Canonical Text Services in CLARIN

Reaching out to the Digital Classics and beyond

Jochen Tiepmar, Thomas Eckart, Dirk Goldhahn and Christoph Kuras
Overview CTS

Canonical Text Services (CTS)

• protocol for a web-based citable text service
• Unique Identifiers (Unique Resource Name, URN) refer to text passages and text parts
• Developed in Homer Multitext Project (www.homermultitext.org), Smith et al. 2009
  http://www.homermultitext.org/hmt-docs/specifications/ctsunr/
  http://www.homermultitext.org/hmt-docs/specifications/cts/
• This implementation was done in Billion Words Project (ESF)
Canonical Citation

Document outer hierarchy

Shakespeare → Sonnets → english → 1st edition

Text passage inner hierarchy

Sonnet 1 → Vers 1

Combined

Shakespeare → Sonnets → english → 1st edition → Sonnet 1 → Vers 1

CTS-URN

urn:cts:demo:shakespeare.sonnets.en.1:1.1
Canonical Citation

urn:cts:demo:shakespeare.sonnets:
urn:cts:demo:shakespeare.sonnets.de:
Canonical Citation

urn:cts:demo:shakespeare.sonnets:35.4
Canonical Citation

urn:cts:demo:shakespeare.sonnets:35
Canonical Citation

urn:cts:demo:shakespeare.sonnets:35.1-35.5
urn:cts:demo:shakespeare.sonnets:35.1-35
Integration in CLARIN

- Integrated in repository of CLARIN center Leipzig
- May function as a template for integration of more CTS servers / instances
- CMDI 1.2 compliant metadata that
  - Allow direct access to services and raw files (here: EpiDoc TEI)
  - Reflects text granularity (currently 3 levels)
    1. Collection (here: Excerpt of *Parallel Bible Corpus*)
    2. Document (here: Bible)
    3. 1 Resource per Book of Bible
Integration in CLARIN

- Presentation in VLO:

  - **urn:cts:pbc:bible**
  - **urn:cts:pbc:bible.parallel.arb.norm:**
  - **urn:cts:pbc:bible.parallel.ceb.bugna:**
  - **urn:cts:pbc:bible.parallel.ces.kralicka:**
  - ...

- Planed: FCS endpoint for content search based on existing fulltext index
The Bible in French. King James Version.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>cts</td>
<td>either</td>
</tr>
<tr>
<td>kingjames.xml</td>
<td>either</td>
</tr>
<tr>
<td>cts</td>
<td>test document</td>
</tr>
<tr>
<td>cts</td>
<td>test document</td>
</tr>
<tr>
<td>cts</td>
<td>test document</td>
</tr>
<tr>
<td>cts</td>
<td>test document</td>
</tr>
<tr>
<td>cts</td>
<td>test document</td>
</tr>
<tr>
<td>cts</td>
<td>test document</td>
</tr>
</tbody>
</table>

Mime type: text/xml

### The Bible in French. King James Version.

<table>
<thead>
<tr>
<th>Resources</th>
<th>Availability</th>
<th>Technical details</th>
<th>Hierarchy</th>
</tr>
</thead>
<tbody>
<tr>
<td>urn:cts:pbc:bible:parallel:fra.kingjames:1</td>
<td>other</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

Mime type: text/xml

Why include support for CTS URNs?

CTS is developed by Humanists & reflects the requirements from certain Digital Humanity communities

   Perseus, Croatiae Auctores Latini, CITE, ...

Creating CTS ready data is a research & project goal in DH

   “pick researchers up, where they are“

Technical benefits

   Normalisation of generic text access, “Outsourcing“ of text content, The Canonical Text Infrastructure, ...
The Canonical Text Infrastructure

(All of the following tools can be tested in the demo session)
CTS Cloning

Server 1

urn:cts:demo:[work]:1.1.1
urn:cts:demo:[work]:1.2.1

Server 2

<passage>
  <div1 n="1" type="song">
    <div2 n="1" type="strophe">
      <div3 n="1" type="line">
        ...
      </div3>
    </div2>
    <div2 n="2" type="strophe">
      <div3 n="1" type="line">
        ...
      </div3>
    </div2>
  </div1>
</passage>
CTS Cloning

http://hdw.eweb4.com/out/1369880.html
Realtime Alignment Tools for CTS

(GUI implemented by Sascha Ludwig)

Alignment based on document structure

Scales very well (several complete bibles in a couple of seconds)
Realtime Alignment Tools for CTS

(GUI implemented by Sascha Ludwig)
Generic Reader

CTRaCE: Canonical Text Reader and Citation Exporter in Proceedings of the Digital Humanities, Krakow, 2016.
CTS-TM (CTS Text Miner)

Server 1

urn:cts:demo:[work]:1.1.1
urn:cts:demo:[work]:1.2.1

<passage>
  <div1 n="1" type="song">
    <div2 n="1" type="strophe">
      <div3 n="1" type="line">
      </div3>
    </div2>
    <div2 n="2" type="strophe">
      <div3 n="1" type="line">
      </div3>
    </div2>
  </div1>
</passage>

(Example Visualizations from work of Stefan Jaenicke)
CTS-TM (CTS Text Miner)

• Raw Data as webservice
CTS-TM (CTS Text Miner)

• Generic Data Visualisations as webservice
CTS-TM (CTS Text Miner)

• Open Text Mining Tool as webservice
<table>
<thead>
<tr>
<th>CTS instance</th>
<th>Tokens</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTA, Deutsches Text Archiv</td>
<td>334'820'482</td>
<td>&gt;1700 German works (literature, scholarly, ...) in 3 editions</td>
</tr>
<tr>
<td>PBC, Parallel Bible Corpus</td>
<td>247’292’629</td>
<td>831 translations of the bible</td>
</tr>
<tr>
<td>Perseus</td>
<td>27’295’030</td>
<td>greekLit, latinLit, farsiLit, pdlrefwk</td>
</tr>
<tr>
<td>German Speeches</td>
<td>6’283’662</td>
<td>German President 1984-2012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>German Chancellery 1998-2011</td>
</tr>
<tr>
<td>Law</td>
<td>851’738</td>
<td>883 german law texts</td>
</tr>
<tr>
<td>TED Subtitle Corpus</td>
<td></td>
<td>51770 documents,105 languages. 1938 English documents, big variety of topics</td>
</tr>
<tr>
<td>Croatia Auctores Latini</td>
<td>5.7 million</td>
<td>Texts written 976-1984, 467 documents, bibliographic data</td>
</tr>
<tr>
<td>Briefe und Texte aus dem intellektuellen Berlin um 1800</td>
<td></td>
<td>German &amp; French letters</td>
</tr>
<tr>
<td>Ali’s monthly journal al-Muqtabas</td>
<td></td>
<td>Arabic Newspaper/Magazin</td>
</tr>
</tbody>
</table>
Future Work

• More data sets
• More tools
  • Text Miner, Touchdevice Reader, Citation Analysis Workflow,...
• Connecting to established existing projects
Questions, Feedback?

[...] would it be ok for you, if this dataset gets referenced in CLARIN? There is a connection between this CTS implementation and CLARIN and the data could be made available in CLARIN using this connection.

(...) political reasons [...] Croatia at the moment seems not to be an official partner of CLARIN, though there are Croatian linguists very much involved with the programme. Therefore it would be great if you could publish our CTS-ready texts in the CLARIN catalog!

(Neven Jovanovic, Croatiae Auctores Latini Project)
Contact

Canonical Text Service

Jochen Tiepmar  
E-Mail: jtiepmar@informatik.uni-leipzig.de

Scalable Data Solutions (ScaDS) Leipzig  
Universität Leipzig  
Ritterstraße 9-13  
04109 Leipzig

CLARIN

Dr. Thomas Eckart  
E-Mail: teckart@informatik.uni-leipzig.de

Dr. Dirk Goldhahn  
E-Mail: dgoldhahn@informatik.uni-leipzig.de

Christoph Kuras  
E-Mail: ckuras@informatik.uni-leipzig.de

NLP - Group  
Universität Leipzig  
Augustusplatz 10  
04109 Leipzig