

# Wie kann der Zugriff, die Wiederverwendung und langfristige Speicherung von linguistischen Korpora realisiert werden?

**LAUDATIO-Repository (Long-term Access and Usage of Deeply Annotated InformaTION)**

Maxi Kindling, Thomas Krause, Carolin Odebrecht, Dennis Zielke, Humboldt-Universität zu Berlin

## 1. Problemstellung

- langfristige Speicherung von historischen Korpora
- Zugang zu historischen Korpora
- Darstellung mittels tief strukturierter Metadaten
- Suche nach Korpora und deren Dokumente
- Nutzung durch externe Forscher (z.B. in externen Forschungsumgebungen)
- Wiederverwendung der Daten in neuen Korpusprojekten
- Bildung einer korpuslinguistischen Infrastruktur für historische Daten

### Herausforderungen

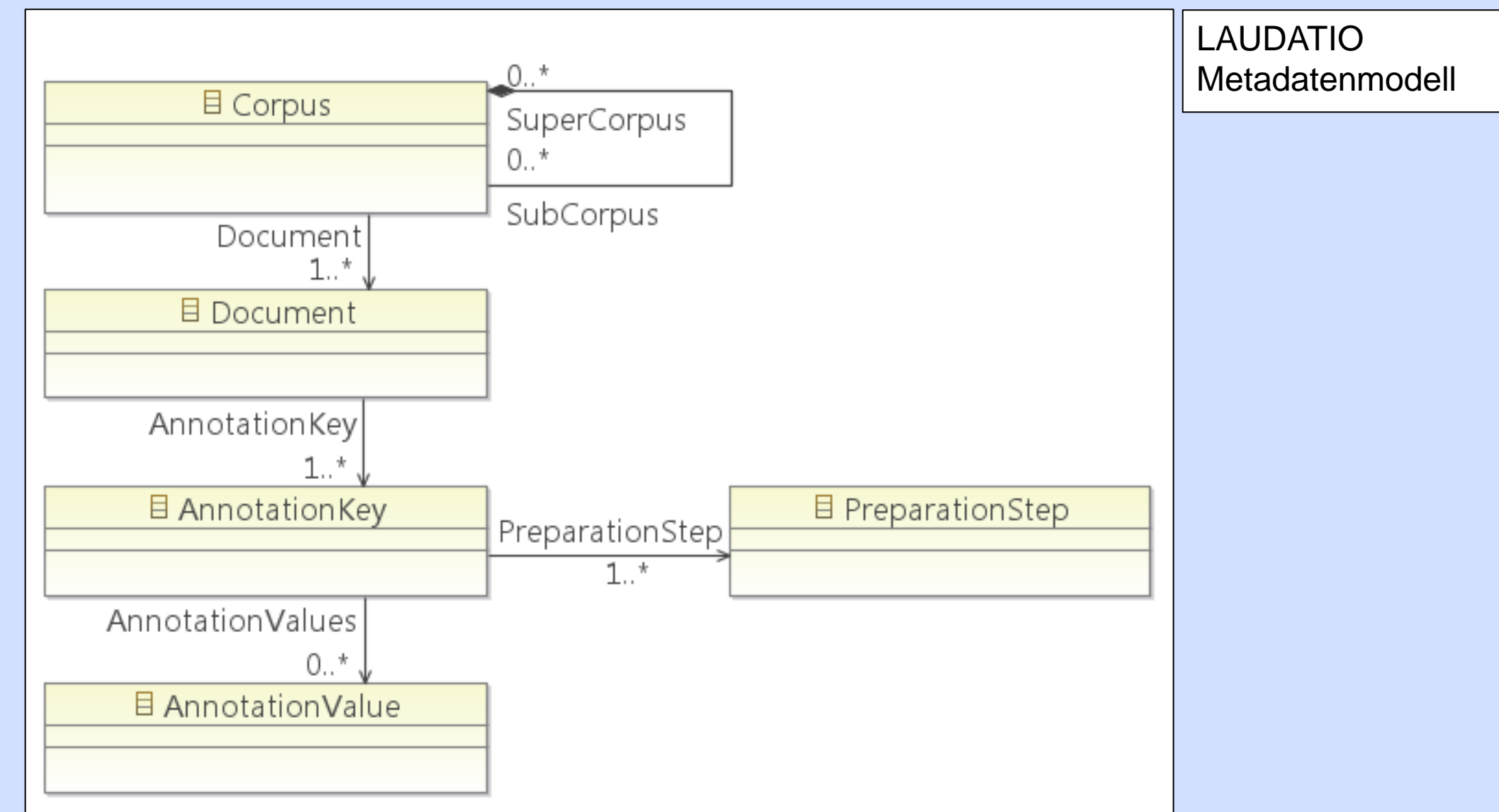
- Heterogene Datenformate
- Geringe Verbreitung von Standards der Datenaufbereitung und Speicherung
- Unzureichende Dokumentation der Annotationen

### Datenformate in der Linguistik

- Verschiedene linguistische Fragestellungen führen zur Verwendung von verschiedenen Annotationstools und Formaten
  - Eingeschränkte Menge an originalen Texten, die unter verschiedenen Aspekten aufbereitet werden
  - Jedes Projekt folgt seiner eigenen Fragestellung
- Infrastruktur muss **formatneutral** arbeiten.

## 2. Metadaten

- Beschreibung der Formate und Inhalte durch gemeinsames modulares Metadatenformat
- Suche und Zugang zu unterschiedlich aufbereiteten Korpora durch konsistente strukturierte Dokumentation
- Modellierung und Umsetzung als TEI P5 Header inkl. Relax-NG-Schema, ODD (Burnard & Bauman 2008, Burnard & Rahtz 2004)
- Komponenten:  
Sammlung von Texten: 'Corpus' ('SubCorpus')  
unterschiedliche Texte – 'Document'  
ling. Markierung – 'AnnotationKey' / 'AnnotationValue'  
komplette Verarbeitungs-Pipeline – 'PreparationStep'



## 3. Repository

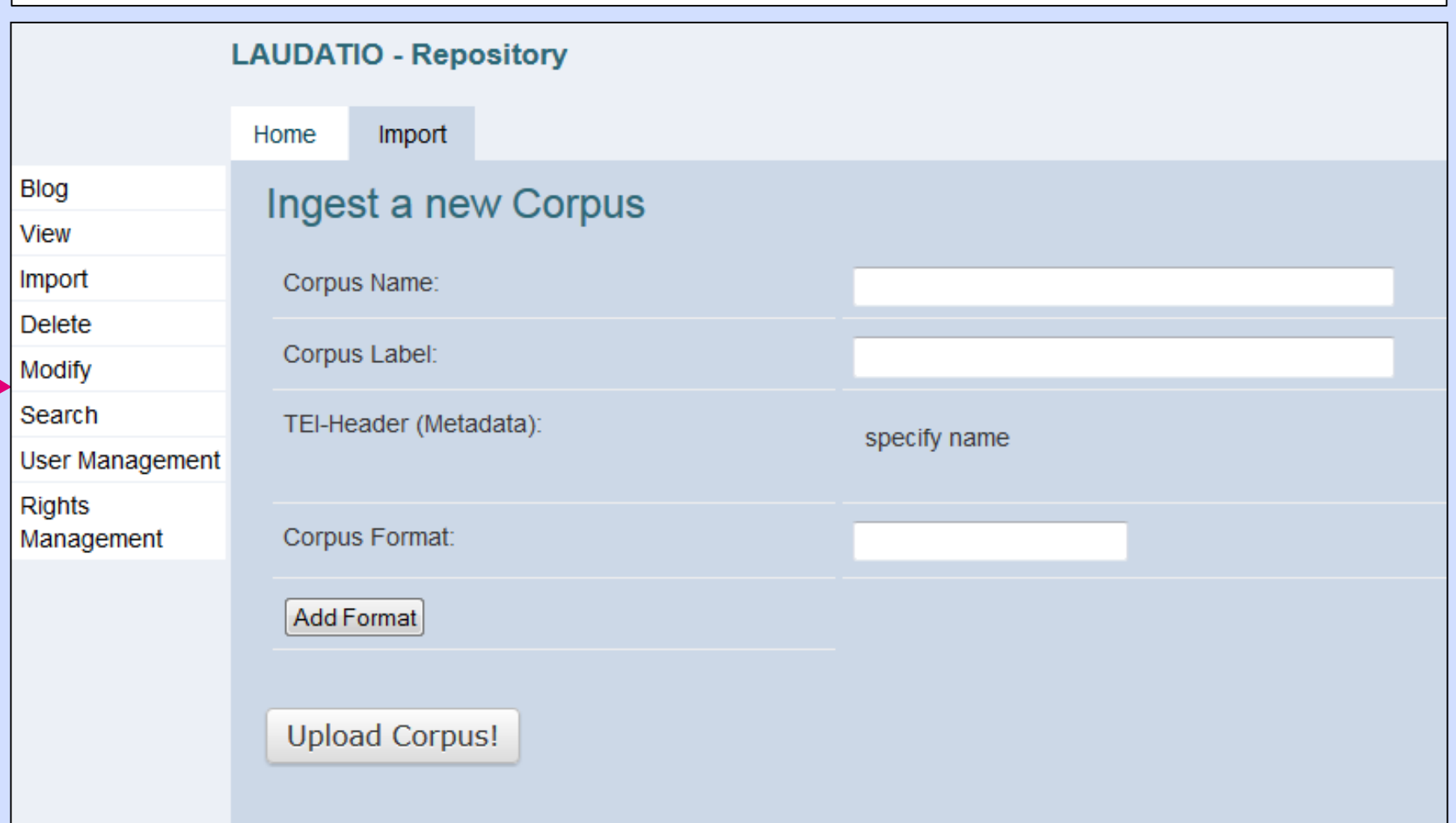
- Sichere langfristige Speicherung der Daten
- Weboberfläche zur Anzeige und Suche
- Schnittstellen zu externen Forschungsanwendungen (u.a. Such- und Visualisierungstool ANNIS, Konvertierungsframework SaltNPepper)
- Suche und Anzeige der Korpora in dem einheitlich strukturierten LAUDATIO - Metadatenmodell

### Technische Komponenten

- Nutzung bestehender Open Source Tools
  - Fedora als Repository-Server
  - Elastic Search als Suchmaschine
- Modularer Aufbau mit der Möglichkeit, einzelne Teile der Software zu ersetzen

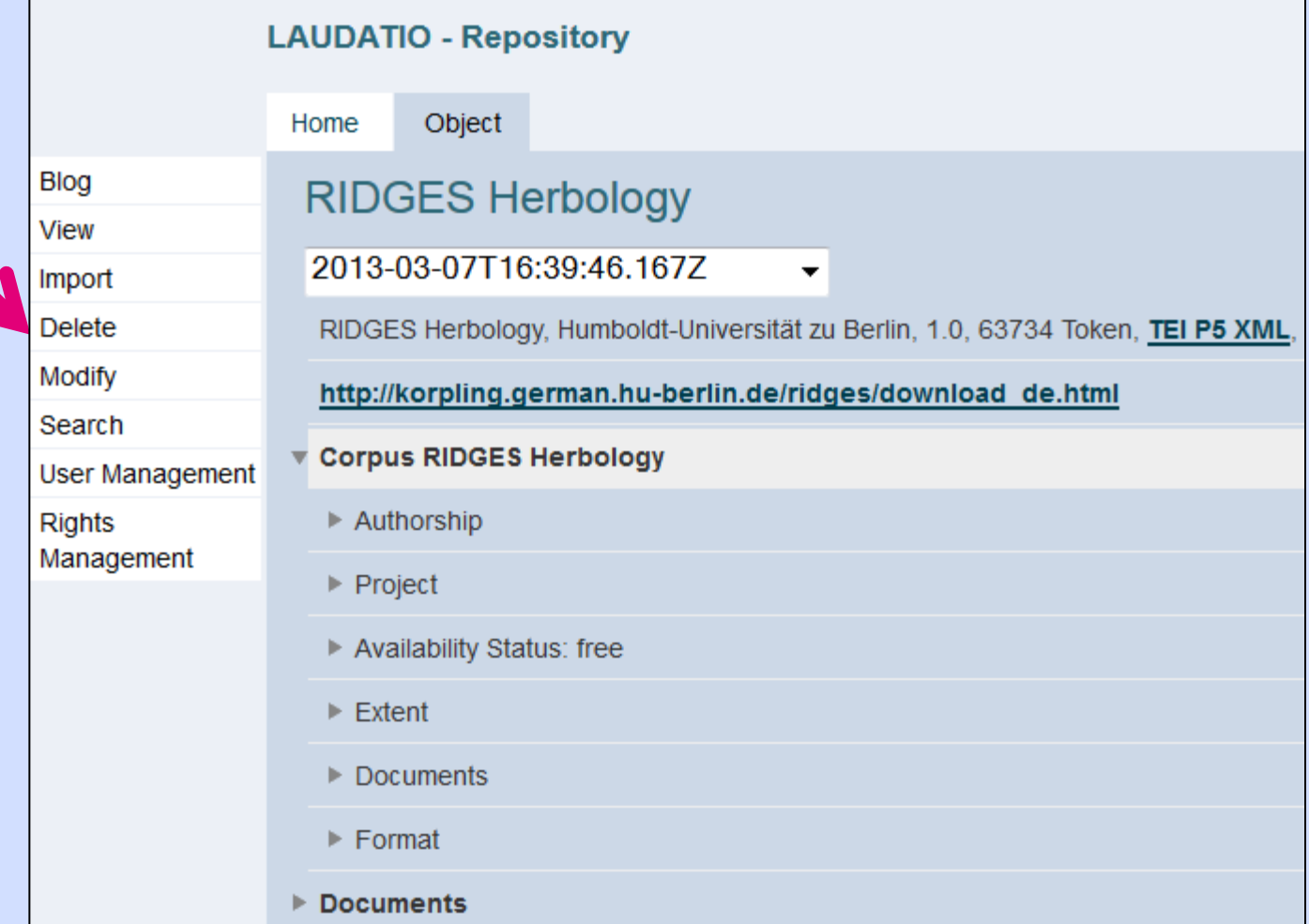
### Import

- Möglichkeit, neue Korpora in verschiedenen Formaten in das System zu importieren



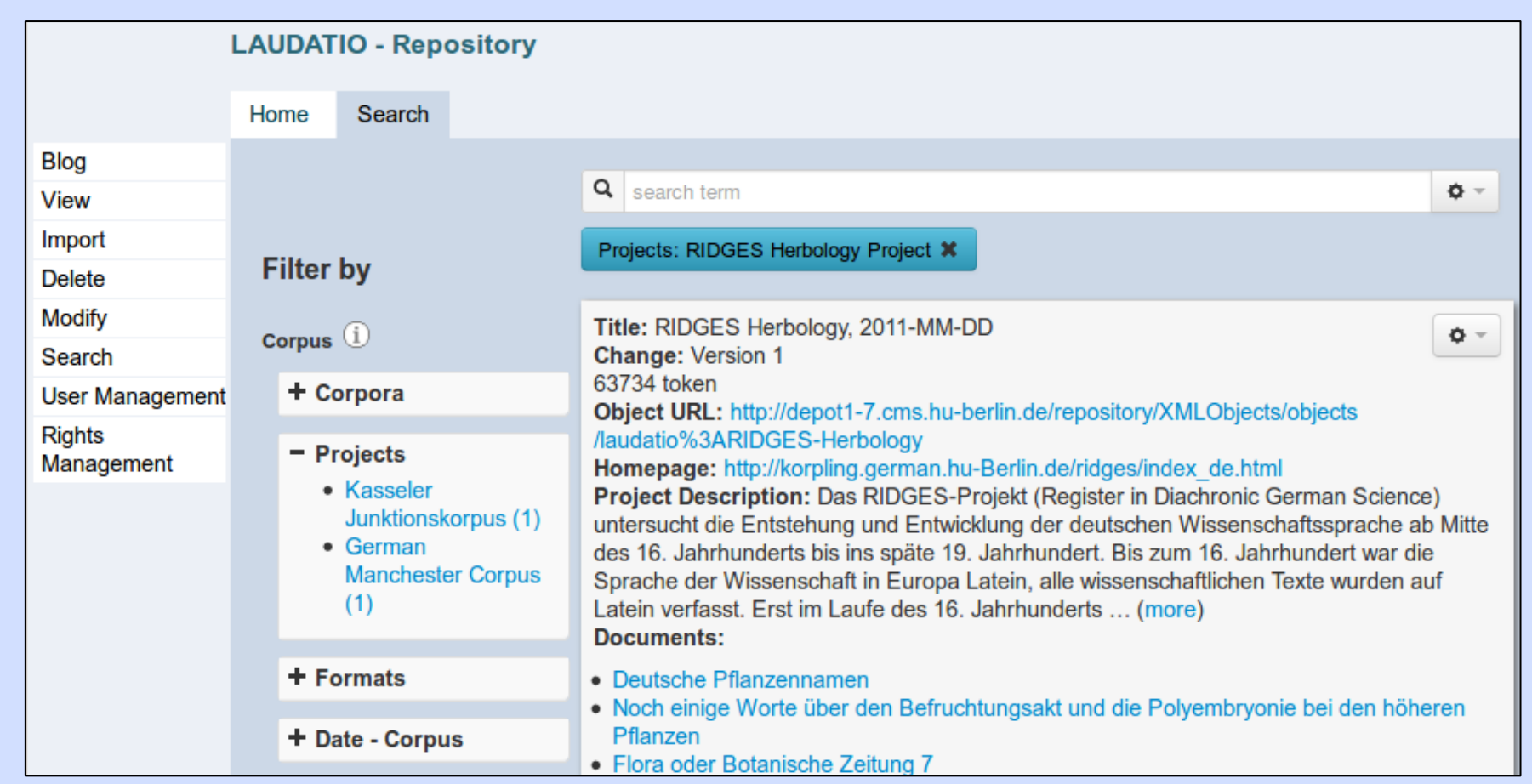
### Anzeige

- Strukturierte Anzeige der Metadaten
- Versionsauswahl
- Download in den verschiedenen Formaten



### Suche und Filtern

- Einschränkung der Suchtreffer durch Filterung nach Kategorien und Werten aus den Metadaten (Facettensuche)
- Freitextsuche
- Anzeige der Treffer (Korpora und Dokumente) mit Verlinkung zur Detailanzeige



### Integration mit externen Forschungsumgebungen

- Linguistische Korpusuche in ANNIS (Zeldes et al. 2009), Verlinkung aus dem Repository
- Konvertierung in neue, noch nicht vorhandene Formate mit SaltNPepper (Zipser & Romary 2010)



## 4. Kooperationen

- Ziel: zentrale Anlaufstelle für historische Korpora (Odebrecht & Zipser 2013)
- Kooperationen mit externen Korpus-Erstellern

### RIDGES Herbology

- Wissenschaftliche Dokumente von 1500 -1900
- Diachrone Forschung zur Entwicklung der Wissenschaftssprache im Deutschen
- Normalisierung und verschiedene Tokenisierung
- im EXMARaLDA-Format (Schmid 2002)

### Kasseler Junktionskorpus (KAJUK)

- private Dokumente des 17. und 19. Jahrhunderts
- ‚Nähetext‘-Forschung, diachrone Grammatik von Frühneuhochdeutsch
- diskontinuierliche Annotationen mit Hilfe von Pointing Relations
- idiosynkratisches XML-Format

### German Manchester Corpus (GerManC)

- geschriebene Dokumente von verschiedenen Registern von 1650 bis 1800
- Forschung zur Entwicklung einer deutschen Standardsprache, vergleichende Studien
- u.a. strukturelle Textkodierung via TEI P5 Lite XML-Format (Burnard & Bauman 2008)

## 5. Ausblick

- Ausbau der Kooperation mit weiteren Projekten wie dem DDD - Referenzkorpus Althochdeutsch (750-1050), Anpassung der Metadaten
- öffentliche Beta-Version der Weboberfläche
- Erweiterung der Funktionalitäten
  - Integration einer PDF-Ansicht, weitere Suchfunktionen
  - TEI Header Formular-Eingabe direkt in der Weboberfläche

### Referenzen:

Ägel, V., Hennig, M. (2007) DFG-Projekt. Explizite und elliptische Junktions in der Syntax des Neuhochochdeutschen. Forschungsnotiz. *Zeitschrift für Germanistische Linguistik* 35. S. 185-189. | Burnard, L., Bauman, S. (Eds.) (2008). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Oxford. | Burnard, L., Rahtz, S. (2004) *RelaxNG with Son of ODD. Extreme Markup Languages*. | Durell, M., Ensslin, A., Bennett, P. (2007) The GerManC project. *Sprache und Datenverarbeitung* 31. S. 71-80. | Elastic Search <http://www.elasticsearch.org/> | Lagoze, C., Payette, S., Shin, E., Wilper, C. (2006). Fedora: an architecture for complex objects and their relationships. *International Journal on Digital Libraries*, 6(2). S.124-138 | Linde, S., Unverzagt, S., Donhauser, K. (erscheint) Old German Reference Corpus. 32. *Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*. Potsdam 2013. | Odebrecht, C., Zipser, F. (2013) LAUDATIO - Eine Infrastruktur zur linguistischen Analyse historischer Korpora. *DTA-/CLARIN-D Konferenz und -Workshops: Historische Textkorpora für die Geistes- und Sozialwissenschaften. Fragestellungen und Nutzungsperspektiven*. Berlin 2013. | RIDGES Herbology [http://korpling.german.hu-berlin.de/ridges/documentation\\_en.html](http://korpling.german.hu-berlin.de/ridges/documentation_en.html). | Schmidt, T. (2002) EXMARaLDA - ein System zur Diskurstranskription auf dem Computer. *Arbeiten zur Mehrsprachigkeit*, Folge B 34: S.1 ff. | Zeldes, A., Ritz, J., Lüdeling, A., Chiarcos, C. (2009). ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009*. Liverpool, UK. | Zipser, F., Romary, L. (2010). A model oriented approach to the mapping of annotation formats using standards. In *Workshop on Language Resource and Language Technology Standards, LREC 2010*. La Valette, Malta.