Software Release and Progress Report

Start date: 2020-06-03 End date: 2020-10-13

Short Summary

At the beginning of this period, Maarten van Gompel moved from CLST, Radboud University Nijmegen to DI, Humanities Cluster, KNAW for 0,6 fte. This means the CLARIAH WP3 tasks under his lead are moved as well. A permanent position at CLST for 0,2 fte is retained for system administration, which focusses on guaranteeing the hosting at CLST of numerous CLARIAH services.

Notable developments:

- Contributed to the start-up of the CLARIAH Interest groups
- Released FoLiA v2.3.0 which supports a new explicit form of XML serialisation
- The tei2folia converter has been extended to support more of TEI
- Deepfrog: first experimental release; featuring transformer-based neural models for PoS, NER and lemmatisation
- LaMachine updates: integrating more software in collaboration with Stichting Open Spraaktechnologie also added deepfrog, significant refactoring and fixes.

I redefined the task plans as requested, defining more deliverables and milestones. The plans themselves also give an indication of progress. I'm trying to get the plans better aligned with the actual reality. Some remarks/concerns:

- Most of the tasks are maintenance & support tasks, intended to run for the entire duration of CLARIAH to ensure users of the software are ensured of support and fixes.
- The current time allocation does not neatly align with contract situation; all time allocations were condensed and HuC contract runs out long before the end of CLARIAH-PLUS, whilst the original idea was for these maintenance & support tasks to be stretched out over the full project duration.
- Because of the rolling nature of the tasks, defining milestones ahead of time is often not possible.
- The LaMachine task was formally allocated far too little time and exceeds this.
- DeepFrog was not planned initially but is a logical continuation of Frog. Its conception was supported by Henk and Antal; I moved it under T139 (Frog). However, time is not formally allocated for this.

CLAM

Project & Task ID: CLARIAH-PLUS WP3 T142

clam v3.0.18

Bugfix release:

- Fix in data API: correctly parse optional attribute from CLAM XML

(Released by Maarten van Gompel on 2020-09-30) https://github.com/proycon/clam/releases/tag/v3.0.18 *(deliverable ID: T142D1)*

clam v3.0.17

- adding all static assets inside the proejct rather than relying on fetching them from other origins, removes dependencies on cloudflare and CDNs

(Released by Maarten van Gompel on 2020-07-20) https://github.com/proycon/clam/releases/tag/v3.0.17 *(deliverable ID: T142D1)*

clam v3.0.16

- Added FORCEHTTPS configuration directive, as a lighter alternative to using the full FORCEURL option.

(Released by Maarten van Gompel on 2020-06-25) https://github.com/proycon/clam/releases/tag/v3.0.16 *(deliverable ID: T142D1)*

clam v3.0.15

- Made it possible to set a maximum number of MBs of data a user can have in running projects. #92
- Made it possible to set a maximum number of running projects per user. #91

(Released by Maarten van Gompel on 2020-06-09) https://github.com/proycon/clam/releases/tag/v3.0.15 *(deliverable ID: T142D1)*

clamservices v2.2.2

Minor update: removed the hard coded preconfigured flat url (it moved) and make things flexible the work with or without flat (let LaMachine handle it)

(Released by Maarten van Gompel on 2020-10-01) https://github.com/proycon/clamservices/releases/tag/v2.2.2

FLAT

Project & Task ID: CLARIAH-PLUS WP3 T062

foliadocserve v0.7.3

automatically convert document directory to an absolute path

(Released by Maarten van Gompel on 2020-06-04) https://github.com/proycon/foliadocserve/releases/tag/v0.7.3 *(deliverable ID: T062D2)*

FoLiA

Project & Task ID: CLARIAH-PLUS WP3 T108

folia v2.3.0

- Added the possibility of serialising FoLiA to *explicit form*. Explicit form is a more verbose XML serialisation that makes assumptions that are usually implicit in FoLiA (such as defaults and element categories) explicit in the output. This facilitates the job for parsers who do not implement the full FoLiA logic. This is meant to be used as an alternative serialisation only in cases where it makes sense (to support such 3rd party parsers). #84
- Documentation and README updates:
  - added the new rust library, amended implementation list
- Added new examples and fixed some existing examples
- Some added flexibility in certain nested of structural elements;
  - allow Word directly under Division
  - allow Linebreaks in tables, figures and lists (outside of items, rows/cells), because these are sometimes used to denote pagebreaks in multi-span tables/figures/lists.

(Released by Maarten van Gompel on 2020-09-02) https://github.com/proycon/folia/releases/tag/v2.3.0 *(deliverable ID: T108D1)*

folia-rust v0.0.5

Release featuring minor fixes/cleanup

(Released by Maarten van Gompel on 2020-09-29) https://github.com/proycon/folia-rust/releases/tag/v0.0.5 *(deliverable ID: T108D5)*

folia-rust v0.0.4

- Updated for FoLiA v2.3 (can handle explicit form now)

(Released by Maarten van Gompel on 2020-09-02) https://github.com/proycon/folia-rust/releases/tag/v0.0.4 *(deliverable ID: T108D5)*

folia-rust v0.0.3

This release adds some essential features to the API: * Implemented a high-level annotate() method analogous to foliapy's add() method. Handles span annotation transparently. * implemented common_ancestors() * implemented get_layer_key() * make sure layers reverse-inherit their set from their children * finished provenance support, an active processor can now be associated with a document * datetime attributes are now chrono::NaiveDateTime structs and properly parsed, rather than a plain String

(Released by Maarten van Gompel on 2020-08-12) https://github.com/proycon/folia-rust/releases/tag/v0.0.3 *(deliverable ID: T108D5)*

foliapy v2.3.0

- Implements FoLiA v2.3
  - Adds support for the explicit form serialisation (proycon/folia#84)
- Bugfixes
  - also serialize when `confidence=0` #23
  - fix missing t-ref/@id attribute #19
  - also perform text validation on string elements #15
  - Implemented extra checks for spurious text
  - added requests library as an explicit dependency
  - fixed tests for new xmldiff version

(Released by Maarten van Gompel on 2020-09-02) https://github.com/proycon/foliapy/releases/tag/v2.3.0 *(deliverable ID: T108D3)*

foliatools v2.3.1

Bugfix release: * [txt2folia] Prevent adding empty text content (#14)

(Released by Maarten van Gompel on 2020-09-11) https://github.com/proycon/foliatools/releases/tag/v2.3.1 *(deliverable ID: T108D6)*

foliatools v2.3.0

- The **tei2folia** converter has been extended to support more of TEI
  - Implements conversion of tokens, sentences and simple linguistic annotation (@pos,@lemma,@join,@msd) (#12 #13)
  - better document ID detection, prefer DOI, then ISSN, then ISBN, then DTADirName (specific to Deutsches Text Archiv), fall back to untyped but check we get something sane out of it. #12
  - implemented conversion of @norm attribute (not sure if this is entirely according to TEI P5 spec but Deutsches Text Archiv uses it.
  - Benefit from some of the newly allowed structural nestings in folia v2.3
  - Implemented handling for tei:trailer and some other elements
  - Ignore styling that is wrapped around structural elements (for now)
  - Added extra sanity checks
- foliavalidator now implements the ability to output to **explicit form** (proycon/folia#84). Explicit form is a more verbose XML serialisation that makes assumptions that are usually implicit in FoLiA (such as defaults and element categories) explicit in the output. This facilitates the job for parsers who do not implement the full FoLiA logic. This is meant to be used as an alternative serialisation only in cases where it makes sense (to support such 3rd party parsers).
- Various fixes for `foliatextcontent`
- implemented a first version of a FoLiA to Salt converetor (proycon/folia#85). This is still in an experimental stage. Salt is a graph based model that acts as an intermediate model in their conversion tool Pepper. This folia2salt convertor in combination with pepper allows users, in theory, to convert FoLiA to formats such as TCF, Paula XML, ANNIS and many others.
- Updated documentation with some more in-depth sections on foliavalidator, tei2folia and folia2salt
- various foliaspec updates

(Released by Maarten van Gompel on 2020-09-02) https://github.com/proycon/foliatools/releases/tag/v2.3.0 *(deliverable ID: T108D6)*

foliautils v0.15

[Maarten van Gompel]

- FoLiA-txt: check if a string is empty after normalisation (fix for #46)

[Ko vd Sloot]

- folia-correct: fix one-off error in hemp handling (when no hemp was found) #45
- some refactoring
  - centralized definition of XML_PARSER_OPTIONS
- bugfix in threading

(Released by Maarten van Gompel on 2020-09-15) https://github.com/LanguageMachines/foliautils/releases/tag/v0.15 *(deliverable ID: T108D7)*

libfolia v2.5.1

[Maarten van Gompel]

- Bugfix: Fixed handling of control characters, strip control characters by default

[Ko vd Sloot]

- fix in date handling (lookup table for month -> integer conversion)
- minor refactoring
- some documentation

(Released by Maarten van Gompel on 2020-09-15) https://github.com/LanguageMachines/libfolia/releases/tag/v2.5.1 *(deliverable ID: T108D4)*

libfolia v2.5

[Maarten van Gompel]

- Adapted to FoLiA v2.3
- Support parsing of the new explicit form

[Ko vd Sloot]

- folialint: updated usage() and man page
- minor refactoring

(Released by Maarten van Gompel on 2020-09-02) https://github.com/LanguageMachines/libfolia/releases/tag/v2.5 *(deliverable ID: T108D4)*

Frog, Ucto & DeepFrog

Project & Task ID: CLARIAH-PLUS WP3 T139

deepfrog v0.2.0

First experimental release

(Released by Maarten van Gompel on 2020-09-29) https://github.com/proycon/deepfrog/releases/tag/v0.2.0 *(deliverable ID: T139bD1)*

frog v0.21

[Ko van der Sloot]

- Fixes a problem with temporary files not being cleaned up properly #92

(Released by Maarten van Gompel on 2020-07-22) https://github.com/LanguageMachines/frog/releases/tag/v0.21 *(deliverable ID: T139aD1)*

python-ucto v0.5.2

- Fixed lowercasing/uppercasing #8
- Removed Python 2.7 support
- Added a notice that sentencedetection is a deprecated parameter, rather than silently ignoring it

(Released by Maarten van Gompel on 2020-10-09) https://github.com/proycon/python-ucto/releases/tag/v0.5.2 *(deliverable ID: T139aD2.3)*

ucto v0.22

[Ko vd Sloot]

- Fix for Byte-order Marker problem #79

(Released by Maarten van Gompel on 2020-10-08) https://github.com/LanguageMachines/ucto/releases/tag/v0.22 *(deliverable ID: T139aD2)*

LaMachine

Project & Task ID: CLARIAH-PLUS WP3 T098

codemetapy v0.3.4

Previous release was a bit premature, there was a bug related to #4 still that has now been fixed.

(Released by Maarten van Gompel on 2020-10-08) https://github.com/proycon/codemetapy/releases/tag/v0.3.4 *(deliverable ID: T098D4)*

codemetapy v0.3.3

- parse dependency versions and store them explicitly; don't stumble over extras (they will be processed as any other dependency, the 'extra' information bit does not get converted. #4
- added a `-no-extras` parameter that disregards all the extras. #4

(Released by Maarten van Gompel on 2020-10-08) https://github.com/proycon/codemetapy/releases/tag/v0.3.3 *(deliverable ID: T098D4)*

LaMachine v2.23

- Added rust #181
  - Added rust-based software sesdiff, ssam
  - Added deepfrog
- Kaldi_NL, g2pservice and forcedalignment2 are obtained from new upstream source
  - various fixes for those webservices
- added asr4J, spreek2schrijf
- Various fixes and improvements for container deployment behind a reverse proxy
  - added custom_flat_settings option
  - use a default CLAM base configuration
  - added force_https option
- Refactoring: Implemented a languagemachine-web-install role and a language-python-link role to refactor some things and remove a lot of unnecessary duplication
- better argument parsing in lamachine-update
- explicitly set ansible python interpeter when running lamachine update
- various fixes

(Released by Maarten van Gompel on 2020-10-01) https://github.com/proycon/LaMachine/releases/tag/v2.23 *(deliverable ID: T098D1)*

LaMachine v2.21

- integrated lamastats

- When calling the virtualenv activate script, an important part of activation wasmissed. Try to automatically call the primary activation script in such a case.
- Adding mitlm to phonetisaurus installation (phonetisaurus-train depends on it), and some CentOS 8 changes
- Docker fix, default interactive shell now starts with non-root and in the proper homedir

(Released by Maarten van Gompel on 2020-08-20) https://github.com/proycon/LaMachine/releases/tag/v2.21 *(deliverable ID: T098D1)*

LaMachine v2.20

- Important fixes for macOS: install ansible through homebrew rather than through pip. This hopefully fixes some recent failures to install on at least macos catalina.
- Java fix: ensure Java is version 11 (versions 13 and 14 are too new for nextflow and probably also for various other tools)
- Automatically start background servers for valkuil and tscan (using uwsgi attach-daemon), and some fixes for existing background servers
- Fixed port number clash when both PICCL and T-scan were enabled
- Adding a sites-extra directory for non-LaMachine managed nginx configuration

(Released by Maarten van Gompel on 2020-07-20) https://github.com/proycon/LaMachine/releases/tag/v2.20 *(deliverable ID: T098D1)*

LaMachine v2.19

- fixes for prebuilt VM flavour
- allow setting ssh key filename rather than forcing id_rsa

(Released by Maarten van Gompel on 2020-06-21) https://github.com/proycon/LaMachine/releases/tag/v2.19 *(deliverable ID: T098D1)*

lamastats v0.2.0

Added support for nginx, changed style

(Released by Maarten van Gompel on 2020-07-24) https://github.com/proycon/lamastats/releases/tag/v0.2.0 *(deliverable ID: T098D6)*

Miscellaneous

Project & Task ID: Dependencies/wrappers and or unforeseen tools (related to CLARIAH projects)

(no releases this period)

Nederlab

Project & Task ID: Nederlab

(no releases this period)

PICCL & TICCL

Project & Task ID: CLARIAH-PLUS WP3 ???

PICCL v0.9.4

Previous release was premature and bugged; this fixes it.

(Released by Maarten van Gompel on 2020-10-01) https://github.com/LanguageMachines/PICCL/releases/tag/v0.9.4

PICCL v0.9.3

Minor update: Added an –outputclass parameter for ticcl.nf to choose the output text class and provide extra flexibility. Set either that or –inputclass.

(Released by Maarten van Gompel on 2020-10-01) https://github.com/LanguageMachines/PICCL/releases/tag/v0.9.3

PICCL v0.9.2

- added a clearer error message with explanation in case the indexNT file is empty (related to LanguageMachines/lexiconenrichment#1)
- removed explicit flat url (let LaMachine handle it)
- minor README update

(Released by Maarten van Gompel on 2020-10-01) https://github.com/LanguageMachines/PICCL/releases/tag/v0.9.2

PICCL v0.9.1

- publish more intermediate output #58
- added a –nofoliacorrect output option to skip the final foliacorrect step

(Released by Maarten van Gompel on 2020-08-19) https://github.com/LanguageMachines/PICCL/releases/tag/v0.9.1

ticcltools v0.7.1

[Ko vd Sloot]

- changed ICU requirement to at least 5.6
- some refactoring
- started implementing a solution for #42
- added error message when the index file is empty.

(Released by Maarten van Gompel on 2020-09-15) https://github.com/LanguageMachines/ticcltools/releases/tag/v0.7.1

Software Quality Guidelines

Project & Task ID: CLARIAH-CORE WP2 Task 54.100

(no releases this period)

Timbl & Mbt

Project & Task ID: Pre-CLARIN

python-timbl v2020.06.08

yet another fix to find boost-python

(Released by Maarten van Gompel on 2020-06-08) https://github.com/proycon/python-timbl/releases/tag/v2020.06.08