

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

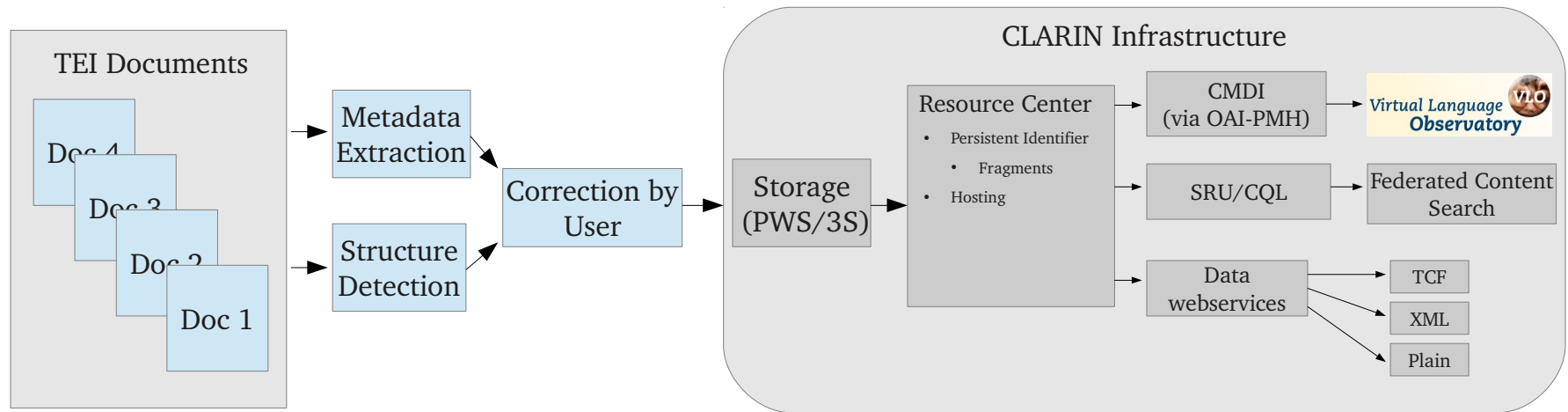


Bundesministerium
für Bildung
und Forschung

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

- „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“



- Adressierbarkeit und Offenlegung verschiedener Granularitäten textueller Ressourcen 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">
```

...

- Beschreibung der Metadaten und Abbildungsinformationen in extra CMDI-Profil

```
Handle system
11858/00-229C-0000-0006-AAD2-A
@/chapter:1
@/chapter:1/dedication
@/chapter:2
@/chapter:3
```

- 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

home upload **tei** tools logout

Insertdate: 2012-10-09T11:08:00.307Z
Authors: Beverwijck, Jan van
Date: 08/2008
Editors: Beverovicus, Gul[ielmus].
Keywords: ...
Licence: ...
Organisation: Camena
Title: Epistolicae Quaestiones, Cum Doctorum Responsis / [Hrsg.: Gul. Beverovicus]. Accedit Ejusdem, Nec non Erasmi, Cardani, Melanchthonis, Medicinae Encomium.

show /Front_matter X
show /Front_matter/motto X
show /Front_matter/dedication X
show /Front_matter/poem:1 X
show /Front_matter/poem:2 X
show /Front_matter/poem:3 X
show /Text_body X
show /Text_body/book X
show /book/treatise X
show /book/letter:2 X
show /book/letter:3 X

Beverwijck_encomium_si mple.xml
Beverwijck_encomium_si mple_utf8.xml
D_026_127.pers.ie.xml
D_026_127.pers.ie_utf8.xml
camcwar.mem.ie.xml
oats.prison.ie.xml
roman.beau1.ie.xml
roman.beau1.ie_utf8.xml
sheridan.mem.ie.xml
soph.el_eng.xml
soph.el_eng_utf8.xml
upz.2.202.xml
upz.2.202_utf8.xml

Depth

load

show meta **show content** **integrate**

Screenshot TEI-Integrator



CLARIN-D

home upload **tei** tools

Private Workspace [logout](#)

Insertdate: 2012-10-09T11:08:00.307Z
Authors: Beverwijck, Jan van
Date: 08/2008
Editors: Beverovicus, Gul[ielmus].
Keywords: ...
Licence: ...
Organisation: Camena
Title: Epistolicae Quaestiones, Cum Doctorum Responsis / [Hrsg.: Gul. Beverovicus]. Accedit Ejusdem, Nec non Erasmi, Cardani, Melanchthonis, Medicinae Encomium.

show /Front_matter X
show /Front_matter/motto X
show /Front_matter/dedication X
show /Front_matter/poem:1 X
show /Front_matter/poem:2 X
show /Front_matter/poem:3 X
show /Text_body X
show /Text_body/book X
show /book/treatise X
show /book/letter:2 X
show /book/letter:3 X

[show meta](#) [show content](#) [integrate](#)

- Beverwijck_encomium_si mple.xml
- Beverwijck_encomium_si mple_utf8.xml
- D_026_127.pers.ie.xml
- D_026_127.pers.ie_utf8.xml
- camcwar.mem.ie.xml
- oats.prison.ie.xml
- roman.beau1.ie.xml
- roman.beau1.ie_utf8.xml
- sheridan.mem.ie.xml
- soph.el_eng.xml
- soph.el_eng_utf8.xml
- upz.2.202.xml
- upz.2.202_utf8.xml

Depth

[load](#)

Screenshot TEI-Integrator

General metadata

home upload **tei** tools logout

Insertdate: 2012-10-09T11:08:00.307Z
Authors: Beverwijck, Jan van
Date: 08/2008
Editors: Beverovicus, Gul[ielmus].
Keywords: ...
Licence: ...
Organisation: Camena
Title: Epistolicae Quaestiones, Cum Doctorum Responsis / [Hrsg.: Gul. Beverovicus]. Accedit Ejusdem, Nec non Erasmi, Cardani, Melanchthonis, Medicinae Encomium.

- [show](#) /Front_matter X
- [show](#) /Front_matter/motto X
- [show](#) /Front_matter/dedication X
- [show](#) /Front_matter/poem:1 X
- [show](#) /Front_matter/poem:2 X
- [show](#) /Front_matter/poem:3 X
- [show](#) /Text_body X
- [show](#) /Text_body/book X
- [show](#) /book/treatise X
- [show](#) /book/letter:2 X
- [show](#) /book/letter:3 X

Beverwijck_encomium_si mple.xml

Beverwijck_encomium_si mple_utf8.xml

D_026_127.pers.ie.xml

D_026_127.pers.ie_utf8.xml

camcwar.mem.ie.xml

oats.prison.ie.xml

roman.beau1.ie.xml

roman.beau1.ie_utf8.xml

sheridan.mem.ie.xml

soph.el_eng.xml

soph.el_eng_utf8.xml

upz.2.202.xml

upz.2.202_utf8.xml

Depth

Screenshot TEI-Integrator

home upload **tei** tools logout

Insertdate: 2012-10-09T11:08:00.307Z
Authors: Beverwijck, Jan van
Date: 08/2008
Editors: Beverovicus, Gul[ielmus].
Keywords: ...
Licence: ...
Organisation: Camena

Document structure

File: *Historiae Quaestiones Sum Doctorum Responsis / [Hrsg.: Gul. Beverovicus]. Accedit Eiusdem, Nec non Erasmi, Cardani, Melanchthonis, Medicinae Encomium.*

- show /Front_matter X
- show /Front_matter/motto X
- show /Front_matter/dedication X
- show /Front_matter/poem:1 X
- show /Front_matter/poem:2 X
- show /Front_matter/poem:3 X
- show /Text_body X
- show /Text_body/book X
 - show /book/treatise X
 - show /book/letter:2 X
 - show /book/letter:3 X

Depth: 3 load

show meta show content integrate

File list:

- Beverwijck_encomium_si mple.xml
- Beverwijck_encomium_si mple_utf8.xml
- D_026_127.pers.ie.xml
- D_026_127.pers.ie_utf8.xml
- camcwar.mem.ie.xml
- oats.prison.ie.xml
- roman.beau1.ie.xml
- roman.beau1.ie_utf8.xml
- sheridan.mem.ie.xml
- soph.el_eng.xml
- soph.el_eng_utf8.xml
- upz.2.202.xml
- upz.2.202_utf8.xml

Virtual Language Observatory

Explore the world of language resources and technology from different perspectives



Browser Resources

- 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...
hildf-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:

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/cmdj/University_of_Leipzig/00-229C-0000-0006-AAD2-A.xml

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:

[Show complete CMDI metadata](#)

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

Danke für die Aufmerksamkeit!

UNIVERSITÄT LEIPZIG

Institute of Computer Science

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung