Metadata in an Infrastructure for Historical Corpora

Carolin Odebrecht*,Thomas Krause*

Computer- und Medienservice HU Berlin

INRIA, France

*Korpuslinguistik und Morphologie HU Berlin

Sprachgeschichte HU Berlin

DFG Deutsche Forschungsgemeinschaft

Outline

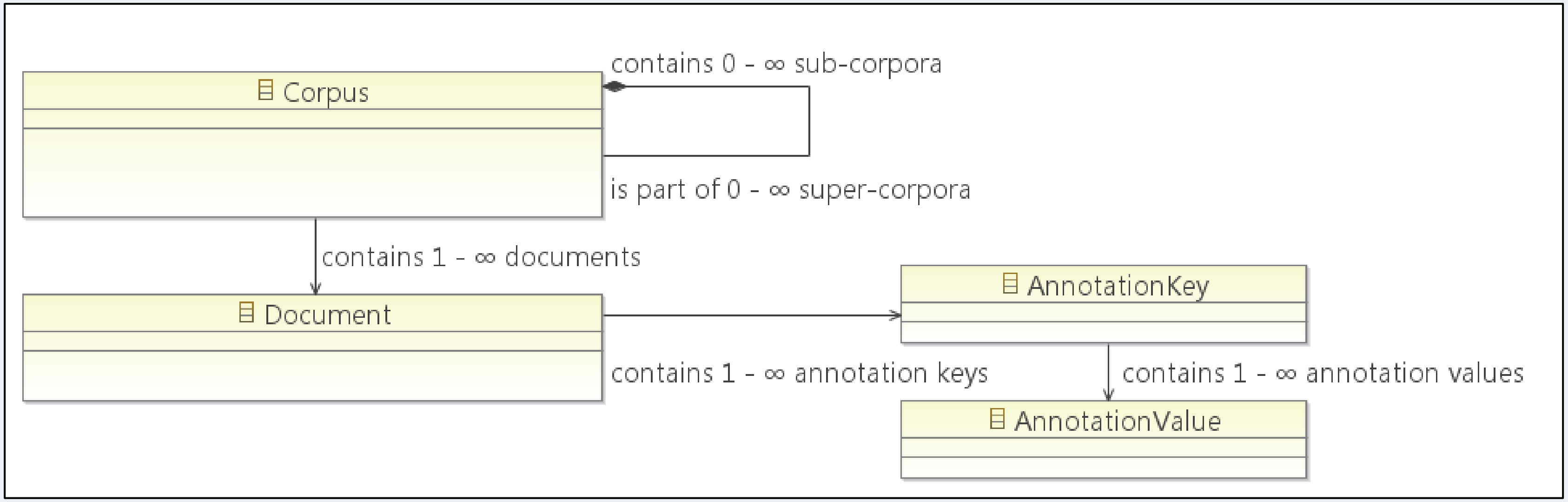
Which functions do metadata need to have for a corpus-linguistics repository?

- corpora
- use of corpora
- metadata
- repository and infrastructure

CORPORA

Corpora Concept

- concept of a corpus
  - corpus
    - set of documents
    - sub-corpora, super-corpora
  - document
    - set of annotations
  - annotation
    - key-value-pairs
    - groups of annotations


Corpus
contains 0 - ∞ sub-corpora
is part of 0 - ∞ super-corpora
contains 1 - ∞ documents
Document
AnnotationKey
contains 1 - ∞ annotation keys
contains 1 - ∞ annotation values
AnnotationValue

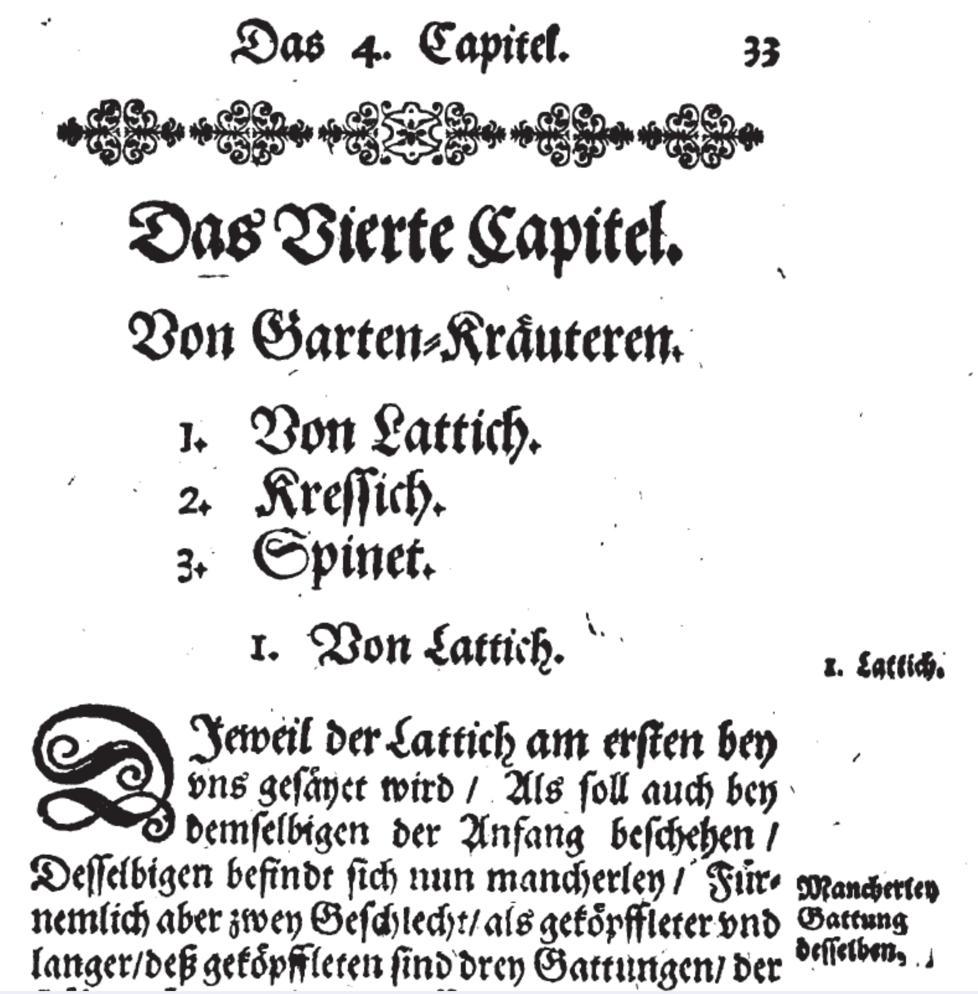
Corpora

- historical corpora
  - small amount of historical texts (900-1900)
  - expensive data preparation
    - manual diplomatic transcription (OCR needs to be corrected as well)
    - manual and semi-automatic normalization
    - manual and semi-automatic annotation
  - multi-layer annotation is necessary (Dipper et al. 2004)

Corpora

- title: Pflantz-Gart
- date: 1639
- place: Bern
- author: Daniel Rhagor

(RIDGES Herbology Corpus)



Das 4. Capitel. 33

Das Vierte Capitel.

Von Garten-Kräuteren.

1. Von Lattich.
2. Kreſſich.
3. Spinet.

1. Von Lattich.

1. Lattich.

DJeweil der Lattich am erſten bey vns geſäyet wird / Als ſoll auch bey demſelbigen der Anfang beſchehen / Deſſelbigen befindt ſich nun mancherley / Fürnemlich aber zwey Geſchlecht / als geköpffleter vnd langer / deß geköpffleten ſind drey Gattungen / der

Mancherley Gattung deſſelben.

Corpora

- for example: research on verbal argument selection
  - word forms, types and normalization
  - identifying verbs and potential arguments

- e.g.

  a) [...] *und da ichs wieder zur Hand nahm*

  ‘and since I took it in (my) hand again’

  b) [...] *vnd wie viel verſtanden / zuerklaͤren. Das magſtu*

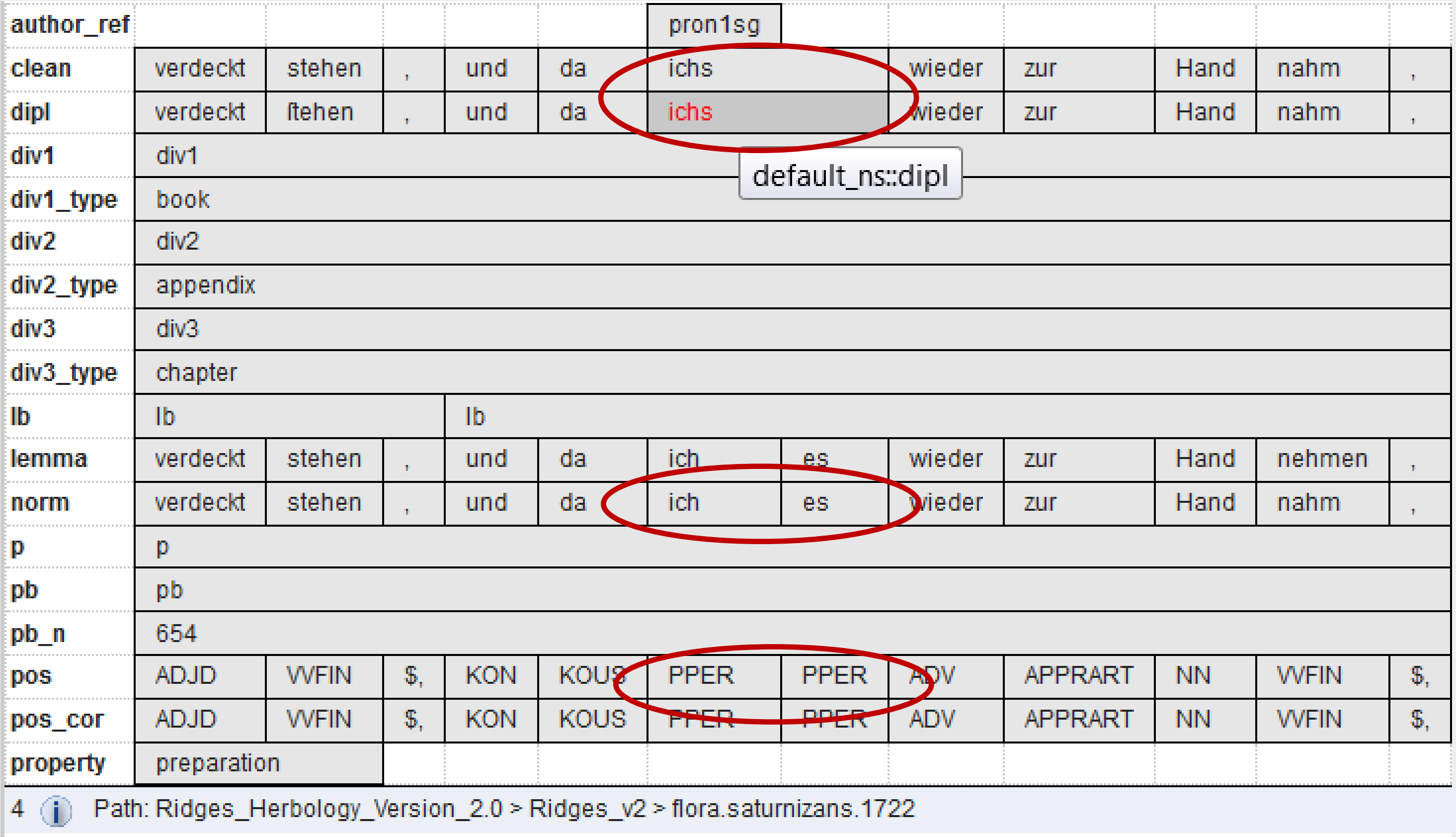
  in Gottes namen zum beſten brauchen.

  ‚and how much to explain comprehensibly. That you may best use in God‘s name.‘

  (a) Flora Saturnizans 1722 b)Alchimistische Praktik 1603, RIGDES Herbology Version 2.0, access: https://korpling.german.hu-berlin.de/annis3/)

Corpora

a) [...] *und da ichs wieder zur Hand nahm*

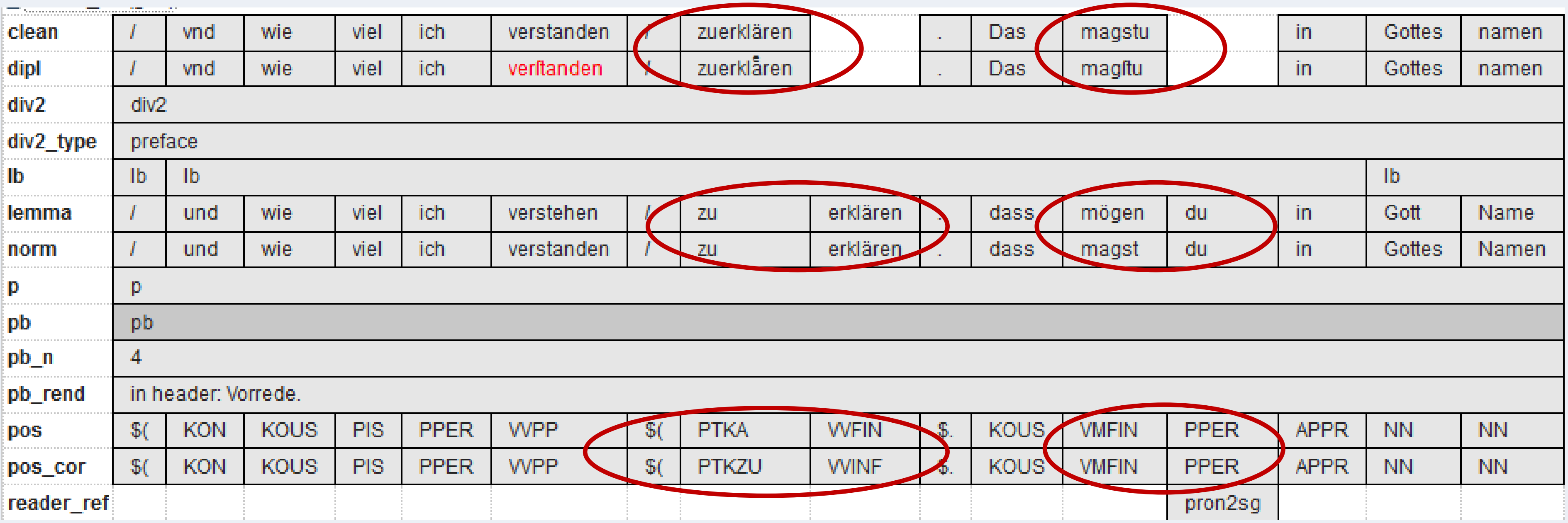


author_ref						pron1sg						
clean	verdeckt	stehen	,	und	da	ichs		wieder	zur	Hand	nahm	,
dipl	verdeckt	ſtehen	,	und	da	ichs		wieder	zur	Hand	nahm	,
div1	div1											
div1_type	book											
div2	div2											
div2_type	appendix											
div3	div3											
div3_type	chapter											
lb	lb			lb								
lemma	verdeckt	stehen	,	und	da	ich	es	wieder	zur	Hand	nehmen	,
norm	verdeckt	stehen	,	und	da	ich	es	wieder	zur	Hand	nahm	,
p	p											
pb	pb											
pb_n	654											
pos	ADJD	VVFIN	$,	KON	KOUS	PPER	PPER	ADV	APPRART	NN	VVFIN	$,
pos_cor	ADJD	VVFIN	$,	KON	KOUS	PPER	PPER	ADV	APPRART	NN	VVFIN	$,
property	preparation											

4 Path: Ridges_Herbology_Version_2.0 > Ridges_v2 > flora.saturnizans.1722

Corpora

b) [...] *vnd wie viel verſtanden / zuerklåren. Das magſtu in Gottes namen zum beſten brauchen.*



clean	/	vnd	wie	viel	ich	verstanden	/	zuerklären		.	Das	magstu		in	Gottes	namen
dipl	/	vnd	wie	viel	ich	verſtanden	/	zuerklåren		.	Das	magſtu		in	Gottes	namen
div2	div2															
div2_type	preface															
lb	lb	lb													lb	
lemma	/	und	wie	viel	ich	verstehen	/	zu	erklären	.	dass	mögen	du	in	Gott	Name
norm	/	und	wie	viel	ich	verstanden	/	zu	erklären	.	dass	magst	du	in	Gottes	Namen
p	p															
pb	pb															
pb_n	4															
pb_rend	in header: Vorrede.															
pos	$(	KON	KOUS	PIS	PPER	VVPP	$(	PTKA	VVFIN	$.	KOUS	VMFIN	PPER	APPR	NN	NN
pos_cor	$(	KON	KOUS	PIS	PPER	VVPP	$(	PTKZU	VVINF	$.	KOUS	VMFIN	PPER	APPR	NN	NN
reader_ref													pron2sg			

Corpora

- technical tokenization
  - smallest unit
- conceptual tokenization/segmentation
  - separate spelling
  - different kinds of normalizations
  - used for search context and specialized precedence queries

➢ infrastructure needs to support multiple tokenizations

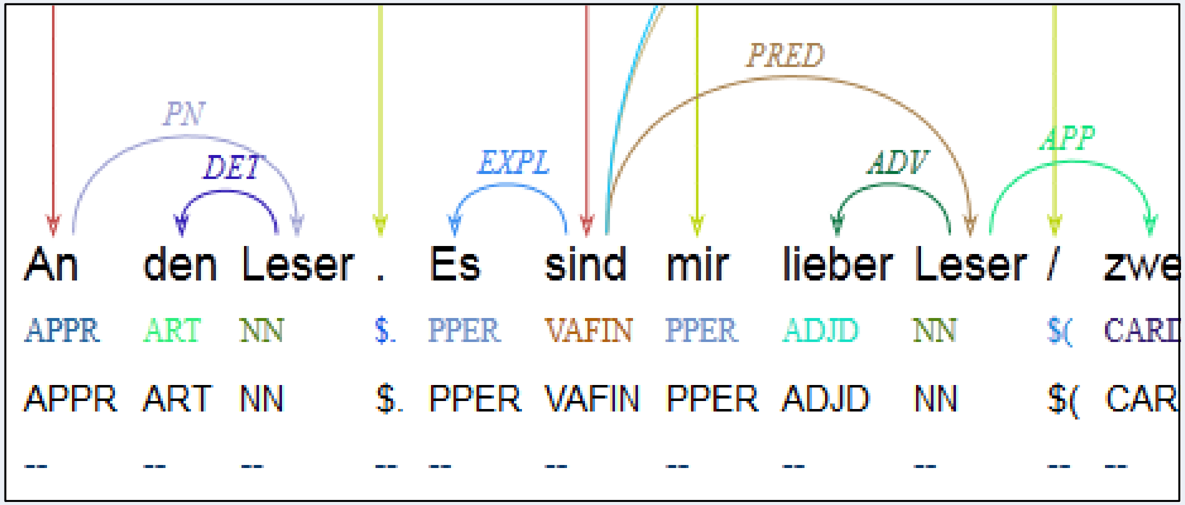
Corpora

- complex corpus structures
  - multiple tokenizations (cf. Krause et al. 2012)
  - normalization (cf. Bollmann et al. 2011)
  - semi-automatic correction (cf. Dickenson & Meurers 2003)
- applying tagging and parsing tools (or further manual annotation) to a normalized annotation

➢ the steps are necessary for further research, e.g. on verbal argument realization

Corpora RIDGES

- e.g. dependency annotations (cf. Nivre 2006) for research on argument selection



An	den	Leser	.	Es	sind	mir	lieber	Leser	/	zwe
APPR	ART	NN	$.	PPER	VAFIN	PPER	ADJD	NN	$(	CARI
APPR	ART	NN	$.	PPER	VAFIN	PPER	ADJD	NN	$(	CAR
--	--	--	--	--	--	--	--	--	--	--

Dependencies in ‚Alchimistische Praktik' (1603), RIDGES Herbology Corpus V2

Corpora

- preparation depends on research question
- different ways to tokenize and normalize
- automatic or semi-automatic preparation does not necessary have good performance

USE OF CORPORA

Use of Corpora

- use of these heterogeneous research data depends on the research question
- workflow
  - annotate data
  - search in data locally
  - publish results

Use of Corpora: LAUDATIO

- (re-) use of these heterogeneous research data depends on the research question
- workflow
  - annotate data
  - search in data locally
  - publish results
  - import data into repository using a CC-BY open source license
  - search data on public server
  - access data

Use of Corpora: LAUDATIO

- research on e.g. diachronic verbal argument selection
  - search corpora for study
    - which texts (document) are in each corpus
    - collect the important texts for the study
  - annotation
    - get information about the annotation in the corpus
    - further annotation
  - publish the extended corpus (CC-BY)

➢ getting necessary information by metadata

➢ access and publishing via repository

METADATA

Metadata

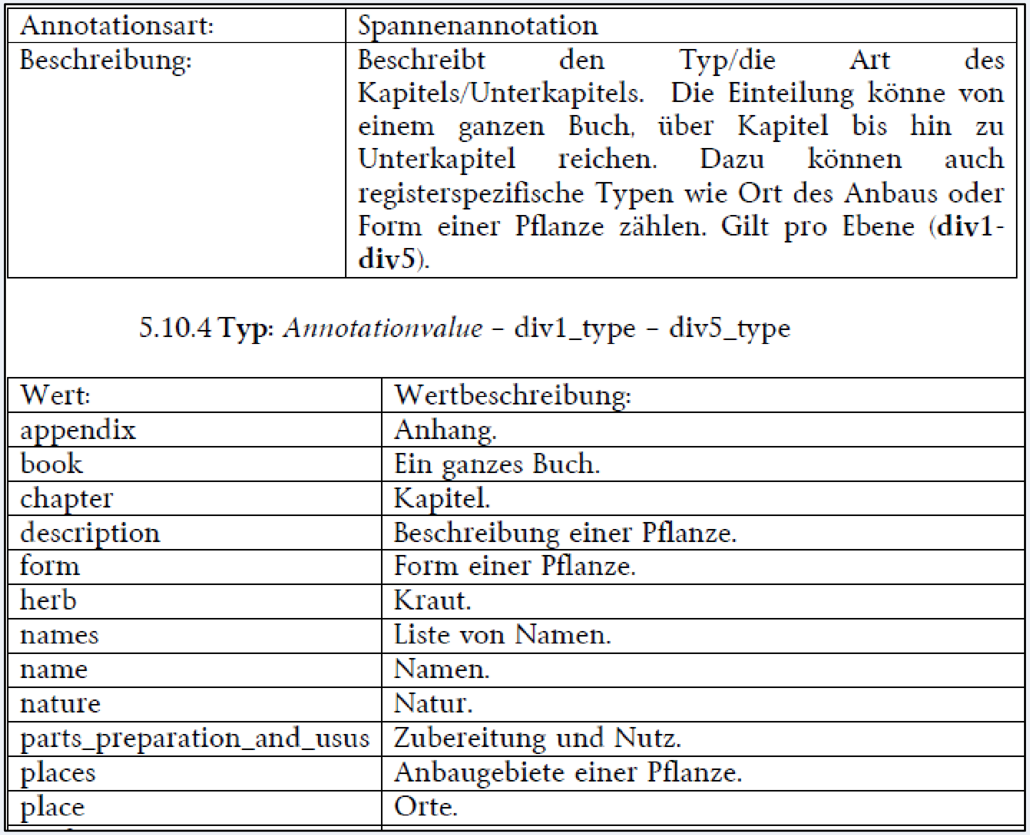
- corpus documentation
  - common documentation of a single corpus
  - different approaches, functions and use cases
    - tagsets
    - project description
    - research
- publications / technical reports which document the corpus

Metadata

- common corpus documentation
  - texts /documents
  - tagset
  - format
  - research
- idiosyncratic structure
  - free texts, tables

Metadata

RIDGES Herbology V2 Corpus Documentation



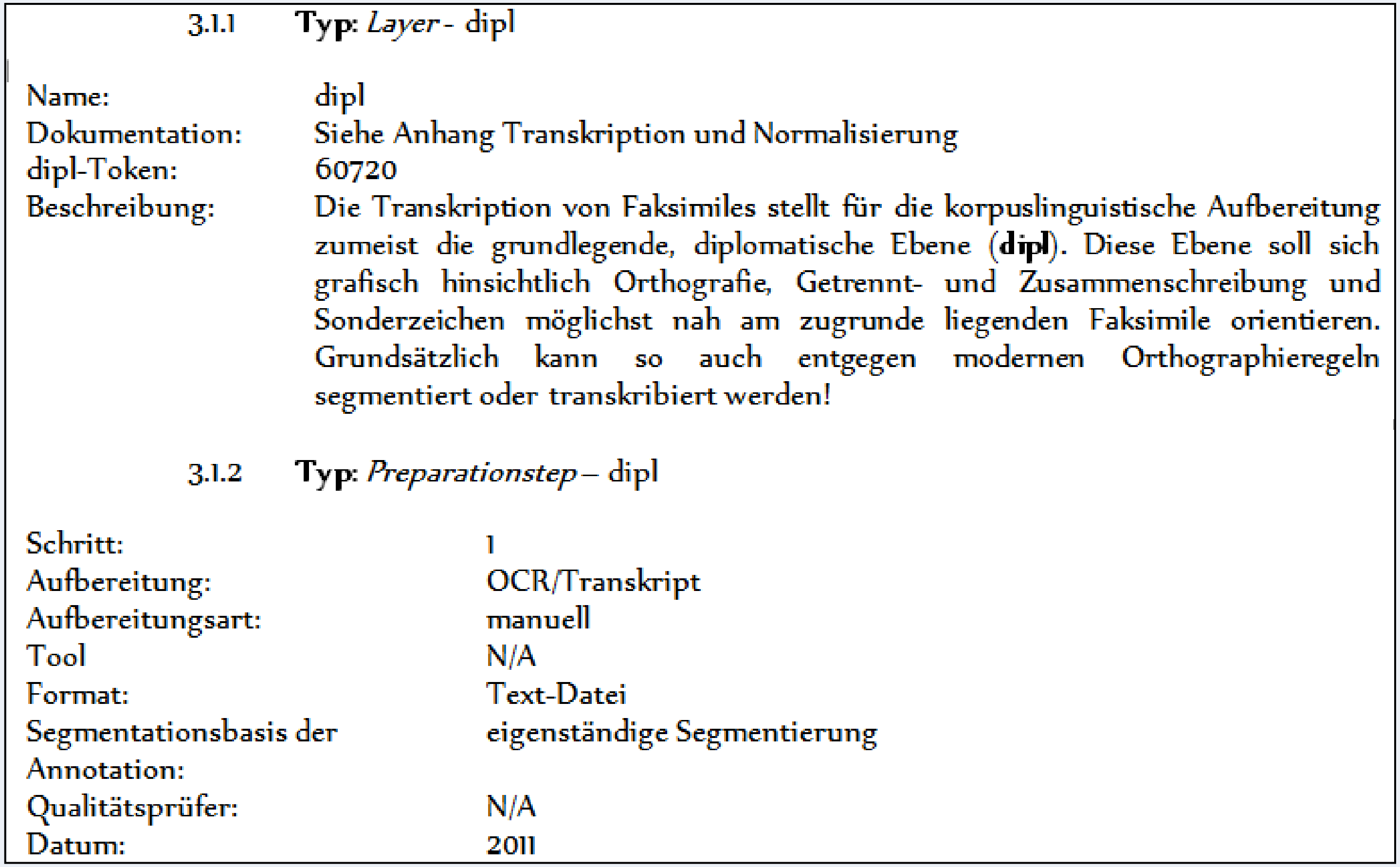
Annotationsart:	Spannenannotation
Beschreibung:	Beschreibt den Typ/die Art des Kapitels/Unterkapitels. Die Einteilung könne von einem ganzen Buch, über Kapitel bis hin zu Unterkapitel reichen. Dazu können auch registerspezifische Typen wie Ort des Anbaus oder Form einer Pflanze zählen. Gilt pro Ebene (**div1-div5**).

5.10.4 **Typ**: *Annotationvalue* – div1_type – div5_type

Wert:	Wertbeschreibung:
appendix	Anhang.
book	Ein ganzes Buch.
chapter	Kapitel.
description	Beschreibung einer Pflanze.
form	Form einer Pflanze.
herb	Kraut.
names	Liste von Namen.
name	Namen.
nature	Natur.
parts_preparation_and_usus	Zubereitung und Nutz.
places	Anbaugebiete einer Pflanze.
place	Orte.

Metadata

RIDGES Herbology V2 Corpus Documentation



3.1.1 Typ: *Layer* - dipl

Name:	dipl
Dokumentation:	Siehe Anhang Transkription und Normalisierung
dipl-Token:	60720
Beschreibung:	Die Transkription von Faksimiles stellt für die korpuslinguistische Aufbereitung zumeist die grundlegende, diplomatische Ebene (**dipl**). Diese Ebene soll sich grafisch hinsichtlich Orthografie, Getrennt- und Zusammenschreibung und Sonderzeichen möglichst nah am zugrunde liegenden Faksimile orientieren. Grundsätzlich kann so auch entgegen modernen Orthographieregeln segmentiert oder transkribiert werden!

3.1.2 Typ: *Preparationstep* – dipl

Schritt:	1
Aufbereitung:	OCR/Transkript
Aufbereitungsart:	manuell
Tool	N/A
Format:	Text-Datei
Segmentationsbasis der Annotation:	eigenständige Segmentierung
Qualitätsprüfer:	N/A
Datum:	2011

Metadata

- common corpus documentation
  - sufficient for one corpus
  - documentation of what is in the corpus

- What do we need to consider if we would like to document **more than one** corpus?
  - structured documentation
  - not idiosyncratic, more general approach
  - extended understanding of metadata

➢ re-use of one or more corpora for your own research

Extended Metadata

- wider scope of corpus documentation
  - metadata for object and content documentation
  - modeling of metadata
    - concept of a corpus; documents, annotation
    - project, research
    - corpus preparation
- uniform metadata model for a heterogeneous field of research data
- structural searching for research data

Extended Metadata

- wider scope of corpus documentation
  - metadata for object and content documentation
    - what is in the corpus (texts, tagsets)
    - what is the corpus made of (formats, tools, pipelines, corpus architecture)
    - who creates the corpus and for which purpose (projects, research)

METADATA META-MODEL

Metadata Meta-Model

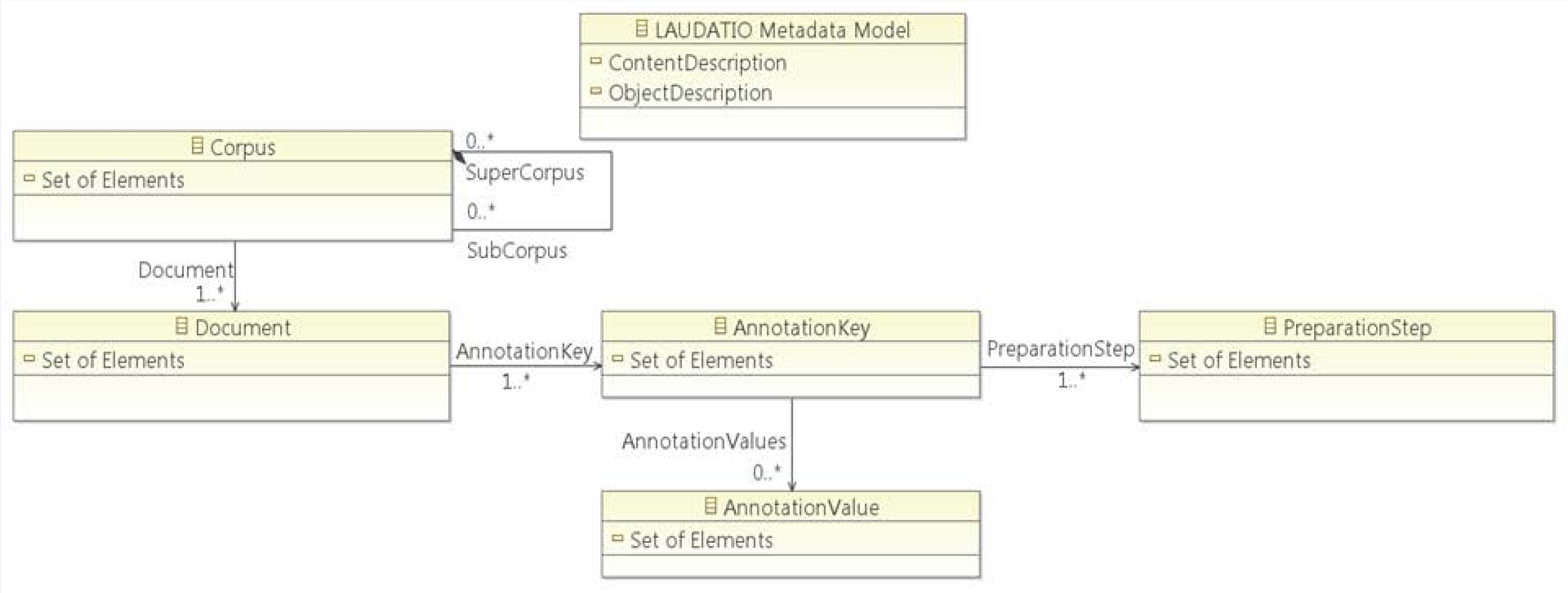
- modelling of metadata
  - defining the function of the documentation
  - defining what needs to be documented
  - defining objects, domains, relations between them

- object- and content documentation for a (re)-use of digital research data (corpus), repository

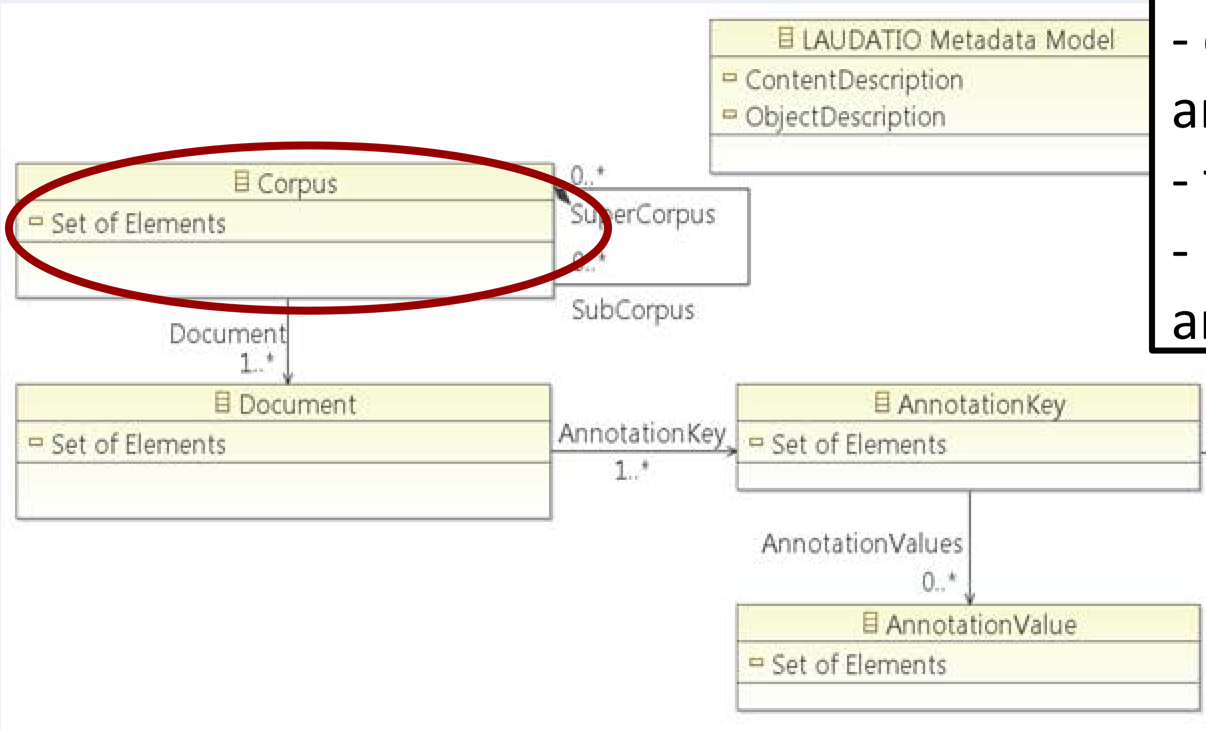
➢ concept of a corpus: corpus, document, annotation,

➢ corpus preparation steps

Metadata Meta-Model



Metadata Meta-Model



contains information:
- project
- corpus editors and annotators
- formats
- list of documents and annotation

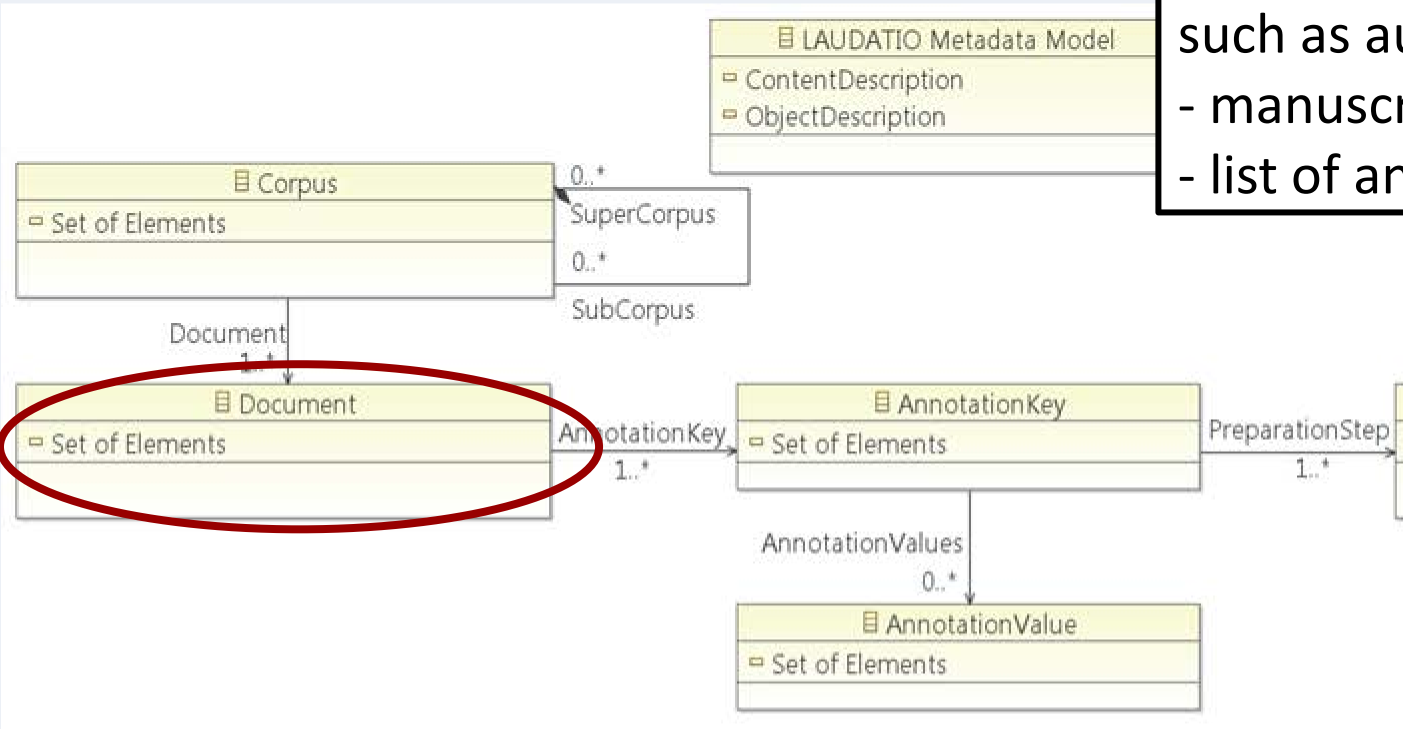
Metadata RIDGES

- **R**egister **in** **D**iachronic **Ge**rman **S**cience
- seminar project initiated by Anke Lüdeling and Amir Zeldes
- TEI xml, EXCEL, PAULA, relANNIS
- for each format a list of annotation

contains information:
- project
- corpus editors and annotators
- formats
- list of documents and annotation

Ridges

Metadata Meta-Model



contains information:
- bibliographic information such as author and year
- manuscript history
- list of annotation keys

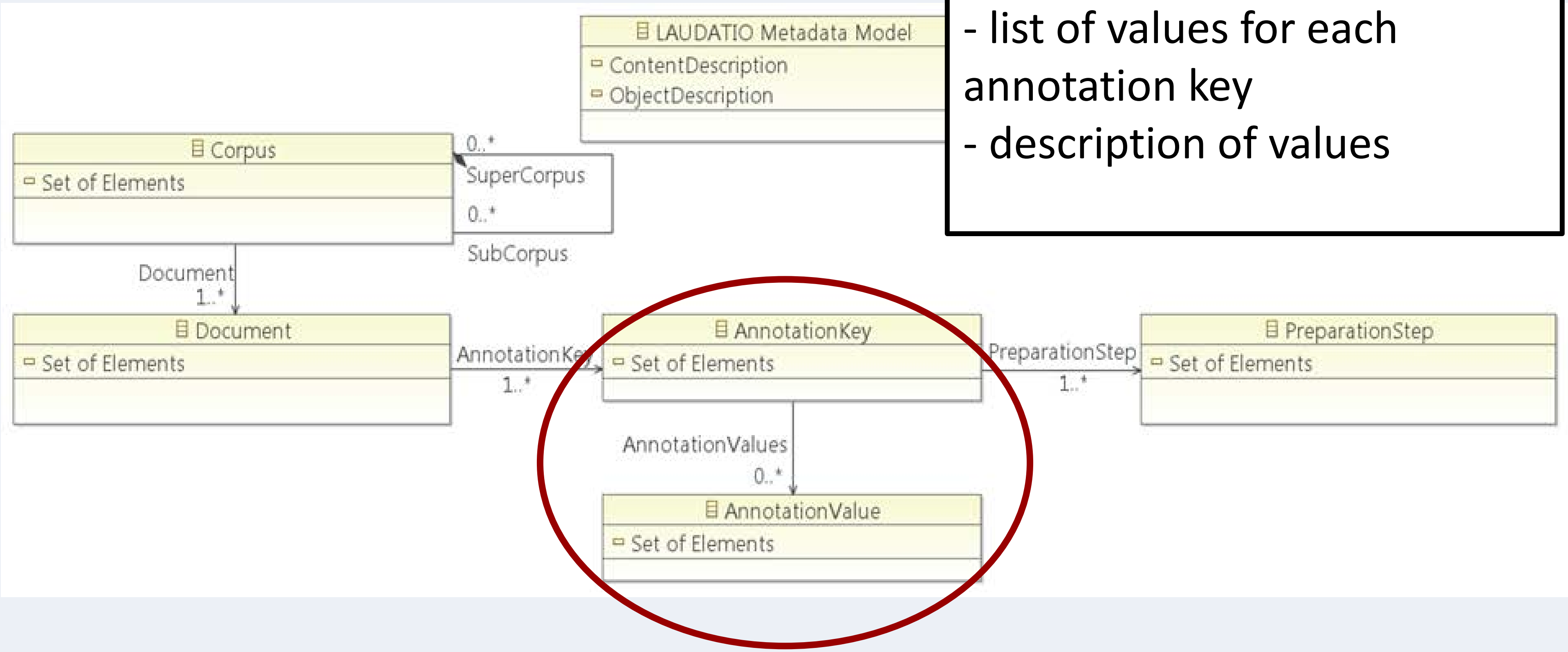
Metadata RIDGES

Title	Deutsche Pflanzennamen
Short	deutsche.pflanzennamen.1870
Author	Graßmann, Hermann
Date	1870
Place	Stettin
Scope	pp. 1-23
Token	10283
Register	Kräuterkunde
Annotation	dipl, clean, pos, norm, lemma, pos_cor, pb, pb_n, lb, [...]

contains information:
- bibliographic information such as author and year
- manuscript history
- list of annotation keys

Metadata Meta-Model

contains information:
- typical annotation guideline
- list of values for each annotation key
- description of values



Metadata RIDGES

- ‘dipl’, ‘clean’ and ‘norm’
  – two-step normalization with a multiple tokenization

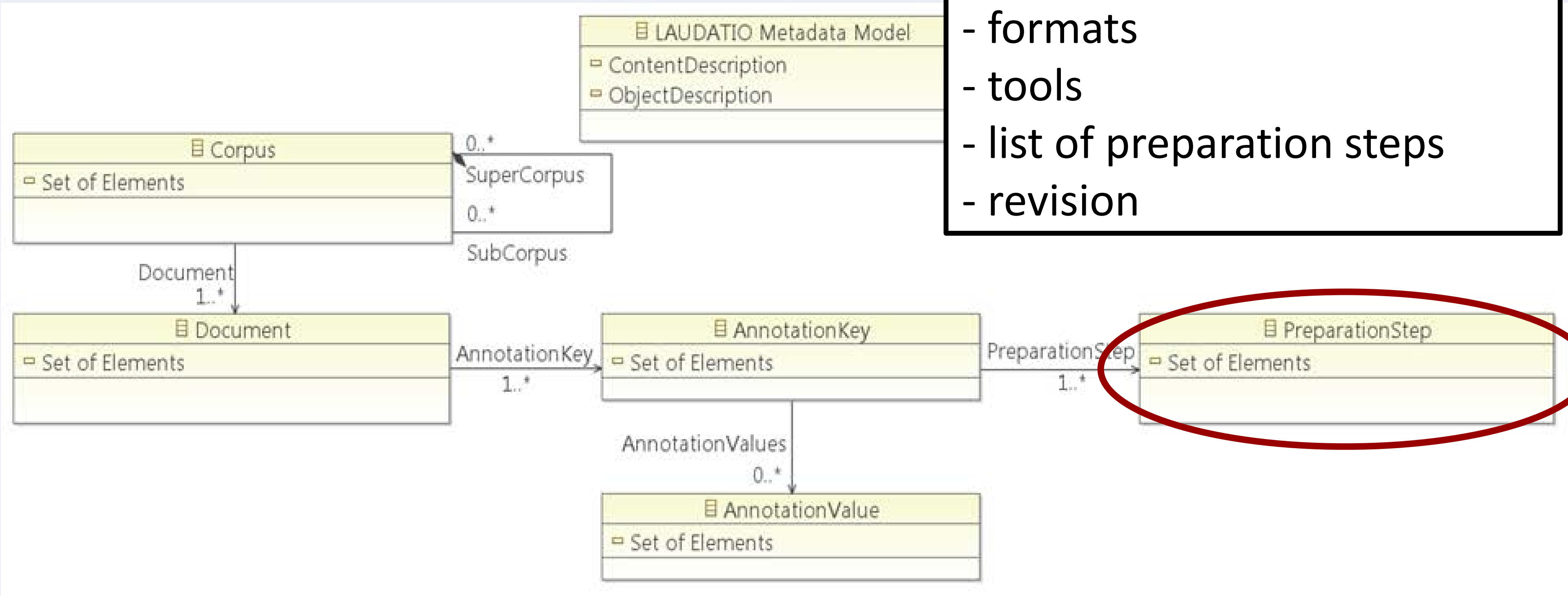
contains information:
- typical annotation guideline
- list of values for each annotation key
- description of values

- ‚hi_rend‘

italics	Text, der kursiv gedruckt ist.
bold	Text, der fett gedruckt ist.
underlined	Text, der unterstrichen gedruckt ist.
red	Text, der rot gedruckt ist.

Metadata Meta-Model

contains information:
- corpus architecture
- formats
- tools
- list of preparation steps
- revision



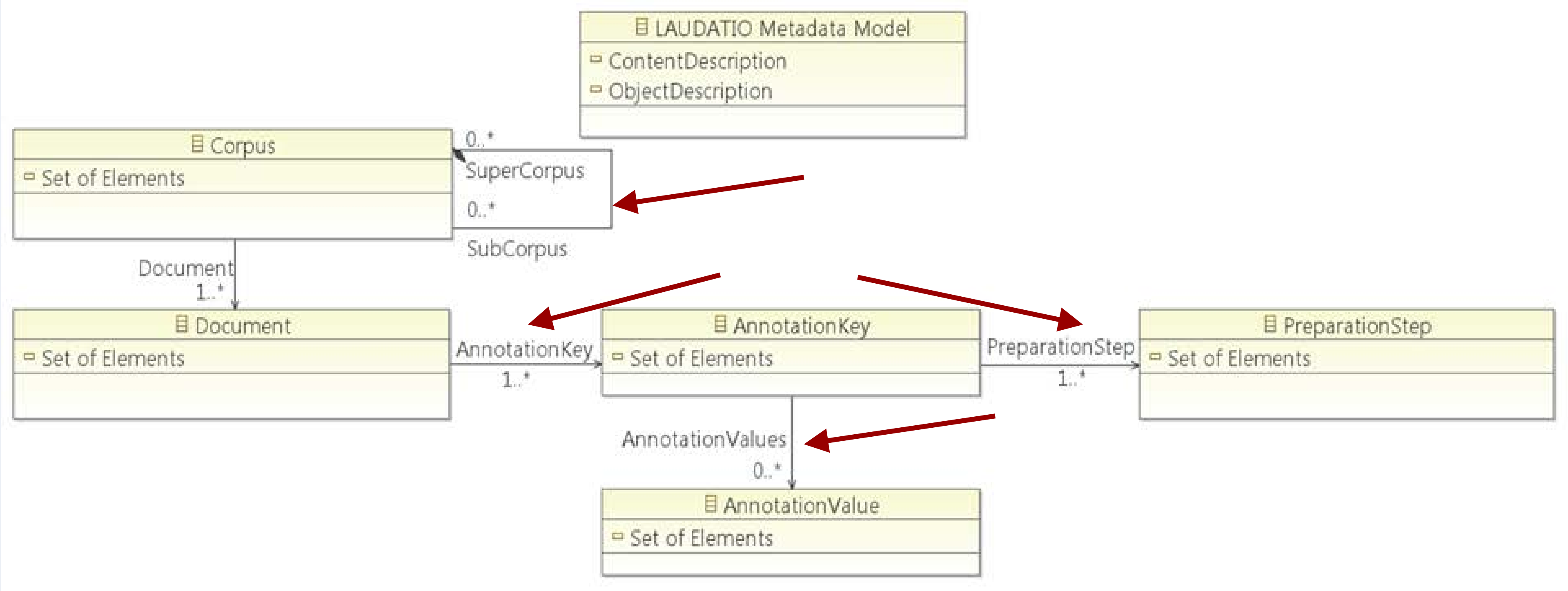
Metadata RIDGES

- ‘dipl’
  - diplomatic transcription, including line breaks, special characters (plain text)
  - tokenized by treetagger
  - import in EXCEL
    - all corrections and changes are applied here
  - conversion from EXCEL to EXMARaLDA via Excel ADDIN
  - conversion from EXMARaLDA to relANNIS via PepperModule
  - revision history

contains information:
- corpus architecture
- formats
- tools
- list of preparation steps
- revision

Metadata Meta-Model

relations between corpus concepts



Metadata Meta-Model

- citation and references of research data
  - the whole corpus
  - documents
  - annotation keys
  - editors and annotators, projects

Metadata Meta-Model

- specificity
  - corpus linguistics vs. other sources-based domains
  - multi-layered corpora
- genericity
  - concept of a corpus
  - precise description of what an annotation level is
  - most text-based projects end up using corpus linguistic methods and tools

Metadata Meta-Model

- existing metadata formats, e.g.
  - specific: TEI subsets such as epidoc
  - generic: CMDI
- serve other functions and therefore do not capture all the required information:
  - concepts of a corpus
  - encoding and annotating historical corpora
  - focus on re-use of corpora

TEI ODD

Double Role

- *target representation format* for primary sources and multi-layered annotations
  - documentation (and transcriptions)
  - focus on metadata (<teiHeader>)
- *modeling tool* for the project
  - ODD as a specification language
  - maximizing compliance with TEI framework
  - introducing a reference customization for annotated corpora

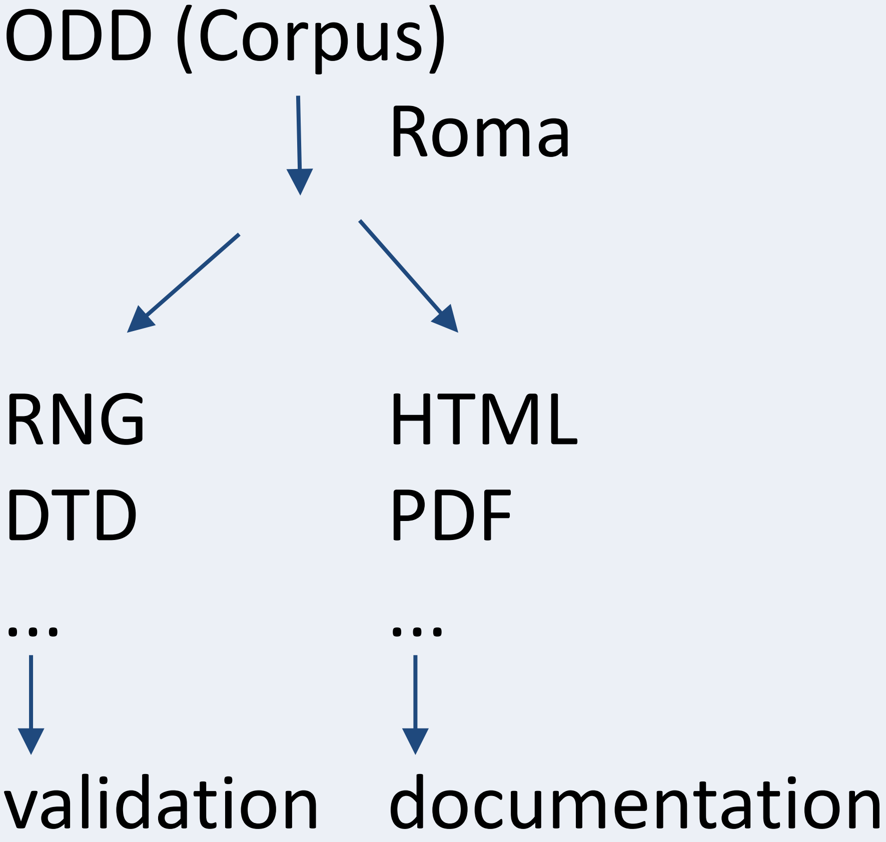
LAUDATIO-ODD

- TEI offers the opportunity for customization
  - "One document does it all“
  - ➢ schema for schema

- using modules and elements of the TEI
- customization of TEI Header for each concept (corpus, document, annotation, preparation)
- versioning for further customization (new corpora)

➢ meta-model defines what is encoded

TEI - ODD



ODD (Corpus)

Roma

RNG

DTD

...

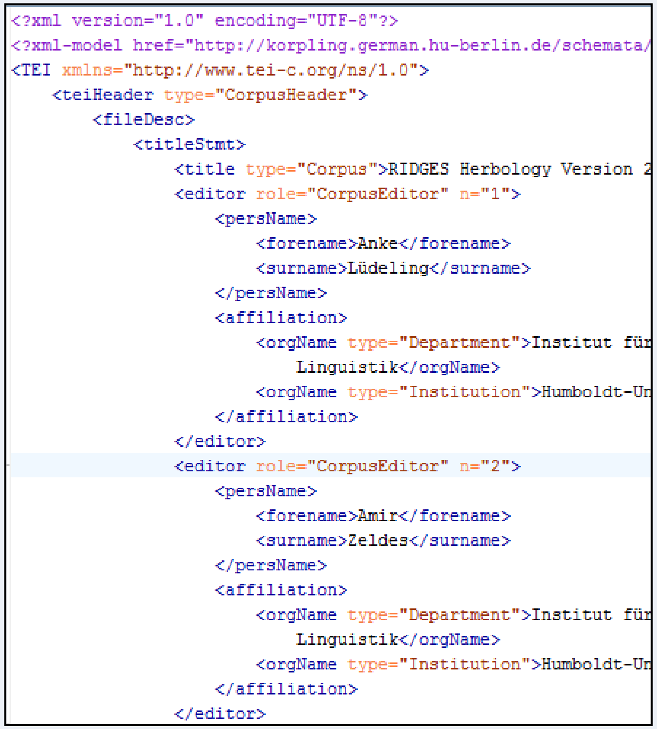
validation

HTML

PDF

...

documentation



```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="http://korpling.german.hu-berlin.de/schemata/
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader type="CorpusHeader">
        <fileDesc>
            <titleStmt>
                <title type="Corpus">RIDGES Herbology Version 2
                <editor role="CorpusEditor" n="1">
                    <persName>
                        <forename>Anke</forename>
                        <surname>Lüdeling</surname>
                    </persName>
                    <affiliation>
                        <orgName type="Department">Institut für
                            Linguistik</orgName>
                        <orgName type="Institution">Humboldt-Un
                    </affiliation>
                </editor>
                <editor role="CorpusEditor" n="2">
                    <persName>
                        <forename>Amir</forename>
                        <surname>Zeldes</surname>
                    </persName>
                    <affiliation>
                        <orgName type="Department">Institut für
                            Linguistik</orgName>
                        <orgName type="Institution">Humboldt-Un
                    </affiliation>
                </editor>
```

TEI Header Corpus, RIDGES V2

PRINCIPLES OF THE LAUDATIO PROJECT

LAUDATIO

Long-term	Computer- und Medienservice HU Berlin
Access	repository
and **U**sage	repository / virtual research environment
of **D**eeply **A**nnotated	token and span annotation, tree bank, dependencies
Informat**io**n	historical texts

“Information” part

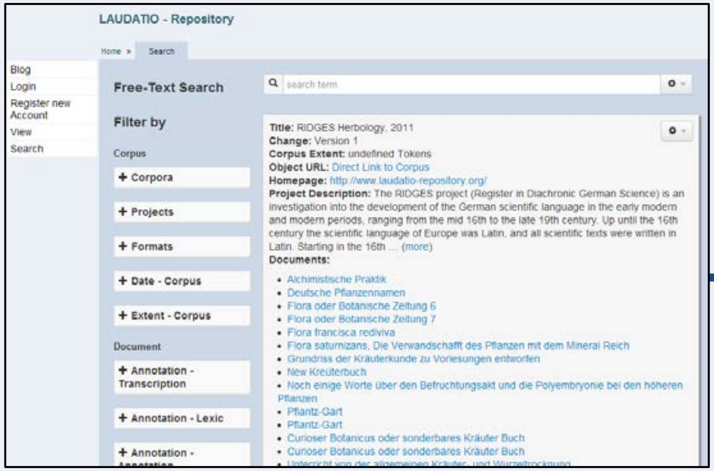
- focus on German historical texts
  - 900-1900; every dialect
  - every register
  - every kind of primary texts, from manual manuscripts to printed publications

- choice depends on research questions and methods

“Deeply Annotated” part

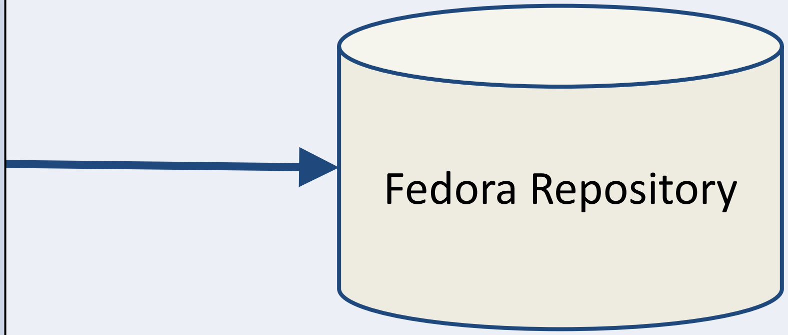
- preparation of the historic corpora
  - transcription
    - encoding, special characters, orthography ...
  - (multiple) tokenization
  - normalization
  - annotation
    - based on the common tokenization

Access the data



Title: RIDGES Herbology, 2011
Change: Version 1
Corpus Extent: undefined Tokens
Object URL: Direct Link to Corpus
Homepage: http://www.laudatio-repository.org/
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 16th ... (more)
Documents:

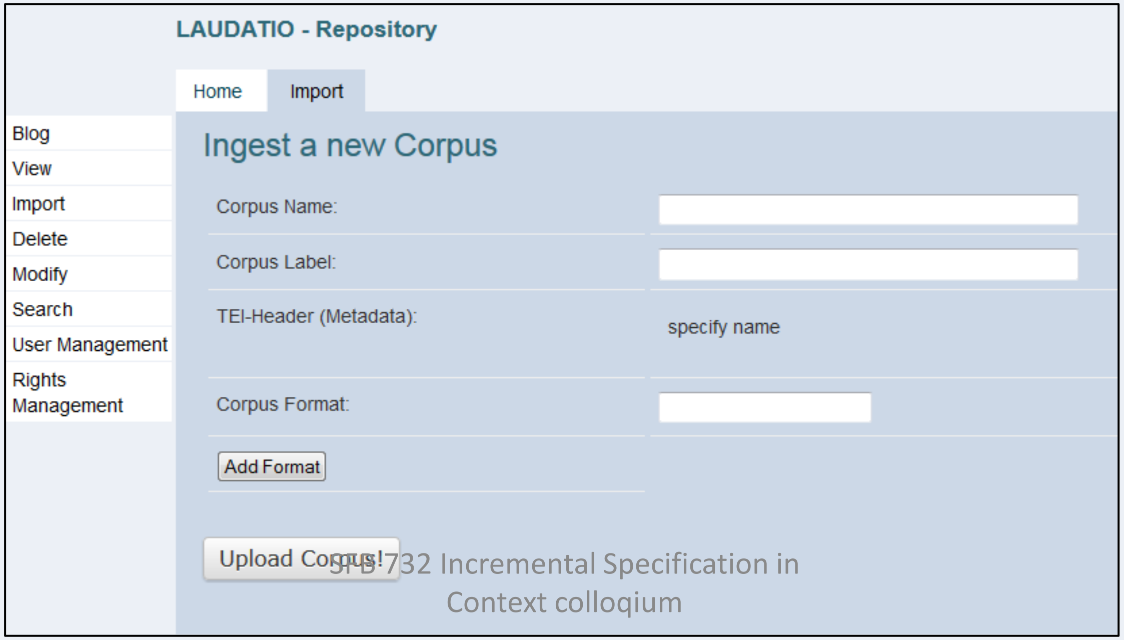
- Alchimistische Praktik
- Deutsche Pflanzennamen
- Flora oder Botanische Zeitung 6
- Flora oder Botanische Zeitung 7
- Flora francisca rediviva
- Flora saturnizans, Die Verwandschafft des Pflanzen mit dem Mineral Reich
- Grundriss der Kräuterkunde zu Vorlesungen entworfen
- New Kreüterbuch
- Noch einige Worte über den Befruchtungsakt und die Polyembryonie bei den höheren Pflanzen
- Pflantz-Gart
- Pflantz-Gart
- Curioser Botanicus oder sonderbares Kräuter Buch
- Curioser Botanicus oder sonderbares Kräuter Buch
- Unterricht von der allgemeinen Kräuter- und Wurzeltrocknung



web interface for access to the repository

Access the data: import

- upload new corpora in different formats
  - each might cover different aspects of the data
- versioning system
- for each upload there is an automatic validation against the metadata scheme



Ingest a new Corpus

Corpus Name:

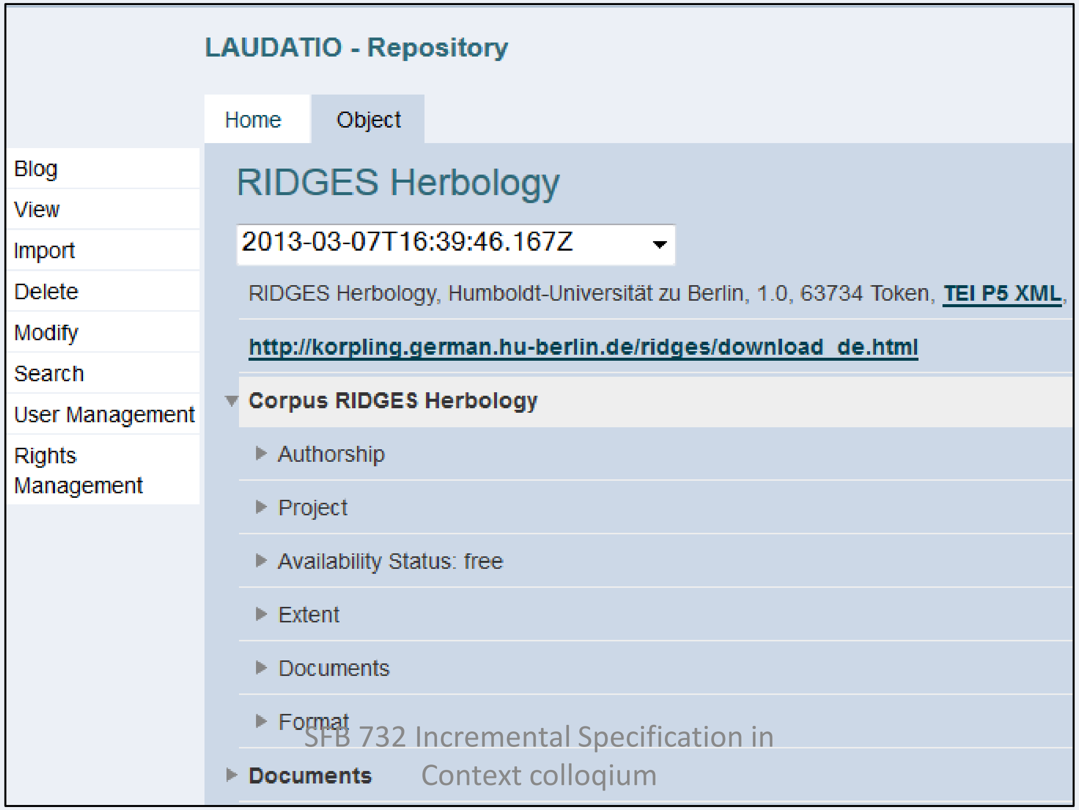
Corpus Label:

TEI-Header (Metadata): specify name

Corpus Format:

Access the data: show

- structured view of the unified metadata
- version selection



RIDGES Herbology

2013-03-07T16:39:46.167Z

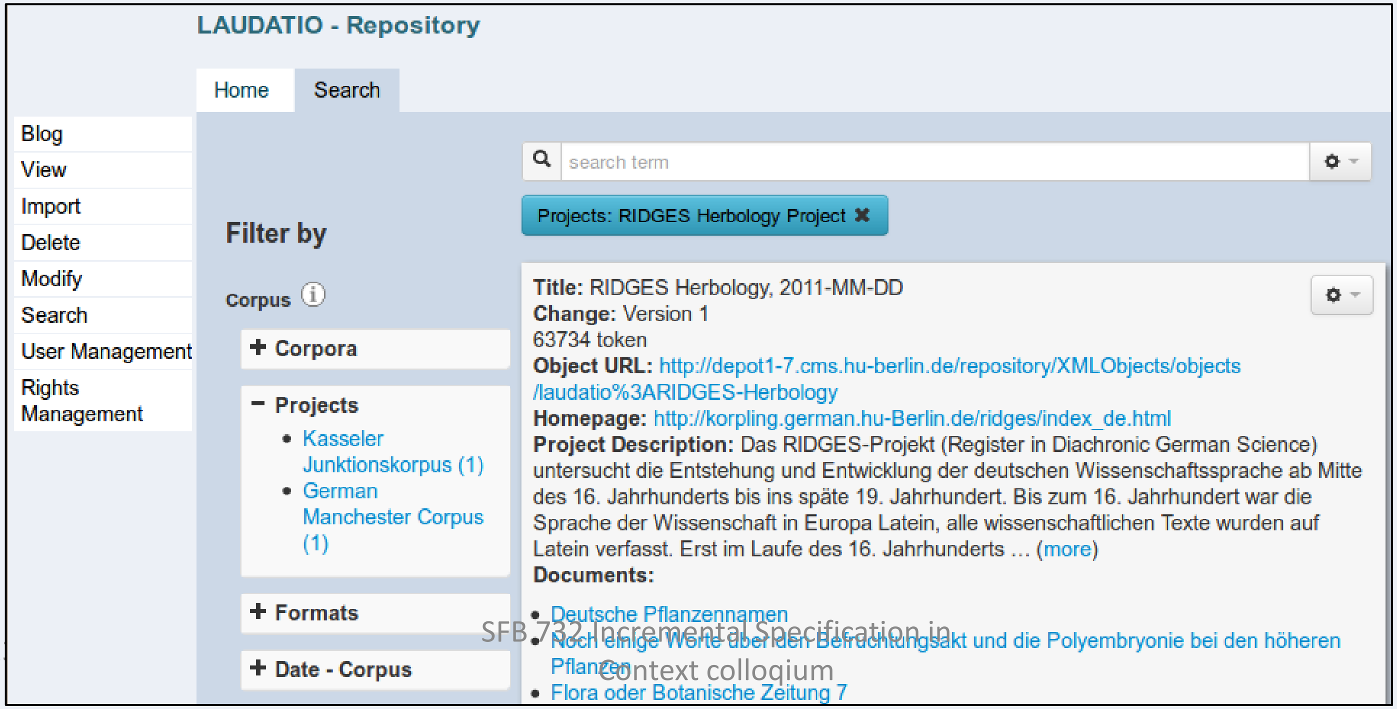
RIDGES Herbology, Humboldt-Universität zu Berlin, 1.0, 63734 Token, TEI P5 XML,

http://korpling.german.hu-berlin.de/ridges/download_de.html

- Corpus RIDGES Herbology
  - Authorship
  - Project
  - Availability Status: free
  - Extent
  - Documents
  - Format
- Documents

Access the data: search for corpora

- filtering the available corpora by selecting values for categories from a list (facet search)
- free-text search
- search for supplemental documents to corpora



Projects: RIDGES Herbology Project ✖

Title: RIDGES Herbology, 2011-MM-DD
Change: Version 1
63734 token
Object URL: http://depot1-7.cms.hu-berlin.de/repository/XMLObjects/objects/laudatio%3ARIDGES-Herbology
Homepage: http://korpling.german.hu-Berlin.de/ridges/index_de.html
Project Description: Das RIDGES-Projekt (Register in Diachronic German Science) untersucht die Entstehung und Entwicklung der deutschen Wissenschaftssprache ab Mitte des 16. Jahrhunderts bis ins späte 19. Jahrhundert. Bis zum 16. Jahrhundert war die Sprache der Wissenschaft in Europa Latein, alle wissenschaftlichen Texte wurden auf Latein verfasst. Erst im Laufe des 16. Jahrhunderts … (more)
Documents:

- Deutsche Pflanzennamen
- Noch einige Worte über den Befruchtungsakt und die Polyembryonie bei den höheren Pflanzen
- Flora oder Botanische Zeitung 7

“Long-term Access and Usage” part

- joint project with the computer and media service (CMS) of our university
- general concept of long term access and use
- virtual language research environment
- who develops, maintains and uses the infrastructure?

→ focus on the historical linguistics

CLARIN-D

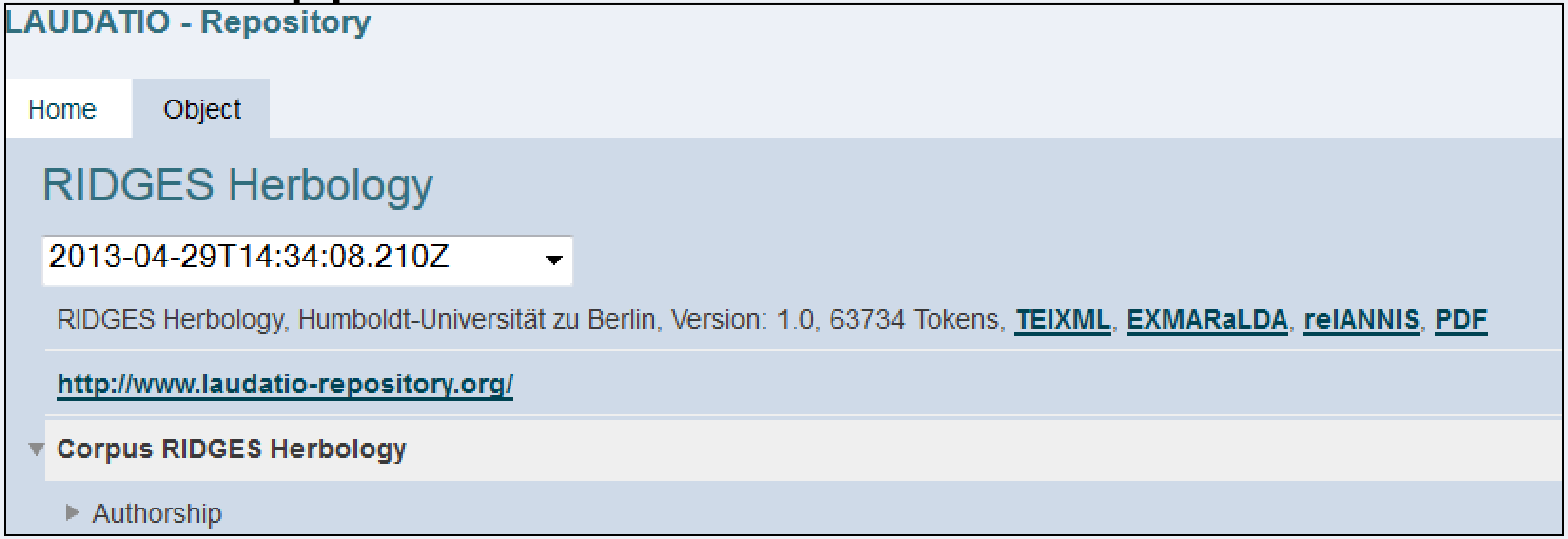
- general linguistic infrastructure project
- developing standards and guidelines for providing sustainability of data
- providing tools and services for the broad linguistic community
  - is there a sustainability concept for these tools?

Use: Corpus sharing

- share your corpus data with a community
- public community is central to search for existing corpora
- re-using the data
  - for search and analysis of already available data
  - getting the raw data a specific paper is based on
  - for adding new data
  - for adopting methods

Format diversity

- users are format-agnostic: tools are important
- LAUDATIO repository is format agnostic
- unified metadata format for documentation and search
- SaltNPepper framework for conversion



RIDGES Herbology

2013-04-29T14:34:08.210Z

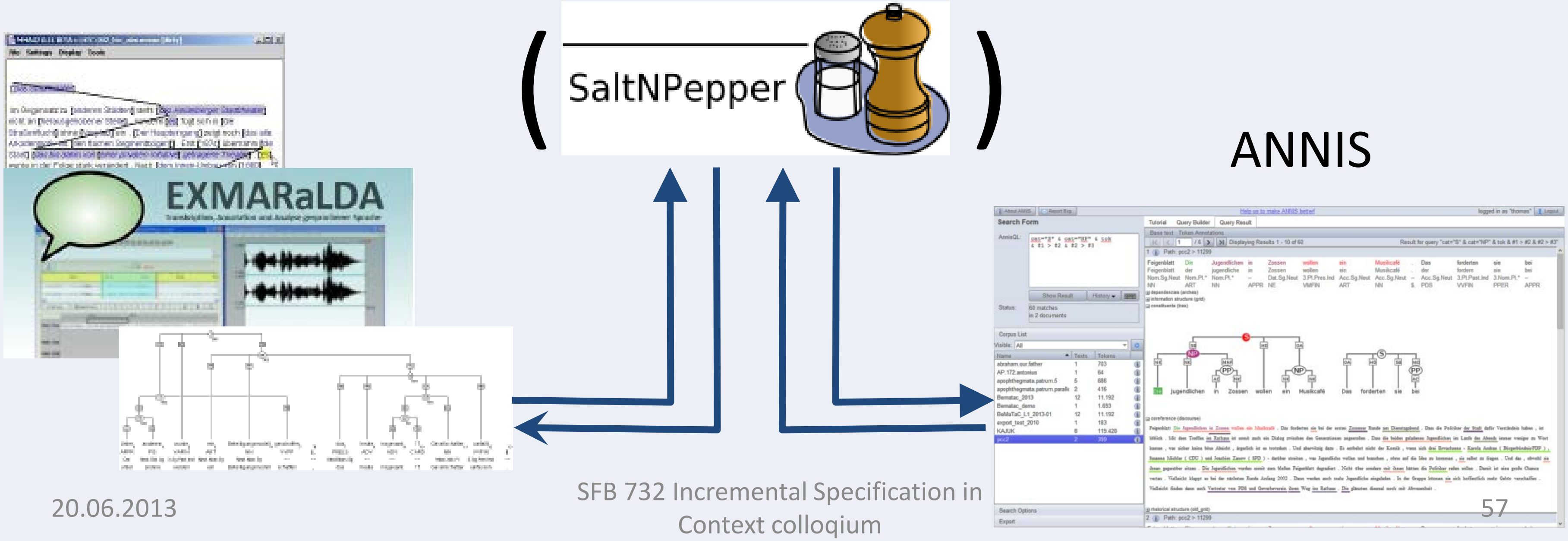
RIDGES Herbology, Humboldt-Universität zu Berlin, Version: 1.0, 63734 Tokens, TEIXML, EXMARaLDA, relANNIS, PDF

http://www.laudatio-repository.org/

Corpus RIDGES Herbology

Authorship

Use: Search the data

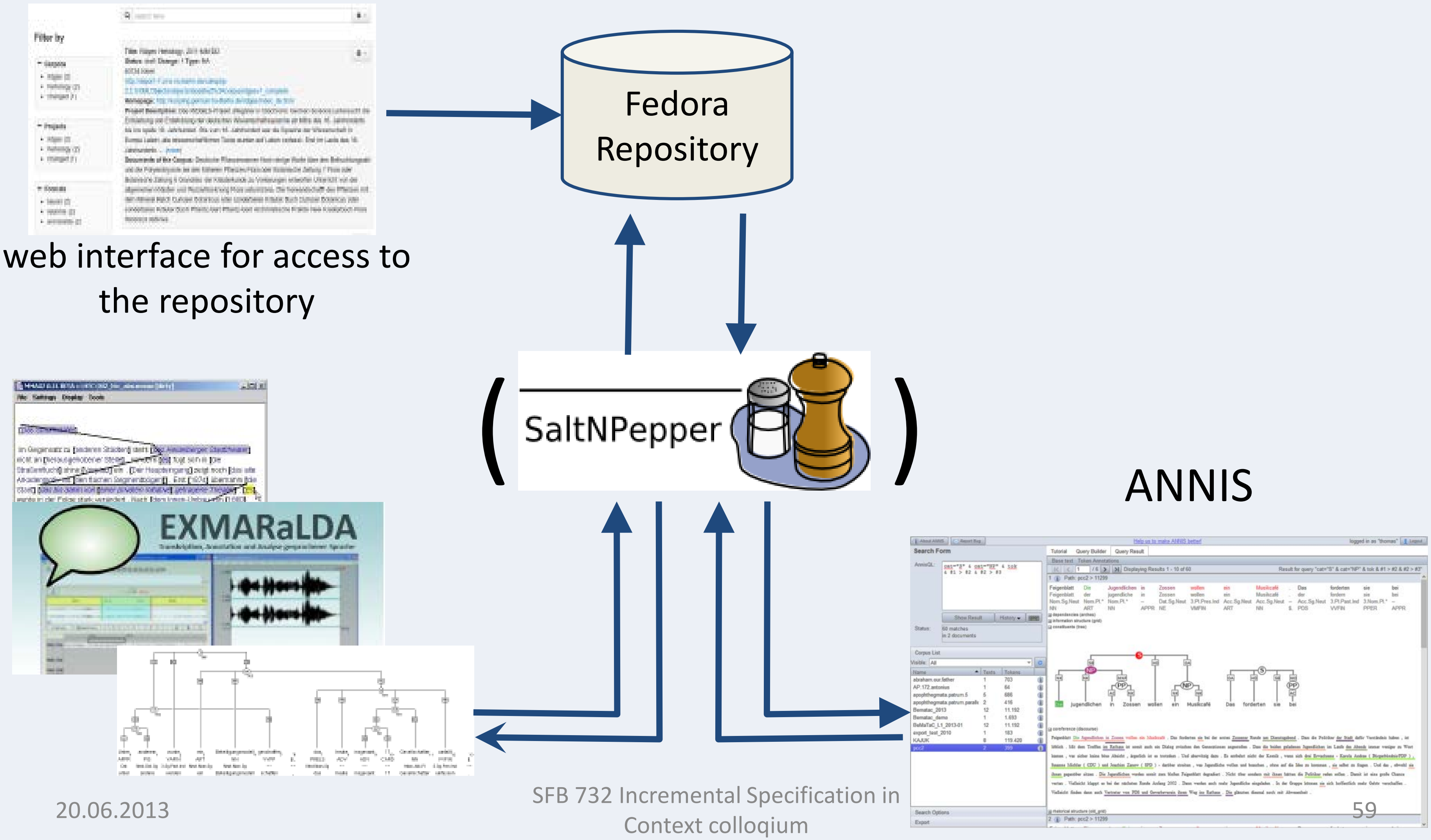


ANNIS

ANNIS

- format- and theory-agnostic
  - no knowledge about possible annotations
  - general key-value strings
  - uses graph model internally (everything is a graph)
- → general search on diverse, multi-layered data
- ANNIS Query Language (AQL)
  - define annotations to search
  - define relations between the annotated nodes

Putting the components together



Fedora Repository

web interface for access to the repository

(SaltNPepper)

EXMARaLDA

ANNIS

OUTLOOK

Outlook

- developing model and TEI ODD
  - cooperation with AHD-DDD-Project
  - corpus projects in other digital Humanities disciplines
- support for structural metadata in search engines like ANNIS
- interoperability with other research environments and formats
- info module for pepper for automatically generating parts of the metadata

Thank you!

LAUDATIO Team:

Malte Belz, Karin Donhauser, Anke Lüdeling,

Laurent Romary, Tino Schernikau,

Peter Schirmbacher, Vivian Voigt,

Benjamin Weißenfels, Dennis Zielke

Supporters:

Maxi Kindling, Tom Ruette, Amir Zeldes, Florian Zipser

DEMONSTRATION

LAUDATIO-Repository

References

- Bollmann, M., Petran, F., & Dipper, S. (2011). **Rule-based normalization of historical texts.** *Language Technologies for Digital Humanities and Cultural Heritage*, 34.
- Burnard, Lou & Bauman, Syd (Eds.) (2008). **TEI P5: Guidelines for Electronic Text Encoding and Interchange.** Oxford.
  http://www.tei-c.org/Guidelines/P5/.
- Burnard Lou., Rahtz, Sebastatian.**: RelaxNG with Son of ODD.** Extreme Markup Languages® 2004
- Dickinson, Markus, and W. Detmar Meurers. "**Detecting errors in part-of-speech annotation.**" Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003.
- Dipper, S., Faulstich, L., Leser, U., & Lüdeling, A. (2004, May). **Challenges in modelling a richly annotated diachronic corpus of german**. In *Workshop on XML-based richly annotated corpora*, Lisbon, Portugal (pp. 21-29).
- Krause, Thomas, Odebrecht, Carolin & Dennis Zielke **Langfristiger Zugang und Nutzung von tief annotierten Korpora: LAUDATIO**. *32. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft.*
  http://www.laudatio-repository.org/
- Krause, Thomas, Lüdeling, Anke, Odebrecht, Carolin & Amir Zeldes (2012) **Multiple Tokenization in a Diachronic Corpus.** *Exploring Ancient Languages through Corpora Conference (EALC),* 14.-16.Juni 2012.
  http://korpling.german.hu-berlin.de/ridges/documentation_en.html
  https://korpling.german.hu-berlin.de/annis3/
- Nivre, J., Hall, J., & Nilsson, J. (2006, May). **Maltparser: A data-driven parser-generator for dependency parsing.** In *Proceedings of LREC* (Vol. 6, pp. 2216-2219).
- Zeldes, Amir, Ritz, Julia, Lüdeling, Anke & Christian Chiarcos (2009) **ANNIS: A Search Tool for Multi-Layer** Annotated Corpora. In: *Proceedings of Corpus Linguistics 2009*, Liverpool, July 20-23, 2009.
- Zipser, Florian & Laurent Romary (2010) **A model oriented approach to the mapping of annotation formats using standards.** In: *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC* 2010. Malta. URL: http://hal.archives-ouvertes.fr/inria-00527799/en/