Showcase

Integration and Annotation of (TEI based) text documents in the CLARIN infrastructure

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

Richard Eckart de Castilho¹, Chris Biemann², Seid Muhie Yimam¹²
UKP Lab (UKP-TUDA)¹ & FG Language Technology²
Technische Universität Darmstadt
eckart@ukp.informatik.tu-darmstadt.de
Elements of the CLARIN-D Infrastructure

(Web) Applications
- TCF Visualiz.
- TLA
- WebLicht
- FCS
- WebAnno
- VLO
- Annex
- WebMaus
- PWS

Protocols and Specifications
- OAI-PMH
- PID
- AAI
- SRU/CQL
- ISOCat

Data formats
- CMDI
- TEI
- TCF

Nine CLARIN-D Data Repositories
Workflow - 1/3 Metadata Creation

- TEI Collection
  - Doc 4
  - Doc 3
  - Doc 2
  - Doc 1
- Metadata Extraction
- Structure Detection
- Correction by User
- Storage (PWS)
Workflow - 2/3 Annotation (optional)

- Annotation tool
- Input: extracted plain text (WS)
- Output: annotated TCF files
Workflow - 3/3 Integration

Permanent Integration in CLARIN Infrastructure

Resource Center
- Persistent Identifier
- Fragments
- Hosting

CMDI (via OAI-PMH)

SRU/CQL

Federated Content Search

Data webservices
- TCF
- XML
- Plain

Virtual Language Observatory
Semi-automatic Resource Curation

• Filling the infrastructure with data as an essential task to attract new users

• Possible sources:
  • Existing resources (e.g. from completed projects)

• Goal:
  • Minimizing effort (creation of CMDI files, services for access, storage at repository etc.)
Resource Curation as a central task

• Arguments to participate:
  • Long term availability
  • Global visibility (search portal/content search)
  • Usage of own resources in CLARIN infrastructure
  • Less effort than manual distribution
• Collections of TEI-encoded text files
Integration at repository Leipzig

Permanent Integration in CLARIN Infrastructure

Resource Center
- Persistent Identifier
- Fragments
- Hosting

 CMDI (via OAI-PMH)

 SRU/CQL

 Federated Content Search

 Data webservices

 TCF

 XML

 Plain

Virtual Language Observatory
Distributed Annotation with WebAnno

- Web-based annotation tool
- Distributed annotation
- Crowdsourcing of annotation tasks
- Open source
- Integration with CLARIN services
- ... under development!

WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations
Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, Chris Biemann
Project setup
- Import documents
- Define guidelines
- Define types/tagsets

Annotation of documents
- Assign workload
- Monitor progress
- Evaluate quality

Curation
- Compare annotations
- Merge annotations
- Correct annotations

Annotators
Crowd

Monitoring

Manager
Curator
DEMO

TEI Integrator

WebAnno

VLO
Thank you for your attention!
Backup Slides
TEI-Integrator – Generate Metadata

Collections:
- This is a test collection.

Collection Resources:
- Beverwijk_encomium_simple.xml
- cancwar_mem.ie.xml
- smith.lectures.ie.xml

Welcome | Metadata | Workspace | Extraction | Tools | Integration

Depth
2

(Re)Generate CMDI | Save CMDI | Show CMDI | Show TCF

Basic

Title
The Soldiers' Monument in Cambridge: Proceedings in relation to the building and dedication of the mor

Author

Licence

Resource Type
- TEI-encoded document

Organisation

Open
Distributive Annotation with WebAnno

Projects Settings

Users:
- anno1 [user]
- anno2 [user]
- anno3 [user]
- anno4 [user]
- anno5 [user]
- richard [admin, curator]
- test [admin, curator]

Add Users | Remove User | Change Permissions
Distributive Annotation with WebAnno

Projects Settings

Tag sets
- [coreference type] BART
- [named entity] NER_WebAnno
- [pos] STTS
- [dependency] Tiger
- [coreference] TuebaDZ

Tag set details
- Name: NER_WebAnno
- Language: en
- Layer: named entity
- Description: Named Entity annotation

Import tag sets
- Format: JSON
- Files: Dateien auswählen

Tag details
- Name: TIME
- Description: Time information

Create project | Create tag | Save tag | Delete tag | Export tag set | Save tag set | Delete tag set
The following pages contain the substance of Lectures on the subject of Domestic Slavery in the United States, which for several years have been delivered to the classes in Moral Science in Randolph Macon College.

Since the year 1844, I have been frequently called on to discuss this subject on various popular occasions in Virginia and North Carolina.

My classes in college were compelled to deal with the subject of domestic slavery.

Not only the popular ideas in regard to African slavery in this country, but the specific treatment of this topic by numerous text authors in Moral Science, rendered this unavoidable.

A deep conviction that the minds of young men were receiving a wrong, and, in the present state of the country, a fatal direction, both as regards the principles of the institution, and the institution itself, induced me to substitute the text authorities on the subject by a course of lectures.

These lectures, therefore, were originally drawn up with a view to oral delivery.

They were modified by the circumstances of their origin.
The following pages contain the substance of Lectures on the subject of Domestic Slavery in the United States, which for several years have been delivered to the classes in Moral Science in Randolph Macon College.

Since the year 1844, I have been frequently called on to discuss this subject on various popular occasions in Virginia and North Carolina.

My classes in college were compelled to deal with the subject of domestic slavery.

Not only the popular ideas in regard to African slavery in this country, but the specific treatment of this topic by numerous text authors in Moral Science, rendered this unavoidable.

A deep conviction that the minds of young men were receiving a wrong, and, in the present state of the country, a fatal direction, both as regards the principles of the institution, and the institution itself, induced me to substitute the text authorities on the subject by a course of lectures.

These lectures, therefore, were originally drawn up with a view to oral delivery.

They were modified by the circumstances of the occasion.
The following pages contain the substance of Lectures on the subject of Domestic Slavery in the United States, which for several years have been delivered to the classes in Moral Science in Randolph Macon College.

Since the year 1844, I have been frequently called on to discuss this subject on various popular occasions in Virginia and North Carolina.

My classes in college were compelled to deal with the subject of domestic slavery.

Not only the popular ideas in regard to African slavery in this country, but the specific treatment of this topic by numerous text authors in Moral Science, rendered this unavoidable.

A deep conviction that the minds of young men were receiving a wrong, and, in the present state of the country, a fatal direction, both as regards the principles of the institution, and the institution itself, induced me to substitute the text authorities on the subject by a course of lectures.

These lectures, therefore, were originally drawn up with a view to oral delivery.

They were modified by the circumstances of their origin.
Distributive Annotation with WebAnno

Projects Settings

WebAnno | Home | User: richard | Log out

Projects

US_collection

WebService

Sending Curated Document to CLARIN

Export curated documents to CLARIN workspace
TEI-Integrator – Push to repository

CLARIN-D Center Leipzig

The CLARIN-D repository at the University of Leipzig offers longterm preservation of digital resources, along with their descriptive metadata. The mission of the repository is to ensure the availability and longterm preservation of resources, to preserve knowledge gained in research, to aid the transfer of knowledge into new contexts, and to integrate new methods and resources into university curricula.

CLARIN-D Center BBAW Berlin

The CLARIN-D repository at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) includes textual resources provided by CLARIN-D member institutions and other institutions and/or organizations that belong to the CLARIN-D extended community.
<table>
<thead>
<tr>
<th>name</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>In a fair country</td>
<td></td>
</tr>
<tr>
<td>LCC data provider &quot;www.faz.net&quot; in resource with identifier 11858/00-229C-0000-0001-B06F-3</td>
<td>Data provider of the Leipzig Corpora Collection: <a href="http://www.faz.net">www.faz.net</a></td>
</tr>
<tr>
<td>Lectures on the Philosophy and Practice of Slavery as exhibited in</td>
<td></td>
</tr>
<tr>
<td>the Institution of Domestic Slavery in the United States: with the</td>
<td></td>
</tr>
<tr>
<td>Duties of Masters to Slaves,</td>
<td></td>
</tr>
<tr>
<td>Testcollection</td>
<td></td>
</tr>
<tr>
<td>The picturesque pocket companion, and visitor's guide, through</td>
<td></td>
</tr>
<tr>
<td>Mount Auburn</td>
<td></td>
</tr>
</tbody>
</table>

Search in this result set:

CQL query: [Input field]  
Send query
### TEI-Integrator – Visibility

#### Table 1: Field Value pairs for the resource "Lectures on the Philosophy and Practice of Slavery as exhibited in the Institution of Domestic Slavery in the United States: with the Duties of Masters to Slaves,"

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>Lectures on the Philosophy and Practice of Slavery as exhibited in the Institution of Domestic Slavery in the United States: with the Duties of Masters to Slaves,</td>
</tr>
<tr>
<td>id</td>
<td>11858/00-229C-0000-0006-AAD2-A@documentid=smith.lectures.ie</td>
</tr>
<tr>
<td>collection</td>
<td>CLARIN-D</td>
</tr>
<tr>
<td>dataProvider</td>
<td>Test Providers</td>
</tr>
<tr>
<td>languages</td>
<td>undetermined, Undefined</td>
</tr>
<tr>
<td>metadataSource</td>
<td><a href="http://catalog.clarin.eu/oai-harvester/test-providers/harvested/results/cmdi/Test_Repository/smith.lectures.ie.cmdi">http://catalog.clarin.eu/oai-harvester/test-providers/harvested/results/cmdi/Test_Repository/smith.lectures.ie.cmdi</a></td>
</tr>
<tr>
<td>organisation</td>
<td>Perseus</td>
</tr>
</tbody>
</table>

Resources:

- http://clarintest.informatik.uni-leipzig.de:8080/fedora/objects/teicoll:1185800229C00000006AAD2Ace3/datastreams/tei/content

Search in this resource:

<table>
<thead>
<tr>
<th>CQL query:</th>
<th></th>
<th>Send query</th>
</tr>
</thead>
</table>

Show CMDI metadata