FoLiA: Maintenance, support, and continued development

ID: T108

Author: Maarten van Gompel

Introduction:

FoLiA provides a rich format for linguistic annotation which serves as an exchange format between tools and services, as well as a format for corpus storage. Alongside the data format, a rich infrastructure of tooling is provided for working with FoLiA documents.

What exists:

FoLiA has been in existence and under continuous active development since 2011, previousy funded by CLARIN-NL and CLARIAH-CORE, and is widely used in CLARIAH and beyond. Tied to FoLiA is a vast infrastructure of libraries and tools to work with the format, other CLARIAH tasks in turn build on these.

The direct infrastructure surrounding FoLiA encompasses the format specification and documentation itself, and the following additional software:

- a FoLiA library for Python: foliapy
- a FoLiA library for C++: libfolia - a FoLiA library for C++
- a FoLiA library for Rust: folia-rust - a FoLiA library for Rust
- A collection of FoLiA command line tools (written in Python) - foliatools; these include (non-exhaustively) converters from/to other formats, validators, and tools to edit/transform or view documents.
- A collection of FoLiA command line tools (written in c++) - foliautils (similar to the above)

What must be adapted / extended / created anew:

This is an ongoing support and development task that assures:

- Support is offered to all users/developers (both within and beyond CLARIAH) with any questions about FoLiA or usage of its various tools and libraries.
- Bugfixes and implementation of feature requests in the various tools and libraries
- When needed and in close contact with users, FoLiA is extended with new annotation types
- The various FoLiA libraries are kept in sync and up to date with the latest FoLiA specification
- Further integration with the the CLARIN/CLARIAH infrastructure where there is a demand

This task is characterised by relatively short development cycles springing from user requests or own insights

Why important for CLARIAH (scientific impact): FoLiA is being used throughout CLARIAH, by both tools and data sets, and is an important facility to maintain for the future.

Targeted/Actual users: tool-developers, data-providers, data scientists

Actual use (quantify!): Used by almost all tools provided by Nijmegen, but also beyond, used by many corpora (SoNaR-500, Basilex, Basiscript, DutchSemCor, OpenCGN, VU-DNC, Nederlab), used by lots of other partners.

Allocation

Lead: Maarten van Gompel (DI, KNAW)

Participants: Maarten van Gompel (DI, KNAW)

Estimated needed PMs: 12PM (spread over the entire duration of the CLARIAH-PLUS project, and originally shared between two participants)

Actually Allocated PMs: 6PM (note: this was 6PM less than initially proposed!)

Deliverables

1. (T108D1) Documentation: FoLiA Documentation and Reference Guide
2. (T108D2) Data: FoLiA Schemas and Specification
3. (T108D3) Software: FoLiA Library for Python: FoLiAPy
    - (T108D3.1) Documentation: Extensive documentation including API reference
4. (T108D4) Software: FoLiA Library for C++: libfolia

    - (T108D4.1) Documentation: API reference
5. (T108D5) Software: FoLiA Library for Rust: folia-rust
    - (T108D5.1) Documentation: API reference
6. (T108D6) Software: FoLiA Tools (Assorted command-line tools for FoLiA, Python-based)
7. (T108D7) Software: FoLiA Utilities (Assorted command-line tools for FoLiA, C++)
8. (T108D8) Software: FoLiA profiler for CLARIN Switchboard/weblicht
9. (T108D9) Software: Piereling Webservice - A webservice to convert various document formats from/to FoLiA. Builds upon foliatools and foliautils.
10. (T108D10) Service: Piereling Webservice deployed at CLST.

Milestones

Due to the ongoing nature of this task, milestones are often defined as the task is ongoing and completed in short development cycles.

1. (T108M1) FoLiA v2.0 - Early 2019 **(COMPLETED)** - FoLiA v2.0 is a major update of the FoLiA format, introducing various changes, most notably support for provenance data and completely renewed documentation.
    - folia#43 Completely revised FoLiA documentation, turn into more formal specification
    - folia#46 Proper support for data provenance logging
    - folia#51 Increases expressivity for multi-word annotations
2. (T108M2) FoLiA v3.0 - This is a hypothetical future milestone that would propose changes to FoLiA for better alignment with the CLARIAH infrastructure; these tasks are not defined yet and need to arise from discussion (in e.g. the CLARIAH Interest Groups) and practical experience. Possible solutions could be a new serialisation form that integrates nicely into the world of linked open data, and that deploy the FoLiA paradigm on top of other existing standards (web annotations perhaps?).

Progress Reports

Detailed progression of this task is logged as part of our regular progress reports .

Changes

Changes with respect to the earlier plan:

- More extensive description and motivation of the task
- More explicit deliverables and milestones
- Marked various earlier deliverables and milestones as completed
- The FoLiA library for Rust is a new addition, and of a slightly more experimental nature. In this library we learn from insights adopted over the years and aim for a more high-performance library.
- Since 14 July 2020 - This task has moved from CLST, Radboud University Nijmegen to Digital Infrastructure, Humanities Cluster, KNAW. The webservice deployments, however, stay at CLST.
- Ko van der Sloot has been an indispensible asset for this task hitherto, but has now retired. Maarten van Gompel takes on maintainance and support of his software developed as part of this task (libfolia, foliautils).

Related tasks

This task is needed for:

- Frog (T139)
- FLAT (T062)
- PICCL

References

- M. van Gompel (2019). FoLiA: Format for Linguistic Annotation - Documentation and Reference Guide. Language and Speech Technoogy Technical Report Series 19-01. Radboud University - https://folia.readthedocs.io
- M. van Gompel, K. van der Sloot, M. Reynaert and A. van den Bosch (2017). FoLiA in Practice: The infrastructure of a Linguistic Annotation Format. In: CLARIN in the Low Countries. pp 71-82 - http://www.jstor.org/stable/j.ctv3t5qjk.13

- M. van Gompel (2014). FoLiA: Format for Linguistic Annotation - Documentation and Reference Guide. Language and Speech Technoogy Technical Report Series 14-01. Radboud University
- M. van Gompel and M. Reynaert (2013). FoLiA: A practical XML Format for Linguistic Annotation - a descriptive and comparative study. Computational Linguistics in the Netherlands Journal. - http://clinjournal.org/sites/clinjournal.org/files/05-vanGompel-Reynaert-CLIN2013.pdf