Integrationsunterstützung für TEI-kodierte Textdokumente in CLARIN-D

Thomas Eckart
Natural Language Processing Group
Institute of Computer Science, University of Leipzig
teckart@informatik.uni-leipzig.de

UNIVERSITÄT LEIPZIG
Institute of Computer Science

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

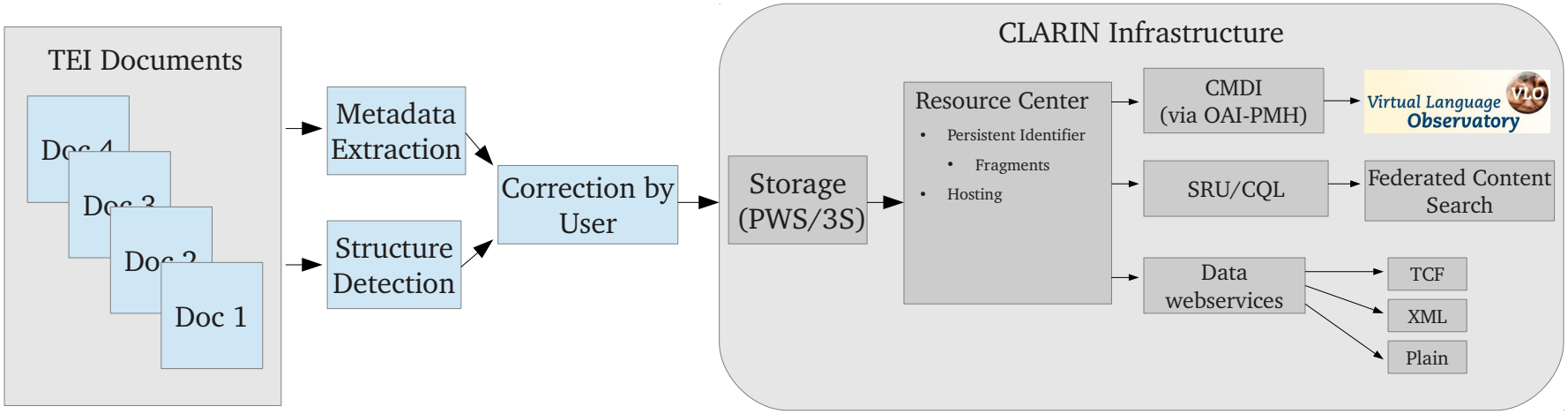
Aufgabe: Ressourcenkuration/-integration

- Befüllen der Infrastruktur mit neuen Daten als zentrale Aufgabe um Attraktivtät zu steigern
- Problem:
  - Aufwand für Integration (Erstellung benutzbarer CMDI Profile, Erstellen Instanzdateien, Distribution, Speicherung etc.)
- Hürden für Beteiligung absenken
- Nutzung vorhandener Ressourcen

Aufgabe: Ressourcenkuration/-integration

- „Pull“-Faktoren:
  - Langzeitarchivierung
  - Globale Sichtbarkeit (Suchportal/Inhaltssuche)
  - Nutzung eigener Ressourcen in der CLARIN Infrastruktur
    - Werkzeuge
    - Vergleich von Ressourcen
    - Virtuelle Aggregationen
    - etc.
  - Geringerer Aufwand als manuelle Verteilung

Übersicht „TEI-Integrator“



Granularität

- Adressierbarkeit und Offenlegung verschiedener Granularitäten textueller Ressources durch die Nutzung von Handles + Part Identifier (inspiriert durch CITE CTS)

```
<TEI>
    <text>
        <div type="chapter" n="1">
            <div type="dedication">
        <div type="chapter" n="2">
        <div type="chapter" n="3">
...
```

Handle system

11858/00-229C-0000-0006-AAD2-A

@/chapter:1

@/chapter:1/dedication

@/chapter:2

@/chapter:3

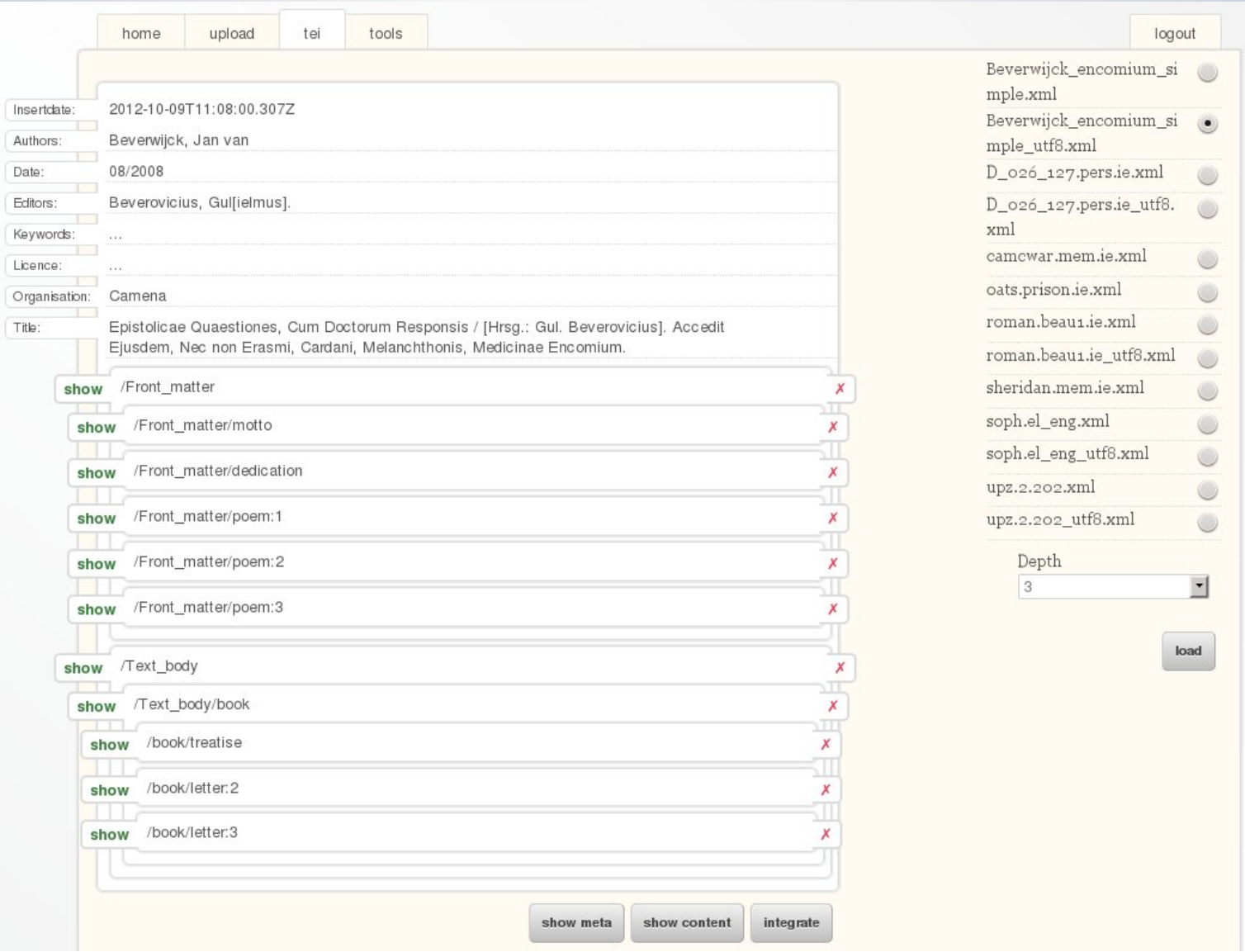
- Beschreibung der Metadaten und Abbildungsinformationen in extra CMDI-Profil

ASV Online Toolbox

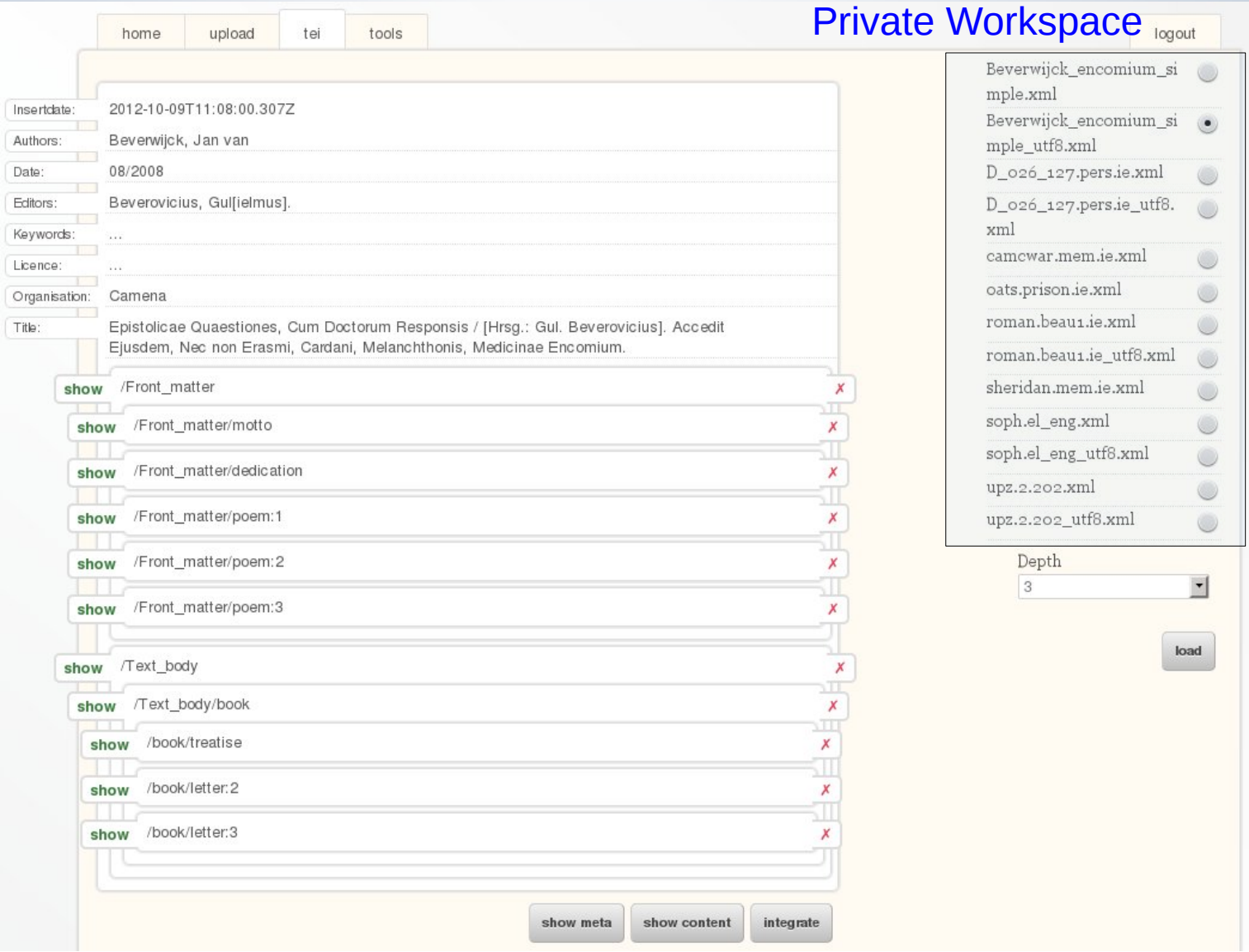
- Einfache Webseite mit Uploadfunktionalität für den Private Work Space (PWS)
- Basiert auf existierender Offline Software
- Bestehende Werkzeuge (Auswahl):
  - LanI – Language Identification
  - Terminologie Extraction
- Vollständig basierend auf RESTful Webservices

DEMO

Screenshot TEI-Integrator



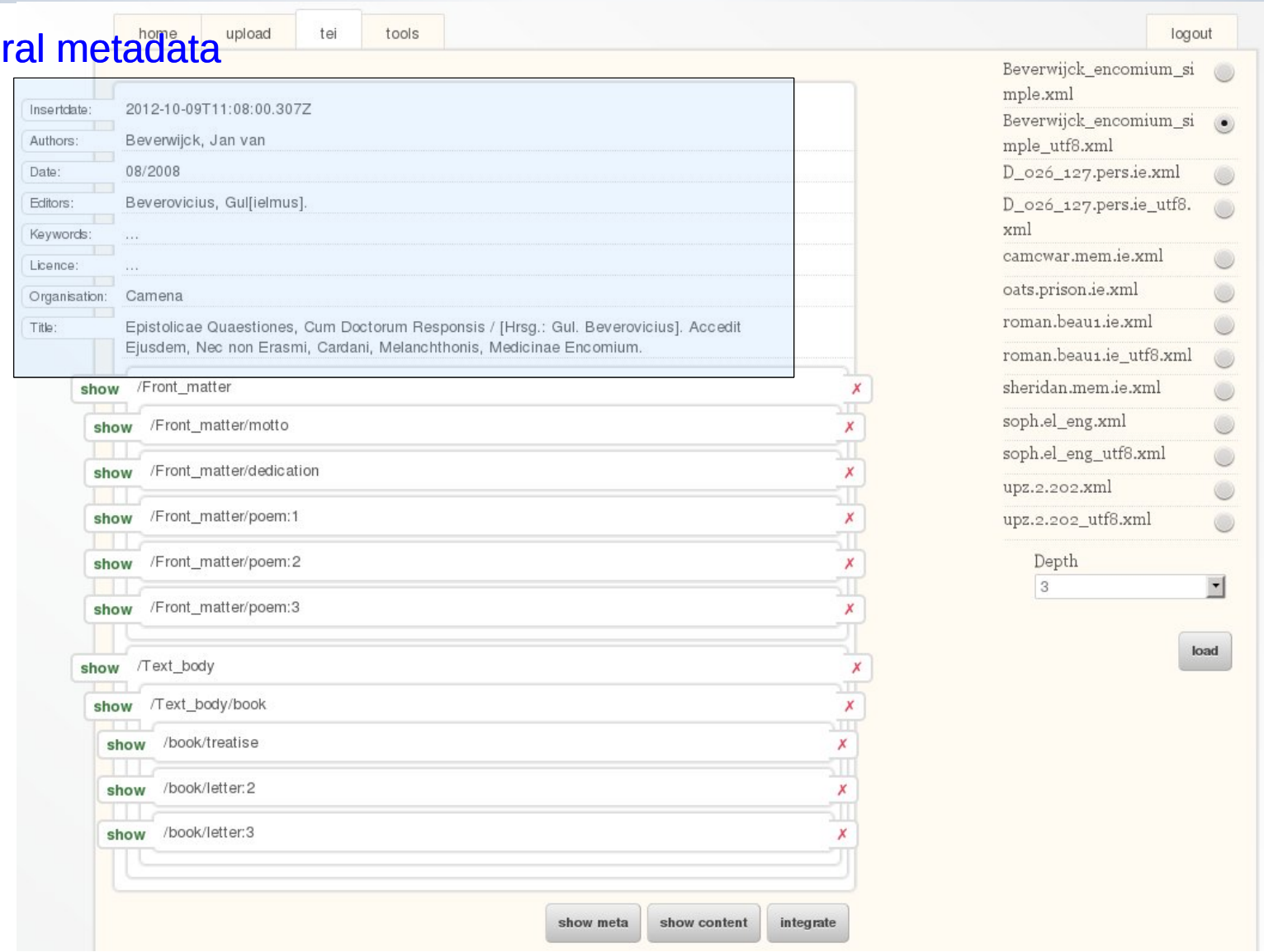
Screenshot TEI-Integrator



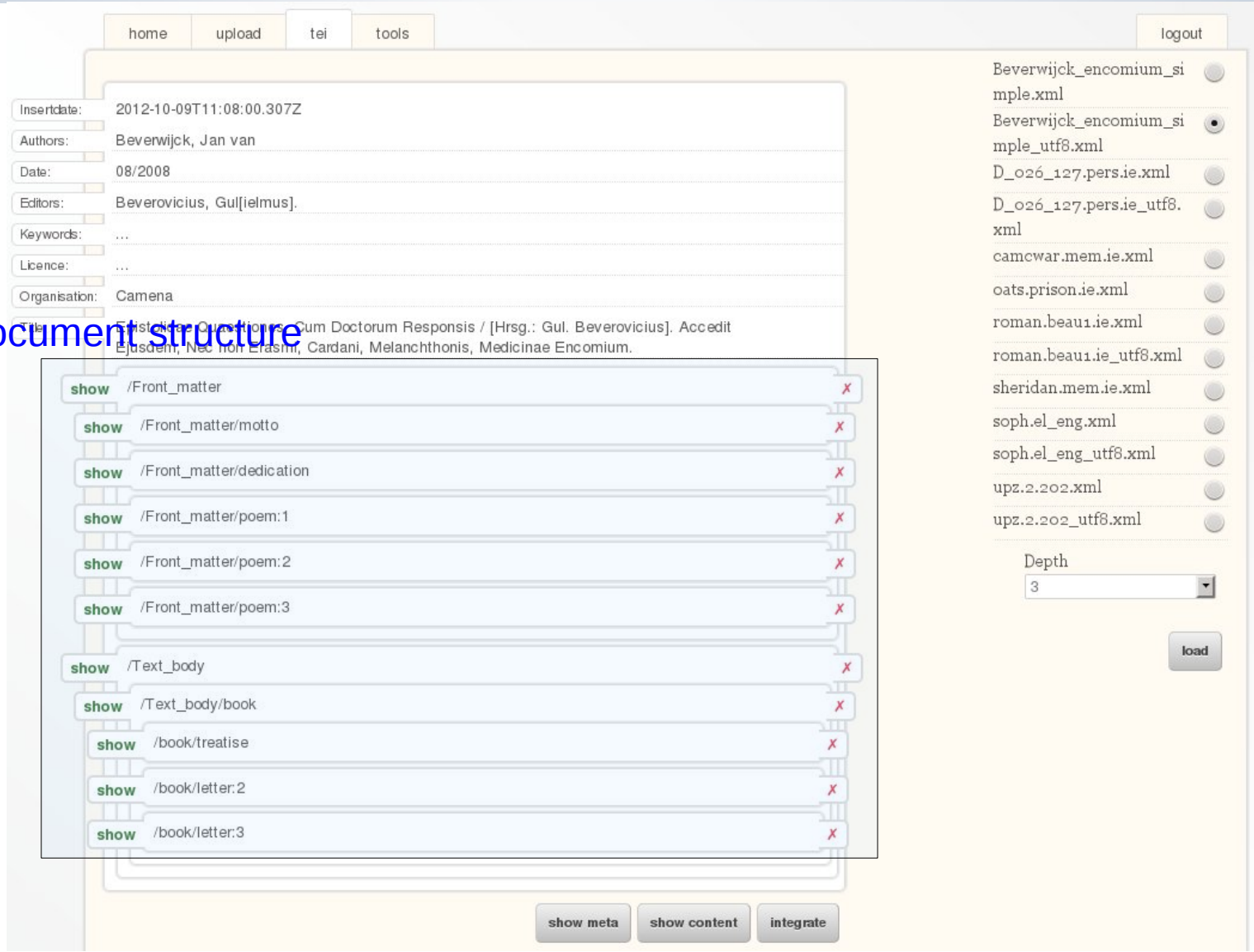
Screenshot TEI-Integrator

CLARIN-D

General metadata



Screenshot TEI-Integrator



d Browser Resources

Electra search

ator (1)
voor Beeld en Geluid Academia collectie (1)

LANGUAGE
English (1)

GENRE
discourse (1)

SUBJECT
beroepssport (1)

name	description
Electra (English)	Document
PAUW & WITTEMAN	Regelmatig terugkerende onderdelen: bespreking van actuele gebeurtenissen en (media)nieuws met de...
hi1of-electra	hearing-impaired children receiving oral education sampled at 12, 18, 24, 30, 36, 42, and 54 months

Search in this result set:
CQL query: Send query

Record 1 out of 3 >

Field	Value
name	Electra (English)
description	Document
id	11858/00-229C-0000-0006-AAD2-A
collection	CLARIN-D Demonstrator
dataProvider	TEST Archive
metadataSource	http://catalog.clarin.eu/oai-harvester/mpi-self-harvest/harvested/results/cmdi/University_of_Leipzig/00-229C-0000-0006-AAD2-A.xm

Resources:
http://clarinws.informatik.uni-leipzig.de:8080/TEI-Integration/getPIDContent/11858/00-229C-0000-0006-AAD2-A
http://clarinws.informatik.uni-leipzig.de:8080/TEI-Integration/GetDocument/11858/00-229C-0000-0006-AAD2-A

Search in this resource:
CQL query: Send query

Show complete CMDI metadata

Zukünftige Arbeiten

- Performanceverbesserungen
- Verbesserte Oberfläche und weitere Funktionalität (Unterstützung von Dokumentkollektionen)
- Verbesserter TCF Export
- Erstellung typischer Workflows (BBAW, TU Darmstadt)

Danke für die Aufmerksamkeit!

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

UNIVERSITÄT LEIPZIG

Institute of Computer Science