UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

ASV Automatische Sprachverarbeitung

European Summer School 2017

Text Mining with Canonical Text Services

Theory Session 1 – Canonical Text Services

Federal Ministry of Education and Research

CLARIN-D

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

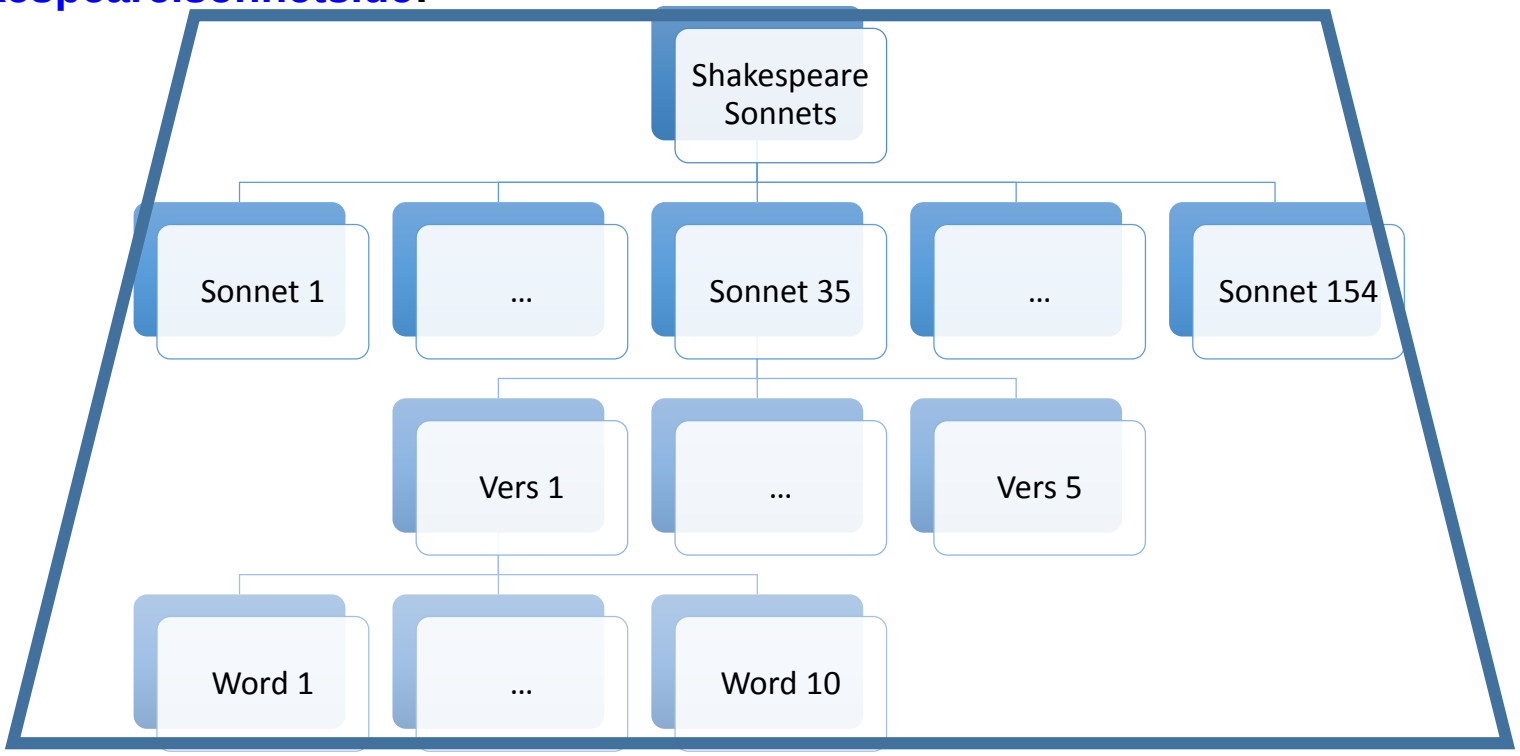
Shakespeare → Sonnets → english → 1st edition → Sonnet 1→ Vers 1

CTS-URN

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

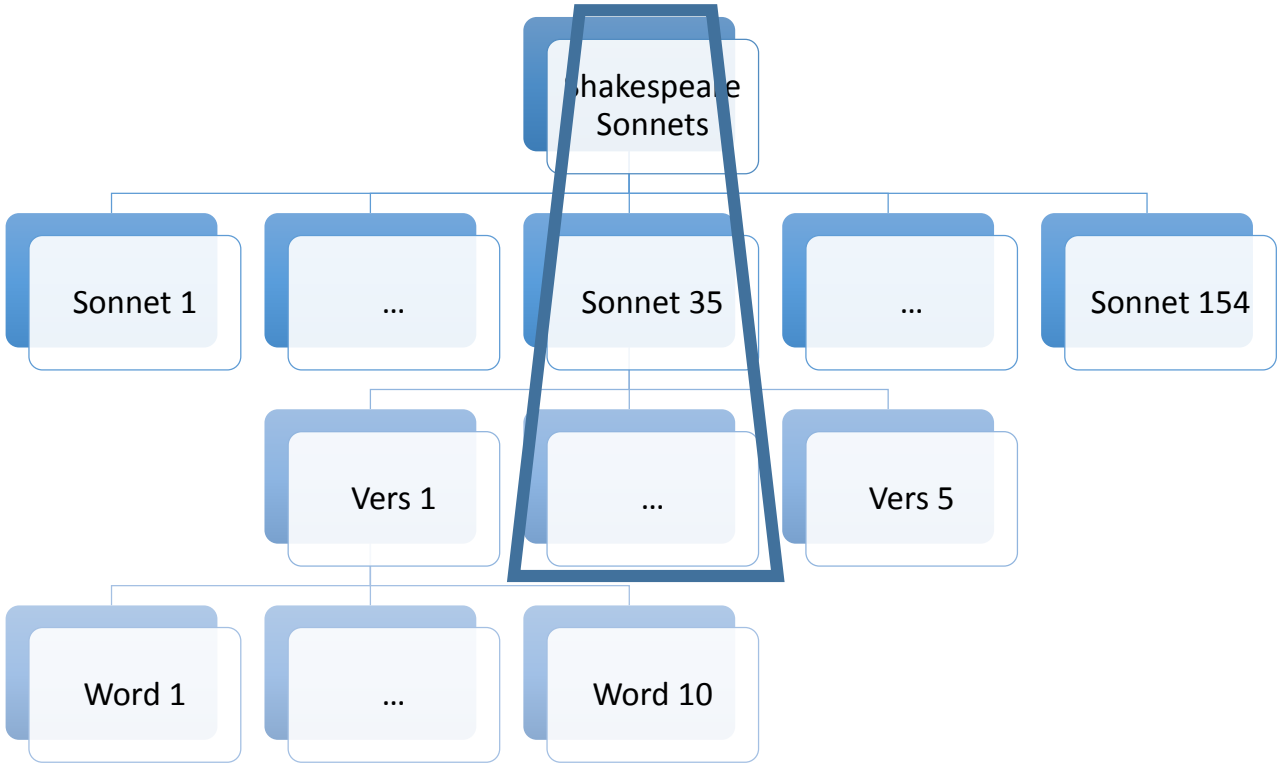
Canonical Citation

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



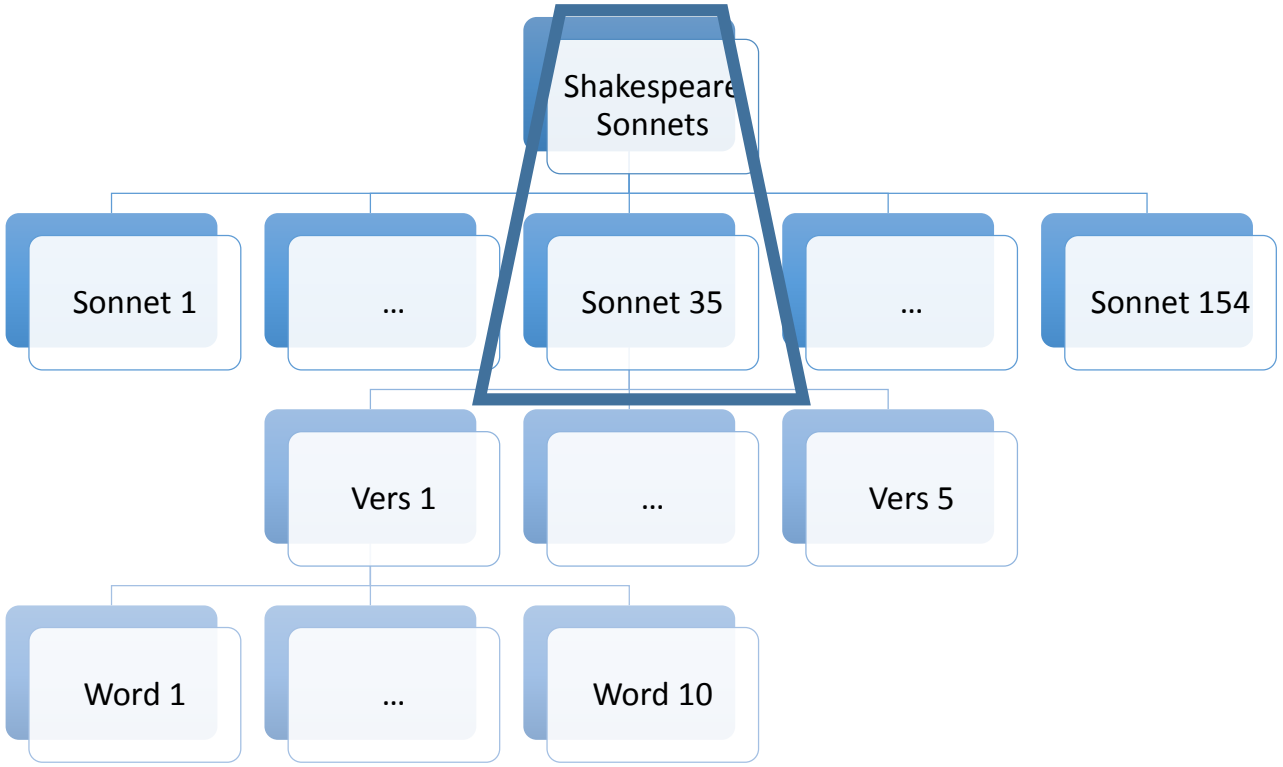
Canonical Citation

urn:cts:demo:shakespeare.sonnets:35.4



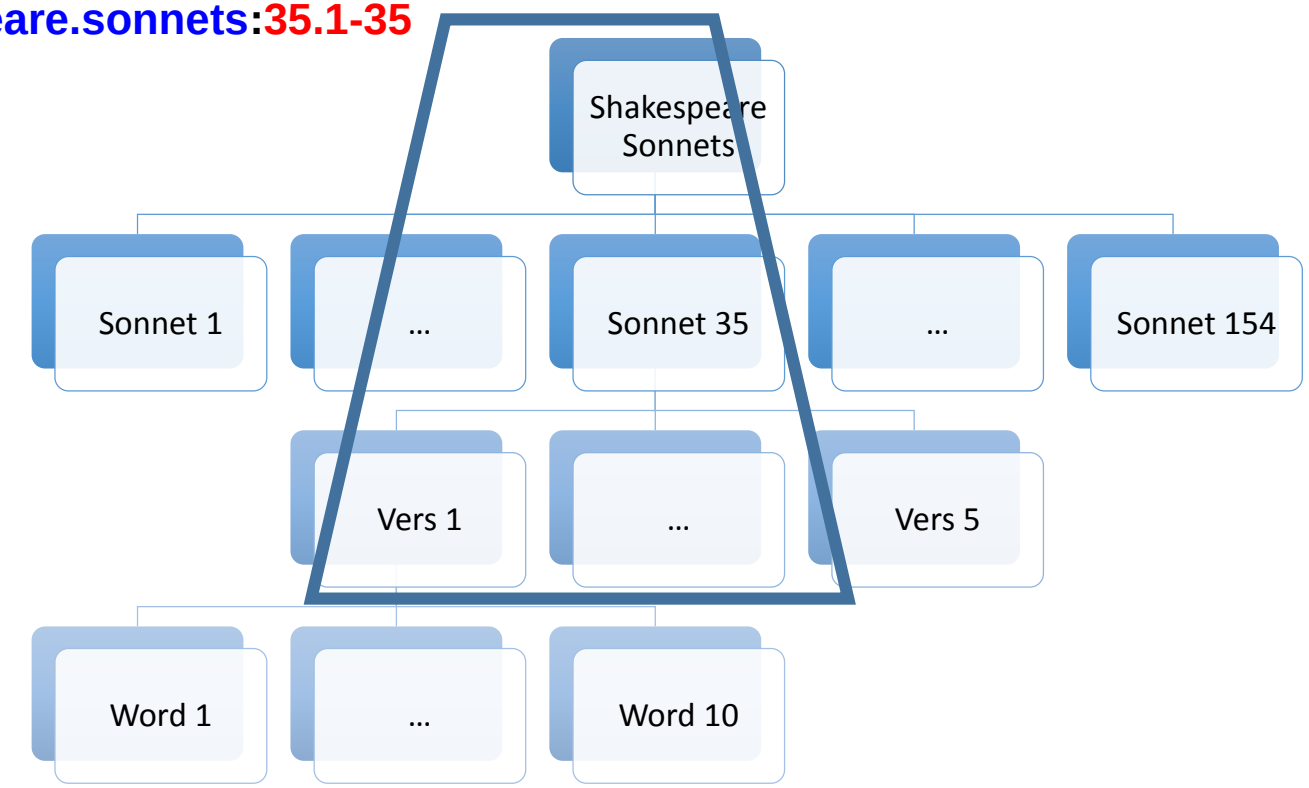
Canonical Citation

urn:cts:demo:shakespeare.sonnets:35



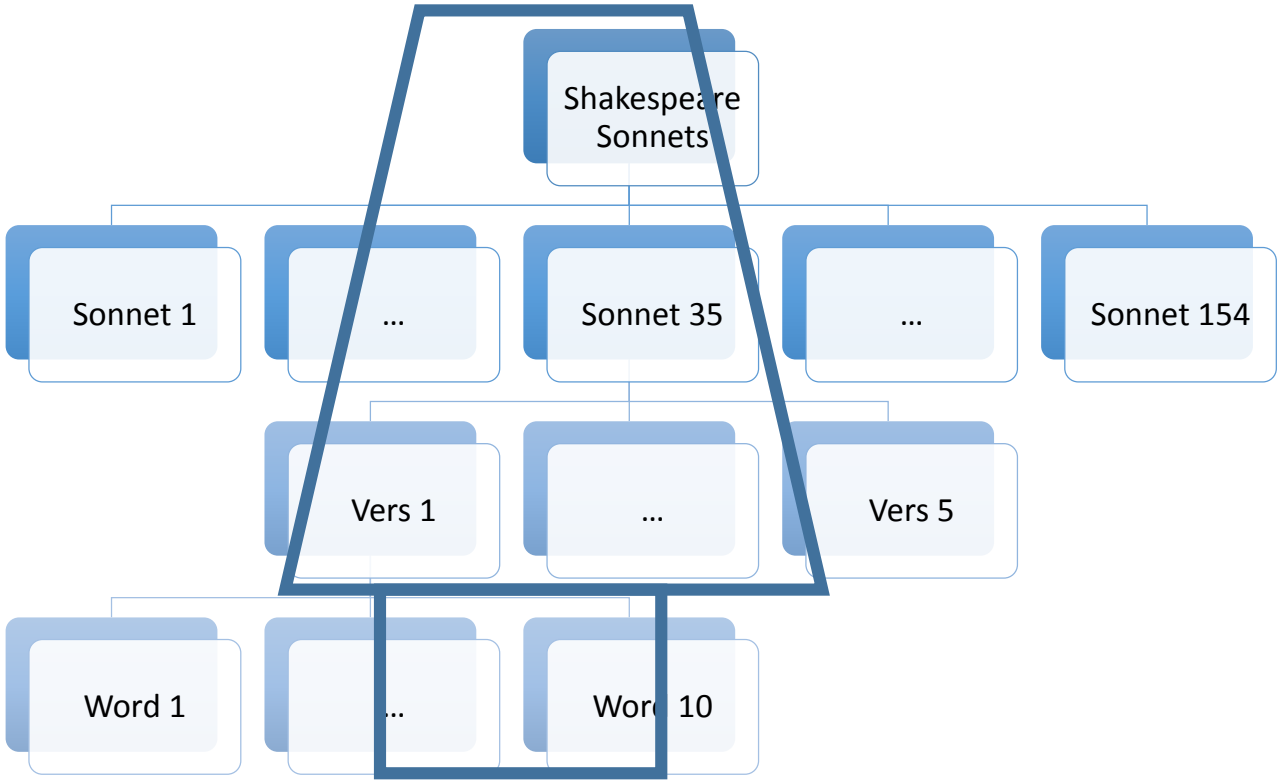
Canonical Citation

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]



CTS URNs

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

urn:cts:[CTSNAMESPACE]:[WORK]:[PASSAGE]

[CTSNAMESPACE] = Namespace (of the data set)

[WORK] = TEXTGROUP.WORK.VERSION.EXEMPLAR

[PASSAGE] = Text passage

Numbers are Letters

-> Numbers do not necessarily reflect document order

CTS URNs / Properties

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

35.5@vaults[1] == 35.5@vaults

urn:cts:, namespace and textgroup required, rest optional

Case sensitive(latinlit vs. LatinLit)

Descriptive parts can contain any character

Korrekt:**urn:cts:demo:shakespeare.sonnets12.2:preface.line1**

chapter2.line4 kann vorkommen vor **chapter2.line1**

Functional characters may not be used in descriptive elements

. : - @ [ ] (complete list http://www.homermultitext.org/hmt-docs/specifications/ctsurn/)

Incorrect:**urn:cts:demo:sha[@kespeare.sonn-ets:35.1[@-**

Static vs Dynamic CTS URNs

Static URNs

Document

urn:cts:pbc:bible.parallel.eng:

urn:cts:pbc:bible.parallel.eng.kingjames:

Text part

urn:cts:pbc:bible.parallel.eng:1

urn:cts:pbc:bible.parallel.eng.kingjames:1.3.2

Dynamic URNs

Text span (From one text part to another)

urn:cts:pbc:bible.parallel.eng:1.2-1.5.6

Sub passage notation

urn:cts:pbc:bible.parallel.eng:1.2@the[2]-1.5.6@five

Functions

GetCapabilities ( )

GetValidReff (urn, level)

GetFirstUrn (urn)

GetPrevNextUrn (urn)

GetLabel (urn)

GetPassage (urn)

GetPassagePlus (urn)

Functions

GetCapabilities ( )

Textinventory

GetValidReff (urn, level)

GetFirstUrn (urn)

GetPrevNextUrn (urn)

GetLabel (urn)

GetPassage (urn)

GetPassagePlus (urn)

Functions

GetCapabilities ( )

GetValidReff (urn, level)

Child URNs on [level] → All URNs in chapter 8

GetFirstUrn (urn)

GetPrevNextUrn (urn)

GetLabel (urn)

GetPassage (urn)

GetPassagePlus (urn)

```
<cts:GetValidReff>
 – <cts:request>
     <cts:requestName>GetValidReff</cts:requestName>
     <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1</cts:urn>
     <cts:level>4</cts:level>
   </cts:request>
 – <cts:reply>
   – <cts:reff>
       <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:</cts:urn>
       <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:1</cts:urn>
       <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:1.pr</cts:urn>
       <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:1.pr.1</cts:urn>
       <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:1.pr.2</cts:urn>
```

Functions

GetCapabilities ( )

GetValidReff (urn, level)

GetFirstUrn (urn)

First „ChildURN“ → 1st URN in chapter 8

GetPrevNextUrn (urn)

GetLabel (urn)

GetPassage (urn)

GetPassagePlus (urn)

```
<cts:GetFirstUrn>
 − <cts:request>
      <cts:requestName>GetFirstUrn</cts:requestName>
      <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:</cts:urn>
   </cts:request>
 − <cts:reply>
      <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:1</cts:urn>
   </cts:reply>
</cts:GetFirstUrn>
```

Functions

GetCapabilities ( )

GetValidReff (urn, level)

GetFirstUrn (urn)

GetPrevNextUrn (urn)

Left and right neighbor URNs → URNs left and right of the URN of line 8

GetLabel (urn)

GetPassage (urn)

GetPassagePlus (urn)

```
<cts:GetPrevNextUrn>
 – <cts:request>
     <cts:requestName>GetPrevNextUrn</cts:requestName>
     <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:1.2.1</cts:urn>
     <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:1.2.1</cts:urn>
   </cts:request>
 – <cts:reply>
   – <cts:prevnext>
     – <cts:prev>
         <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:1.2</cts:urn>
       </cts:prev>
     – <cts:next>
         <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:1.2.2</cts:urn>
       </cts:next>
     </cts:prevnext>
   </cts:reply>
</cts:GetPrevNextUrn>
```

Functions

GetCapabilities ( )

GetValidReff (urn, level)

GetFirstUrn (urn)

GetPrevNextUrn (urn)

GetLabel (urn)

Informal Description of [urn] → „Shakespears Sonnet 36 Line 9“

Get

Get


```
<GetLabel>
  <request>
    <requestName>GetLabel</requestName>
    <requestUrn>urn:cts:pbc:deu.luther1912:1.2.10-2.2.3</requestUrn>
  </request>
  <reply>
    <label>
      "Die Bibel in Deutsch. Luther von 1912. The Bible in German" from book "1", chapter "2", sentence "10" to book "2", chapter "2", sentence "3"
    </label>
  </reply>
</GetLabel>
```

Functions

GetCapabilities ( )

GetValidReff (urn, level)

GetFirstUrn (urn)

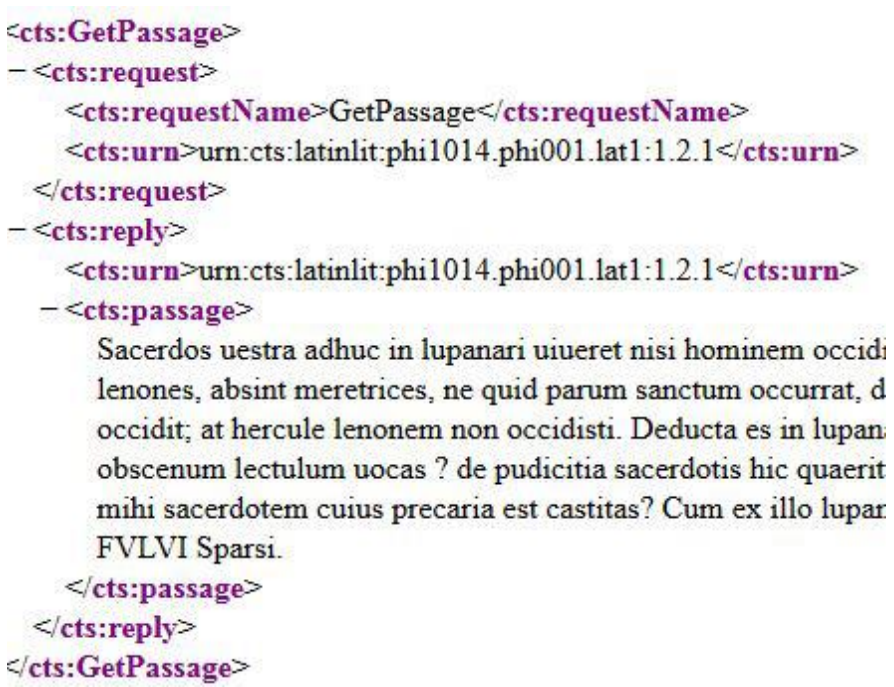
GetPrevNextUrn (urn)

GetLabel (urn)

GetPassage (urn)

Text passage for [urn]

GetPassagePlus (urn)



Functions

GetCapabilities ( )

GetValidReff (urn, level)

GetFirstUrn (urn)

GetPrevNextUrn (urn)

GetLabel (urn)

GetPassage (urn)

GetPassagePlus (urn)

Combined information without text inventory

```
<cts:GetPassagePlus>
 − <cts:request>
     <cts:requestName>GetPassagePlus</cts:requestName>
     <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:1.2.1</cts:urn>
   </cts:request>
 − <cts:reply>
     <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:1.2.1</cts:urn>
   − <cts:passage>
       Sacerdos uestra adhuc in lupanari uiueret nisi hominem occidisset.
       lenones, absint meretrices, ne quid parum sanctum occurrat, dum sa
       occidit; at hercule lenonem non occidisti. Deducta es in lupanar, ac
       obscenum lectulum uocas ? de pudicitia sacerdotis hic quaeritur. "N
       mihi sacerdotem cuius precaria est castitas? Cum ex illo lupanari cr
       FVLVI Sparsi.
     </cts:passage>
     <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:1.2.1</cts:urn>
     <cts:groupname>phi1014</cts:groupname>
     <cts:label>Edition: Controversiae</cts:label>
   − <cts:prev>
       <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:1.2</cts:urn>
     </cts:prev>
   − <cts:next>
       <cts:urn>urn:cts:latinlit:phi1014.phi001.lat1:1.2.2</cts:urn>
     </cts:next>
     <cts:validreff/>
   </cts:reply>
</cts:GetPassagePlus>
```

Requests

http://cts.informatik.uni-leipzig.de/pbc/cts/?request=GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1.3.5-1.3.8@the[2]

Requests

http://cts.informatik.uni-leipzig.de/pbc/cts/?request=GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1.3.5-1.3.8@the[2]

Server, Endpoint, Access Point

Requests

http://cts.informatik.uni-leipzig.de/pbc/cts/?request=GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1.3.5-1.3.8@the[2]

Function name

GetPassage, GetPassagePlus, GetLabel,
GetCapabilities, GetValidReff, GetFirstUrn,
GetPrevNextUrn

Requests

http://cts.informatik.uni-leipzig.de/pbc/cts/?request=GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1.3.5-1.3.8@the[2]

Function parameters

URN, (citation)level,…

Response

http://cts.informatik.uni-leipzig.de/pbc/cts/?request=GetPassage&urn=urn:cts:pbc:bible.parallel.eng.kingjames:1.3.5-1.3.8@the[2]


```
<GetPassage>
 − <request>
    <requestName>GetPassage</requestName>
   − <requestUrn>
      urn:cts:pbc:bible.parallel.eng.kingjames:1.3.5-1.3.8@the[2]
    </requestUrn>
  </request>
 − <reply>
   − <urn>
      urn:cts:pbc:bible.parallel.eng.kingjames:1.3.5-1.3.8@the[2]
    </urn>
   − <passage>
      For God doth know that in the day ye eat thereof , then your eyes shall be opened , and ye shall be as gods , knowing good and evil . And when
      the woman saw that the tree was good for food , and that it was pleasant to the eyes , and a tree to be desired to make one wise , she took of the
      fruit thereof , and did eat , and gave also unto her husband with her ; and he did eat . And the eyes of them both were opened , and they knew
      that they were naked ; and they sewed fig leaves together , and made themselves aprons . And they heard the
    </passage>
    <license>Public Domain</license>
   − <source>
      Retrieved via Canonical Text Service http://cts.informatik.uni-leipzig.de/pbc/cts/ with CTS URN
      urn:cts:pbc:bible.parallel.eng.kingjames:1.3.5-1.3.8@the[2]
    </source>
  </reply>
</GetPassage>
```

Example Requests

GetCapabilities	The text inventory with every CTS URN on document level and the corresponding meta information.
GetPassage	The text passage that is referenced by a given CTS URN.
GetLabel	An informal description for the text passage that is referenced by a given CTS URN. For example useful for tooltips.
GetValidReff	All the CTS URNs that "belong to" a given CTS URN on a given citation level. For example, every *sentence CTS URN* that "belongs to" a given *chapter CTS URN*.
GetPrevNextUrn	The previous and next static CTS URNs for a given CTS URN in document order.
GetFirstUrn	The first CTS URN that "belongs to" a given CTS URN. For example, the first *sentence CTS URN* that "belongs to" a given *chapter CTS URN*.
GetPassagePlus	The combined information for a given CTS URN.

Selected 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

Contact

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

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