek**. **I** hotel opposite. My stretcher was placed in a hall at the side of the hotel. The stretchers were so close that the biscuit, sweet and cup of tea. Every now and again the firing ceased outside. I believe this was the respect for to the one opposite with operating theatre. An enemy tank put a shell through the roof of our hotel and late was a hospital. Eventually German troops filtered in and we realised the end was coming.

An SS Sergeant Major came in and passed round German cigarettes while smoking **British** from our supply dr saying it was the fault of the **Jews** that we were fighting. **The Dutchman, Steve**, I think his name was – agreed been acting as a guide for us, he was also a Jew and a very brave one too. **The Medics** eventually convinced t fighting and had been wounded. They released him much to our joy.

In the final stages of the battle we were transported in **German Army** ambulances to the **King William III Barr** earliest paras moved were taken by cattle trains but I was one of the lucky ones and was taken by ambulance **Germans** were unloading tanks from trains in the sidings. **I** presume we were left there to deter the **RAF** or **Yai** train and I remember the old German **Matron** in charge ruled the German **Medics** with a rod of iron but was ki

10 October 1944

Text
undefined

Named Entity
<https://raw.githubusercontent.com/pragoc/folia/master/sets/definitions/namedentities/foiaaset.xml>

Gun

Product

Bren Gun

Correction: wrongclass

Original **Organisation**

Named Entity: **Bren Gun**

Original Text: **Bren**

Original Text: **Bren**

Original Text: **Gun**

Original Text: **Gun**

Discussion: Interoperability

Aims for the future

- Continued work to extend the format and libraries on a need-to basis
- Continued support to all parties using FoLiA
- Continued work on FLAT
- Finish the FoLiA/NAF conversion tools
- Migrate set definitions to RDF?
- What is the desirability of more RDF support? I.e. RDF schema for FoLiA and conversion to RDF for annotated documents. Use cases?



Discussion: Questions?

