

Year report 2020

Maarten van Gompel

Start date: 2020-01-01 End date: 2020-12-31

All tasks mentioned in this report initially started at CLST, Radboud Universiteit Nijmegen, but the development of these tasks has since summer 2020 been largely moved to KNAW Humanities Cluster, where I am currently employed. This doesn't make any difference content-wise, nor will it have any negative implications for our progression. Continued hosting and maintenance of the services that spring from this software is furthermore still guaranteed at CLST, Radboud University.

CLAM: Maintenance and Support (T142)

Task description and plan: [CLARIAH-PLUS WP3 T142](#)

The last explicit milestone regarding CLAM was finished in October 2019 and introduced a brand-new web interface. Nevertheless, a lot has been done in 2020 with regard to bug fixes and minor improvements. An extensive overview can be found in the appendix. CLAM has been used once more to create various new webservice, mostly services focus on speech recognition at CLST, Radboud University. In the scope of this task, support has been offered to the developers of these services.

FoLiA (T108)

Task description and plan: [CLARIAH-PLUS WP3 T108](#)

FoLiA remains an important pillar under many of our tasks. This WP3 task provides maintenance and support on a wide infrastructure of libraries and tools that deal with the FoLiA ata format, often in collaboration with end-users. We celebrate our 10-year anniversary this year. In 2020 we have done twjo major releases of the format itself, and a number of smaller intermediate releases. The appendix will show the full overview.

A limited excerpt of new features this period has been, amongst others, an implementation for modality annotation, improved text validation, improved flexibility in metadata, and the ability to split a document into multiple files.

On the front of the FoLiA-tools, we report significant improvements to the TEI-to-FoLiA converter, the ReStructuredText-to-FoLiA converter, the AbbyXML-to-FoLiA converter and the PageXML-to-FoLiA converter. Work has also been done on experimental conversion to Salt XML, which provides opportunities for linking with the German infrastructure around Salt, Pepper and the ANNIS annotation environment. A `foliasplit` tool has been introduced that makes it possible to split a FoLiA document on certain criteria.

FLAT (T062)

Task description and plan: [CLARIAH-PLUS WP3 T062](#)

This task has become mainly a maintenance and support task. There was also a lack of clarity what parts of the FLAT plan were desired from a wider WP3-perspective and the time we got allocated for the task was limited so choices had to be made. There appears a clear interest in FLAT from the research community, and I support requests from these take precedence above any other plans we may have had. The task description (linked above) will further explain this. We currently have two prominent use-cases for FLAT: [1](#) [2](#)).

In 2020, new releases have been done for both the frontend and the backend, see the appendix for a complete overview. The most notable release addresses an issue that had long been planned; the correct visualisation of PICCL/TICCL output (another CLARIAH task).

Frog, Ucto en DeepFrog (T139)

Task description and plan: [CLARIAH-PLUS WP3 T139](#)

The lead developer of Frog (and also co-developer of FoLiA) has retired in spring 2020. Before that, a last major Frog version was released that included various improvement. In doing so, we have delivered a stable tool with a long development history. Frog is a popular tool used by many in the Dutch and Flemish NLP communities. Continued support is guaranteed despite the retirement of the lead developer, because I have taken over maintenance and support.

This also applies to the tokeniser ucto, a dependency which we include as part of this task. This WP3 task now focusses primarily on bugfixes and maintenance, further innovative development of Frog has come to an end.

Work has also been started on a possible successor for Frog that employs more state-of-the-art deep learning techniques. We call it DeepFrog. This, however, exceeds the allocated time and funding for this task. Nevertheless, we have delivered a first experimental version of DeepFrog.

LaMachine (T098)

Task description and plan: [CLARIAH-PLUS WP3 T098](#)

LaMachine is the umbrella under which we make available a lot of our software and the shape in which we offer it as a bundle to the larger community both within and outside of CLARIAH. It is being used both by various individual researchers and developers, as well as by hosting partners to offer our tools as a service. We notice an increasing interest in LaMachine also from within CLARIAH and other work packages within it, as well as by external organization, such as for example Stichting Open Spraaktechnologie. LaMachine is used [relatively often](#).

This task is a support and maintenance task that takes up significantly more time than originally anticipated. Moreover, significantly less time was allocated than we originally asked for. The VRE project, described in the next section, has close relations with this task and compensates for this to a certain extent.

The appendix once again provides a detailed overview of the ten or so LaMachine releases that have seen the light in this year. These releases mainly consist of additions of new tools and services, upgrades of existing ones, bugfixes, and expansions to other platform. Keeping everything up to date with the perpetually changing reality is one of the main challenges for LaMachine. It is by definition a system that needs constant maintenance, as it is the system that provides the integration for various underlying components that all move in their own way and pace.

WP3 Virtual Research Environment (T137)

Task description and plan: [CLARIAH-PLUS WP3 T137](#)

In December 2020 I was asked to take on the development and coordination of the WP3 Virtual Research Environment, after I proposed a new plan for continuation of this task. This is a change of course with respect to earlier attempts in CLARIAH-CORE to shape the VRE project. Earlier attempts were highly ambitious but failed to come to any delivered results. The renewed plan is more modest in its aspirations and attempts first and foremost to focus on interconnecting existing CLARIAH software, rather than develop any new overarching infrastructure. In practice, LaMachine is the closest thing in WP3 we have to a Virtual Research Environment and was therefore chosen as the foundation for further integration.

In this task, we are also looking to connect to related projects like the CLARIN switchboard. Work has already been done to this end in the past. An important objective of the revised VRE plan is also to disseminate the results of our efforts to end-user (researchers), in the form of for example demonstration videos that show the capabilities of the research environment.

PICCL

This task is not my responsibility, it's led by Martin Reynaert, but I would like to give a small update on the progression here. With the retirement of lead-developer Ko van der Sloot, an important last release of TICCL-tools has been done and is the culmination of more than a year's worth of work, TICCL-tools forms the main dependency of PICCL, and PICCL itself has also been updated accordingly.

Appendix: Software Releases

Below you can find an elaborate overview of all release notes of all our WP3-related software released in 2020.

CLAM

Project & Task ID: [CLARIAH-PLUS WP3 T142](#)

clam v3.0.23

Minor update: * Adapted some default mime types to comply to e.g. RFC 2361 for compatibility with the CLARIN switchboard

([Released](#) by Maarten van Gompel on 2020-12-11) <https://github.com/proycon/clam/releases/tag/v3.0.23> (deliverable ID: T142D1)

clam v3.0.22

Bugfix release: * Do not serialise empty/unfilled input parameters in CLAM metadata #103 * Added utility functions isncname and makencname to check if a string is a valid XML NCName and to make it so if not.

([Released](#) by Maarten van Gompel on 2020-11-30) <https://github.com/proycon/clam/releases/tag/v3.0.22> (deliverable ID: T142D1)

clam v3.0.21

Bugfix release: Fixes for shortcut method, important authentication fix and handle url-encoded URLs.

([Released](#) by Maarten van Gompel on 2020-11-19) <https://github.com/proycon/clam/releases/tag/v3.0.21> (deliverable ID: T142D1)

clam v3.0.20

Bugfix release: Fixed error message when downloading remote files and implemented proper URL decoding (#100)

([Released](#) by Maarten van Gompel on 2020-11-10) <https://github.com/proycon/clam/releases/tag/v3.0.20> (deliverable ID: T142D1)

clam v3.0.19

Bugfix release: * Fixed files not being added to list after upload in web interface #94

([Released](#) by Maarten van Gompel on 2020-10-27) <https://github.com/proycon/clam/releases/tag/v3.0.19> (deliverable ID: T142D1)

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 project 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)

clam v3.0.14

Bugfix release: many plain text error messages had the wrong mimetype

(Released by Maarten van Gompel on 2020-04-29) <https://github.com/proycon/clam/releases/tag/v3.0.14> (deliverable ID: T142D1)

clam v3.0.13

Bugfix release: Fixes for clamnewproject #90

(Released by Maarten van Gompel on 2020-04-10) <https://github.com/proycon/clam/releases/tag/v3.0.13> (deliverable ID: T142D1)

clam v3.0.12

Various fixes for improved scalability (many projects or a project with many files): * Better caching of the the project index and more granular updating of the project cache (less regenerating) * Improved reporting of project size, was too often out of date #89 * Narrowed the scope of the provenance data generated by CLAM, it was too verbose, had too much duplication per file, and led to scalability issues #79 * Implemented a 'quick' mode that skips loading metadata (can be triggered by the user or is automatically enabled if a timeout value (90s by default) is reached #79

(Released by Maarten van Gompel on 2020-04-03) <https://github.com/proycon/clam/releases/tag/v3.0.12> (deliverable ID: T142D1)

clam v3.0.11

Bugfix release: * Username was not passed properly for actions #88 * Reworked optional authentication for actions (was inconsistent and implemented too inelegantly)

(Released by Maarten van Gompel on 2020-02-27) <https://github.com/proycon/clam/releases/tag/v3.0.11> (deliverable ID: T142D1)

clam v3.0.10

Bugfix release: * provide a filename suggestion for download archives (adding Content-Disposition in HTTP response) #87

(Released by Maarten van Gompel on 2020-01-24) <https://github.com/proycon/clam/releases/tag/v3.0.10> (deliverable ID: T142D1)

clamservices v2.2.3

Various fixes for uto webservice; was broken

(Released by Maarten van Gompel on 2020-11-30) <https://github.com/proycon/clamservices/releases/tag/v2.2.3>

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>

clamservices v2.2.1

Minor bugfix release

(*Released* by Maarten van Gompel on 2020-04-22) <https://github.com/proycon/clamservices/releases/tag/v2.2.1>

clamservices v2.2.0

- Added FoLiA input support for SpaCy service

(*Released* by Maarten van Gompel on 2020-04-22) <https://github.com/proycon/clamservices/releases/tag/v2.2.0>

clamservices v2.1.0

Added SpaCy service (requires spacy and spacy2folia)

(*Released* by Maarten van Gompel on 2020-01-14) <https://github.com/proycon/clamservices/releases/tag/v2.1.0>

FLAT

Project & Task ID: [CLARIAH-PLUS WP3 T062](#)

flat v0.9.4

Minor bugfix release:

- fixes space rendering bug introduced in #139 (#166)

(*Released* by Maarten van Gompel on 2020-12-12) <https://github.com/proycon/flat/releases/tag/v0.9.4> (*deliverable ID: T062D1*)

flat v0.9.3

This releases mainly fixes/improves various issues when visualising TICCL/PICCL output:

- Raise a proper error when a correction set is missing #163
- Visualise substrings as used in TICCL output (requires them to be present in markup form) #92
 - Major interface improvements in substring visualisation
- Properly display documents with a non-default textclass #139
- Show feedback on currently selected mode #164
- Added KNAW HuC affiliation Note: adding substrings/markup annotation is still not supported yet

(*Released* by Maarten van Gompel on 2020-12-07) <https://github.com/proycon/flat/releases/tag/v0.9.3> (*deliverable ID: T062D1*)

flat v0.9.2

- Updated FLAT to the latest FoLiA version (v2.4), adds support for modality annotation #161

(*Released* by Maarten van Gompel on 2020-11-17) <https://github.com/proycon/flat/releases/tag/v0.9.2> (*deliverable ID: T062D1*)

foliadocserve v0.7.5

Bugfix release:

- Automatically fix documents with unassigned processors (fixunassignedprocessor)

(*Released* by Maarten van Gompel on 2021-02-10) <https://github.com/proycon/foliadocserve/releases/tag/v0.7.5> (*deliverable ID: T062D2*)

foliadocserve v0.7.4

- also propagate/serialize annotations inside str (proycon/flat#92)

(*Released* by Maarten van Gompel on 2020-12-07) <https://github.com/proycon/foliadocserve/releases/tag/v0.7.4> (*deliverable ID: T062D2*)

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 v2.4.2

- Predefine some subsets for style annotation #90
- Allow features in markup annotation #89
- Allow features in text content
- Added extra documentation for handling leading/trailing whitespace #88
- Allow for multiple foreign metadata nodes in FoLiA, even in 'native' mode #91

(Released by Maarten van Gompel on 2021-01-07) <https://github.com/proycon/folia/releases/tag/v2.4.2> (deliverable ID: T108D1)

folia v2.4.1

- Ignore all leading/trailing whitespace in text content #88

(Released by Maarten van Gompel on 2020-12-11) <https://github.com/proycon/folia/releases/tag/v2.4.1> (deliverable ID: T108D1)

folia v2.4.0

- Added modality annotation (#86) this is now preferred also for sentiment annotation (the dedicated sentiment annotation type is deprecated but remains for backward compatibility) as well as other modalities such as negations, truthfulness, doubt.
- Added a simple set definition for geolocation and an example to the documentation (using metric annotation)
- Minor backward-compatibility breaking change: renamed modalityfeature in coreference links to mod so it doesn't conflict with the new modality element, I've never seen anybody use this aspect of coreference linking in FoLiA yet so it's a small risk I'm taking. Let me know if it causes issues for anybody.
- Reintroduced and documented External annotation (#87), allowing you to separate child documents from parent documents whilst maintaining links.

(Released by Maarten van Gompel on 2020-11-16) <https://github.com/proycon/folia/releases/tag/v2.4.0> (deliverable ID: T108D1)

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 v0.0.6

Updated for FoLiA v2.4: includes modality annotation and external annotation

([Released](#) by Maarten van Gompel on 2020-11-16) <https://github.com/proycon/folia-rust/releases/tag/v0.0.6> (deliverable ID: T108D5)

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 v2.4.6

Bugfix release:

- the `fixunassignedprocessor` procedure should assign the first annotator rather than the last (it's more likely that the bug occurred where only one annotator existed)

([Released](#) by Maarten van Gompel on 2021-02-10) <https://github.com/proycon/foliapy/releases/tag/v2.4.6> (deliverable ID: T108D3)

foliapy v2.4.5

Bugfix release:

- Implemented important backward compatibility for text consistency validation prior to FoLiA v2.4.1, fixes the regression in issue #92, relates to #88

([Released](#) by Maarten van Gompel on 2021-02-03) <https://github.com/proycon/foliapy/releases/tag/v2.4.5> (deliverable ID: T108D3)

foliapy v2.4.4

Updated for FoLiA v2.4.2:

- Extra predefined features on style annotation `proycon/olia#90`
- Allow mixing `ForeignMetadata` and `NativeMetadata` (`proycon/olia#91`)

([Released](#) by Maarten van Gompel on 2021-01-07) <https://github.com/proycon/foliapy/releases/tag/v2.4.4> (deliverable ID: T108D3)

foliapy v2.4.3

Re-release after minor fix in test suite; previous release was a bit premature.

(*Released* by Maarten van Gompel on 2020-12-11) <https://github.com/proycon/foliapy/releases/tag/v2.4.3> (*deliverable ID: T108D3*)

foliapy v2.4.2

- Adapted for FoLiA v2.4.1: strip whitespace left and right if there is only a sole string (proycon/folia#88)

(*Released* by Maarten van Gompel on 2020-12-11) <https://github.com/proycon/foliapy/releases/tag/v2.4.2> (*deliverable ID: T108D3*)

foliapy v2.4.1

Implemented a `move()` method alongside `copy()`, which does no deep copy.

(*Released* by Maarten van Gompel on 2020-11-18) <https://github.com/proycon/foliapy/releases/tag/v2.4.1> (*deliverable ID: T108D3*)

foliapy v2.4.0

Updated for FoLiA v2.4.0:

- Implemented modality annotation (proycon/folia#86)
- Revised external annotation (proycon/folia#87)
- Properly handle removal of markup annotation (proycon/foliatools#21)

(*Released* by Maarten van Gompel on 2020-11-16) <https://github.com/proycon/foliapy/releases/tag/v2.4.0> (*deliverable ID: T108D3*)

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*)

foliapy v2.2.5

Minor fix:

- no hard fail on missing version, but assume an old version instead #17

(*Released* by Maarten van Gompel on 2020-01-13) <https://github.com/proycon/foliapy/releases/tag/v2.2.5> (*deliverable ID: T108D3*)

foliapy v2.2.4

Minor fix in `setup.py` so it installs even when not on a proper utf-8 locale. (#16)

(*Released* by Maarten van Gompel on 2020-01-03) <https://github.com/proycon/foliapy/releases/tag/v2.2.4> (*deliverable ID: T108D3*)

foliatools v2.4.6

- folia2html: Implemented support for render superscript/subscript #26
- folia2html: implemented the ability to add custom external CSS stylesheets #26
- updated help info for fixunassignedprocessor procedure

(Released by Maarten van Gompel on 2021-02-10) <https://github.com/proycon/foliatools/releases/tag/v2.4.6> (deliverable ID: T108D6)

foliatools v2.4.4

Minor bugfix release:

- Fixes an issue in folia2salt (thanks to @parkervg)

(Released by Maarten van Gompel on 2021-01-07) <https://github.com/proycon/foliatools/releases/tag/v2.4.4> (deliverable ID: T108D6)

foliatools v2.4.3

- [foliatextcontent] Fixed and improved substring linking, adding markup elements that reference substrings, also supports corrections #23

(Released by Maarten van Gompel on 2020-12-07) <https://github.com/proycon/foliatools/releases/tag/v2.4.3> (deliverable ID: T108D6)

foliatools v2.4.2

Minor update:

- [tei2folia] Better handling, detection and validation of IDs #22

(Released by Maarten van Gompel on 2020-11-30) <https://github.com/proycon/foliatools/releases/tag/v2.4.2> (deliverable ID: T108D6)

foliatools v2.4.1

Major performance improvement in foliasplit.

(Released by Maarten van Gompel on 2020-11-18) <https://github.com/proycon/foliatools/releases/tag/v2.4.1> (deliverable ID: T108D6)

foliatools v2.4.0

- [rst2folia] implemented rubric handling
- [foliasplit] Implemented a new tool to split a FoLiA document into multiple documents, based on a user's selection criteria. Also allows for linking from a parent document to external child documents. #20
- [foliaerase] Fixed the inability to properly handle markup elements #21

(Released by Maarten van Gompel on 2020-11-16) <https://github.com/proycon/foliatools/releases/tag/v2.4.0> (deliverable ID: T108D6)

foliatools v2.3.2

Bugfix release:

- [rst2folia] Made more robust against failures #17
- [rst2folia] support for conversion of containers (divs) from html #18

(Released by Maarten van Gompel on 2020-11-10) <https://github.com/proycon/foliatools/releases/tag/v2.3.2> (deliverable ID: T108D6)

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 converter (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 converter 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 v0.16

[Ko vd Sloot]

- requires libfolia 2.7 or above
- provenance data is better for a lot of modules
- added better checking on invalid NCnames in some modules.
- FoLiA-abby:
 - a lot of refactoring and additions to handle font/style information
- FoLiA-pm:
 - Notes are handled correctly now
 - fixed error in xlink attributes
- FoLiA-page:
 - more types of Page files are handled now
 - fixed annotation declarations
 - fixed offset calculation (due to change in FoLiA's opinion on those)
 - page number is added as a node and in the metadata
 - added a –trusttokens option. This means that Word items in the Page file are added as Word's in the FoLiA, embedded in Sentences.
 - added a –norefs option to avoid adding references to the original texts
- FoLiA-correct:
 - make sure that the default is to run on 1 thread
 - added a –rebase-inputclass option
- FoLiA-alto:
 - the -t option was not always handled correctly

[Maarten van Gompel]

- FoLiA-benchmark: guard against compiler optimisation #48

(*Released* by Ko van der Sloot on 2021-01-07) <https://github.com/LanguageMachines/foliautils/releases/tag/v0.16>
(deliverable ID: T108D7)

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)

foliautils v0.14

[Martin Reynaert]

- updated man pages

[Ko vd Sloot]

- added man pages
- revised usage() in many modules
- the default separator in FoLiA-stats is '_' now
- fix for: <https://github.com/LanguageMachines/foliautils/issues/37>
- fix for: <https://github.com/LanguageMachines/foliautils/issues/41>
- adapted to changes in libfolia
- many small code refactorings
- FoLiA-correct is improved a lot, allowing ngram corrections in FoLiA
- FoLiA-stats accepts a 'word_in_doc' mode now
- FoLiA-alto by default created nodes now. use -oldstring to get
- improved a lot in tests/
- many small fixes

(*Released* by Ko van der Sloot on 2020-04-15) <https://github.com/LanguageMachines/foliautils/releases/tag/v0.14>
(deliverable ID: T108D7)

libfolia v2.7

- implemented a more relaxed MetaData scheme, allowing mixing 'foreign' and 'native' MetaData
- bumped the .so version to 15
- features may be present in and <t-*> nodes now

(*Released* by Ko van der Sloot on 2021-01-07) <https://github.com/LanguageMachines/libfolia/releases/tag/v2.7>
(deliverable ID: T108D4)

libfolia v2.6.1

[Maarten van Gompel]

- Updated for FoLiA v2.4.1: strip leading/trailing whitespace in text content (proycon/folia#88)

[Ko vd Sloot]

- Fixed problem with text-consistency errors for within

(*Released* by Maarten van Gompel on 2020-12-11) <https://github.com/LanguageMachines/libfolia/releases/tag/v2.6.1>
(deliverable ID: T108D4)

libfolia v2.6

[Maarten van Gompel]

- Updated for FoLiA v2.4
- Revised external implementation
- Implemented Modality annotation

[Ko vd Sloot]

- cleanup and extra sanity tests
- Implemented an 'explicit' mode for Document (FoLiA v2.3) and in folialint

(*Released* by Maarten van Gompel on 2020-11-17) <https://github.com/LanguageMachines/libfolia/releases/tag/v2.6>
(deliverable ID: T108D4)

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)

libfolia v2.4

- comment in Doxygen format added
- bumped the library version to 14
- fix for <https://github.com/proycon/folia/issues/82>
- fix for <https://github.com/proycon/folia/issues/42>
- fixed problem with using new tag names on pre 1.6 documents
- better checks in folia_engine on text inconsistencies and such (<https://github.com/LanguageMachines/libfolia/issues/43>)
- confidence output is more consistent now
- removed the folia_builder (was not used)
- code refactorings and cleanup, removing unused functions

(*Released* by Ko van der Sloot on 2020-04-15) <https://github.com/LanguageMachines/libfolia/releases/tag/v2.4>
(deliverable ID: T108D4)

libfolia v2.3.2

Bug fix release * fix for <https://github.com/LanguageMachines/foliautils/issues/37> * fix for <https://github.com/LanguageMachines/foliautils/issues/38> * fixes in Correction handling. * fixed a Multi-Threading problem with the static reverse_old map

(*Released* by Ko van der Sloot on 2020-01-13) <https://github.com/LanguageMachines/libfolia/releases/tag/v2.3.2> (deliverable ID: T108D4)

piereling v0.2.1

Minor update: changed mimetype for TEI in accordance with the CLARIN switchboard

(*Released* by Maarten van Gompel on 2020-11-30) <https://github.com/proycon/piereling/releases/tag/v0.2.1> (deliverable ID: T108D9)

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 v0.22

[Ko vd Sloot]

- start using the tmp_stream() class from ticcutils 0.25

[Maarten van Gompel]

- Require libfolia 2.6

(*Released* by Maarten van Gompel on 2020-11-17) <https://github.com/LanguageMachines/frog/releases/tag/v0.22> (deliverable ID: T139aD1)

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)

frog v0.20.1

Bug fix release. - added missing Doxygen.cfg to the tarball

(*Released* by Ko van der Sloot on 2020-04-15) <https://github.com/LanguageMachines/frog/releases/tag/v0.20.1> (deliverable ID: T139aD1)

frog v0.20

[Ko vd Sloot]

- added Doxygen to the build
- added a lot of comment in Doxygen format
- adapted to the newest ticcutils version
- adapted to latest libfolia
- adapted to latest ucto
- lots of code refactorings

- implemented `-JSONin` option (server only)
- implemented `-JSONout` option
- added a `-allow-word-correction` option which allows ucto to correct FoLiA Word nodes

[Iris Hendrix]

Documentation updates

(*Released* by Ko van der Sloot on 2020-04-15) <https://github.com/LanguageMachines/frog/releases/tag/v0.20>
(deliverable ID: T139aD1)

frogdata v0.18

[Ko van der Sloot]

- added some comment to `nld/frog.cfg`
- added some test files to test using external Mblem/MBMA and Mbt servers
- updated NER data

[Maarten van Gompel]

- added a `nld-vnn` configuration

(*Released* by Ko van der Sloot on 2020-04-15) <https://github.com/LanguageMachines/frogdata/releases/tag/v0.18>
(deliverable ID: T139aD1.2)

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)

ucto v0.21.1

- fix for <https://bugs.debian.org/cgi-bin/bugreport.cgi?bug=941498>

(*Released* by Ko van der Sloot on 2020-04-15) <https://github.com/LanguageMachines/ucto/releases/tag/v0.21.1>
(deliverable ID: T139aD2)

ucto v0.21

- Adapted to newest libfolia 2.4
- adapted some tests
- added an `-allow-word-corrections` option
- improved handling of odd FoLiA

(*Released* by Ko van der Sloot on 2020-04-15) <https://github.com/LanguageMachines/ucto/releases/tag/v0.21>
(deliverable ID: T139aD2)

LaMachine

Project & Task ID: [CLARIAH-PLUS WP3 T098](#)

codemetapy v0.3.5

Added the ability to detect multiple authors #5

(*Released* by Maarten van Gompel on 2020-10-15) <https://github.com/proycon/codemetapy/releases/tag/v0.3.5>
(deliverable ID: T098D4)

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)

codemetapy v0.3.2

Minor bugfix release: do add duplicate entrypoints

(*Released* by Maarten van Gompel on 2020-02-03) <https://github.com/proycon/codemetapy/releases/tag/v0.3.2>
(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 interpreter 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 was missed. 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)

LaMachine v2.17

- LaMachine Release after final releases by @kosloot of full frog, ucto, libfolia, timbl, mbt software stack
- Fix for alpino
- Added fame_align (frisian aligner) to LaMachine
- Added forced alignment service
- Allow python-timbl to fail on macOS for now #175

(*Released* by Maarten van Gompel on 2020-04-16) <https://github.com/proycon/LaMachine/releases/tag/v2.17>
(deliverable ID: T098D1)

LaMachine v2.16

- Added huggingface transformers library
- Fixed kaldilib to atlas3 by default instead of openblas #172
- Various fixes for speech webservices (oral history and others)
- Allow setting ssh keypairs from the LaMachine configuration (needed for restricted access to e.g. our gitlab server)
- Added an option to skip nvm activation things
- Make lxc profile configurable in bootstrap.sh

(*Released* by Maarten van Gompel on 2020-03-11) <https://github.com/proycon/LaMachine/releases/tag/v2.16>
(deliverable ID: T098D1)

LaMachine v2.15

- ansible deprecated some features, adapting accordingly (#167)
- explicitly set ALPINO_HOME and alpino-specific lib path in alpino/tscan uwsgi #153 (should already be set properly in venv though)
- completed tscan integration #153
- Rerun all activation scripts after update so the shell is up to date for any added environment variables immediately #153
- some documentation updates

(*Released* by Maarten van Gompel on 2020-01-26) <https://github.com/proycon/LaMachine/releases/tag/v2.15>
(deliverable ID: T098D1)

LaMachine v2.14

- Added julia #165
- Added spaCy webservice (a CLAM webservice with spacy2folia for FoLiA output support)
- Docker: expose more ports by default (8080, 8888 and 9999 are now exposed by default in the Dockerfile)
- Adding kaldi-nl resource (#106) and oral history (+webservice) (#140) [requires private access]
- Minor updates to contributing guidelines **v2.13**:
- Fix for alpino, it is linked against extra libraries which are in a non-default location
- Continue even if shared data path can't be created (due to permission issues sometimes)
- Various fixes for improved jupyter lab integration
 - auto-starting it
 - store jupyter notebooks in `www_data_path` and allow this to be stored on the shared data volume (relevant for docker/VM)
- Implemented new `www_data_path` option and a `shared_www_data` boolean that by defaults sets the data path to be on the shared volume (for VM/container)
 - explicitly ask during bootstrap whether to put web data on the shared volume
- **LaMachine now specifically provides support status indications for different Linux distributions/operating systems**, we distinguish four categories (gold, silver, bronze and deprecated). Support indications are given both in the bootstrap script as in the documentation.
- CentOS 8 support (silver support status), CentOS 7 support is being deprecated.
- Added `clam_include` directive to include a base CLAM configuration from the configuration of each clam service
- adding various “maintained by lamachine” comments to configuration files, warning the user not to edit them.
- Fixes and updated instructions for for LXD containers
- Fixes in `lamachine-start-webserver` and cleanup of `lamachine-start-webserver` output

(Released by Maarten van Gompel on 2020-01-14) <https://github.com/proycon/LaMachine/releases/tag/v2.14> (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)

ticcutils v0.25

[Ko vd Sloot]

- added new 3.9.1 version of nlohmann JSON library
- added `tmp_stream` class, removed `tempname` and `tempdir` from the API.
- cleanup (typos and small modernisations)

[Maarten van Gompel]

- removing a const qualifier that caused problems on older libxml2 (CentOS 7) #23

(Released by Maarten van Gompel on 2020-11-17) <https://github.com/LanguageMachines/ticcutils/releases/tag/v0.25>

ticcutils v0.24

- added documentation in Doxygen format
- removed dependency on Boost
- renamed `TimblServer` namespace to `TiCCServer`. breaks all builds that use it!
- bumped library version to 8.0.0
- updated `ClientSocket` and `ServerSocket` classes.

- removed the Lexicon class from Treehash
- cleaned up LogStream/LogBuffer classes. removing unused stuff
- updated json to <https://github.com/nlohmann/json/releases/tag/v3.7.3>
- added a tempdir() member to FileUtils
- many small code refactorings everywhere

(*Released* by Ko van der Sloot on 2020-04-15) <https://github.com/LanguageMachines/ticcutils/releases/tag/v0.24>

Nederlab

Project & Task ID: Nederlab

nederlab-pipeline v0.9.4

Bugfix release for fix pipeline: Acts can also be independent chapters

(*Released* by Maarten van Gompel on 2021-02-23) <https://github.com/proycon/nederlab-pipeline/releases/tag/v0.9.4>

nederlab-pipeline v0.9.3

Minor release update; better gz compression in fix pipeline

(*Released* by Maarten van Gompel on 2021-02-22) <https://github.com/proycon/nederlab-pipeline/releases/tag/v0.9.3>

nederlab-pipeline v0.9.2

Further bugfix release for dbnl fix pipeline

(*Released* by Maarten van Gompel on 2021-02-02) <https://github.com/proycon/nederlab-pipeline/releases/tag/v0.9.2>

nederlab-pipeline v0.9.1

bugfix release for dbnl fix pipeline

(*Released* by Maarten van Gompel on 2021-02-02) <https://github.com/proycon/nederlab-pipeline/releases/tag/v0.9.1>

nederlab-pipeline v0.9.0

Recent changes: * Implemented a script that fixes the DBNL FoLiA v2 documents as delivered in 2019. This script fixes the IDs and adds the necessary (sub)metadata. Discussed in Jira ticket: <https://jira.socialhistoryservices.org/browse/TT-709> Older changes (2019): * enable ignore option for wikiente * implement support for language constrain in modernisation * added resources (migrated from inl/nederlab-linguistic-enrichment) * only do language identification on sentences! * simplifying the pipeline, do not run frog in batches anymore but one frog per file (at cost of init time and extra memory, but easier to handle potential errors) * replacing folialangid with colibri-lang, use --subcodes for colibri-lang * do language detection before tokenization

(*Released* by Maarten van Gompel on 2021-02-02) <https://github.com/proycon/nederlab-pipeline/releases/tag/v0.9.0>

PICCL & TICCL

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

PICCL v0.9.5

Added a string linking stage to ticcl, this adds extra markup information (t-str/t-correction) using the foliatextcontent tool, this is in turn needed by FLAT for proper visualisation.

(*Released* by Maarten van Gompel on 2020-12-11) <https://github.com/LanguageMachines/PICCL/releases/tag/v0.9.5>

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>

PICCL v0.9.0

This PICCL release builds upon the long awaited TICCLtools v0.7:

Ticcl:

- Fixed chaining
- Implemented `chainclean` and made it optional
- Changed default separator to underscore
- TICCL-rank invocation changed
- changed `skipcols`
- added `–low` `–high` and `–ngrams` parameter
- added alphabet file to TICCL-unk

General:

- Migrated to nextflow process selectors, solved deprecation warnings ([#57](#))
- verify output files have non-zero size
- Added schematic figures to document the architecture of the pipelines

Webservice:

- Added inputtemplate for custom lexicon [#56](#)

(Released by Maarten van Gompel on 2020-04-15) <https://github.com/LanguageMachines/PICCL/releases/tag/v0.9.0>