Jaarrapportage 2020

Maarten van Gompel

Startdatum: 2020-01-01 Einddatum: 2020-12-31

De in deze rapportage genoemde taken zijn aanvankelijk begonnen bij CLST, Radboud Universiteit Nijmegen, maar de ontwikkeling daarvan is sinds de zomer van 2020 grotendeels met mijzelf verhuisd naar het KNAW Humanities Cluster. Inhoudelijk maakt dit geen verschil en heeft het ook geen nadelige gevolgen gehad voor de voortgang. Doorlopende hosting en beheer van deze de uit deze software voortkomende services bij CLST, Radboud Universiteit is ook gegarandeerd.

CLAM: Maintenance and Support (T142)

Taakbeschrijving: CLARIAH-PLUS WP3 T142

De laatste grote milestone wat betreft CLAM is in October 2019 afgerond, met de introductie van een nieuwe web-interface. Toch is er in 2020 ook veel gebeurd op het gebied van bugfixes, een uitgebreid overzicht is in de appendix te vinden. Ook is CLAM wederom gebruikt om een aantal nieuwe webservices in te richten, met name spraakservices bij CLST aan de Radboud Universiteit. In het kader van deze taak is hiertoe ondersteuning geboden aan de bouwers van deze services.

FoLiA (T108)

Taakbeschrijving en planning: CLARIAH-PLUS WP3 T108

FoLiA blijft een belangrijke pijler onder heel veel van onze taken. Deze taak behelst onderhoud en supportwerk aan een uitgebreide infrastructuur van libraries en tools rondom het FoLiA dataformaat, vaak in samenspraak met gebruikers. We vieren inmiddels al ons 10-jarig bestaan. Er zijn in de periode twee grote releases van het formaat zelf geweest, en een aantal kleinere tussenreleases. De appendix toont weer het volledige overzicht.

Een beperkte greep uit de nieuwe features in FoLiA deze periode zijn o.a. ondersteuning voor modaliteitsannotatie, verbeterde tekstvalidatie, verbeterde metadata mogelijkheden, de mogelijkheid een document op te breken naar meerdere bestanden.

Wat betreft de FoLiA tools en zijn de TEI-naar-FoLiA, de ReStructuredText-naar-FoLiA, de AbbyXML-naar-FoLiA en de PageXML-naar-FoLiA converters sterk verbeterd. Er is ook gewerkt aan experimentele conversie naar Salt XML wat mogelijkheden biedt om beter aan te sluiten bij de Duitse infrastructuur rondom Salt, Pepper en de ANNIS annotatieomgeving. Er is een `foliasplit` tooltje bijgekomen die het mogelijk maakt een document op bepaalde criteria te splitsen in meerdere documenten.

FLAT (T062)

Taakbeschrijving en planning: CLARIAH-PLUS WP3 T062

Deze taak is grotendeels een onderhoud en support taak geworden, want er heerste in eerste instantie ook wat onduidelijk welk deel van de plannen voor FLAT gehonoreerd kon worden en waaruit vanuit WP3 behoefte was. Vanuit verschillende onderzoekers blijkt er wel degelijk behoefte aan FLAT, en deze worden dan ook met prioriteit ondersteund. Het taakplan ligt dit nader toe. Er lopen momenteel twee voorname use-cases (1 2).

Er zijn zowel van de frontend als de backend in 2020 een drietal releases uitgekomen, zie de appendix voor een volledig overzicht. De voornaamste release richtte zich op een punt want al langer op de planning stond; namelijk het correct visualiseren van PICCL/TICCL output (een andere CLARIAH taak).

Frog, Ucto en DeepFrog (T139)

Taakbeschrijving en planning: CLARIAH-PLUS WP3 T139

De hoofdontwikkelaar van Frog (en tevens medeontwikkelaar van FoLiA) is in het voorjaar van 2020 met pensioen gegaan. Daarvoor is er nog een grote Frog versie opgeleverd met laatste verbeteringen en hebben we stabiele tool die veelvuldig gebruikt wordt in de Nederlandse en Vlaamse NLP gemeenschap. Doorlopende ondersteuning is toch gewaarborgd, want ik heb deze taak overgenomen. Dit geldt ook voor de tokeniser ucto, een dependency van Frog die we ook onder deze taak scharen. Deze taak richt zich dan nu ook voornamelijk op onderhoud en bugfixes, verdere innovatieve doorontwikkeling van Frog an sich komt hiermee wel ten einde.

Er is ook gewerkt aan een mogelijke opvolger DeepFrog die op nieuwe deep-learning technieken stoelt. Dit valt eigenlijk buiten de begrootte en gebudgeteerde tijd voor deze taak, want die is hiervoor niet toereikend en reeds verbruikt. Toch hebben we een eerste experimentele versie van DeepFrog opgeleverd.

LaMachine (T098)

Taakbeschrijving en planning: CLARIAH-PLUS WP3 T098

LaMachine fungeert als de kapstok waaraan we veel van onze software ophangen en de vorm waarin we het aanbieden aan de brede gemeenschap binnen en buiten CLARIAH. Het wordt gebruikt zowel door individuele onderzoekers als door hosting-partners om onze oplossingen als service aan te bieden. We merken ook een toenemende interesse hierin vanuit andere werkpakketten binnen CLARIAH en organisaties als de Stichting Open Spraaktechnologie. LaMachine wordt relatief veel gebruikt. Al met al is de taak een ondersteunings- en ontwikkelingstaak die veel meer tijd in beslag neemt dan aanvankelijk gedacht, ook was er al minder tijd gehonoreerd dan gevraagd en noodzakelijk. Het VRE project, in de volgende sectie beschreven, heeft nauwe banden met deze taak en compenseert dit grotendeels weer gelukkig.

De appendix geeft een gedetailleerde overzicht van de ca. tien LaMachine releases die afgelopen jaar zijn gedaan. Hier moet je denken aan toevoegingen van allerlei nieuwe tools en services, vele verbeteringen en bugfixes, maar ook simpelweg aan het geheel up to date houden met de constant verderlopende realiteit. LaMachine is per definitie iets wat lopend aan onderhoud onderhevig is want het stemt allerlei verschillende onderdelen op elkaar af die zich elk in hun eigen tempo voortbewegen.

WP3 Virtual Research Environment (T137)

Taakbeschrijving en planning: CLARIAH-PLUS WP3 T137 In december 2020 ben ik gevraagd de VRE ontwikkeling op me te nemen en te coördineren nadat ik daarvoor een hernieuwd plan had opgesteld. Dit is een koerswijziging ten opzichte van eerdere pogingen in CLARIAH-CORE om het VRE project vorm te geven, die erg ambitieus van karakter waren maar helaas niet tot een bruikbaar resultaat hebben geleid. Kernpunt bij de nieuwe insteek is om vanuit bestaande CLARIAH software te werken en ons puur te richten op deze met elkaar te verbinden, er wordt in die zin geen nieuwe overkoepelende infrastructuur ontwikkeld. LaMachine zat praktisch gezien al het dichtste bij een Virtual Research Environment en zal daarom de basis vormen voor verdere integratie.

We zoeken in dit project ook aansluiting bij met name het CLARIN Switchboard, waartoe sowieso al werk is verzet in het verleden. Belangrijk deel van dit herziene project is ook de disseminatie van de resultaten naar eindgebruikers (onderzoekers), in de vorm van o.a. demo video's die de mogelijkheden van de onderzoeksomgeving laten zien.

PICCL

Deze taak valt niet onder mijn verantwoordelijkheid maar was er één onder die van Martin Reynaert. Ik wil hem toch even kort noemen omdat er dit jaar, met het pensioen van de hoofdontwikkelaar Ko van der Sloot, een belangrijke laatste release is gedaan van PICCL en de onderliggende TICCLtools waarin meer dan een jaar aan ontwikkeling zit.

Appendix: Software Releases

Hieronder een uitgebreid technische overzicht van de release notes van alle software versies die in het kader van bovenstaande taken opgeleverd zijn 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 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)*

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 ucto 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 occured 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/folia#90
- Allow mixing ForeignMetadata and NativeMetadata (proycon/folia#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: mplemented 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 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 v0.16

[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 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)*

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 kaldi lib 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 boostrap 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 strict 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 "maintainted 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 scripts 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