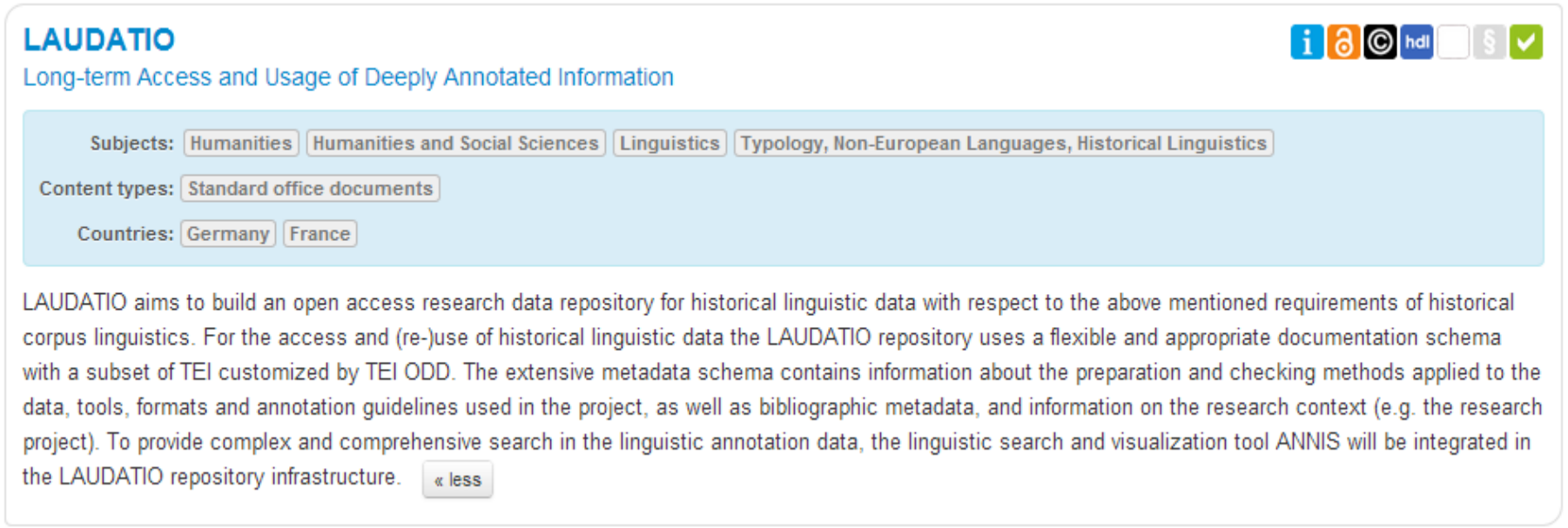


LAUDATIO – Open-Access-Forschungsdatenrepository für historische Korpuslinguistik

103. Bibliothekartag 2014 in Bremen

Session „Informationsinfrastrukturen für Forschungsdaten“

Maxi Kindling
Institut für Bibliotheks- und Informationswissenschaft, Humboldt-Universität zu Berlin, 3. Juni 2014

LAUDATIO

Long Term Access and Usage of Deeply Annotated Information

Ziel:

Forschungsdaten (historische Textkorpora)
für eine Fachdisziplin (historische Linguistik)
langfristig und nutzerorientiert zu speichern,
nach den Prinzipien von Open Access bereitzustellen
und (wieder)-nutzbar zu machen.

OA-Forschungsdatenrepository für historische Linguistik

Projektförderung „Infrastrukturen für Forschungsdaten“ der DFG seit 2012 (bis 2014)

Projektpartner

- Computer- und Medienservice der HU Berlin
- Institut für deutsche Sprache und Linguistik der HU Berlin
  - Lehrstuhl Korpuslinguistik
  - Lehrstuhl Historische Linguistik
- INRIA, France

Begleitung durch

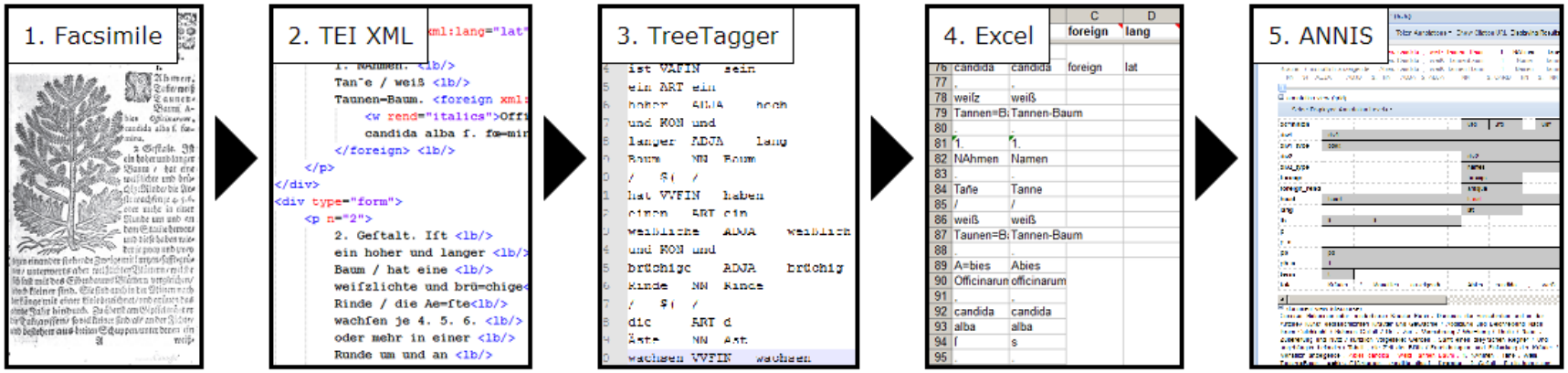
- Institut für Bibliotheks- und Informationswissenschaft der HU Berlin

Forschungsdaten: deutsche historische Texte und deren linguistische Annotationen

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

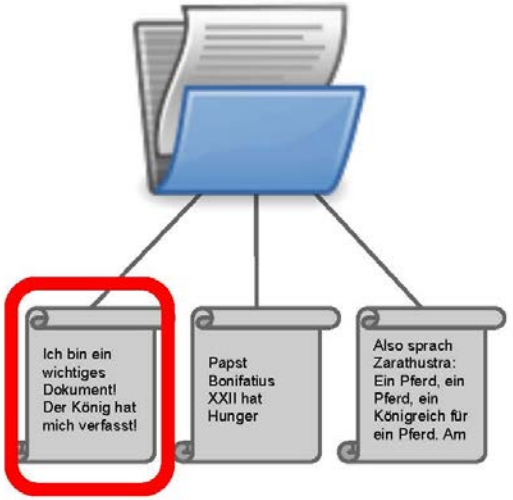


Komplexe Datenstruktur

- Ein Korpus besteht aus einem oder mehreren Dokumenten.
- Ein Dokument besteht aus einer oder mehreren Annotationen.
- Es gibt einen oder mehrere Autoren, die das Dokument verfasst haben.
- Es gibt einen oder mehrere Editoren, die das Korpus zusammenstellen.
- Es gibt einen oder mehrere Annotatoren

  die das Korpus bearbeiten.

→ komplexe Datenaufbereitung



Funktionalitäten des Forschungsdatenrepositories

Import

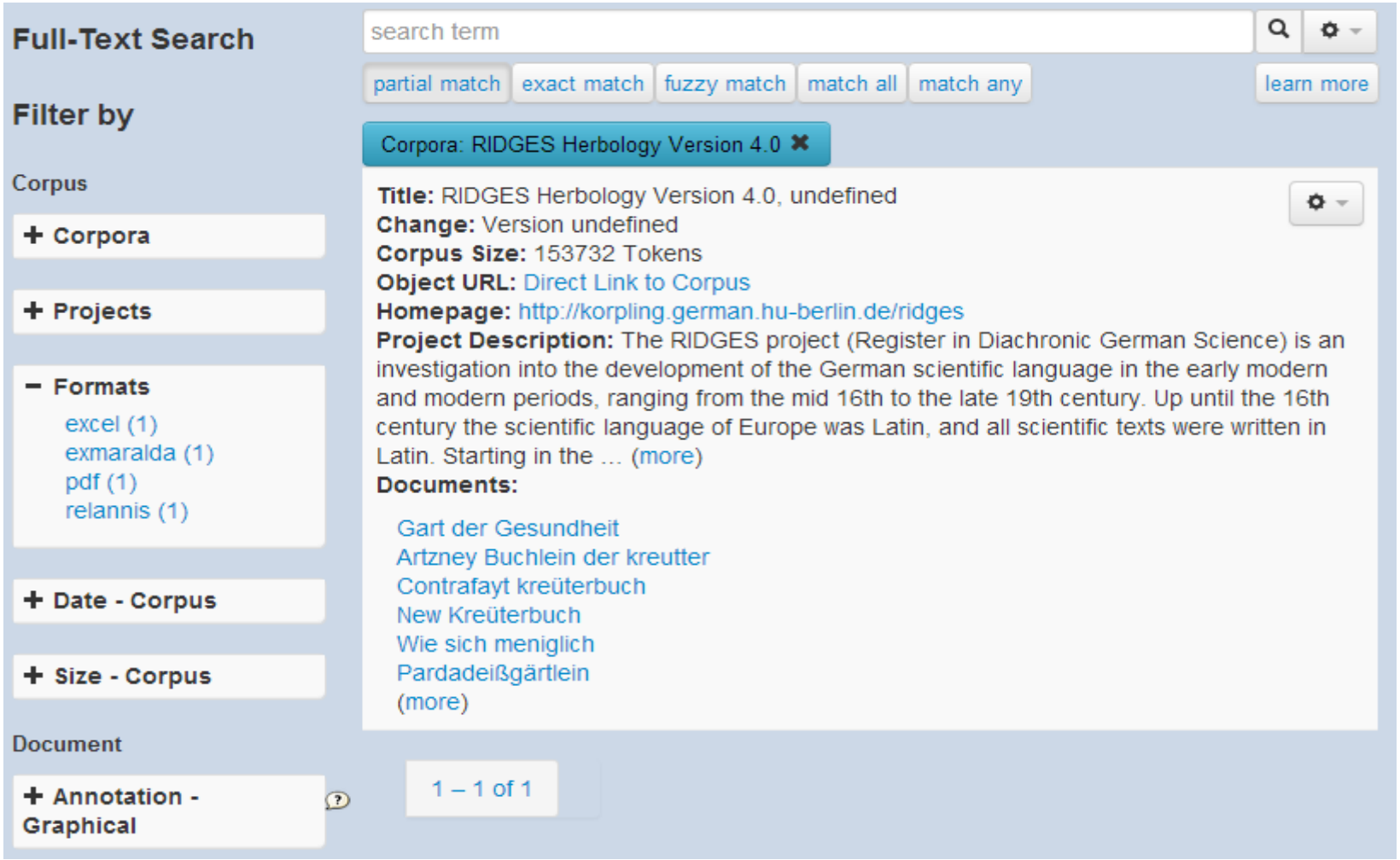
Search

View

Modify

(Configuration)


Home » View » RIDGES-Herbology
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):
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


Full-Text Search
search term
partial match
exact match
fuzzy match
match all
match any
learn more
Filter by
Corpora: RIDGES Herbology Version 4.0
Corpus
+ Corpora
+ Projects
− Formats
excel (1)
exmaralda (1)
pdf (1)
relannis (1)
+ Date - Corpus
+ Size - Corpus
Document
+ Annotation - Graphical
Title: RIDGES Herbology Version 4.0, undefined
Change: Version undefined
Corpus Size: 153732 Tokens
Object URL: Direct Link to Corpus
Homepage: http://korpling.german.hu-berlin.de/ridges
Project Description: The RIDGES project (Register in Diachronic German Science) is an investigation into the development of the German scientific language in the early modern and modern periods, ranging from the mid 16th to the late 19th century. Up until the 16th century the scientific language of Europe was Latin, and all scientific texts were written in Latin. Starting in the … (more)
Documents:
Gart der Gesundheit
Artzney Buchlein der kreutter
Contrafayt kreüterbuch
New Kreüterbuch
Wie sich meniglich
Pardadeißgärtlein
(more)
1 – 1 of 1

Software und Standards

- ✓ Modularer Aufbau auf der Basis der Fedora Repository-Software
- ✓ Source Code verfügbar über GitHub
- ✓ Einbindung des Forschungswerkzeugs ANNIS (Suche und Visualisierung)
- ✓ Beschreibung der Korpora auf Basis von TEI P5
- ✓ Handle-Identifier für Korpora
- ✓ Lizenzierung der Korpora mit Creative Commons

- – Flat-Design auf Basis von aktuellen Webstandards in Vorbereitung
- – Zertifizierung/Orientierung an Standards/Guidelines für Forschungsdatenrepositories (z.B. Data Seal of Approval) in Vorbereitung

Herausforderungen

Projektbezogen, aber auch im Hinblick auf HU-Aktivitäten und Folgeprojekt(e):

Begriffsverständnis von Forschungsdaten

Komplexe Dokumentation (Aufbereitungsschritte)

Dokumentationsaufwand

Geringe Verbreitung von Standards der Datenaufbereitung und Speicherung

Open-Access-Zugang zu historischen Korpora noch nicht etabliert

Herausforderungen

Angebot für ein Teilgebiet einer Disziplin

Einbindung weiterer Fachdisziplinen wie z.B. Musik- , Geschichts- und Literaturwissenschaftler, die auf altdeutschen Texten arbeiten

Integration in Forschungsinfrastrukturen

Metadaten-Mapping

Veröffentlichung von Metadaten als Linked Open Data

Versionierung / Zitation von Forschungsdaten

Übertragbarkeit von Know-how aus fachspezifischen Projekten

Integration von Infrastruktur-Angeboten für HU-ForscherInnen sowie Fachdisziplinen

Forschungsdatenmanagement-Initiative an der HU

- Umfrage
- Forschungsdaten-Policy
- Guidelines
- Konzeption eines umfassenden Infrastruktur- und Serviceangebots

Weiterentwicklung zu einem / Aufbau eines disziplinübergreifenden Forschungsdatenrepositories?

Vielen Dank für Ihre Aufmerksamkeit!

Fragen/Feedback: maxi.kindling@hu-berlin.de