Open Access Research Data Repository for Corpus Linguistic Data – A Modular Approach

Open Repositories Conference 2014 in Helsinki

Session „IG3F: Interest Group Session 3F

(Fedora / Islandora)“

Dennis Zielke
Computer and Media Service, Humboldt-Universität zu Berlin, June 13 2014

Agenda

Project information

Challenges

Complex data structure

Functionalities and research data access

Technologies

Outlook

Dennis Zielke
Computer and Media Service, Humboldt-Universität zu Berlin, June 13 2014

Project information

LAUDATIO

Long Term Access and Usage of Deeply Annotated Information

Long-term preservation, user-oriented storage, and re-use of research data (historical text corpora) for a sub discipline of linguistics (hictorical linguistics) according to the Open Access principles

Dennis Zielke
Computer- and Media Science, Humboldt-Universität zu Berlin, June 13 2014

Project information

Funded by German Research Foundation (DFG)

Funding program „Research data infrastructures“

Funding period from 2011 to 2014

Project partners: Computer- and Media Service, Department of Historical Linguistics and Corpus Linguistics (all HU Berlin) and INRIA, France

Supported by: Berlin School of Library and Information Science (BSLIS) HU Berlin

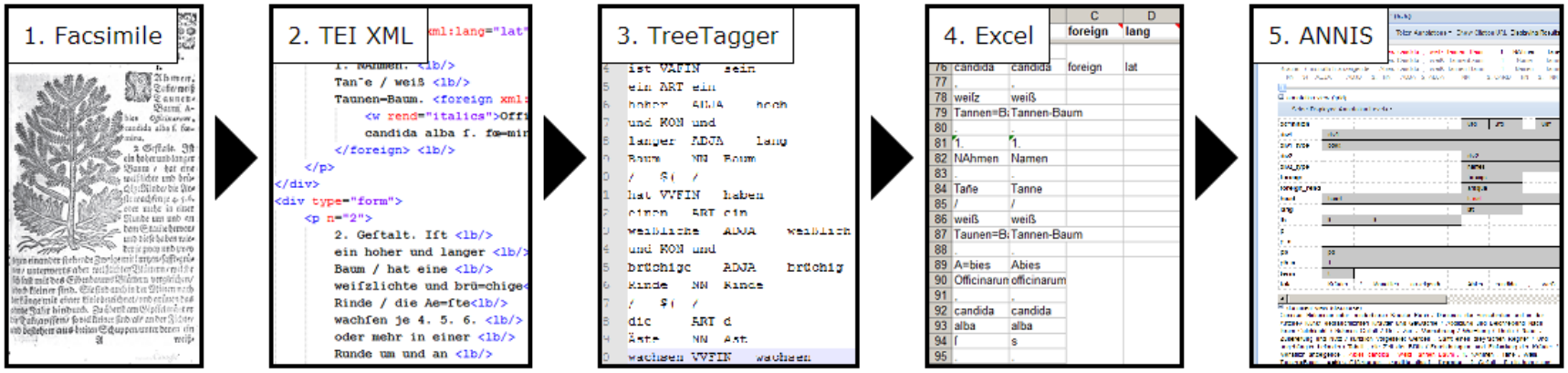
Research data

German historical texts and linguistic annotation including all dialects from 9th to the 19th century

Corpus Pipeline

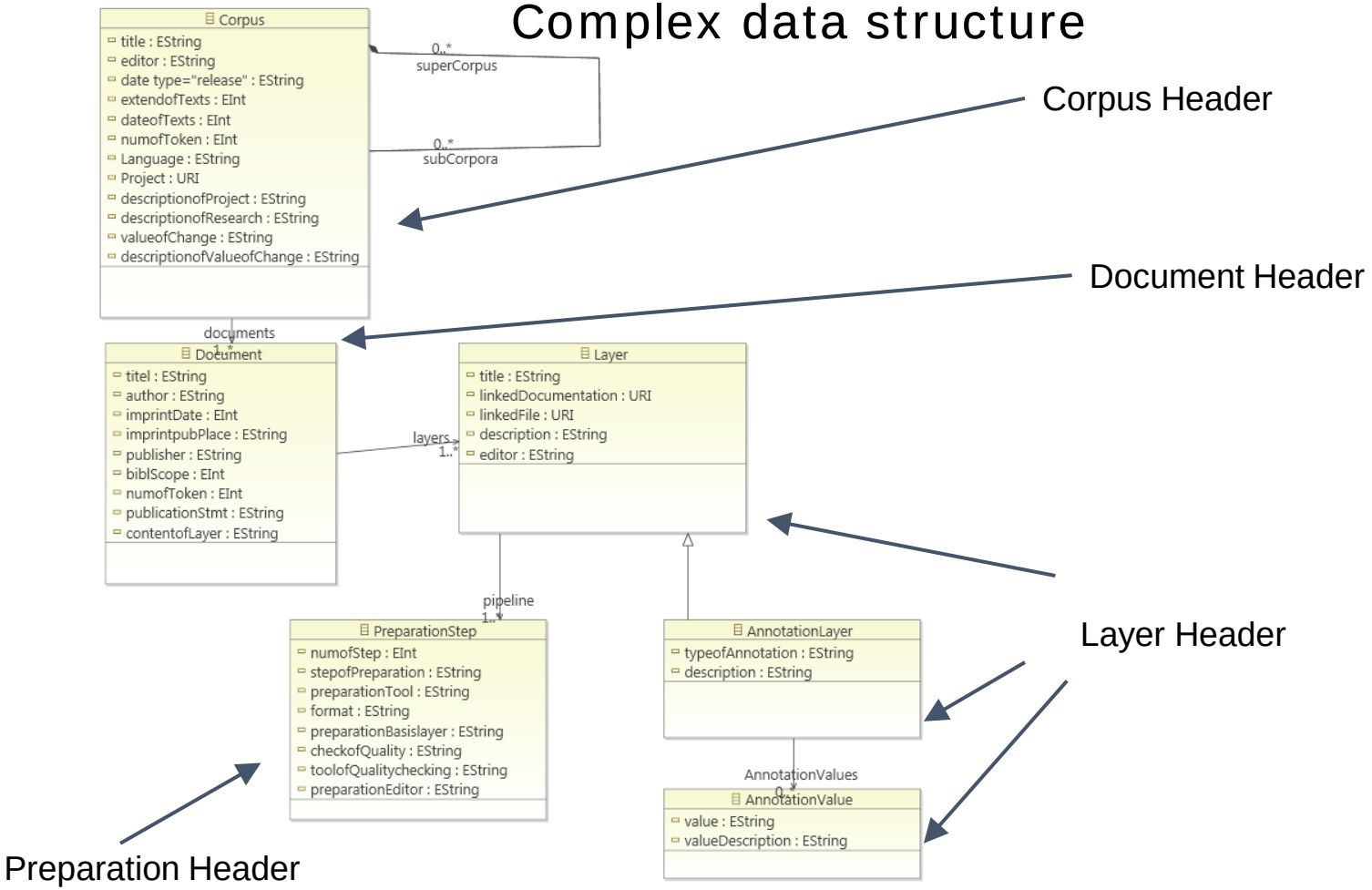
Corpora are collected in several stages:

1. Obtain facsimile, usually from Google Books
2. Correct OCR or transcribe text, marking up structure with TEI
3. Tokenize, part-of-speech tag and lemmatize with TreeTagger.
4. Add corpus specific manual annotations using Excel
5. Export the merged corpus to persistent formats and the ANNIS search tool



http://www.laudatio-repository.org

Complex data structure



Dennis Zielke
Computer and Media Service, Humboldt-Universität zu Berlin, June 13 2014

TEI XML P5

```
− <teiCorpus>
  + <teiHeader type="CorpusHeader"></teiHeader>
  − <teiCorpus>
    + <teiHeader type="DocumentHeader" style="Herbology"></teiHeader>
    + <teiHeader type="DocumentHeader" style="Herbology"></teiHeader>
    + <teiHeader type="DocumentHeader" style="Herbology"></teiHeader>
    − <teiHeader type="DocumentHeader" style="Herbology">
      − <fileDesc xml:id="V4-1543-NewKraeuterbuch">
        − <titleStmt>
            <title>New Kreüterbuch</title>
          − <author>
            − <persName>
                <forename>Leonhart</forename>
                <surname>Fuchs</surname>
              </persName>
            </author>
          − <editor>
            − <persName>
                <forename>NA</forename>
                <surname>NA</surname>
              </persName>
```

Import

Search

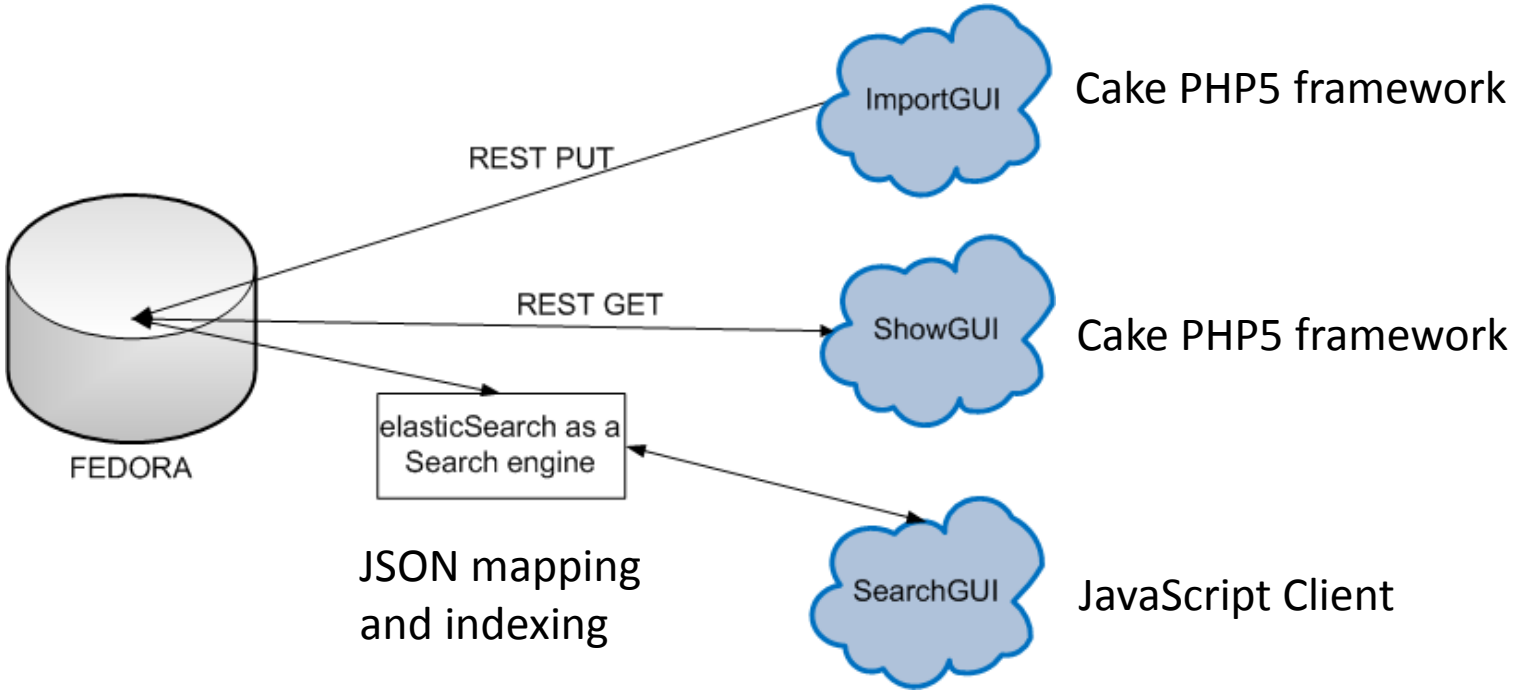
Browse

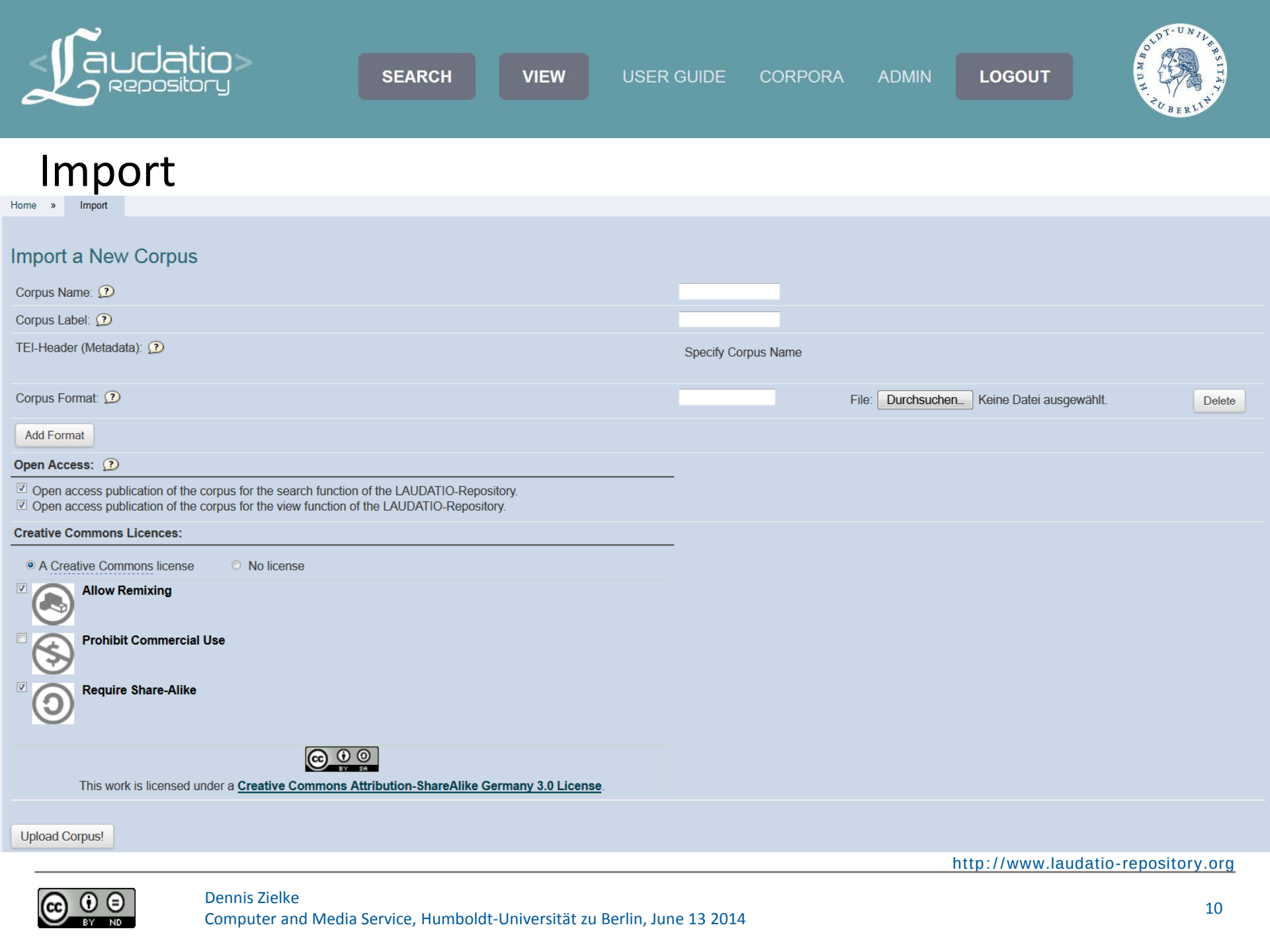
Modify

(Configuration)

Analyze

Fedora Repository Architecture




Laudatio repository
SEARCH
VIEW
USER GUIDE
CORPORA
ADMIN
LOGOUT
HUMBOLDT-UNIVERSITÄT ZU BERLIN
Import
Home » Import
Import a New Corpus
Corpus Name:
Corpus Label:
TEI-Header (Metadata):
Specify Corpus Name
Corpus Format:
File: Durchsuchen... Keine Datei ausgewählt.
Delete
Add Format
Open Access:
Open access publication of the corpus for the search function of the LAUDATIO-Repository.
Open access publication of the corpus for the view function of the LAUDATIO-Repository.
Creative Commons Licences:
A Creative Commons license
No license
Allow Remixing
Prohibit Commercial Use
Require Share-Alike
This work is licensed under a Creative Commons Attribution-ShareAlike Germany 3.0 License.
Upload Corpus!

Browse

ANNIS

RIDGES Herbology Version 4.0

2014-05-13 12:49:50 ▾

RIDGES Herbology Version 4.0, Humboldt-Universität zu Berlin, 4.0, 153732 Tokens, Fourth version. Extension of the corpus. Licence (for corpus and related documents): CC BY

Formats: EXCEL, EXMARaLDA, relANNIS, PDF

Always quote when using this data!
Lüdeling, Anke; Odebrecht, Carolin; Zeldes, Amir; RIDGES-Herbology (Fourth version. Extension of the corpus.) Version: 4.0. Humboldt-Universität zu Berlin. http://korpling.german.hu-berlin.de/ridges/.
http://hdl.handle.net/11022/0000-0000-2106-4

▼ **Corpus RIDGES Herbology Version 4.0**

- ▶ Authorship
- ▶ Project
- ▶ Publication

Size: 153732 Tokens

▼ Documents

Gart der Gesundheit
Artzney Buchlein der kreutter
Contrafayt kreüterbuch
New Kreüterbuch

Search

Full-Text Search

search term

partial match | exact match | fuzzy match | match all | match any | learn more

Filter by

Corpus

+ Corpora

+ Projects

+ Formats

+ Date - Corpus

+ Size - Corpus

Document

+ Annotation - Graphical

+ Annotation - Lexical

+ Annotation - Transcription

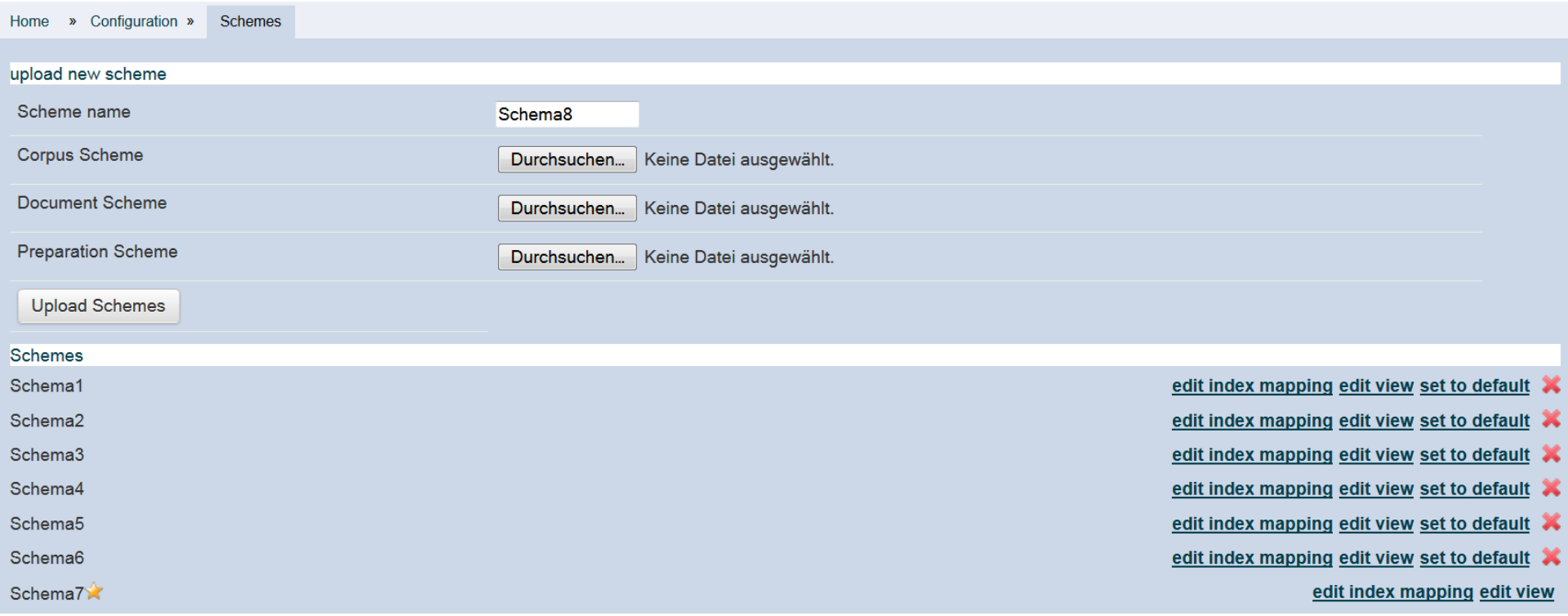
+ Annotation - Syntactical

Title: GerManC, 2007-04
Change: Version 1.0
Corpus Size: 900000 Words
Object URL: Direct Link to Corpus
Homepage: http://www.llc.manchester.ac.uk/research/projects/germanc/
Project Description: The ultimate aim of the project is to compile a representative historical corpus of written German for the years 1650-1800. This is a crucial period in the development of the language, as the modern standard was formed during it, and competing regional norms were finally eliminated. A central aim of the project is to provide a basis for comparative studies of ... (more)
Documents:

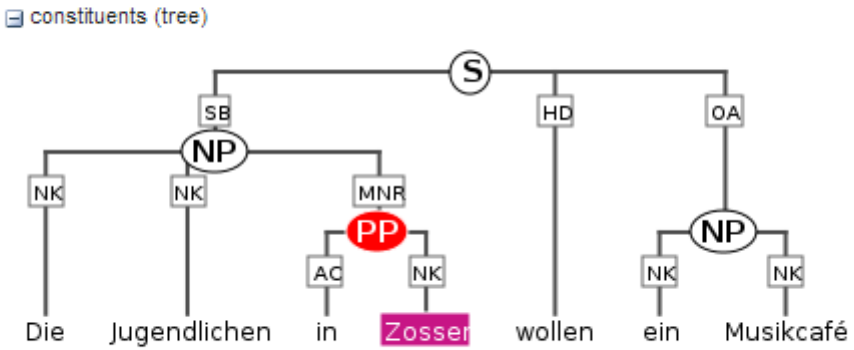
Newes Historisch-Politisches Schau-Spiel/ Genandt Die Teutsche Groß-Königin Leonilda
Die sterbende EURIDICE [...]
Der in seiner Freyheit vergnügte ALCIBIADES, [...]
Cardenio vnd Celinde, Oder Unglücklich Verliebete
Cleopatra
Trauer-Spiel Von dem Neapolitanischen Haupt-Rebellen Masaniello
(more)

Title: Deutsche Diachrone Baumbank, 2013
Change: Version 1.0
Corpus Size: 8580 Tokens
Object URL: Direct Link to Corpus
Homepage: http://korpling.german.hu-berlin.de/ddb-doku/index.htm
Project Description: Deutsche Diachrone Baumbank. Das durch den Berliner Senat geförderte Projekt "Interdisziplinärer Forschungsverbund Linguistik - Bioinformatik zur Berechnung von Verwandtschaft und Abstammung" hat angestrebt, Wege zu finden, wie bioinformatische Methoden dazu verwendet werden können, die Verwandtschaft zwischen (schriftlichen) Sprachdaten automatisch messbar zu ... (more)

Configuration



Search and visualization of annotations via ANNIS



http://www.sfb632.uni-potsdam.de/annis/gallery.html

Technologies

- CakePHP 5
- Fedora 3.6 as backbone
- Fedora REST interface
- ElasticSearch (JSON)
- External EPIC PID-Webservice v2 for PID assignments (handle)
- Third party Open Source libraries on Github
- Flat-Design (html5, CSS3) work in progress

Outlook

- Integration of other humanities disciplines e.g. musicology, history, literature studies working with historical (German) texts
- Building a multidisciplinary repository infrastructure at HU Berlin?
- Compliance with Guidelines/Certificates (e.g. Data Seal of Approval)
- Metadata editor
- Metadata as Linked Open Data
- ...

Thank you for your attention!

Source code is available on Github:
https://github.com/DZielke/laudatio

Any questions/feedback: zielkede@cms.hu-berlin.de