Modellierung von linguistischen Forschungsdaten

Kolloquium Korpuslinguistik

13.11.2013

Carolin Odebrecht

Humboldt-Universität zu Berlin

Überblick

1. Forschungskontext
2. Forschungsfrage
3. Anwendungsbereich
4. Metamodell
5. Problemstellung

1. Forschungskontext

Korpora

Ridges

MANCHESTER 1824
The University of Manchester

GerManC

...

Formate

TEI XML
EXMARaLDA
XML
CoNLL
relANNIS
PAULA
...

Daten über Daten

Herausgeber
Titel
Projekte
Annotationen
Fragestellung
Datenaufbereitung
Dokumente
...

Arbeiten mit Korpora

ANNIS
CQP
SaltNPepper

Treetagger / RF Tagger
Malt Parser / Berkeley Parser
ELAN / EXMARaLDA Partitur Editor / Excel

1. Forschungskontext

- Korpuslinguistik

  Forschungsfrage bedingt:

  - Auswahl und Umfang der Datengrundlage
  - Art der Annotation (damit auch Datenverarbeitung)
  - Form des Zugriffs und der Auswertung
  - Metadaten

➢ Diversität durch Forschung

➢ kaum einheitliche Dokumentation

➢ unterschiedliches Verständnis von Forschungsdaten

2. Forschungsfrage

- Wie können solche linguistische Forschungsdaten modelliert werden, um eine
  - einheitliche
  - formatunabhängige
  - generische

  Dokumentation zu ermöglichen?

➢ Mehrwert für die Forschung
  - ➢ Entwicklung eines allgemein anwendbaren Konzepts von Korpora (historische Text-Korpora)
  - ➢ Entwicklung eines Modells für Forschungsdaten-Repositories
  - ➢ Dokumentation
  - ➢ ‘best practice’
  - ➢ besseres Verständnis von Daten und dadurch auch von Methoden der (historischen) Korpuslinguistik

2. Forschungsfrage

…
…: …
…: …

Meta-Metamodell

Auto
Farbe: RGB
Räder: Integer

Metamodell

Ferrari
Farbe: rot
Räder:4

Modell

Instanz

http://www.ferrari.com/Site_Collection_Image_430x140/120308_1200065_CAR_F12berlinetta_430x140.jpg

2. Forschungsfrage

- Korpora sind Daten
  - Modelle von Daten
  - Daten = RIDGES, GerManC, KAJUK etc.
- Metadaten sind Daten über Daten
  - Modelle von Metadaten
  - Metadaten = Autor, Titel, Erscheinungsjahr etc.

2. Forschungsfrage

Weiterhin zu klären:

- Wie schaut man auf Forschungsdaten? (Kann nur für jeweils eine wissenschaftliche Disziplin geklärt werden)
- Funktion und Anwendungsbereich des Modells
- Modelle und ihre Anwendungen, u.a.:
  - Europeana Data Model (Linked Data)
  - Salt (Metamodel für linguistische Annotationen, Funktion u.a. Entwicklung eines Konverterframeworks)
  - ‚One Document Does it all' - ODD (Text Encoding, TEI Customization)
  - Component MetaData Infrastructure - CMDI (CLARIN)

2. Forschungsfrage

- interdisziplinäre Arbeit – Schnittmengen mit
  - Informationswissenschaft
    - Forschungsdaten
  - Informatik
    - Modellierung
  - Linguistik
    - Korpuslinguistik
    - Historische Linguistik

Diskussion in einem größeren Rahmen

Methode

Forschungsgegenstand

2. Forschungsfrage

Methode (1. Schritt):

- Entwicklung eines Klassendiagramms
  - Analyseebene: Konzepte der fachlichen Anwendungsdomäne
  - Verstehen und Dokumentieren des Problembereichs
- mit Hilfe der Unified Modeling Language (UML)
- Modellierung

  formalisierte Repräsentation der mentalen Idee
  - Abbildung der „realen Welt“
  - unterschiedliche Abstraktionsebenen

2. Forschungsfrage

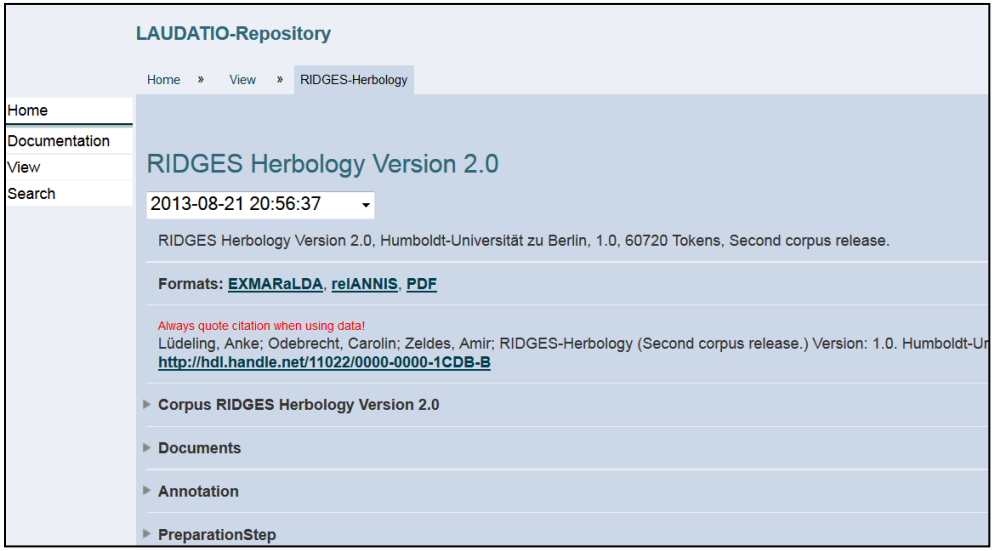
Methode (möglicher 2. Schritt):

- Entwicklung eines Entwurfsdiagramms
  - Lösung also das „Wie“, dass sich aus dem Problembereich („Was“) des Klassendiagramms ergibt
  - hierfür aber Implementationssprache notwendig
    - Wahl dieser (Java, Python etc.)
    - Machbarkeit

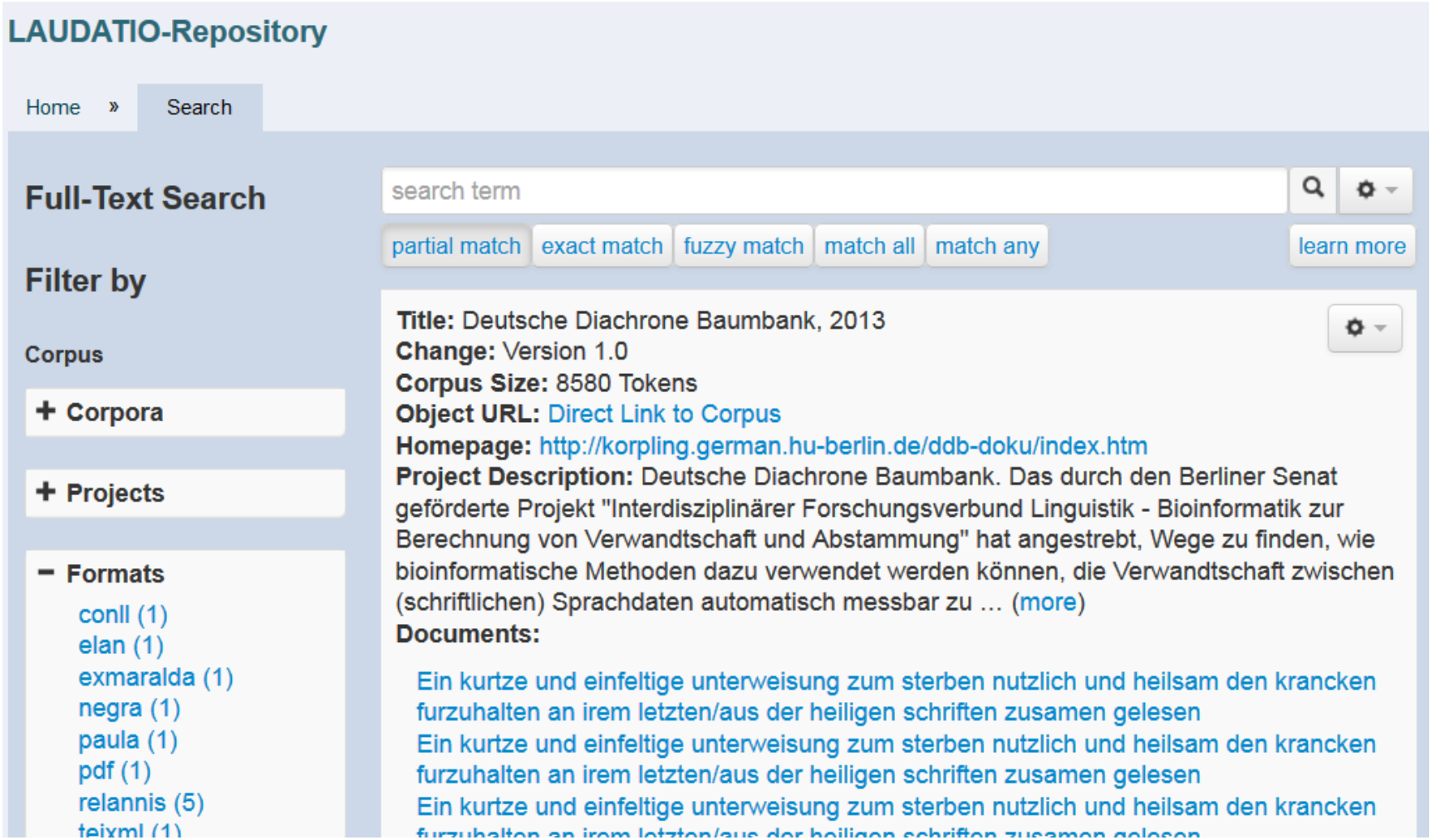
➢ Nutzen eines Klassenmodells ohne Implementierung?

3. Anwendungsbereich

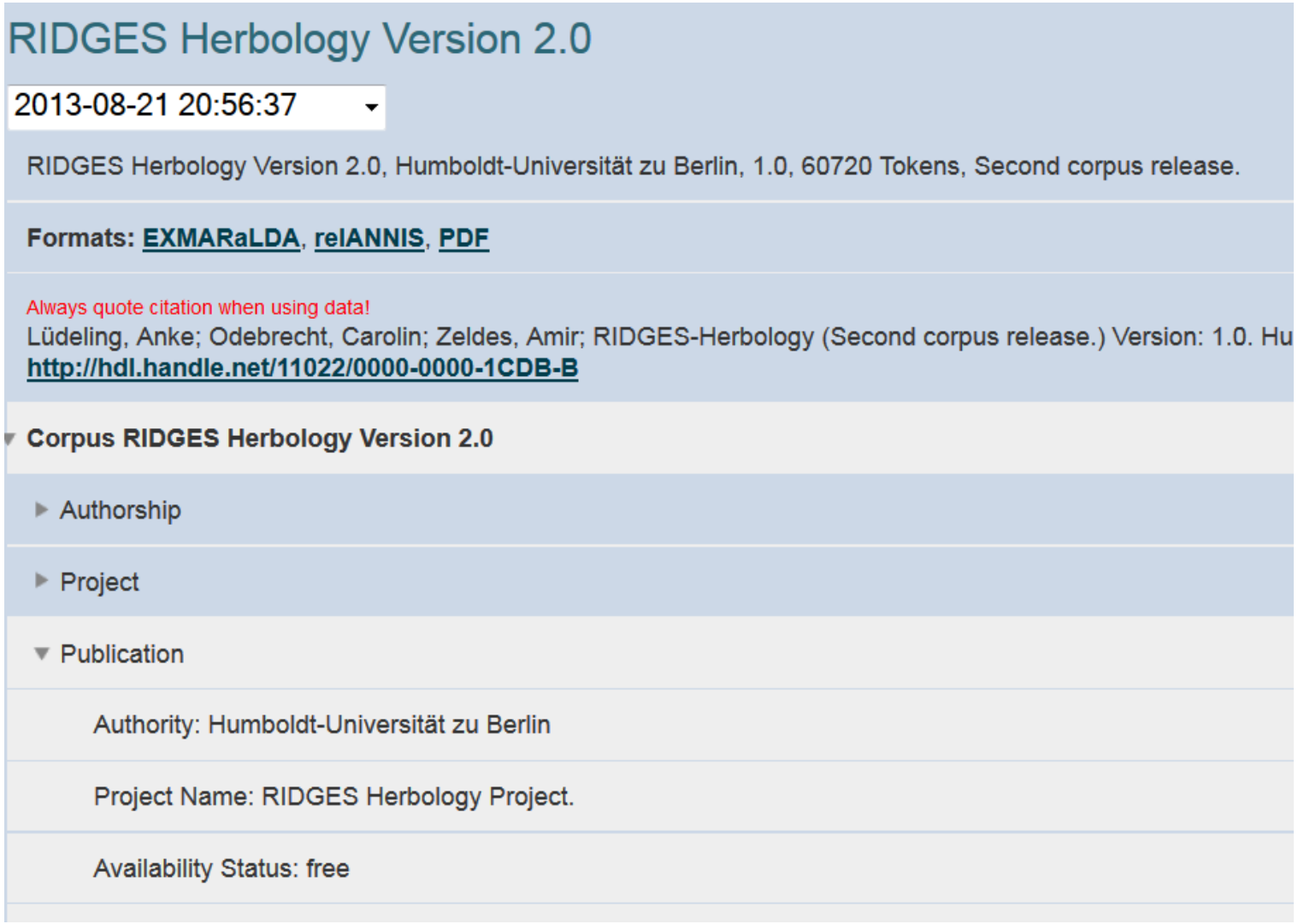
- LAUDATIO-Repository
  - Speicherung von und Zugang zu historischen Korpora aller Art
  - Frei-Text- und Facetten-Suche
  - Informationen über Korpora (Metadaten)
  - Download
  - Upload
  - Referenzieren



3. Anwendungsbereich



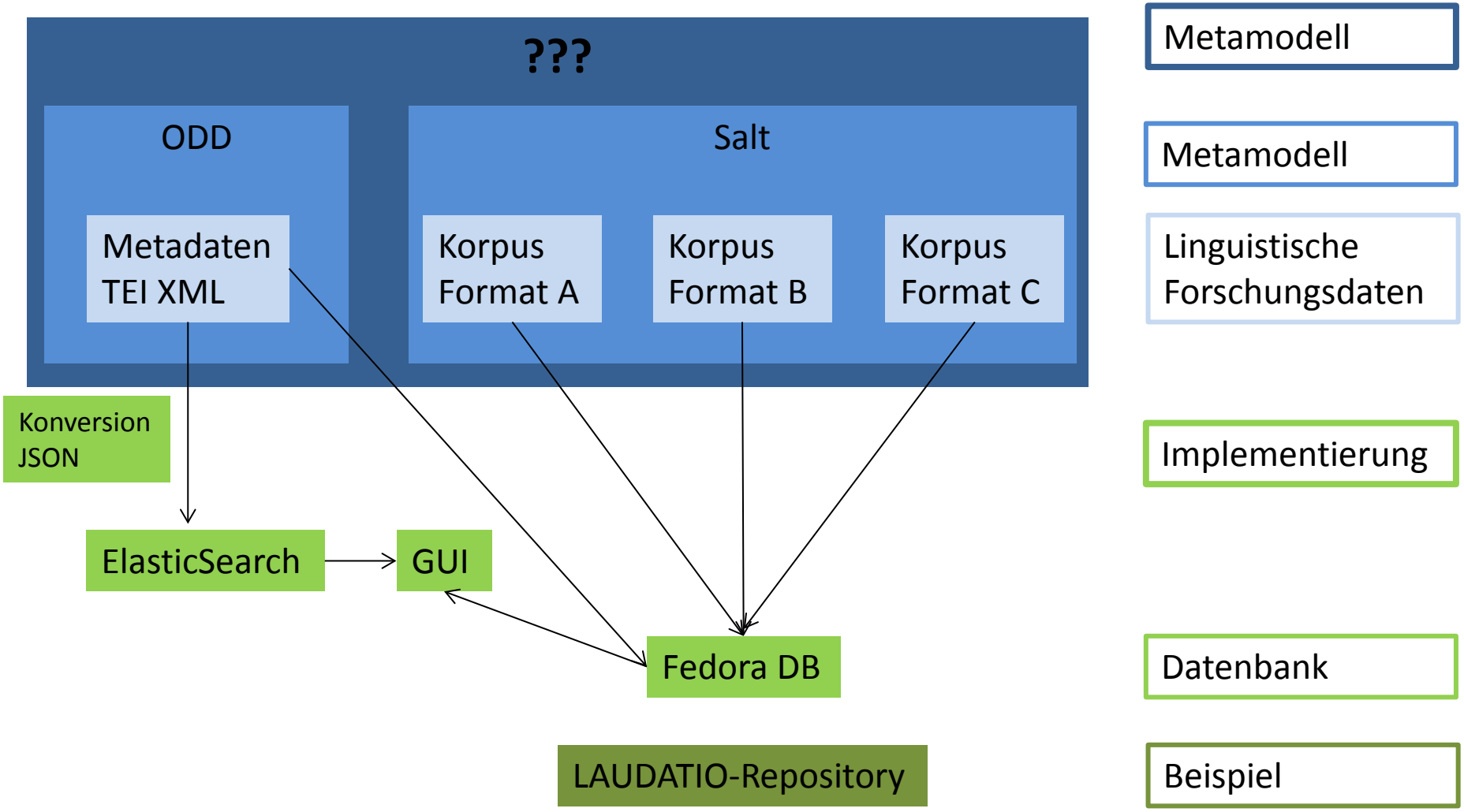
3. Anwendungsbereich



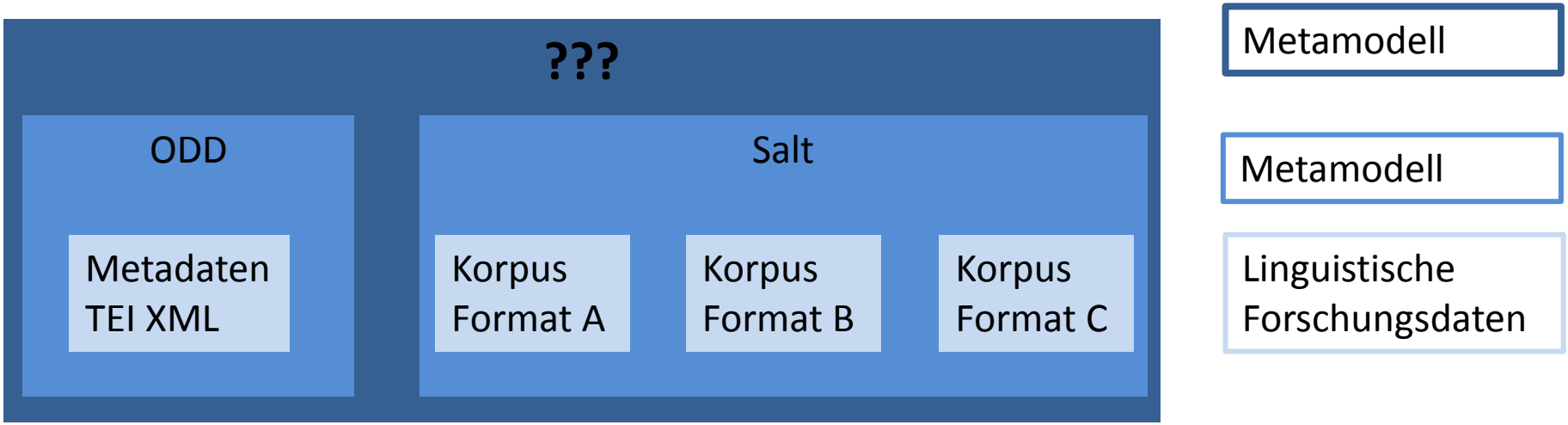
3. Anwendungsbereich

- Konkret:
  - Korpus RIDGES Herbology 2.0
  - Formate: EXMARaLDA, relANNIS, PDF
  - Metadaten für Suche und Anzeige zu u.a.:
    - wer hat das Korpus wann erstellt, annotiert und publiziert
    - welche historische Texte sind im Korpus enthalten
    - welche Arten von Annotationen wurden mit welchen Tools/Verfahren wie erstellt
    - welches Tagset wurde bei welcher Annotation verwendet
  - download, upload

3. Anwendungsbereich



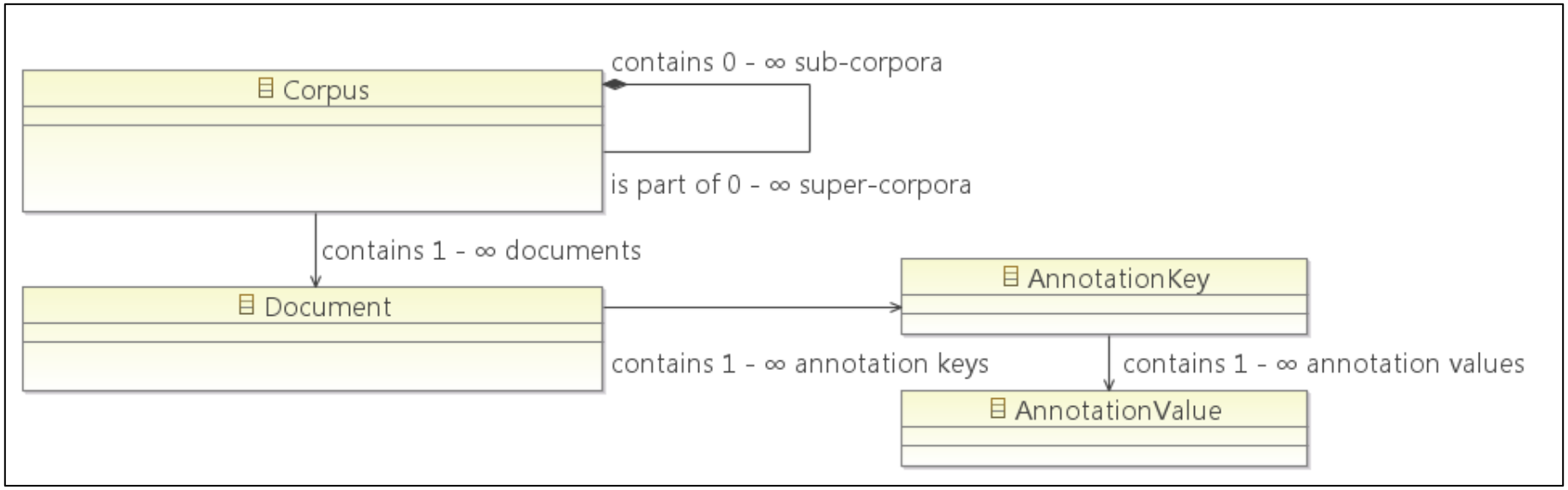
3. Anwendungsbereich



- Linguistische Forschungsdaten
  - Korpora (schriftsprachlich, historisch)
- Salt als Metamodell für Korpora
  - Format unabhängig, frei gewählte Klassen
  - für die Konvertierung von Formaten
- ODD als Metamodell für Metadaten
  - Format abhängig (TEI XML)
  - für Text Encoding

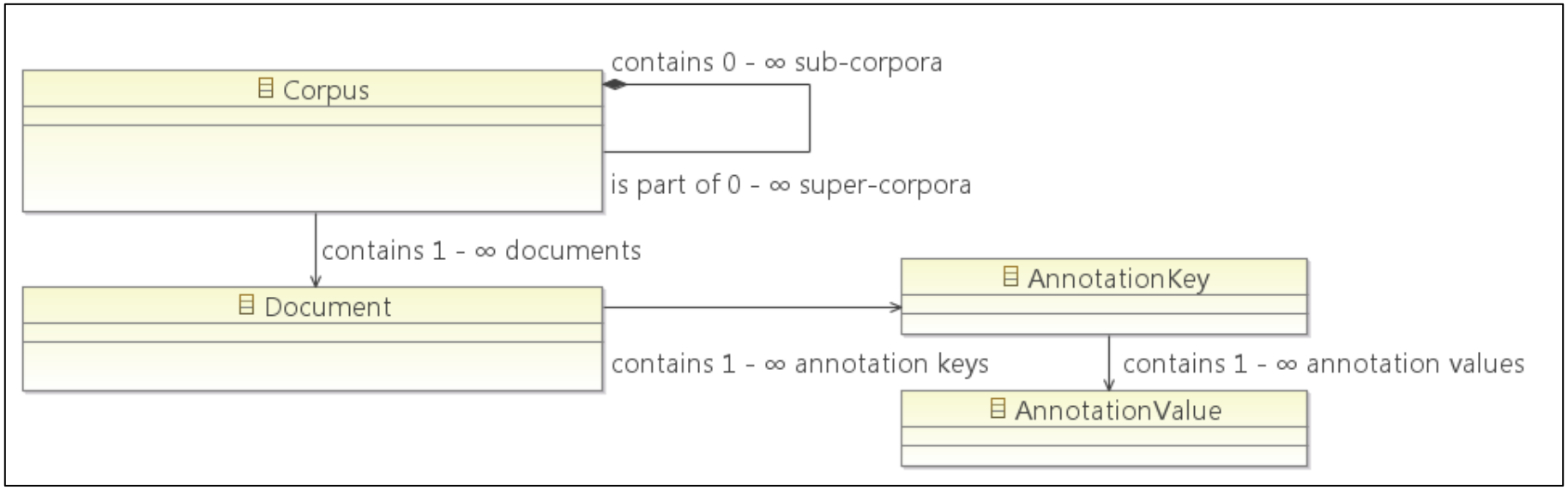
4. Metamodell

- Konzept eines Korpus für den Bereich der (historischen) Korpuslinguistik (kurz gefasst):
  - ‚Corpus' = Summe aller Dokumente
  - ‚Document' = Summe aller Annotationsschlüssel
  - ‚Annotation' = Annotationsschlüssel, mindestens 1 Annotationswert

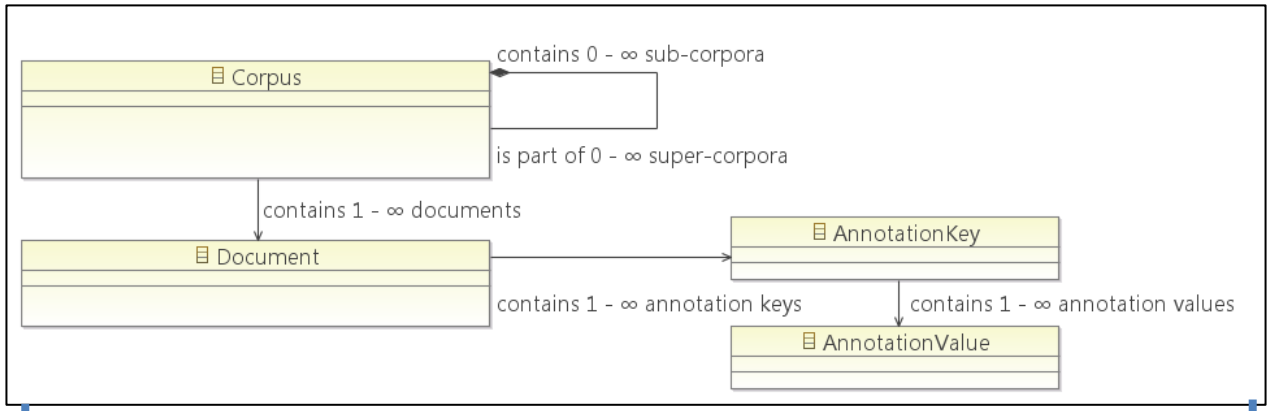


4. Metamodell

- Metadaten für alle Konzepte z.B.:
  - Corpus: Autor (Hrsg., Annotator), Format, Liste aller Dokumente & Annotationen, Datum, Revision, Projekt
  - Document: Autor, Annotationsliste, Ort, Verlag, Datum, Revision
  - Annotation: Autor (Hrsg., Annotator), Format, Tools, Datum, Revision, Projekt



4. Metamodell



Wie kommt das Metamodell in das Repository?

RIDGES Herbology Version 2.0

2013-08-21 20:56:37

RIDGES Herbology Version 2.0, Humboldt-Universität zu Berlin, 1.0, 60720 Tokens, Second corpus release.

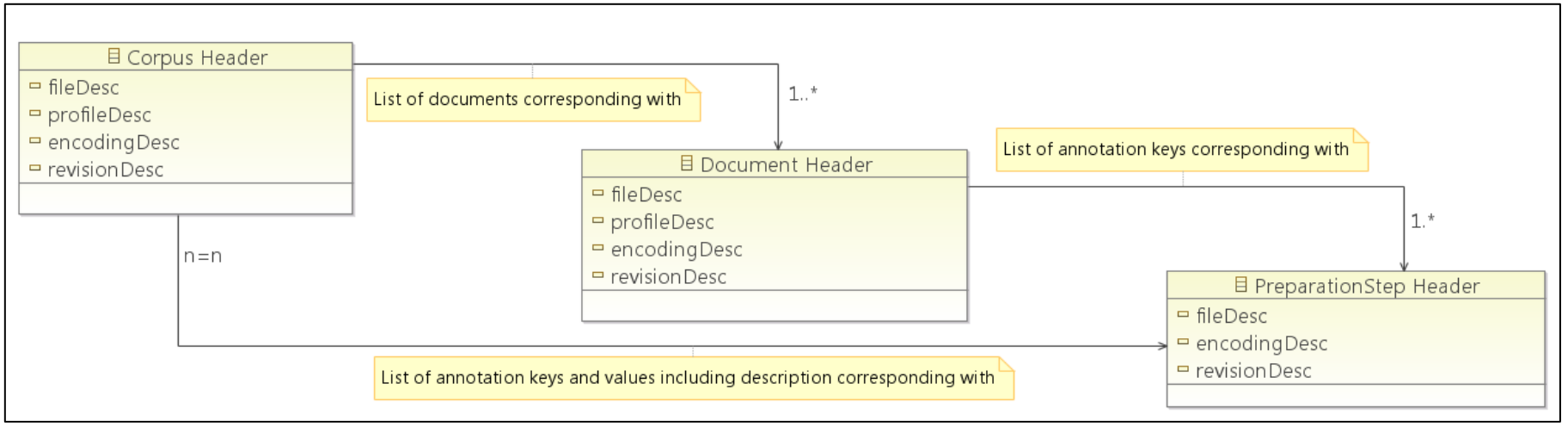
Formats: EXMARaLDA, relANNIS, PDF

Always quote citation when using data!
Lüdeling, Anke; Odebrecht, Carolin; Zeldes, Amir; RIDGES-Herbology (Second corpus release.) Version: 1.0. Humboldt-Universität zu Berlin. http://
http://hdl.handle.net/11022/0000-0000-1CDB-B

- **Corpus RIDGES Herbology Version 2.0**
  - Authorship
    - Corpus Editor
      - Anke Lüdeling
      - Amir Zeldes
      - Carolin Odebrecht

4. Metamodell

- technische Realisierung mit dem Metamodell ODD
  - Abbildung des Metamodells in TEI XML Header Struktur
  - Generierung von Schemata zur Erstellung von TEI XML Dateien
  - drei TEI-Header mit jeweils einer Spezifikation durch die ODD
  - alle Header referenzieren auf einander
  - bilden die Grundlage für die GUI im Repository



4. Metamodell

ODD in TEI XML

- gibt vor, was wo wie stehen soll
- erzeugt Schemata
- Dokumentation der Modelle

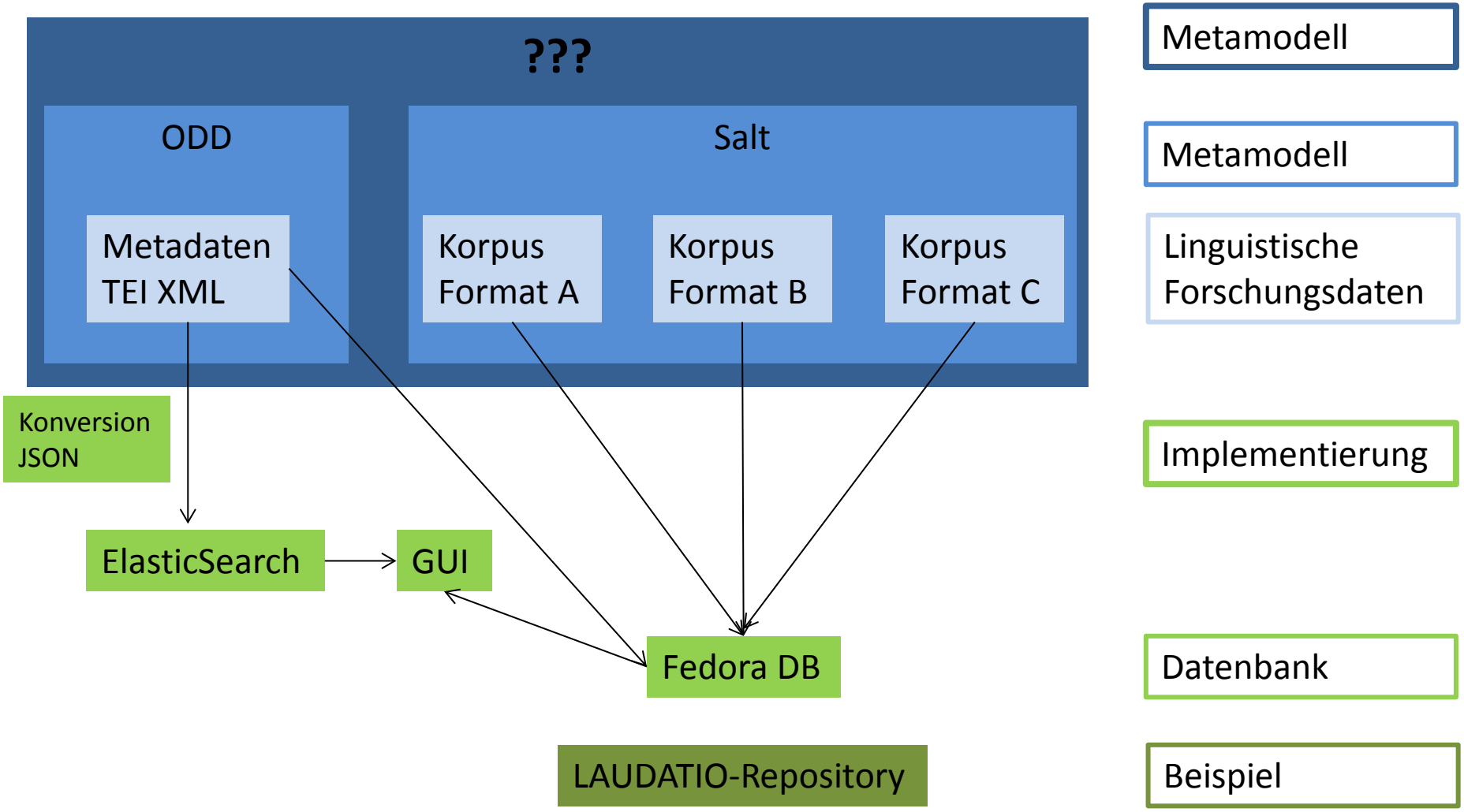
‚Corpus'-Header in TEI XML

- hier für das Korpus RIDGES 2.0
- wird nach Schema validiert
- Übergabe ans Repository

```
      <specDesc key="titleStmt"/>
      <specDesc key="extent" atts="type"/>
      <specDesc key="publicationStmt"/>
      <specDesc key="sourceDesc"/>
      <specDesc key="profileDesc"/>
      <specDesc key="encodingDesc" atts="n"/>
      <specDesc key="revisionDesc"/>
    </specList>
  </p>
  <p>
    <egXML xmlns="http://www.tei-c.org/ns/Examples">
      <TEI>
        <teiHeader type="CorpusHeader">
          <fileDesc>
            <titleStmt>
              <title>...</title>
              <editor>...</editor>
              <author>...</author>
            </titleStmt>
            <extent>...</extent>
            <publicationStmt>
              <authority>...</authority>
              <idno>...</idno>
              <availability>...</availability>
```

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader type="CorpusHeader">
      <fileDesc>
         <titleStmt>
            <title type="Corpus">RIDGES Herbology Version 2.0</title>
            <editor role="CorpusEditor" n="1">
               <persName>
                  <forename>Anke</forename>
                  <surname>Lüdeling</surname>
               </persName>
               <affiliation>
                  <orgName type="Department">Institut für deutsche Sprache und
                     Linguistik</orgName>
                  <orgName type="Institution">Humboldt-Universität zu Berlin</orgName>
               </affiliation>
            </editor>
            <editor role="CorpusEditor" n="2"> [10 lines]
            <editor role="CorpusEditor" n="3"> [10 lines]
            <author role="Annotator" n="1"> [10 lines]
            <author role="Annotator" n="2"> [10 lines]
            <author role="Annotator" n="3"> [10 lines]
            <author role="Annotator" n="4"> [10 lines]
            <author role="Annotator" n="5"> [10 lines]
            <author role="Annotator" n="6"> [10 lines]
            <author role="Annotator" n="7"> [10 lines]
            <author role="Annotator" n="8"> [10 lines]
            <author role="Annotator" n="9"> [10 lines]
            <author role="Annotator" n="10"> [10 lines]
            <author role="Annotator" n="11"> [10 lines]
```

5. Problemstellung



5. Problemstellung

- Modelle „wissen“ nichts von einander
- haben unterschiedliche Blickwinkel auf Korpora
- bilden nur Teile ab (motiviert durch Funktion)

- ➢ Wenn Korpora wiederverwendet werden,
- ➢ wenn Metadaten auch Teil der Korpora sind, also auch als Annotationen verstanden werden (auch Austausch dieser)
- ➢ wenn ein Dokument nicht mehr die Summe aller Annotationen ist

... wie modelliert man das?

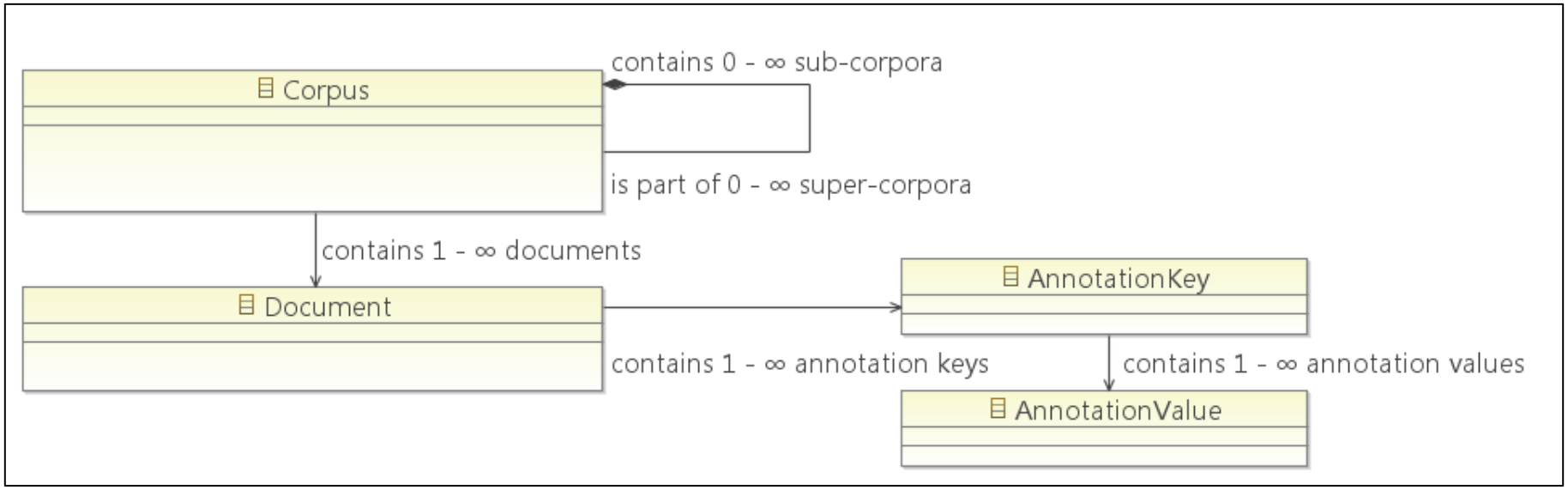
5. Problemstellung

- Korpora sind Daten
  - Modelle von Daten
  - Daten = RIDGES, GerManC, KAJUK etc.
- Metadaten sind Daten über Daten
  - Modelle von Metadaten
  - Metadaten = Autor, Titel, Erscheinungsjahr etc.

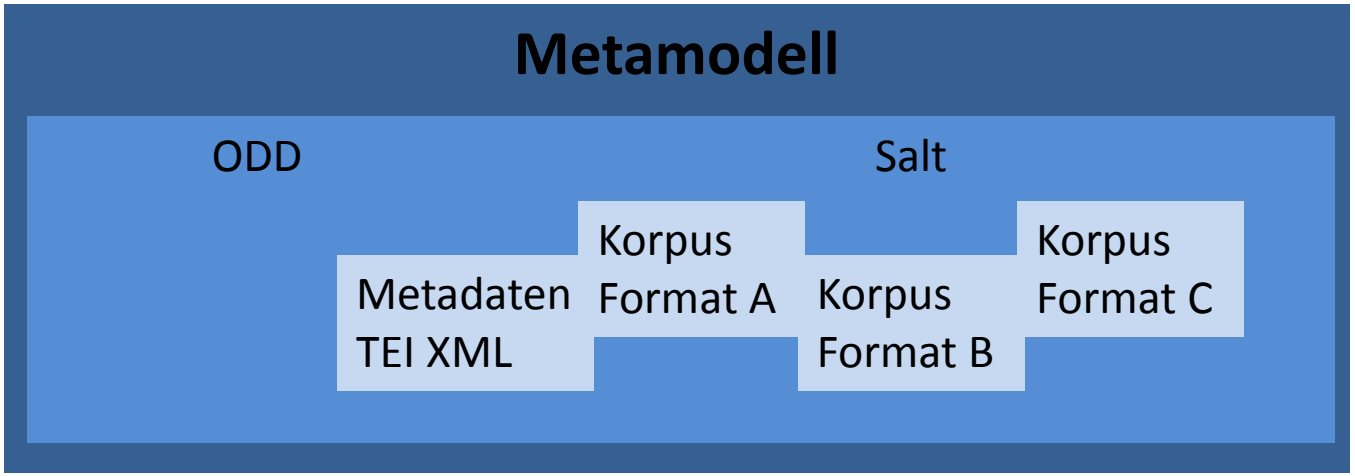
  oder
- Metadaten sind Daten!
  - wann sind Metadaten Daten und wann Metadaten?

5. Problemstellung

- Annotationen zwischen Dokumenten
  - so nicht im Metamodell
  - Idee: Auflösung des Konzeptes ‚Document‘, ‚Annotation‘ mit dem Label ‚Document‘?



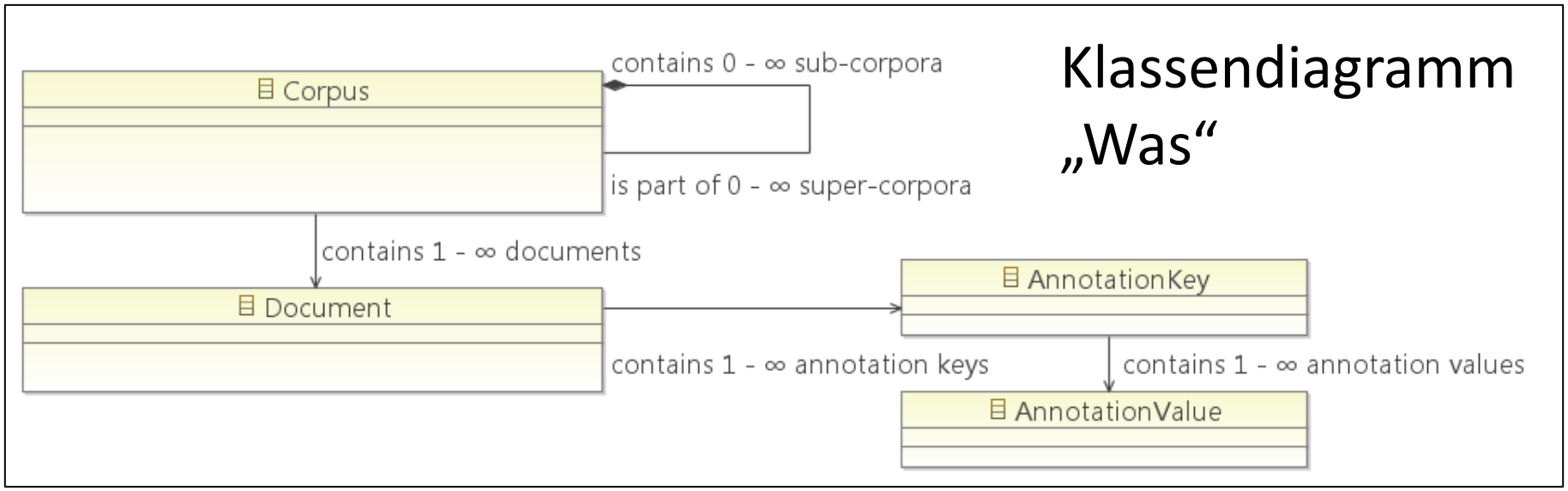
5. Problemstellung



Metamodell

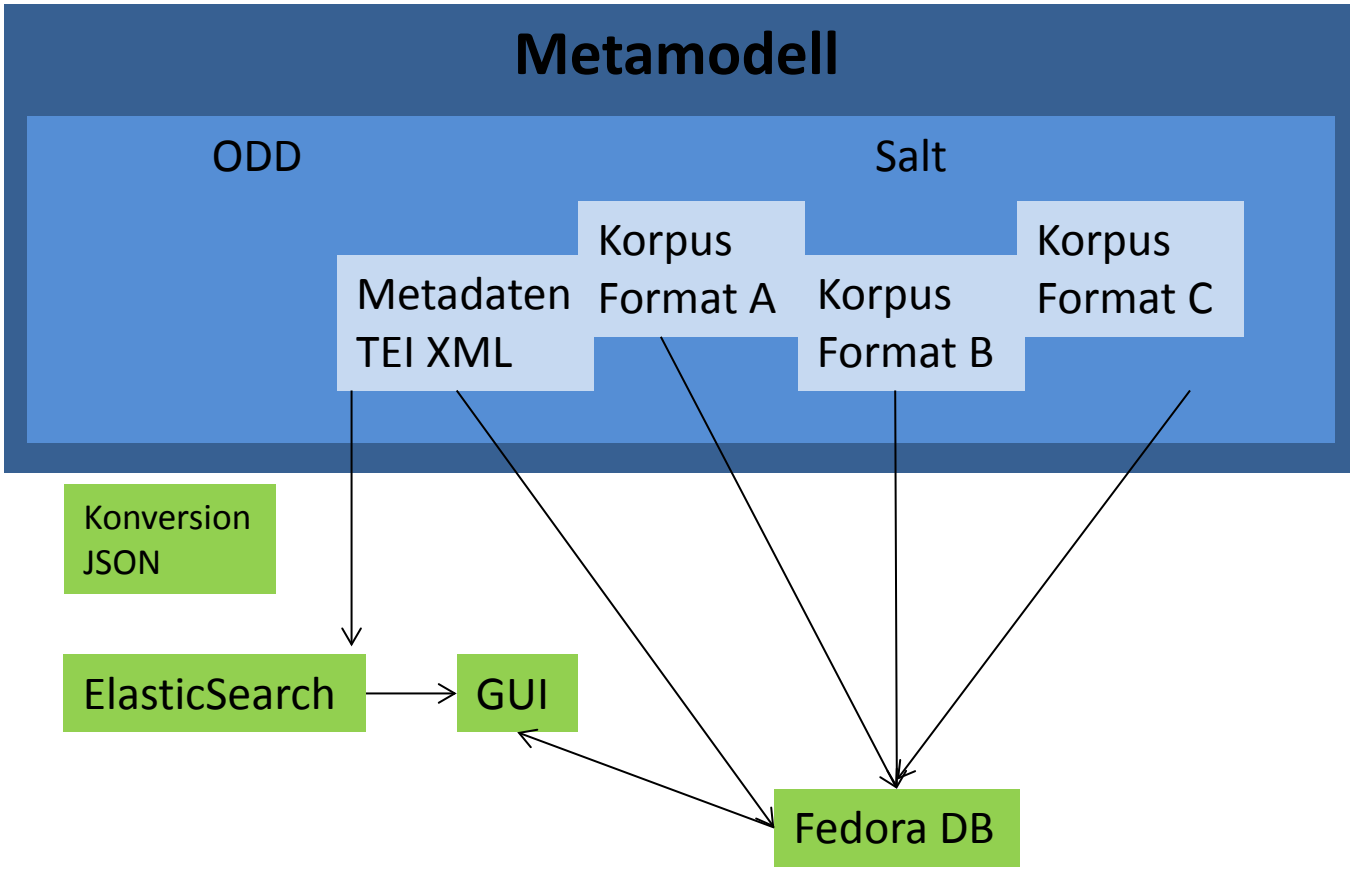
Metamodell

Linguistische
Forschungsdaten



Klassendiagramm
„Was“

5. Problemstellung



Metamodell

Metamodell

Linguistische
Forschungsdaten

Implementierung

Datenbank

Beispiel

Entwurfsdiagramm
„Wie“

LAUDATIO-Repository

Vielen Dank

Florian Zipser, Thomas Krause & Laurent Romary

für die tollen Diskussionen

und die große Hilfe!

Referenzen

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, Sebastian (2004) RelaxNG with Son of ODD*. Extreme Markup Languages Proceedings 2004.* Montréal, Québec.

CMDI: http://www.clarin.eu/node/3219

Durrell, Martin, Ensslin, Astrid, Bennett, Paul (2207) The GerManC project In *Sprache und Datenverarbeitung* 31 (2007), pp. 71-80.

Europeana Data Model: http://pro.europeana.eu/edm-documentation

Krause, Th., Odebrecht, C., Zielke, D. (2013) Langfristiger Zugang und Nutzung von tief annotierten Korpora: LAUDATIO. 32. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft. Berlin. http://www.laudatio-repository.org/

Nivre, J., Hall, J., Nilsson, J. (2006) Maltparser: A data-driven parser-generator for dependency parsing. *Proceedings of LREC .Vol. 6*. pp. 2216-2219.

Schmid, H. (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing.* 2013-06-08.

Schmidt, Th., Wörner, K. (2009) EXMARaLDA - Creating, analysing and sharing spoken language corpora for pragatic research*. Pragmatics 19/4. pp.565-582.*

Zeldes, A., Ritz, J., Lüdeling, A. Chiarcos, Ch. (2009) ANNIS: A Search Tool for Multi-Layer Annotated Corpora. *Proceedings of Corpus Linguistics 2009*, Liverpool. July 20-23, 2009.

Zipser, F., Romary, L. (2010) A model oriented approach to the mapping of annotation formats using standards. *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010* Malta. URL: http://hal.archives-ouvertes.fr/inria-00527799/en/