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 webbased citable text service
- Unique Identifiers(**U**nique **R**esource **N**ame, **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/ctsurn/
  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

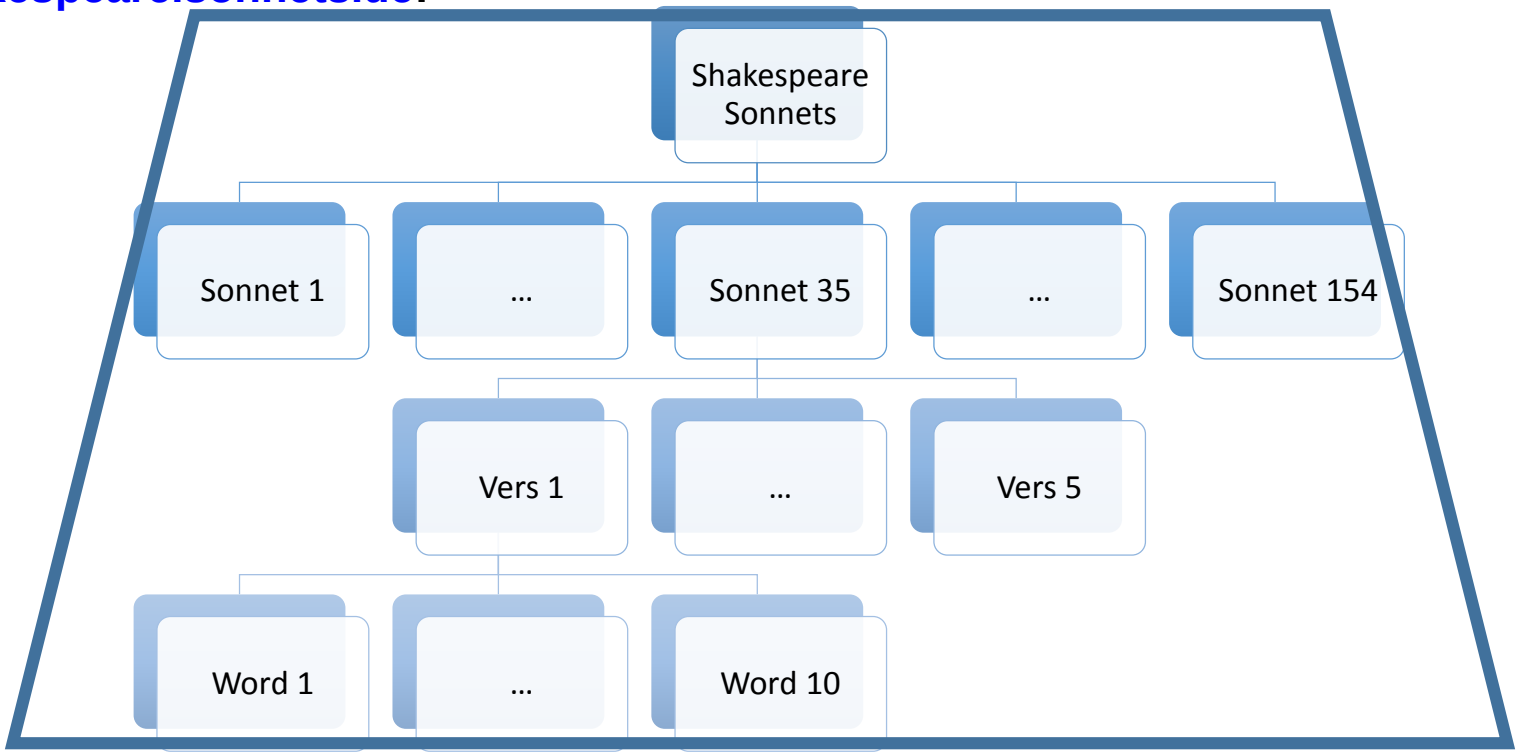
Shakespeare → Sonnets → english → 1st edition → Sonnet 1→ Vers 1

CTS-URN

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

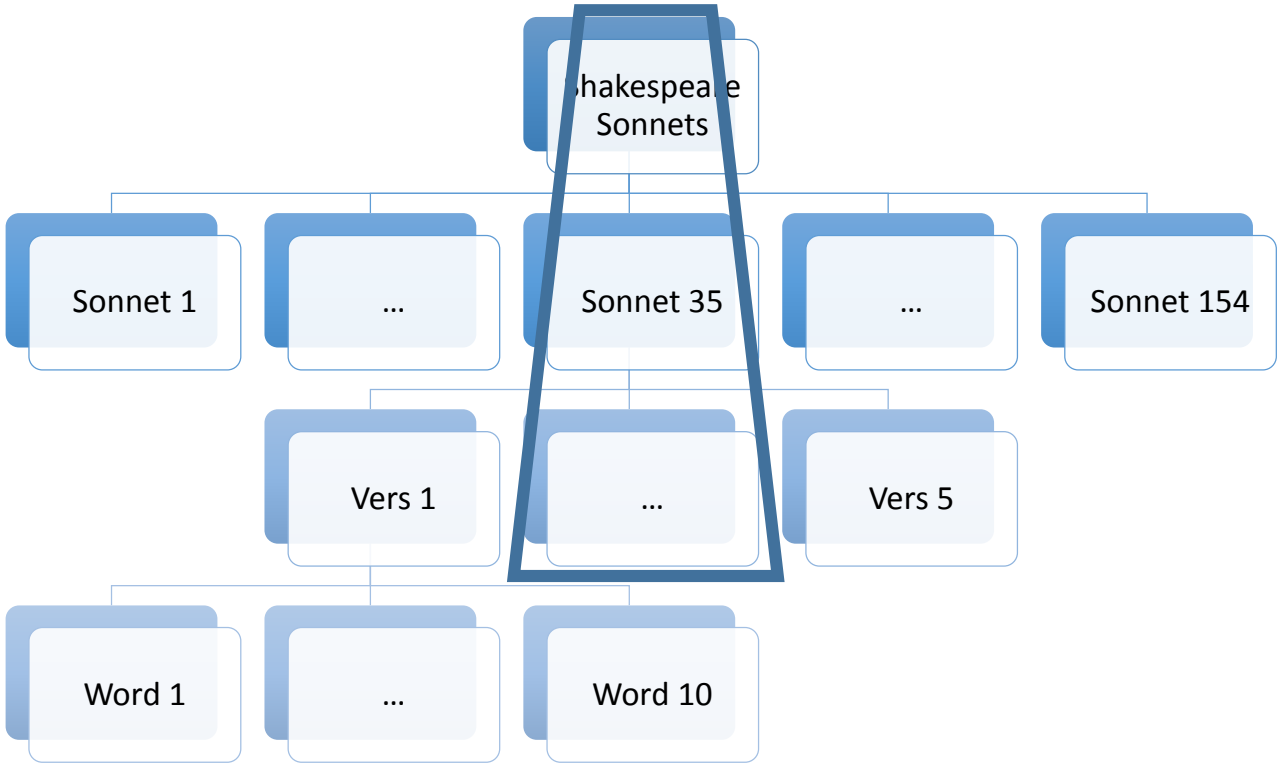
Canonical Citation

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



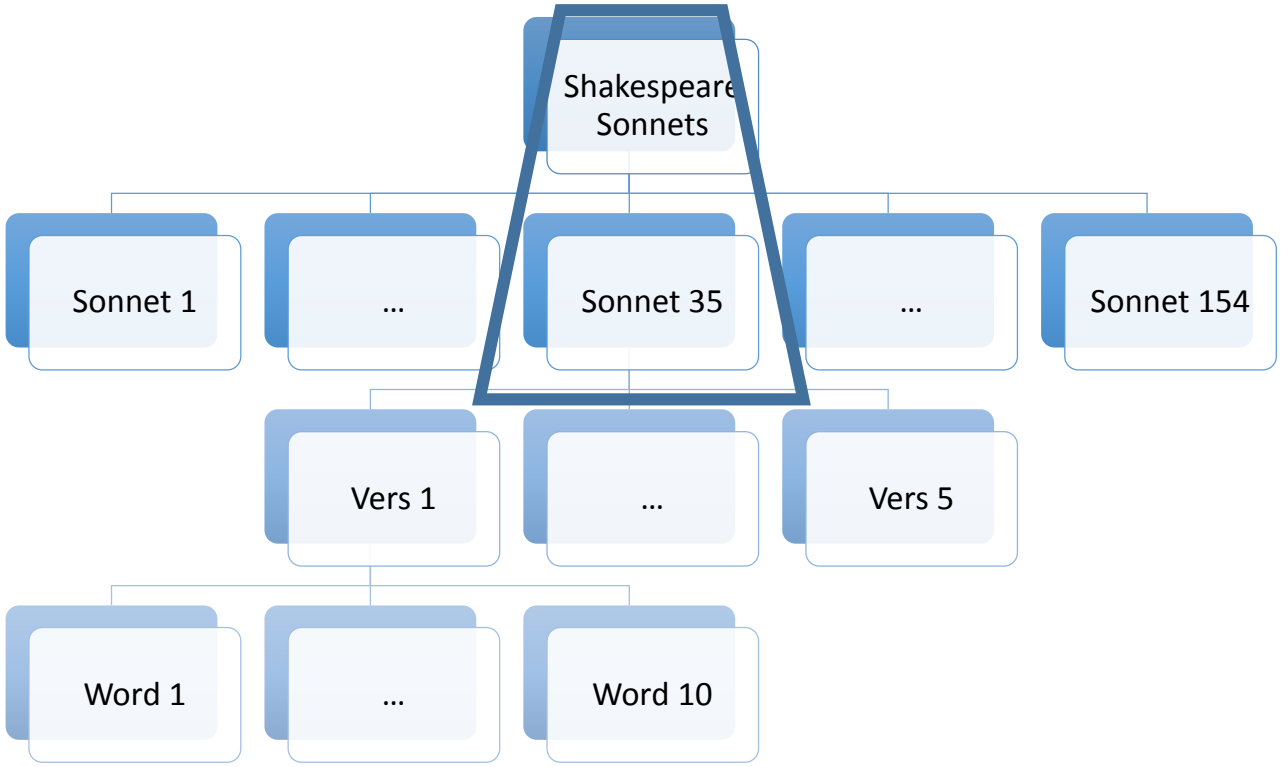
Canonical Citation

urn:cts:demo:shakespeare.sonnets:35.4



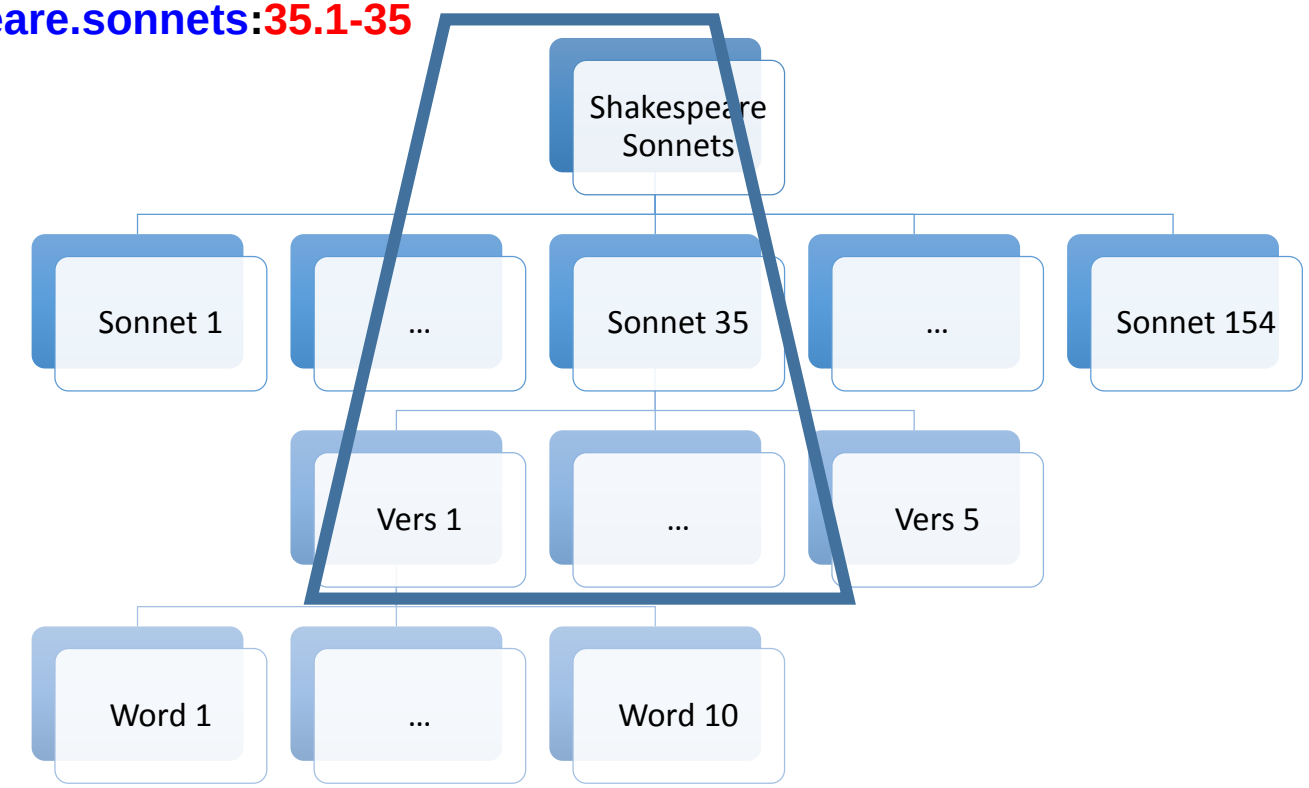
Canonical Citation

urn:cts:demo:shakespeare.sonnets:35



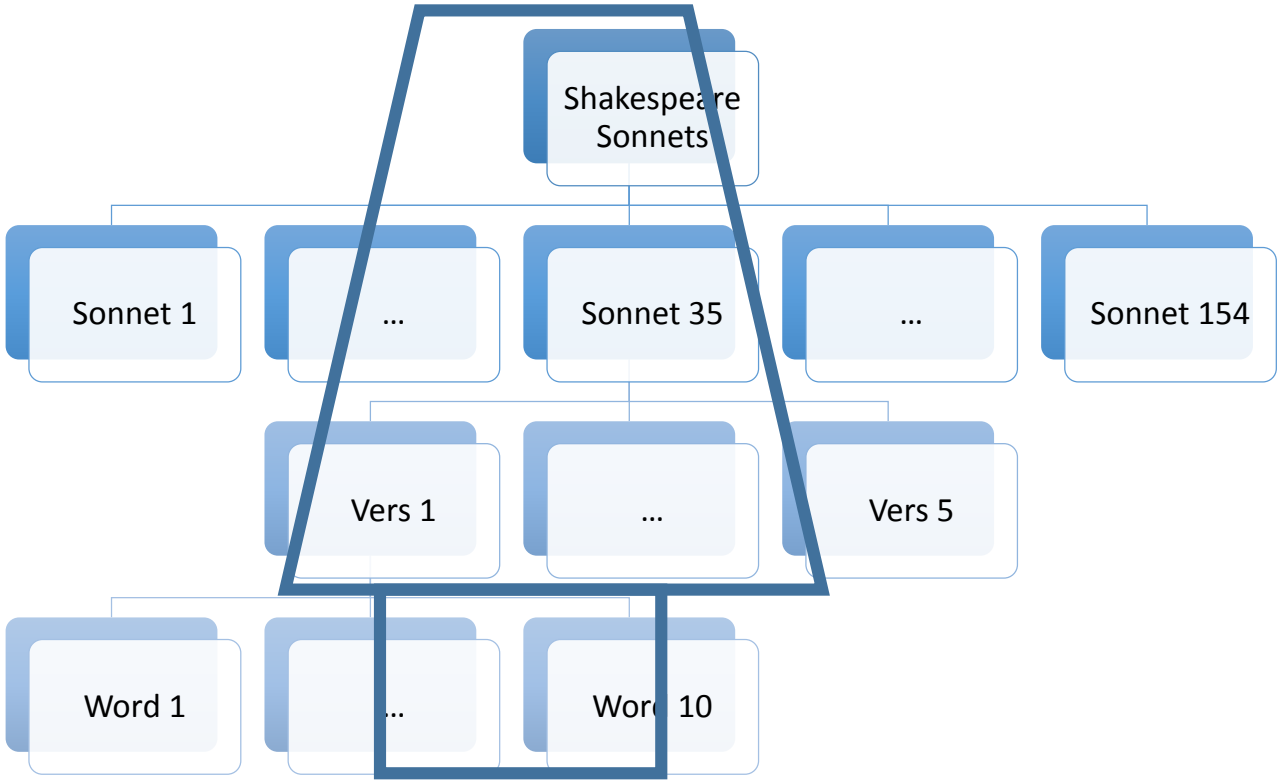
Canonical Citation

urn:cts:demo:shakespeare.sonnets:35.1-35.5
urn:cts:demo:shakespeare.sonnets:35.1-35



Canonical Citation

urn:cts:demo:shakespeare.sonnets:35.1@grieved-35.5@faults[1]



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:

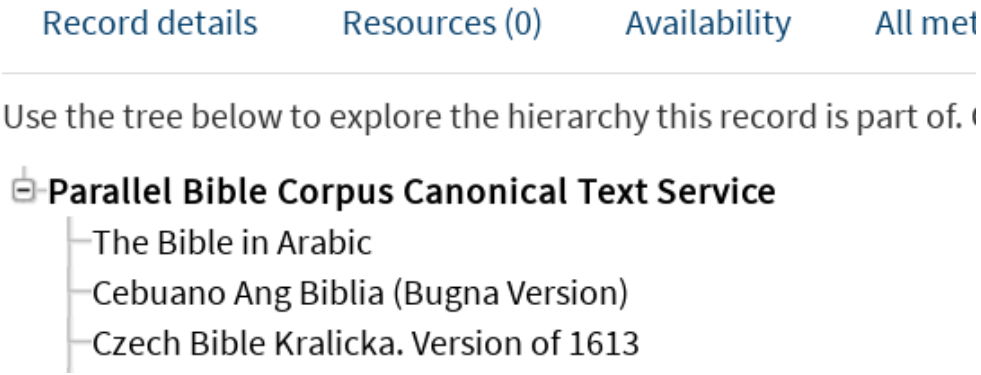
urn:cts:pbc:bible

urn:cts:pbc:bible.parallel.arb.norm:

urn:cts:pbc:bible.parallel.ceb.bugna:

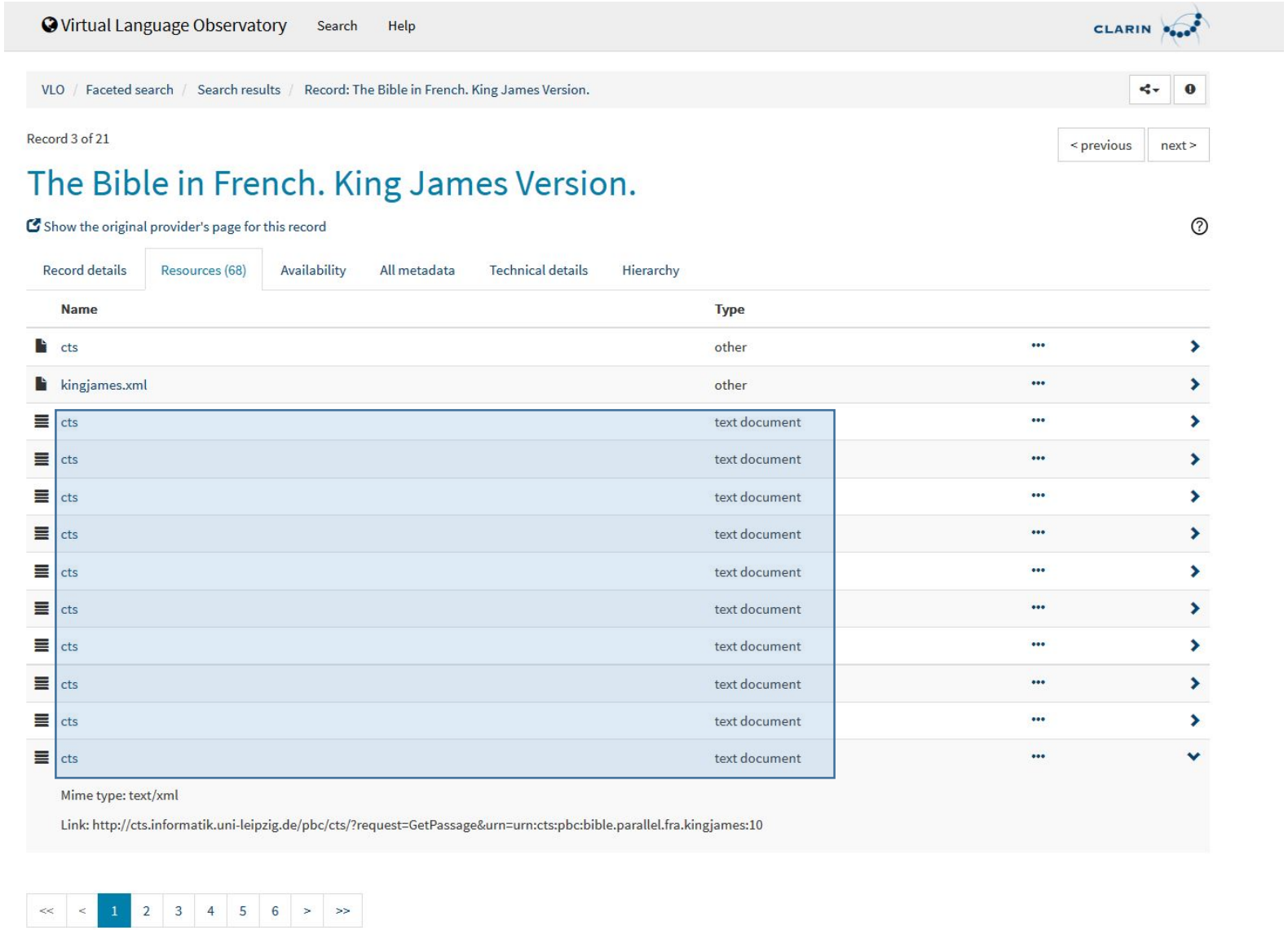
urn:cts:pbc:bible.parallel.ces.kralicka:

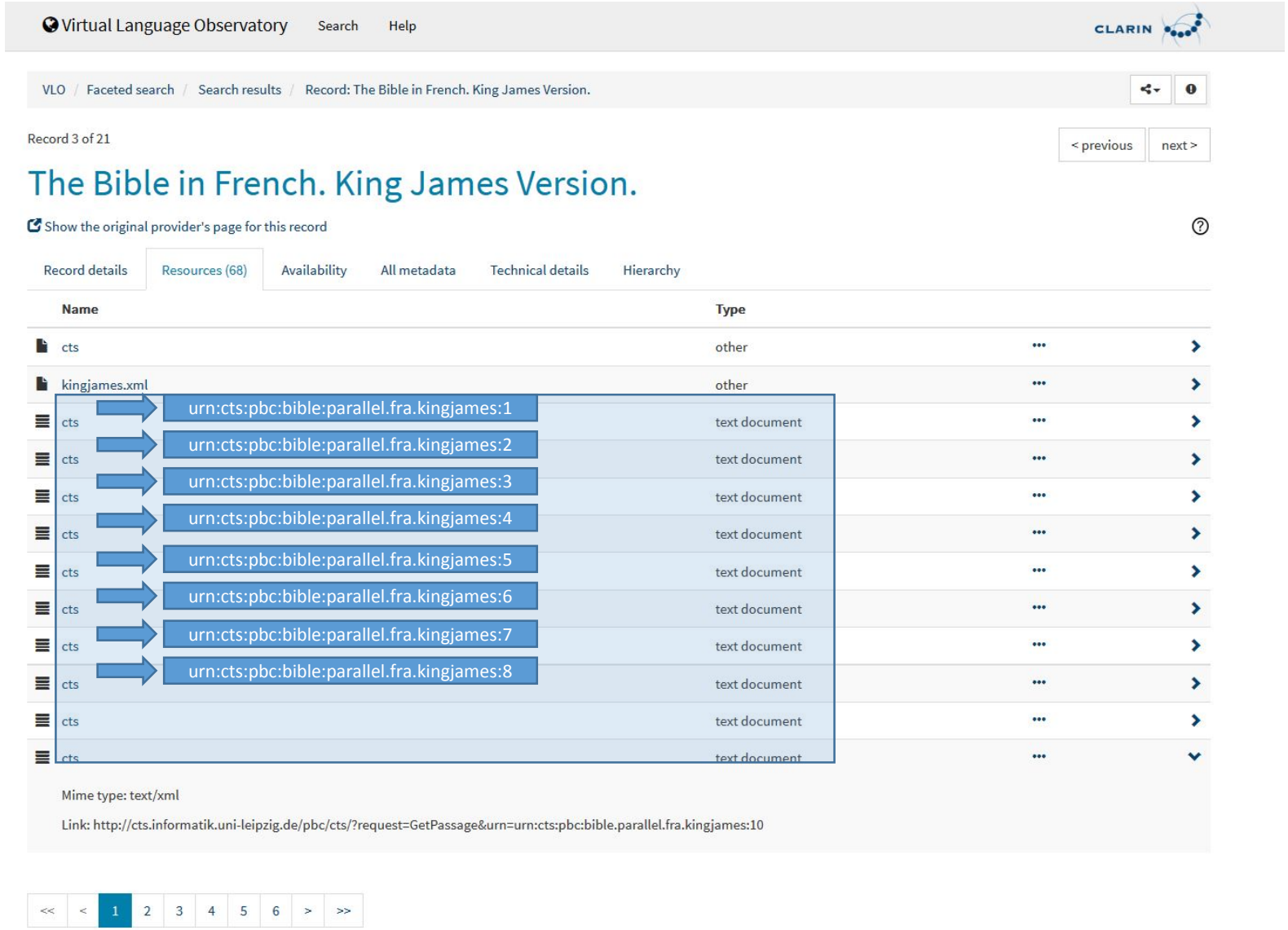
...



- Planed: FCS endpoint for content search based on existing fulltext index

http://aspra3.informatik.uni-leipzig.de:8080/vlo/?fq=collection:Canonical+Text+Services+NLP+Leipzig




Virtual Language Observatory
Search
Help
CLARIN
VLO / Faceted search / Search results / Record: The Bible in French. King James Version.
Record 3 of 21
< previous
next >
The Bible in French. King James Version.
Show the original provider's page for this record
Record details
Resources (68)
Availability
All metadata
Technical details
Hierarchy
Name
Type
cts
other
kingjames.xml
other
cts
text document
urn:cts:pbc:bible:parallel.fra.kingjames:1
urn:cts:pbc:bible:parallel.fra.kingjames:2
urn:cts:pbc:bible:parallel.fra.kingjames:3
urn:cts:pbc:bible:parallel.fra.kingjames:4
urn:cts:pbc:bible:parallel.fra.kingjames:5
urn:cts:pbc:bible:parallel.fra.kingjames:6
urn:cts:pbc:bible:parallel.fra.kingjames:7
urn:cts:pbc:bible:parallel.fra.kingjames:8
Mime type: text/xml
Link: http://cts.informatik.uni-leipzig.de/pbc/cts/?request=GetPassage&urn=urn:cts:pbc:bible.parallel.fra.kingjames:10
<< < 1 2 3 4 5 6 > >>

Why include support for CTS URNs?

CTS is developed by Humanists & reflects the

requirements from certain Digital Humanity

communites

- Perseus, Croatiae Auctores Latini, CITE, …

Creating CTS ready data is a research & project goal in

DH

- “pick researchers up, where they are“

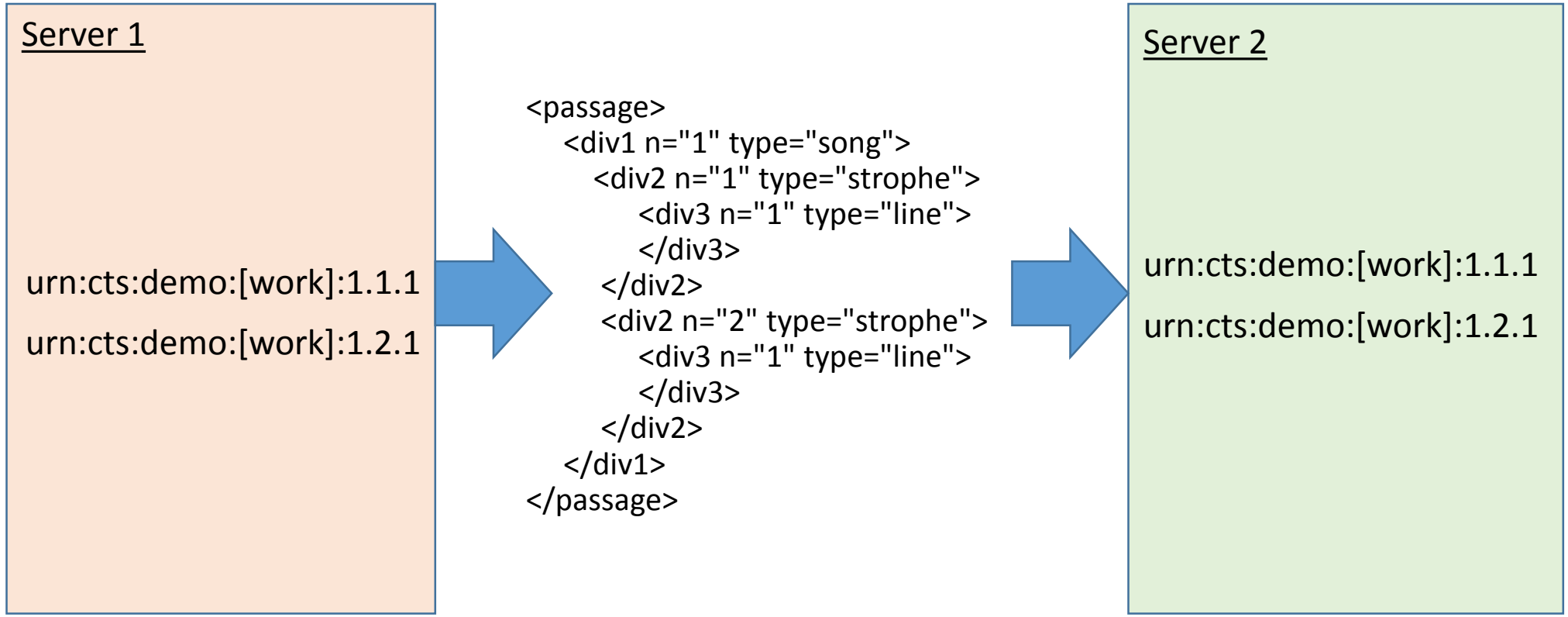
Technical benefits

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

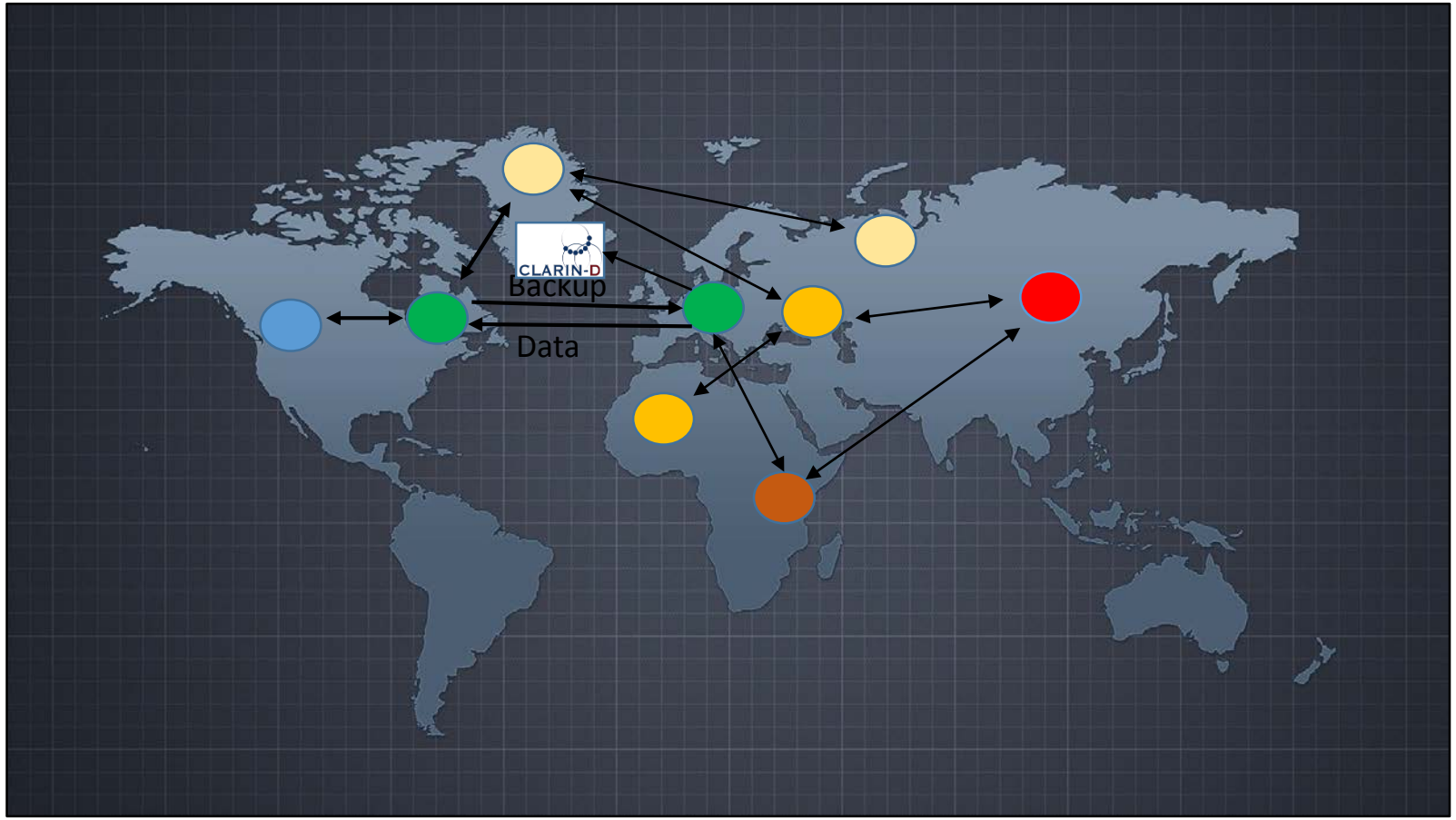
The Canonical Text Infrastructure

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

CTS Cloning

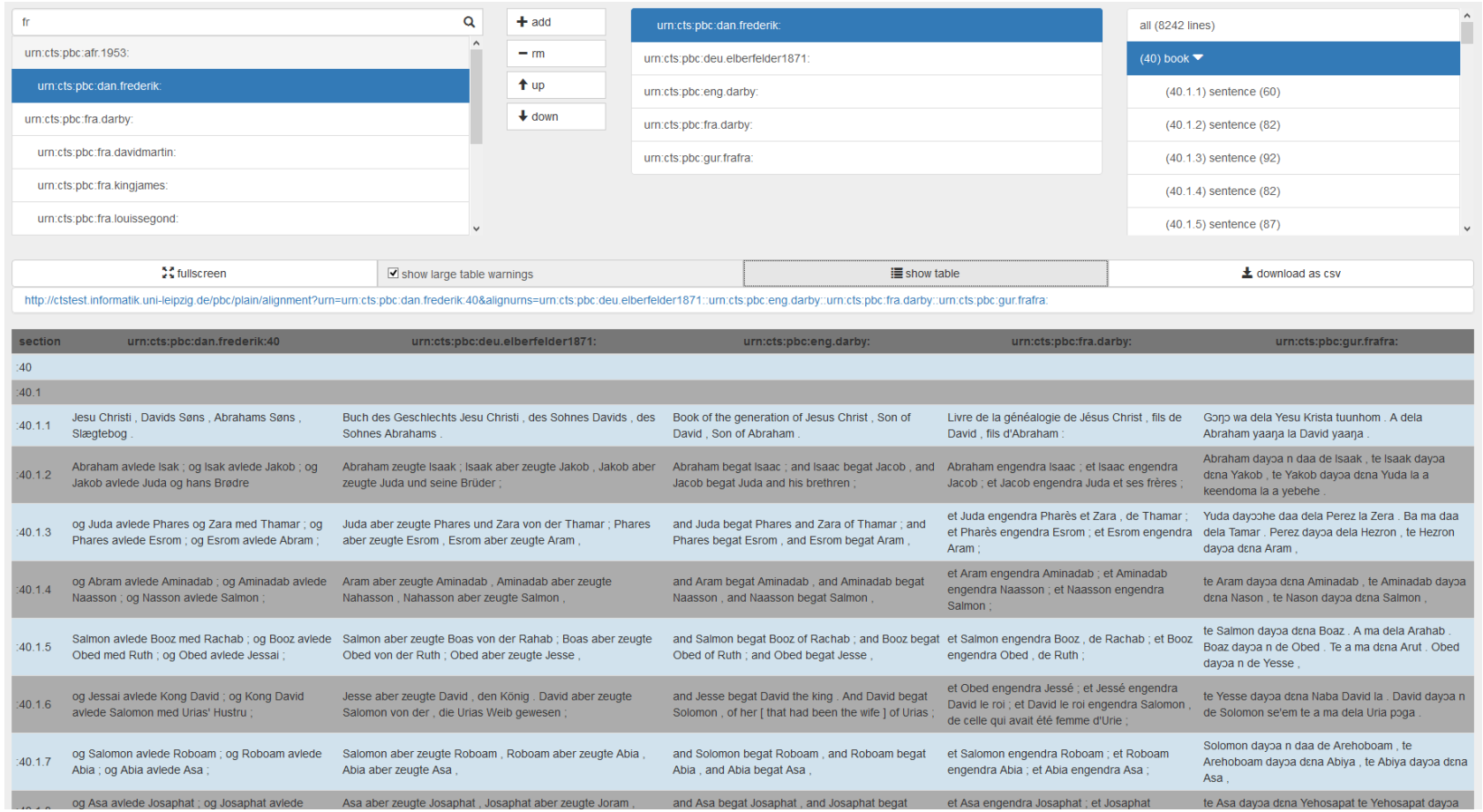

Server 1
urn:cts:demo:[work]:1.1.1
urn:cts:demo:[work]:1.2.1
<passage>
<div1 n="1" type="song">
<div2 n="1" type="strophe">
<div3 n="1" type="line">
</div3>
</div2>
<div2 n="2" type="strophe">
<div3 n="1" type="line">
</div3>
</div2>
</div1>
</passage>
Server 2
urn:cts:demo:[work]:1.1.1
urn:cts:demo:[work]:1.2.1

CTS Cloning



http://hdw.eweb4.com/out/1369880.html

Realtime Alignment Tools for CTS



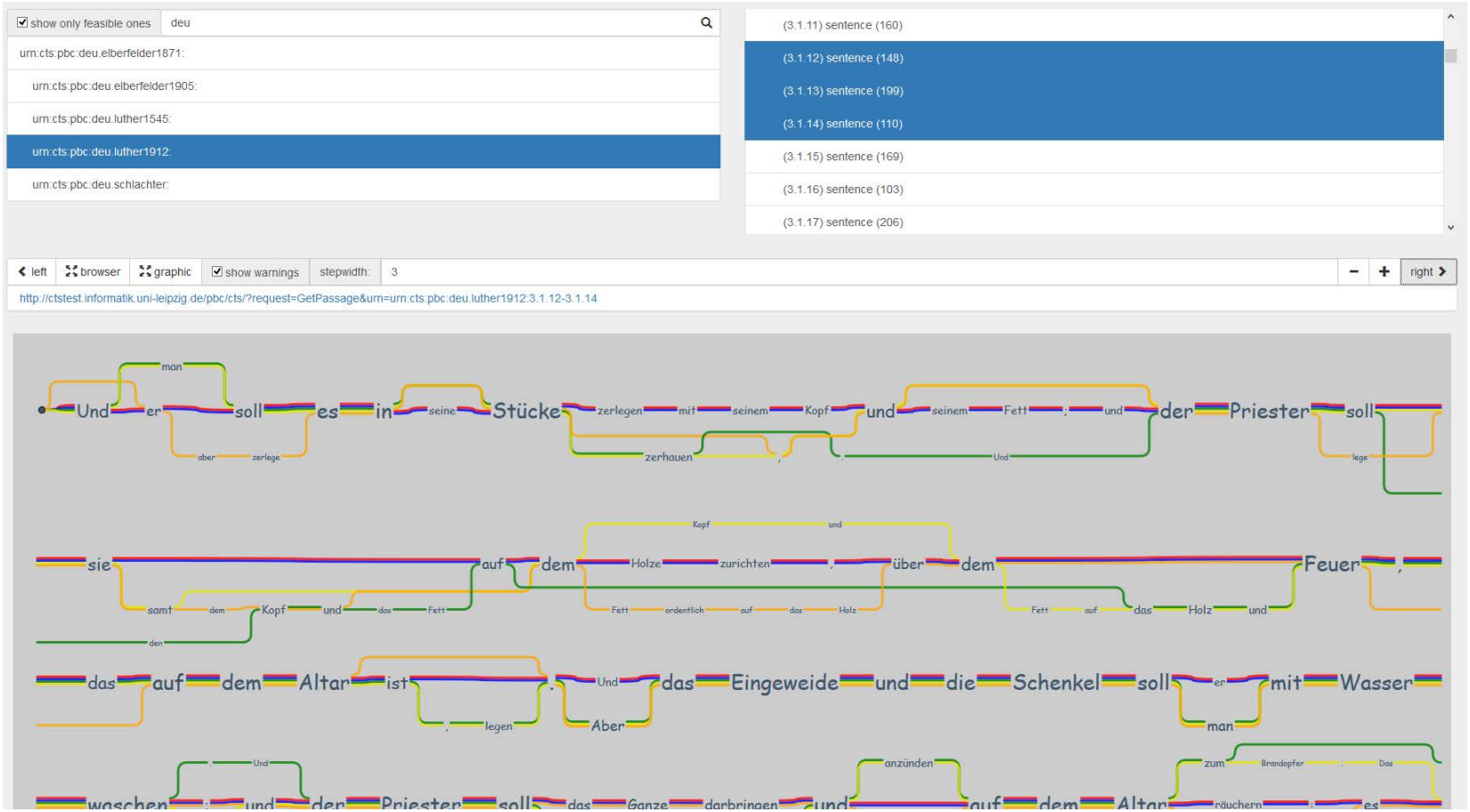
section	urn:cts:pbc:dan.frederik:40	urn:cts:pbc:deu.elberfelder1871:	urn:cts:pbc:eng.darby:	urn:cts:pbc:fra.darby:	urn:cts:pbc:gur.frafra:
:40					
:40.1					
:40.1.1	Jesu Christi , Davids Søns , Abrahams Søns , Slægtebog .	Buch des Geschlechts Jesu Christi , des Sohnes Davids , des Sohnes Abrahams .	Book of the generation of Jesus Christ , Son of David , Son of Abraham .	Livre de la généalogie de Jésus Christ , fils de David , fils d'Abraham :	Gɔŋɔ wa dela Yesu Krista tuunhom . A dela Abraham yaaŋa la David yaaŋa .
:40.1.2	Abraham avlede Isak ; og Isak avlede Jakob ; og Jakob avlede Juda og hans Brødre	Abraham zeugte Isaak ; Isaak aber zeugte Jakob , Jakob aber zeugte Juda und seine Brüder ;	Abraham begat Isaac ; and Isaac begat Jacob , and Jacob begat Juda and his brethren ;	Abraham engendra Isaac ; et Isaac engendra Jacob ; et Jacob engendra Juda et ses frères ;	Abraham dayɔa n daa de Isaak , te Isaak dayɔa dɛna Yakob , te Yakob dayɔa dɛna Yuda la a keendoma la a yebehe .
:40.1.3	og Juda avlede Phares og Zara med Thamar ; og Phares avlede Esrom ; og Esrom avlede Abram ;	Juda aber zeugte Phares und Zara von der Thamar ; Phares aber zeugte Esrom , Esrom aber zeugte Aram ,	and Juda begat Phares and Zara of Thamar ; and Phares begat Esrom , and Esrom begat Aram ,	et Juda engendra Pharès et Zara , de Thamar ; et Pharès engendra Esrom ; et Esrom engendra Aram ;	Yuda dayɔɔhe daa dela Perez la Zera . Ba ma daa dela Tamar . Perez dayɔa dela Hezron , te Hezron dayɔa dɛna Aram ,
:40.1.4	og Abram avlede Aminadab ; og Aminadab avlede Naasson ; og Nasson avlede Salmon ;	Aram aber zeugte Aminadab , Aminadab aber zeugte Nahasson , Nahasson aber zeugte Salmon ,	and Aram begat Aminadab , and Aminadab begat Naasson , and Naasson begat Salmon ,	et Aram engendra Aminadab ; et Aminadab engendra Naasson ; et Naasson engendra Salmon ;	te Aram dayɔa dɛna Aminadab , te Aminadab dayɔa dɛna Nason , te Nason dayɔa dɛna Salmon ,
:40.1.5	Salmon avlede Booz med Rachab ; og Booz avlede Obed med Ruth ; og Obed avlede Jessai ;	Salmon aber zeugte Boas von der Rahab ; Boas aber zeugte Obed von der Ruth ; Obed aber zeugte Jesse ,	and Salmon begat Booz of Rachab ; and Booz begat Obed of Ruth ; and Obed begat Jesse ,	et Salmon engendra Booz , de Rachab ; et Booz engendra Obed , de Ruth ;	te Salmon dayɔa dɛna Boaz . A ma dela Arahab . Boaz dayɔa n de Obed . Te a ma dɛna Arut . Obed dayɔa n de Yesse ,
:40.1.6	og Jessai avlede Kong David ; og Kong David avlede Salomon med Urias' Hustru ;	Jesse aber zeugte David , den König . David aber zeugte Salomon von der , die Urias Weib gewesen ;	and Jesse begat David the king . And David begat Solomon , of her [that had been the wife] of Urias ;	et Obed engendra Jessé ; et Jessé engendra David le roi ; et David le roi engendra Salomon , de celle qui avait été femme d'Urie ;	te Yesse dayɔa dɛna Naba David la . David dayɔa n de Solomon se'em te a ma dela Uria pɔga .
:40.1.7	og Salomon avlede Roboam ; og Roboam avlede Abia ; og Abia avlede Asa ;	Salomon aber zeugte Roboam , Roboam aber zeugte Abia , Abia aber zeugte Asa ,	and Solomon begat Roboam , and Roboam begat Abia , and Abia begat Asa ,	et Salomon engendra Roboam ; et Roboam engendra Abia ; et Abia engendra Asa ;	Solomon dayɔa n daa de Arehoboam , te Arehoboam dayɔa dɛna Abiya , te Abiya dayɔa dɛna Asa ,
	og Asa avlede Josaphat ; og Josaphat avlede	Asa aber zeugte Josaphat , Josaphat aber zeugte Joram ,	and Asa begat Josaphat , and Josaphat begat	et Asa engendra Josaphat ; et Josaphat	te Asa dayɔa dɛna Yehosapat te Yehosapat dayɔa

(GUI implemented by Sascha Ludwig)

Alignment based on document structure

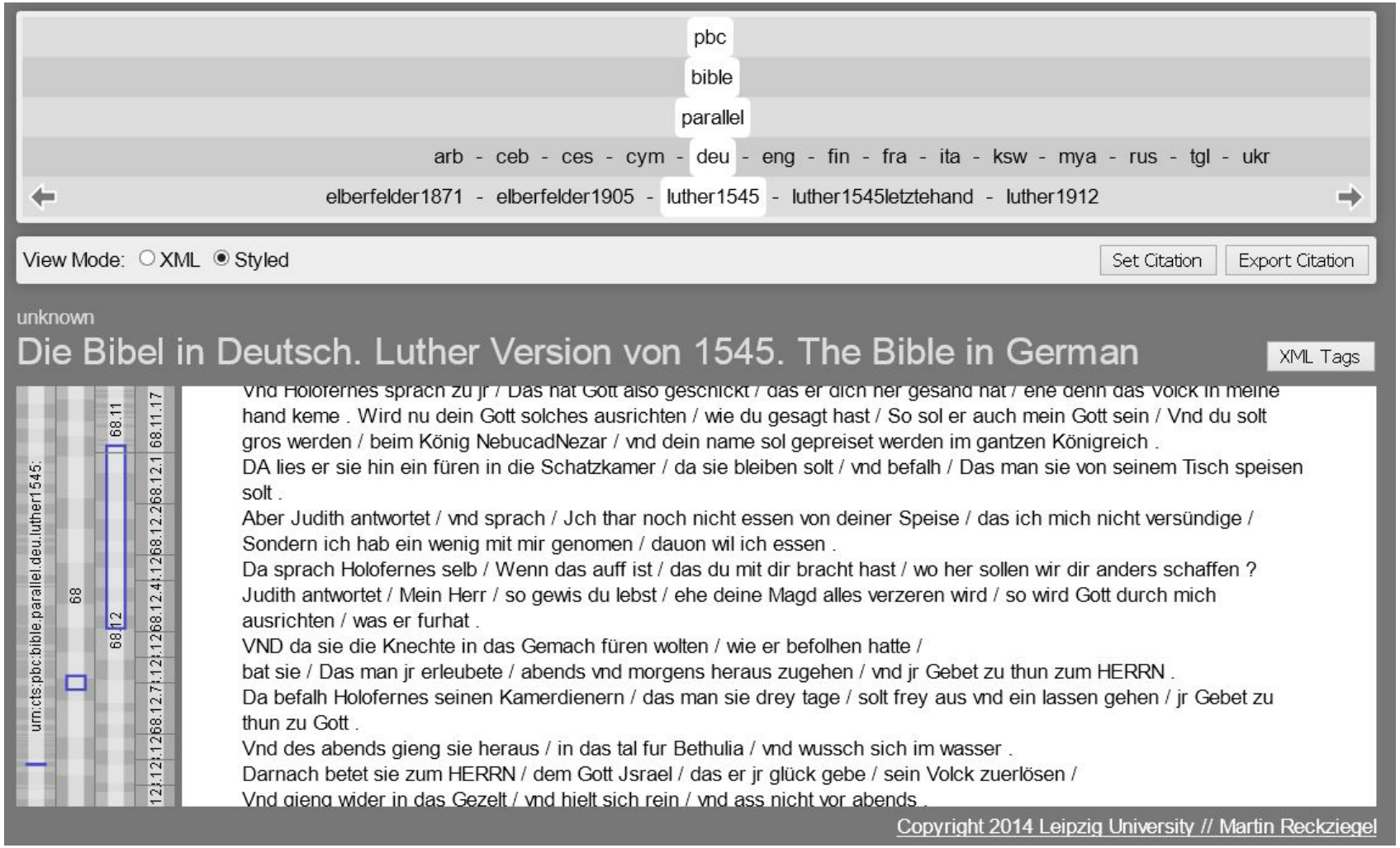
Scales very well (several complete bibles in a couple of seconds)

Realtime Alignment Tools for CTS



(GUI implemented by Sascha Ludwig)

Generic Reader



Reckziegel M., Jaenicke S. & Scheuermann G. 2016.
CTRaCE: Canonical Text Reader and Citation Exporter in Proceedings of the Digital Humanities, Krakow, 2016.

CTS-TM (CTS Text Miner)

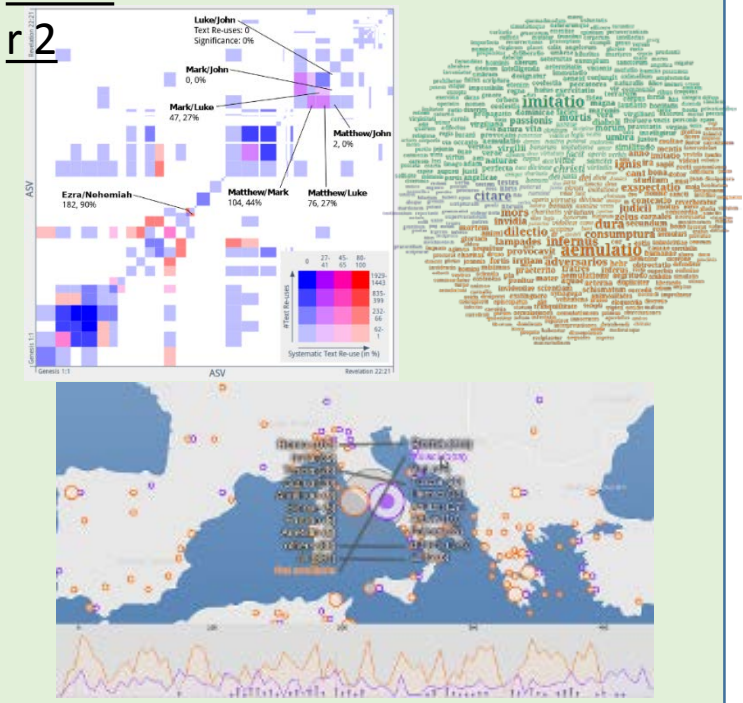
Server 1

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>
```

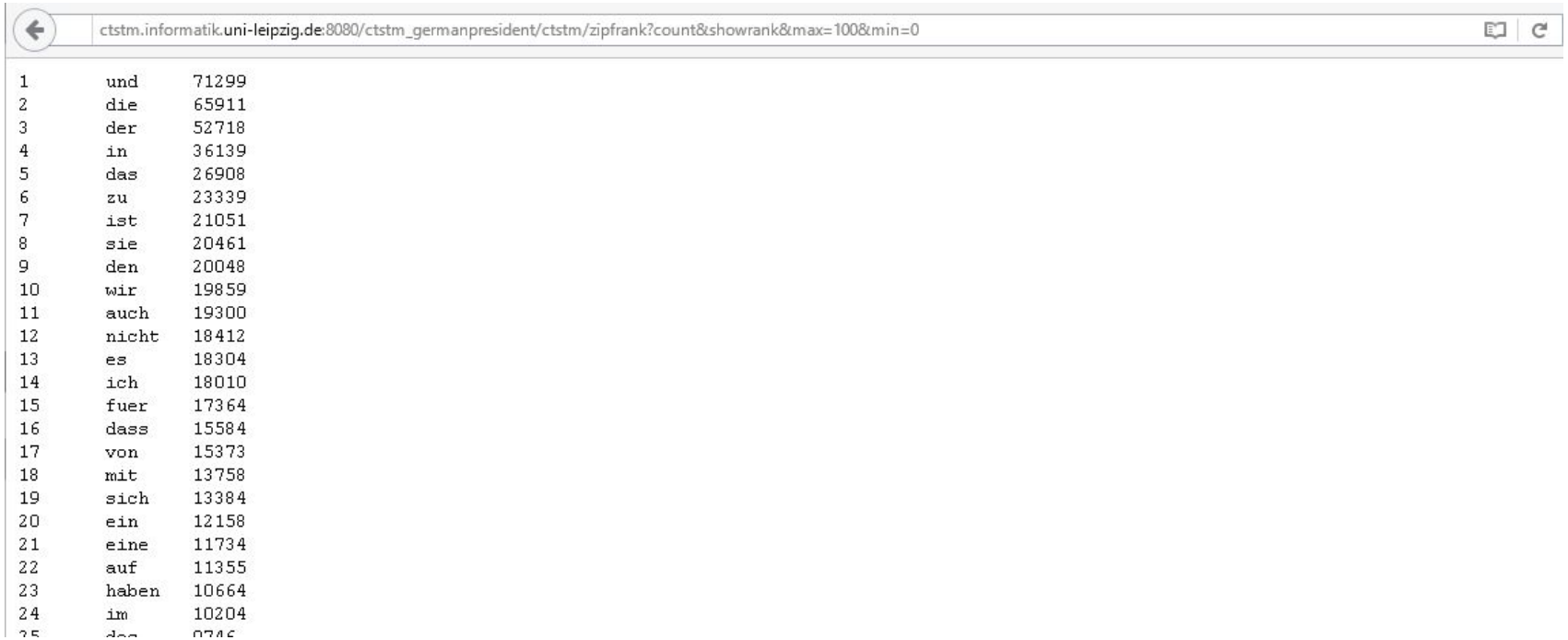
Server 2



(Example Visualizations from work of Stefan Jaenicke)

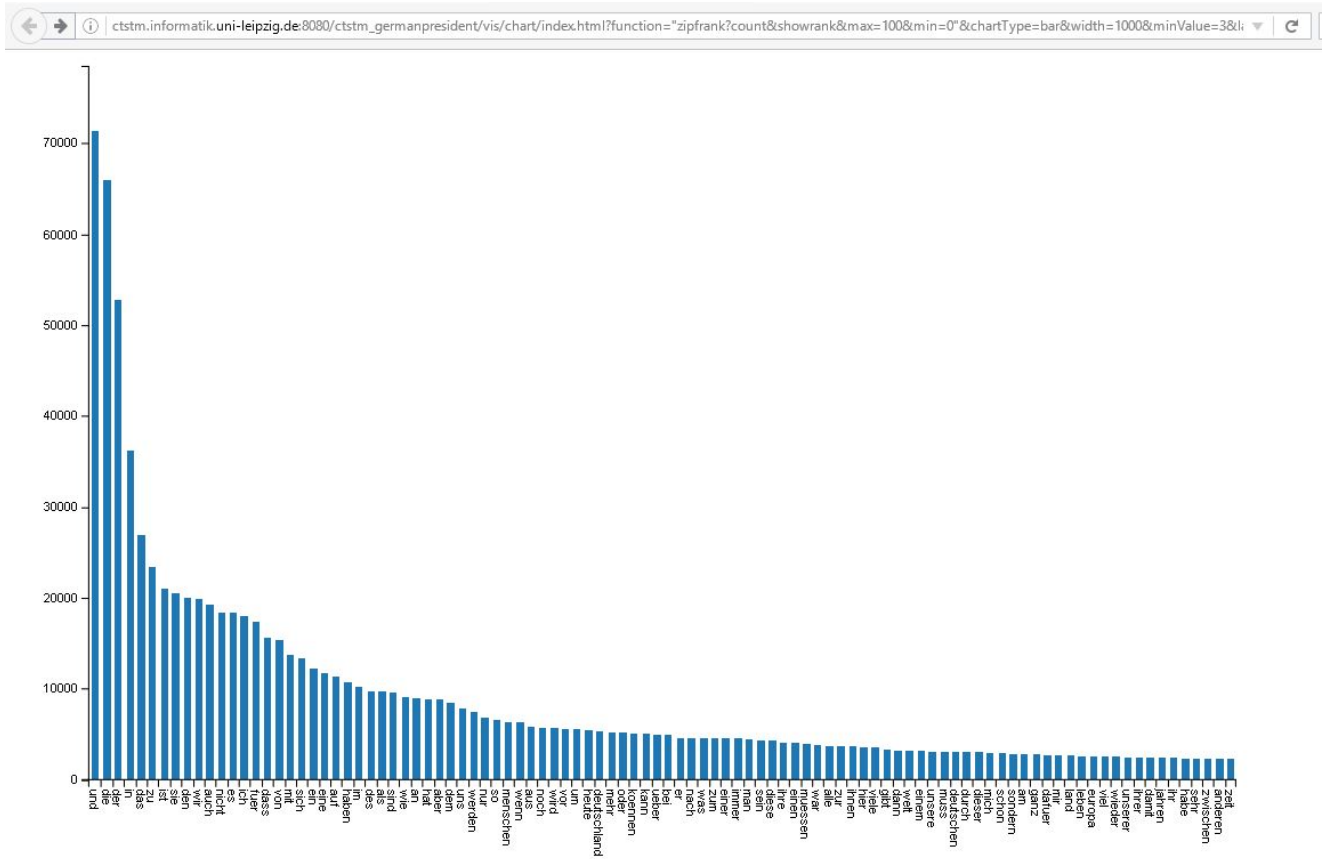
CTS-TM (CTS Text Miner)

- Raw Data as webservice



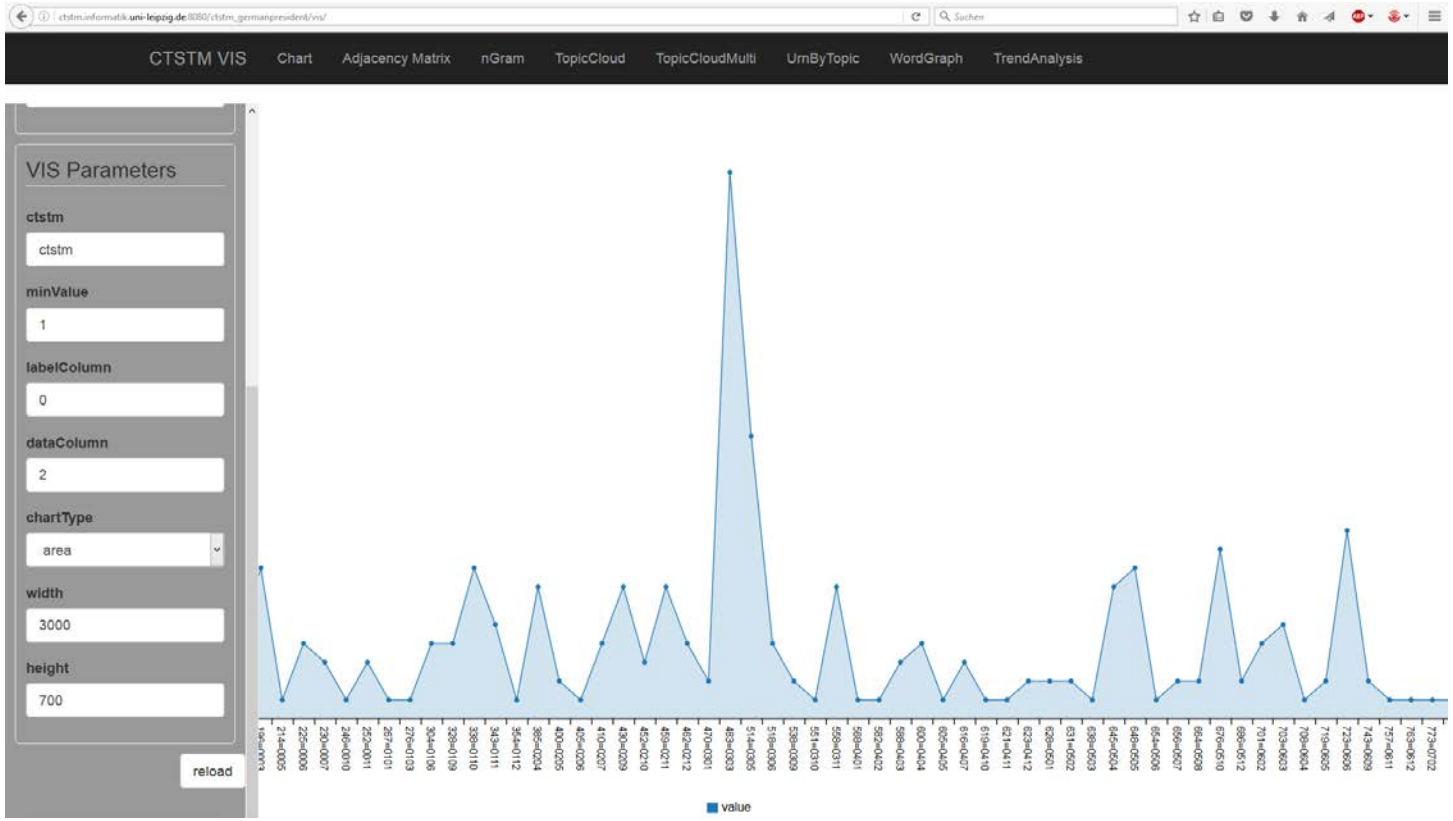
CTS-TM (CTS Text Miner)

- Generic Data Visualisations as webservice



CTS-TM (CTS Text Miner)

- Open Text Mining Tool as webservice



Datasets

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

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.

> (…), for [...] 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