

Canonical Text Services in CLARIN

Reaching out to the Digital Classics and beyond

Jochen Tiepmar, Thomas Eckart, Dirk Goldhahn and Christoph Kuras



UNIVERSITÄT LEIPZIG



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/ctsur/>
<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

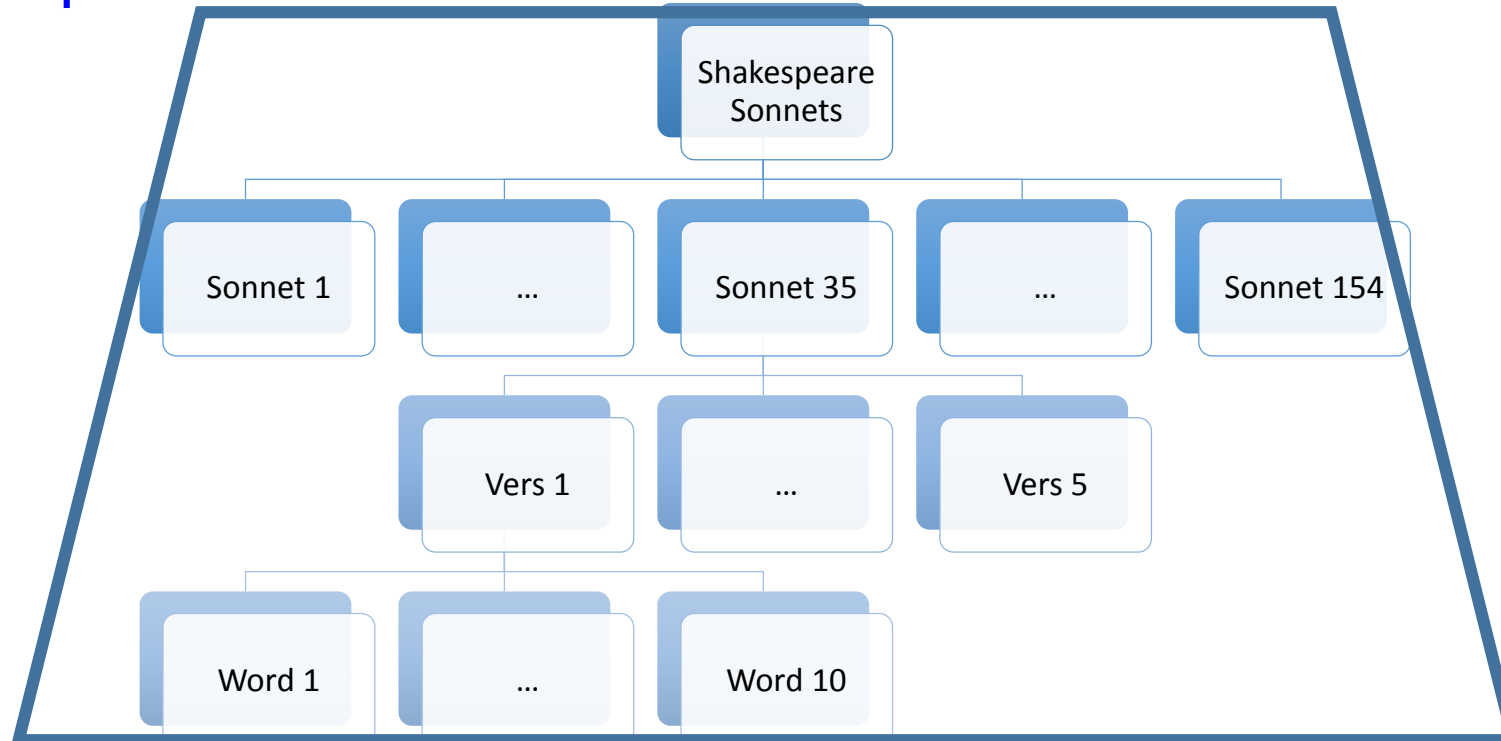
CTS-URN

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

Canonical Citation

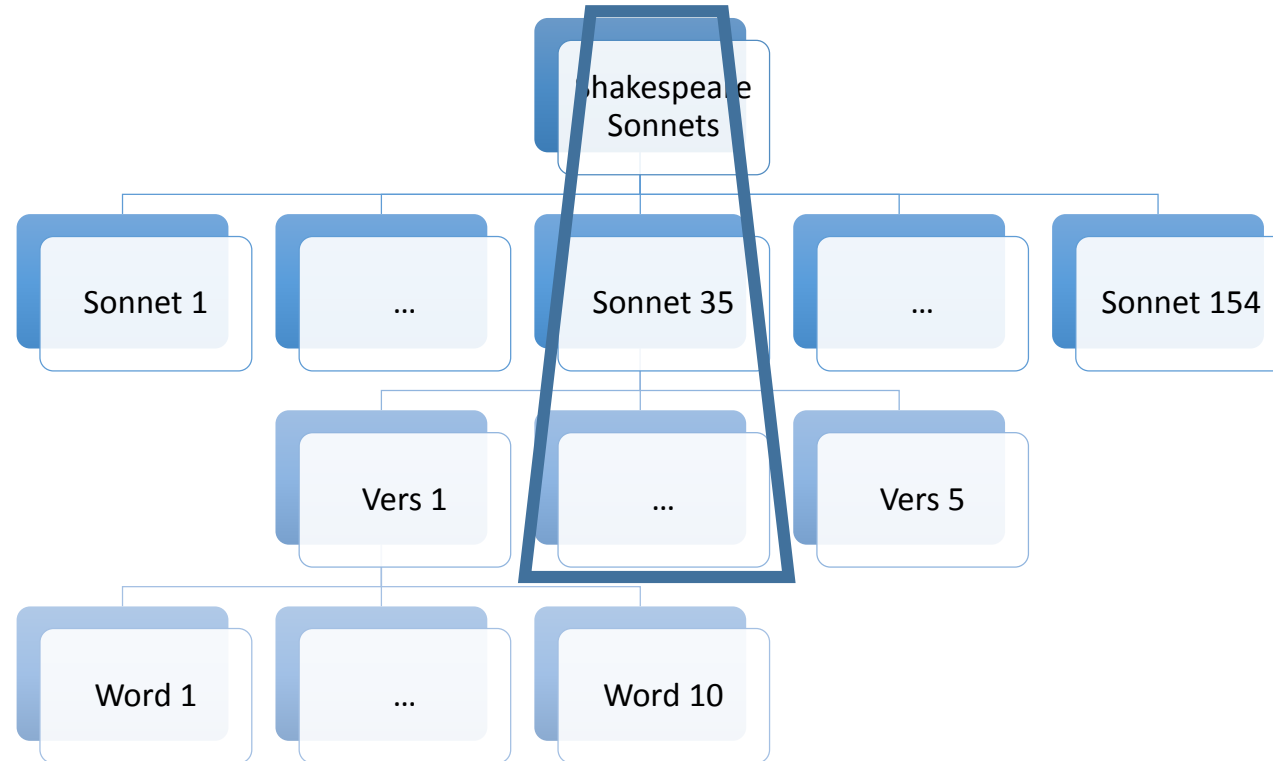
urn:cts:demo:shakespeare.sonnets:

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



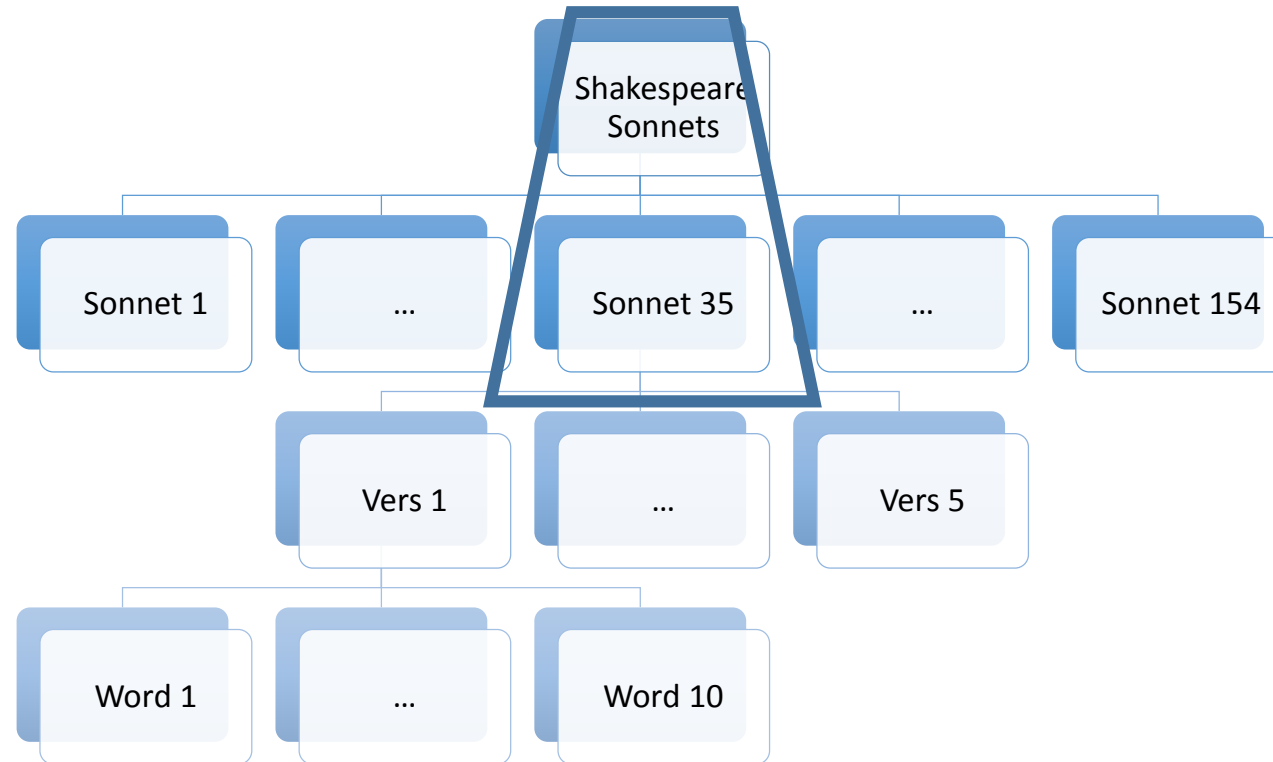
Canonical Citation

urn:cts:demo:shakespeare.sonnets:35.4



Canonical Citation

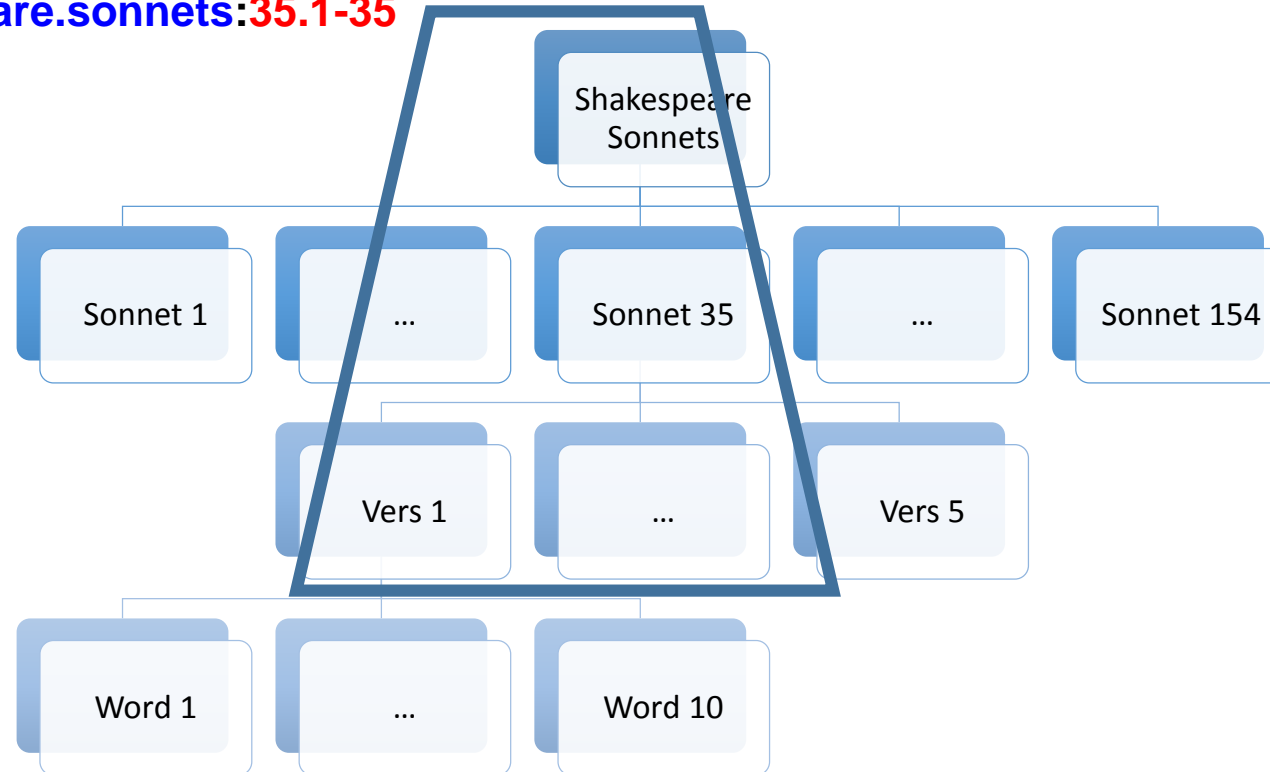
urn:cts:demo:shakespeare.sonnets:35



Canonical Citation

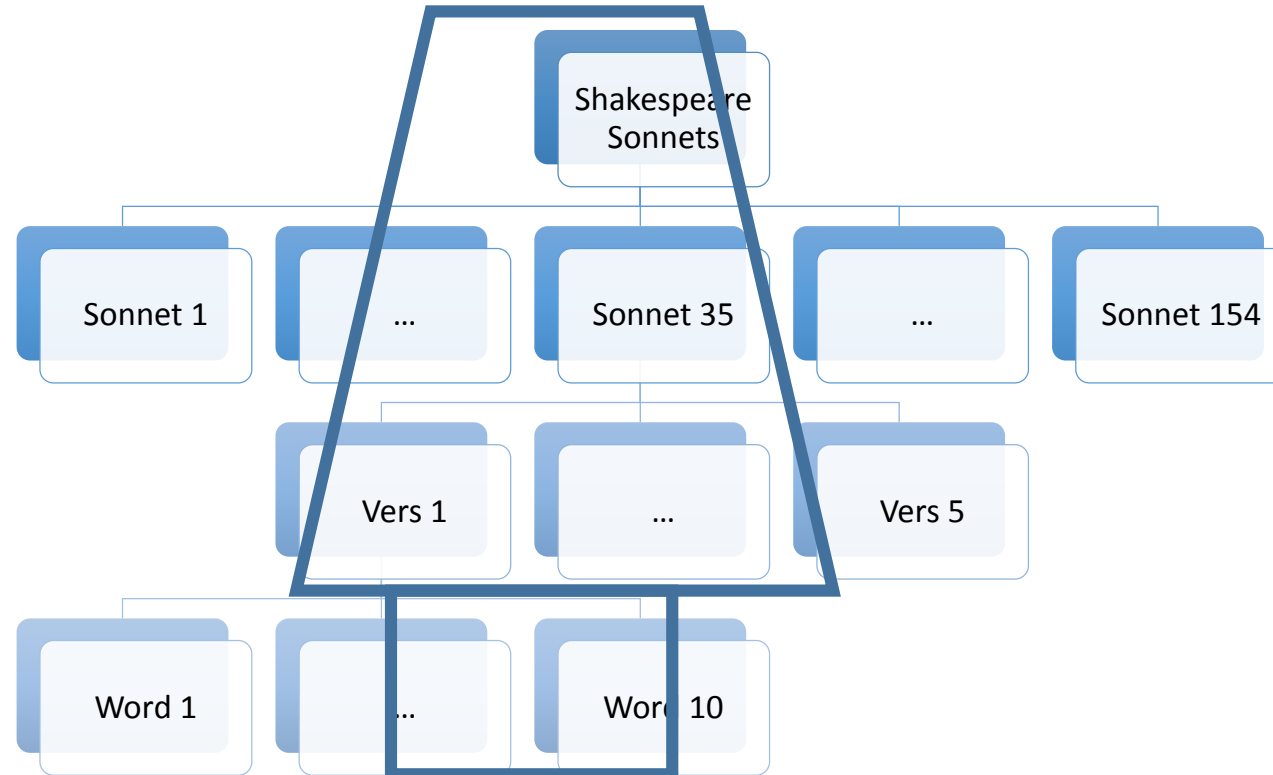
urn:cts:demo:shakespeare.sonnets:35.1-35.5

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:pbcbible

urn:cts:pbcbible.parallel.arb.norm:

urn:cts:pbcbible.parallel.ceb.bugna:

urn:cts:pbcbible.parallel.ces.kralicka:

...



[Record details](#) [Resources \(0\)](#) [Availability](#) [All met](#)

Use the tree below to explore the hierarchy this record is part of.

[-] **Parallel Bible Corpus Canonical Text Service**

├── The Bible in Arabic

├── Cebuano Ang Biblia (Bugna Version)

└── Czech Bible Kralicka. Version of 1613

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
















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	...	>
 cts	text document	...	>
 cts	text document	...	>
 cts	text document	...	>
 cts	text document	...	>
 cts	text document	...	>
 cts	text document	...	>
 cts	text document	...	>
 cts	text document	...	>
 cts	text document	...	>
 cts	text document	...	▼

Mime type: text/xml

Link: <http://cts.informatik.uni-leipzig.de/psc/cts/?request=GetPassage&urn=urn:cts:psc:bible.parallel.fra.kingjames:10>

<< < 1 2 3 4 5 6 > >>

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 → urn:cts:pbcbible:parallel.fra.kingjames:1	text document	...	>
cts → urn:cts:pbcbible:parallel.fra.kingjames:2	text document	...	>
cts → urn:cts:pbcbible:parallel.fra.kingjames:3	text document	...	>
cts → urn:cts:pbcbible:parallel.fra.kingjames:4	text document	...	>
cts → urn:cts:pbcbible:parallel.fra.kingjames:5	text document	...	>
cts → urn:cts:pbcbible:parallel.fra.kingjames:6	text document	...	>
cts → urn:cts:pbcbible:parallel.fra.kingjames:7	text document	...	>
cts → urn:cts:pbcbible:parallel.fra.kingjames:8	text document	...	>
cts	text document	...	>
cts	text document	...	>
cts	text document	...	>

Mime type: text/xml

Link: <http://cts.informatik.uni-leipzig.de/pbc/cts/?request=GetPassage&urn=urn:cts:pbcbible:parallel.fra.kingjames:10>

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“

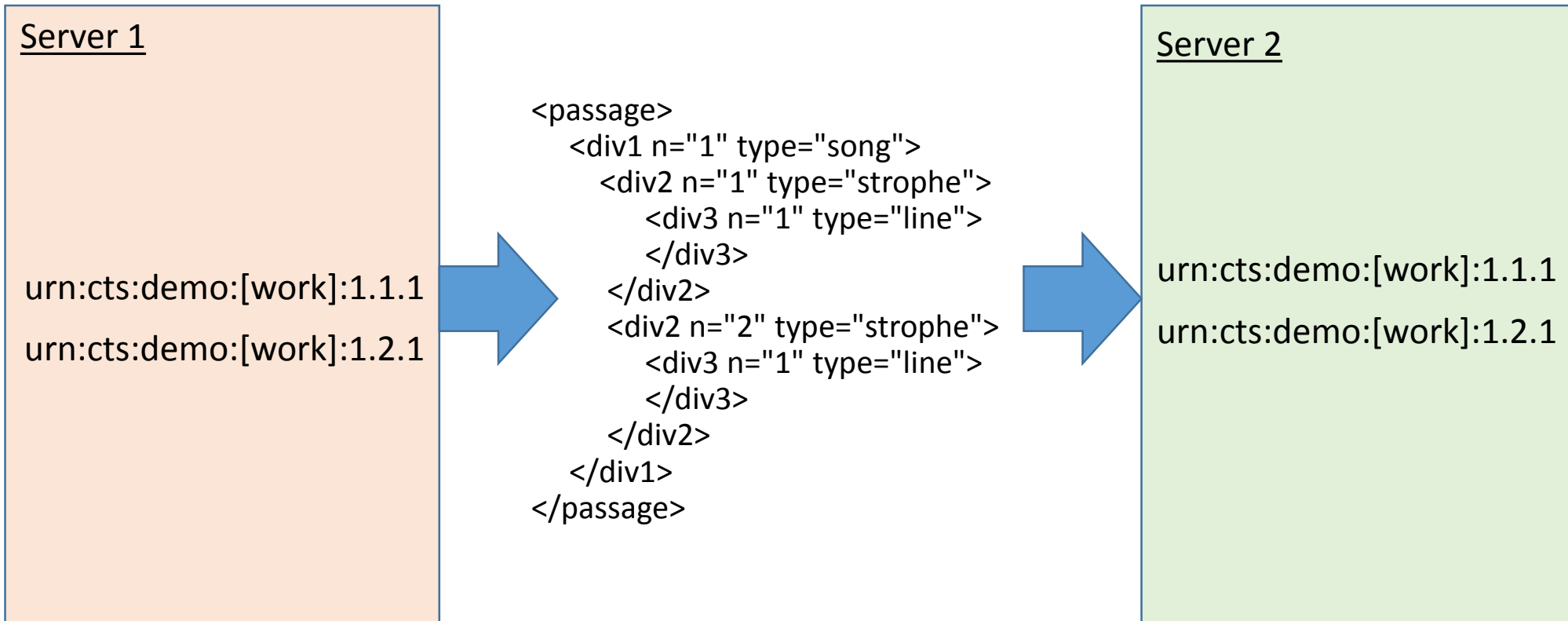
Technical benefits

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



CTS Cloning



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

Realtime Alignment Tools for CTS

The screenshot displays the Realtime Alignment Tools for CTS interface. At the top, there is a search bar containing 'fr'. Below it, a list of source documents is shown, with 'urn.cts.pbc.dan.frederik.' selected. To the right, a list of target documents is shown, with 'urn.cts.pbc.deu.elberfelder1871:' selected. Below these lists are controls for adding, removing, and moving documents. The main area shows a detailed alignment table for the first 17 sections of the Bible, comparing the source document (urn.cts.pbc.dan.frederik:40) with the target document (urn.cts.pbc.deu.elberfelder1871:). The table columns are labeled with the source and target document URNs. The rows show the alignment of text segments, with the source text on the left and the target text on the right. The alignment is based on document structure, and the table shows that the source text is aligned with the target text in a 1:1 ratio for the first 17 sections.

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 ;	Gorjo wa dela Yesu Krista tuunhom . A dela Abraham yaanja la David yaanja .
: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 dayoa n daa de Isaak , te Isaak dayoa dna Yakob , te Yakob dayoa dna 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 dayoche daa dela Perez la Zera . Ba ma daa dela Tamar , Perez dayoa dela Hezron , te Hezron dayoa dna Aram ,
:40.1.4	og Abram avlede Aminadab ; og Aminadab avlede Naasson ; og Naasson 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 dayoa dna Aminadab , te Aminadab dayoa dna Nason , te Nason dayoa dna Salmon ,
:40.1.5	Salmon avlede Booz med Rachab ; og Booz avlede Obed med Ruth ; og Obed avlede Jessal ;	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 dayoa dna Boaz . A ma dela Arahab . Boaz dayoa n de Obed . Te a ma dna Arut . Obed dayoa n de Yesse ,
:40.1.6	og Jessal 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 dayoa dna Naba David la . David dayoa n de Solomon se'em te a ma dela Uria poga .
: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 dayoa n daa de Arehoboam , te Arehoboam dayoa dna Abiya , te Abiya dayoa dna Asa ,
:40.1.8	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 dayoa dna Yehosapat te Yehosapat dayoa

(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

The screenshot displays a web-based interface for aligning Canonical Text Services (CTS). At the top left, there is a search bar with the text 'deu' and a search icon. Below it, a list of URNs is shown, with 'urn:cts:pb:deu.luther1912:' selected. To the right, a list of sentence IDs is displayed, with '(3.1.12) sentence (148)' highlighted in blue. Below these lists, there are navigation controls including 'left', 'browser', 'graphic', 'show warnings', and 'stepwidth: 3'. The main area shows a complex alignment diagram between two text segments. The top segment is: 'Und er soll es in seine Stücke zerlegen mit seinem Kopf und seinem Fett, und der Priester soll'. The bottom segment is: 'sie auf dem Holze zurichten über dem Feuer, samt dem Kopf und das Fett Fett ordentlich auf das Holz und Fett auf das Holz und'. Colored lines (red, blue, green, orange) connect words and phrases between the two segments, illustrating the alignment. For example, 'Kopf' in the bottom segment aligns with 'Kopf' in the top segment. The diagram also shows how phrases like 'auf dem Holze' and 'über dem Feuer' are aligned with 'mit seinem Kopf' and 'und seinem Fett' respectively.

(GUI implemented by Sascha Ludwig)

Generic Reader

The screenshot shows the Generic Reader interface. At the top, there are navigation buttons for 'pbc', 'bible', and 'parallel'. Below these are language codes: 'arb - ceb - ces - cym - deu - eng - fin - fra - ita - ksw - mya - rus - tgl - ukr'. A list of document identifiers is shown: 'elberfelder1871 - elberfelder1905 - luther1545 - luther1545letzehand - luther1912'. The 'View Mode' is set to 'Styled'. There are 'Set Citation' and 'Export Citation' buttons. The main content area displays the text 'Die Bibel in Deutsch. Luther Version von 1545. The Bible in German' with a 'XML Tags' button. The text is presented in a parallel view with a vertical scrollbar on the left. The text content is as follows:

Vnd Holofernes sprach zu jr / Das hat Gott also geschickt / das er dich her gesand hat / ehe denn das volck in meine hand keme . Wird nu dein Gott solches ausrichten / wie du gesagt hast / So sol er auch mein Gott sein / Vnd du solt gros werden / beim König NebucadNezar / vnd dein name sol gepreiset werden im gantzen Königreich .
DA lies er sie hin ein führen in die Schatzkamer / da sie bleiben solt / vnd befahl / Das man sie von seinem Tisch speisen solt .
Aber Judith antwortet / vnd sprach / Jch thar noch nicht essen von deiner Speise / das ich mich nicht versündige / Sondern ich hab ein wenig mit mir genomen / dauon wil ich essen .
Da sprach Holofernes selb / Wenn das auff ist / das du mit dir bracht hast / wo her sollen wir dir anders schaffen ?
Judith antwortet / Mein Herr / so gewis du lebst / ehe deine Magd alles verzeren wird / so wird Gott durch mich ausrichten / was er furhat .
VND da sie die Knechte in das Gemach führen wolten / wie er befohlen hatte / bat sie / Das man jr erleubete / abends vnd morgens heraus zugehen / vnd jr Gebet zu thun zum HERRN .
Da befahl Holofernes seinen Kamerdienern / das man sie drey tage / solt frey aus vnd ein lassen gehen / jr Gebet zu thun zu Gott .
Vnd des abends gieng sie heraus / in das tal fur Bethulia / vnd wussch sich im wasser .
Darnach betet sie zum HERRN / dem Gott Jsrael / das er jr glück gebe / sein Volck zuerlösen /
Vnd gieng wider in das Gezelt / vnd hielt sich rein / vnd ass nicht vor abends .

Copyright 2014 Leipzig University // Martin Reckziegel

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)

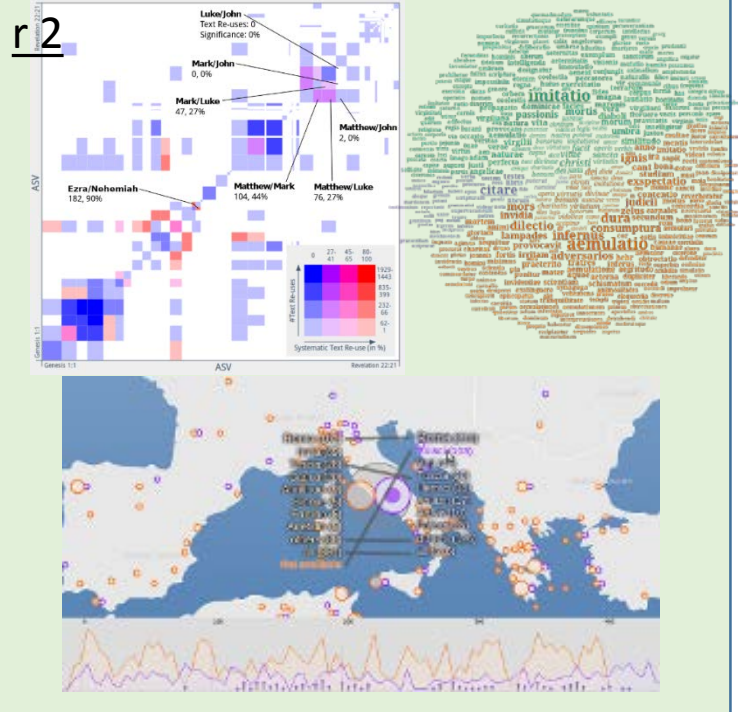
Server 1

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

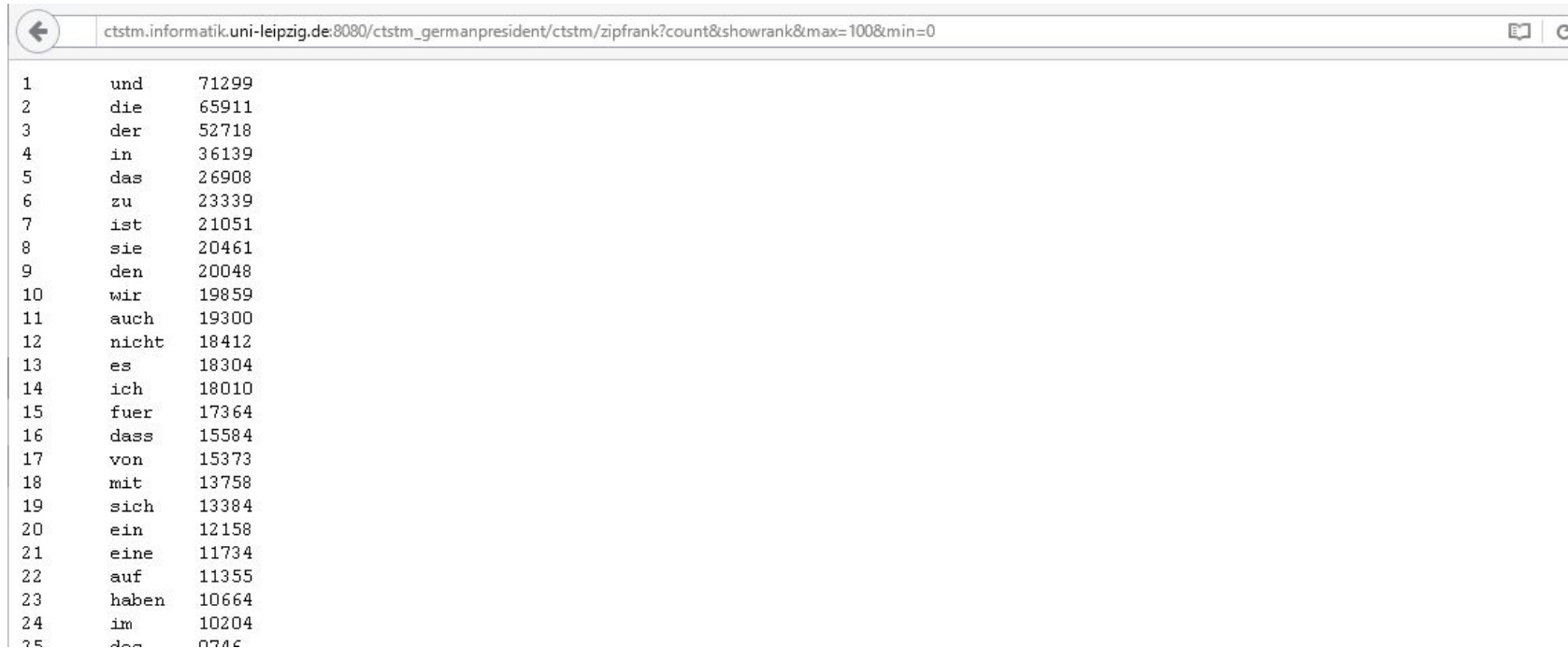
r2



(Example Visualizations from work of [Stefan Jaenicke](#))

CTS-TM (CTS Text Miner)

- Raw Data as webservice

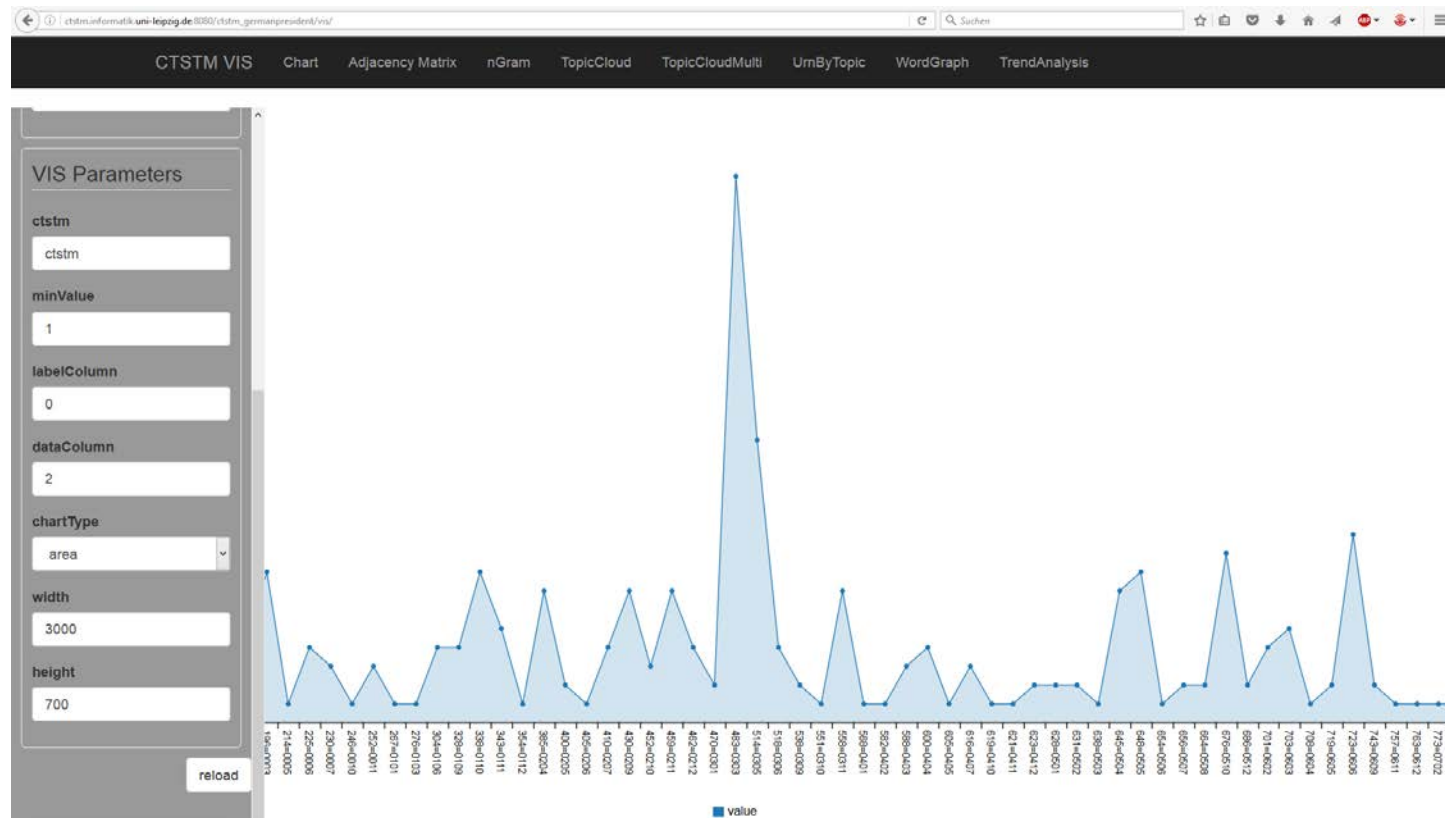


The screenshot shows a web browser window with the URL `ctstm.informatik.uni-leipzig.de:8080/ctstm_germanpresident/ctstm/zipfrank?count&showrank&max=100&min=0`. The browser displays a list of words and their frequencies, ordered by frequency in descending order. The list is as follows:

1	und	71299
2	die	65911
3	der	52718
4	in	36139
5	das	26908
6	zu	23339
7	ist	21051
8	sie	20461
9	den	20048
10	wir	19859
11	auch	19300
12	nicht	18412
13	es	18304
14	ich	18010
15	fuer	17364
16	dass	15584
17	von	15373
18	mit	13758
19	sich	13384
20	ein	12158
21	eine	11734
22	auf	11355
23	haben	10664
24	im	10204
25

CTS-TM (CTS Text Miner)

- Open Text Mining Tool as webservice



Canonical Text Services in CLARIN (2016)

Datasets

CTS instance	Tokens	Description
DTA, Deutsches Text Archiv	334'820'482	>1700 German works (literature, scholarly, ...) 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



UNIVERSITÄT 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

