Software Release and Progress Report

Start date: 2021-01-19 End date: 2021-04-06

Short Summary

One of the main things I have been working on the past period is the VRE task. There's been significant work on CLAM and LaMachine to implement the functionality promised in the VRE plan. All of this is still in development pending release so not included in this report yet. The actual status for these VRE milestones can be tracked here:

- T137M1 - CLAM - https://github.com/proycon/clam/milestone/11
- T137M2 - LaMachine - https://github.com/proycon/clam/milestone/11
- T137M3 - FLAT - https://github.com/proycon/flat/milestone/17

A major point that I added to the VRE plan was to finally finish the proper federated authentication solution (using OpenID Connect). This has been promised for years but now I moved to KNAW I finally have access to the relevant authentication infrastructure and contact with the maintainers thereof to finally realize this plan. It did take up a considerable portion of the VRE time. The existing plan is well on track but will need to be stretched over the entire year as I have some other priorities coming up that will hinder my ability to work on this for some months.

Right now the VRE execution is still a one-man effort but I had productive discussions with Hennie Brugman (KNAW HuC) and Hayco de Jong, albeit after a long delay, on what their team can contribute to the VRE plan. Rather than delegate a portion of the existing plan, we think it would be better to take the opportunity to integrate further tools their Team Text is working on, the first focus being on TextRepo. A clear plan on how to proceed with HuC's Team Text still has to be formulated.

I spoke with Jan, Gijsbert and Daan about partially succeeding Daan as technical coordinator for WP3, as he left us per April first. I volunteered because of the overlap with the things I'm already doing in the Interest Groups and Technical Committee.

I have spent a lot of time on work surrounding the Interest Groups and Technical Committee, most notably on coordinating the collection of the use cases (see https://github.com/CLARIAH/usecases) and setting up guidelines for the interest groups as a whole. The actual interest groups I coordinate (Text Processing & Workflows) haven't really seen much member activity yet. I've been in contact with Roeland (who's succeeding GertJan as CTO) and with Susanne as a programme coordinator. Getting people to actually participate in the interest groups and in submitting use cases proves to be a great deal more challenging than I anticipated. We do have a decent foundation of use cases now. For WP3 we have use cases from DANS (1), INT (covering mostly WP6), RUN and KNAW (the latter consisting of my own contributions). UU and MPI have not submitted anything yet (neither have VU and RUG, but I don't think they have a formal role in WP3 nowadays?). FA expressed interest and is probably still deliberating.

A major new FoLiA release (v2.5) is also pending and will soon be published. It strictly specifies the role of whitespace in FoLiA and solves all kind of issues that arose in that regard. Most of this work was prompted by the use case Retrodigitization of Text-critical Editions (Bayerische Akademie der Wissenschaften), which I'm providing support for in the scope of my CLARIAH FoLiA & FLAT maintenance and support tasks.

Susanne and I compiled an overview of my tasks and their status, I'll keep that one up to date here: https://github.com/LanguageMachines/clariah-plus-tasks/blob/master/clariah-plus-workplan.fods

My time is getting a bit fragmented over all the various tasks, especially considering there are also big non-CLARIAH tasks that I'm also participating and also due to the large increase in overhead because of the Interest Groups, Technical Committee and soon possibly also WP3 Technical Coordination.

CLAM

Project & Task ID: CLARIAH-PLUS WP3 T142

(no releases this period)

FLAT

Project & Task ID: CLARIAH-PLUS WP3 T062

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)*

FoLiA

Project & Task ID: CLARIAH-PLUS WP3 T108

foliapy v2.4.8

Minor bugfix release: do not serialize metadata attribute if the submetadata element doesn't exist (prevents invalid FoLiA)

(Released by Maarten van Gompel on 2021-03-10) https://github.com/proycon/foliapy/releases/tag/v2.4.8 *(deliverable ID: T108D3)*

foliapy v2.4.7

Bugfix release for handling very large (or many) documents: Enable XML_HUGE_TREE option by default

(Released by Maarten van Gompel on 2021-03-01) https://github.com/proycon/foliapy/releases/tag/v2.4.7 *(deliverable ID: T108D3)*

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)*

foliatools v2.4.9

Bugfix release, previous release was premature.

(Released by Maarten van Gompel on 2021-03-03) https://github.com/proycon/foliatools/releases/tag/v2.4.9 *(deliverable ID: T108D6)*

foliatools v2.4.8

folia2html: Implemented support for outputting based on other text classes #30

(Released by Maarten van Gompel on 2021-03-03) https://github.com/proycon/foliatools/releases/tag/v2.4.8 *(deliverable ID: T108D6)*

foliatools v2.4.7

- folia2html: translate t-hbr as a soft-hyphen
- folia2html: translate features on structural elements to css classes
- folia2html: fix in translating t-str to css classes
- foliasplit: prevent duplicate IDs in the root element

(Released by Maarten van Gompel on 2021-03-03) https://github.com/proycon/foliatools/releases/tag/v2.4.7 *(deliverable ID: T108D6)*

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)*

Frog, Ucto & DeepFrog

Project & Task ID: CLARIAH-PLUS WP3 T139

(no releases this period)

LaMachine

Project & Task ID: CLARIAH-PLUS WP3 T098

(no releases this period)

Miscellaneous

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

(no releases this period)

Nederlab

Project & Task ID: Nederlab

nederlab-pipeline v0.9.11

Bugfix release and added extra check to guard against a certain corruption in the metadata database

(Released by Maarten van Gompel on 2021-03-10) https://github.com/proycon/nederlab-pipeline/releases/tag/v0.9.11

nederlab-pipeline v0.9.10

Bugfix release

(Released by Maarten van Gompel on 2021-03-09) https://github.com/proycon/nederlab-pipeline/releases/tag/v0.9.10

nederlab-pipeline v0.9.9

Further fixes in DBNL fix pipeline

(Released by Maarten van Gompel on 2021-03-09) https://github.com/proycon/nederlab-pipeline/releases/tag/v0.9.9

nederlab-pipeline v0.9.8

Incorporated extra validation step in fix pipeline

(Released by Maarten van Gompel on 2021-03-03) https://github.com/proycon/nederlab-pipeline/releases/tag/v0.9.8

nederlab-pipeline v0.9.7

Minor update for dbnl fix pipeline: Reverted unnecessary patch from previous release and ensure output files can't overwrite input

(Released by Maarten van Gompel on 2021-03-01) https://github.com/proycon/nederlab-pipeline/releases/tag/v0.9.7

nederlab-pipeline v0.9.6

Minor update for dbnl fix pipeline: added sync in an attempt to tackle an elusive issue

(Released by Maarten van Gompel on 2021-02-26) https://github.com/proycon/nederlab-pipeline/releases/tag/v0.9.6

nederlab-pipeline v0.9.5

Bugfix release in fix pipeline: Unassign unused metadata

(Released by Maarten van Gompel on 2021-02-24) https://github.com/proycon/nederlab-pipeline/releases/tag/v0.9.5

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