SIS and the Wittgenstein Advanced Search Tools (WAST)

Daniel Bruder, M.A.

Wittgenstein Summer School 2014

Retrospect

What went on in the last year?

- Strengthen Digital Humanities @ CIS
- Deploy WAST Technology "Landscape"
- Cambridge Cooperation
- Presentation of CIS and WAST in Passau and Madrid: Digital Humanities Conference
  - great success, good feedback
- application for Open Humanities Awards
  - http://openhumanitiesawards.org/
- Work on existing components:
  - wf, SIS, highlighting, reader, website, helppage, ...

What went on in the last year? (cont'd)

- New components:
  - Feedback app
    - make bug reporting available to externals
    - http://wastfeedback.cis.uni-muenchen.de/
  - wab2cis
    - Work in progress: WAB-XML -> XSL-Transformations -> CIS-XML / Raw text
  - Graph Editor
- more . . .

Follow-Up: SIS

- Symmetric Index Structures
- Finite State Automata for ultra-fast symmetric search:
  - "Symmetric full-text-indexing and deterministic autocomplete / suggestion search by using SCDAWGs (Symmetric Compacted Directed Acyclic Word Graphs)"
  - Master Thesis (Magister Artium) with Prof. Klaus U. Schulz
    - Daniel Bruder, 2012
    - `http://www.cip.ifi.lmu.de/~bruder/ma/MA/sis/`
- Technology Draft
  - Request for comments

SIS – Current State of the Art

Last year: Goals for the Wittgenstein-Project (related to SIS)

- ▶ (symmetric) autocomplete / suggestion search for the Wittgenstein-corpus
    - ▶ BACK TO RAW TEXT (Oyvind++)
- ▶ full compliance with WAB-XML (TEI)
    - ▶ BACK TO RAW TEXT (Oyvind++)
- ▶ full UTF-8 capability
    - ▶ DONE (Estelle++)
- ▶ UI (user interface design)
    - ▶ NO COMMENTS
- ▶ full serialization of indexed document data
    - ▶ DONE (Flo++)
- ▶ hard-to-track bug where retrieval hits disappeared:
    - ▶ FIXED (Estelle++)

Request for comments!

- Please use SIS ...
    - `http://sis.cis.lmu.de`
- ... and file your requests, improvement ideas, etc...
    - `http://wastfeedback.cis.uni-muenchen.de/`
- Thanks!

Wittgenstein Advanced Search Tools – WAST

Software Architecture and Project Management

- Technology "landscape"
  - collect unbound tools and components under one roof
  - establish solid project structure
    - collect components
    - add new components easily into existing landscape
  - establish project workflow
    - streamline development
    - establish software development "best practices"

```
<#include resources/wast-components-structure.ditaa>
```

Establish Industry-like Software Development Standards

Software Development Best practices

- everything under version control
    - git
- self-hosted gitlab instance
    - central web service
    - code review
    - `https://gitlab.cis.uni-muenchen.de/`
    - Stefan++ Thomas++
- gitlab-groups and permissions
    - easy collaboration with external people
    - project management and access control
    - `https://gitlab.cis.uni-muenchen.de/groups/wast`

Software Development Best practices (cont'd, #1)

- ▶ git-versioned website: development and stable branch
  - ▶ "unified deployment", build systems
  - ▶ controlled deploy / update
  - ▶ rollback-functionality
  - ▶ simplify development on localhost
  - ▶ Flo++

- ▶ Test Driven Development (TDD)
  - ▶ intensive testing
  - ▶ avoid regressions
  - ▶ also shows the API and usage to future maintainers / developers

Software Development Best practices (cont'd, #2)

- ▶ Continuous Integration (CI)
  - ▶ automated testing of new features and functionality
  - ▶ transparent test results
  - ▶ https://gitlabci.cis.lmu.de/
  - ▶ Stefan++ Thomas++
- ▶ extensive documentation
  - ▶ make know-how transparent and transitive
  - ▶ use as means for education
  - ▶ http://www.cip.ifi.lmu.de/~bruder/wast/
- ▶ work on XSL-Transformations
  - ▶ Oyvind++

Software Development Best practices (cont'd, #3)

- wiki
- mailinglists
- Education:
    - Theses
    - Courses
    - Practical Work

- bug tracking best practices
    - resolve bugs transparently
    - and in ordered fashion (priorities, components, maintainers)

Bug Tracking best practices

```
<#include resources/bug-tracking-workflow-status.plantuml>
```

Courses taught

- "WAST – Wittgenstein Advanced Search Tools"
- based on WAST documentation
- ˜12 attendees
- raise new talent

Next Steps / Goals

- Integration-Testing
- End-to-End (E2E)-Testing
- more Test Driven Development (TDD)
- Incorporation of *new data*
- Adaptation to *new editions*
    - open source to other projects
    - WAST --> *AST
- explore non-XML, flat-file approaches
    - "Matrix-Implementation"
    - Neo4J: Graph Database

Questions?

Thank you!

- attendees . . .
  - for your attention
  - and your visit
- collaborators . . .
  - for your bug fixes, ideas, commitment, "free time" . . .
  - Flo, Estelle, Stefan, Thomas, Max, Angela, Matthias, Oyvind, etc. etc.
- Max
  - for all your efforts
  - and organization of this workshop!

Fin.