UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

ASV Automatische Sprachverarbeitung

European Summer School 2017

Text Mining with Canonical Text Services

Theory Session 8 – Text Reuse

Federal Ministry of Education and Research

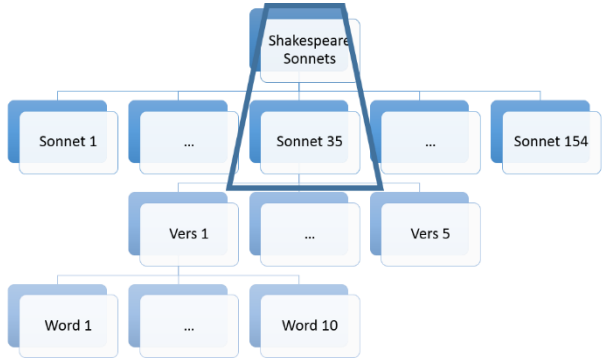
CLARIN-D

Full Text Search

Find low level CTS URNs for a given text passage

Trivial inside text parts

urn:cts:german_speeches:Bundespraesident.1990.10.3:1.2.3
urn:cts:german_speeches:Bundespraesident.1990.10.3:
Simply query for text instead of CTS URN

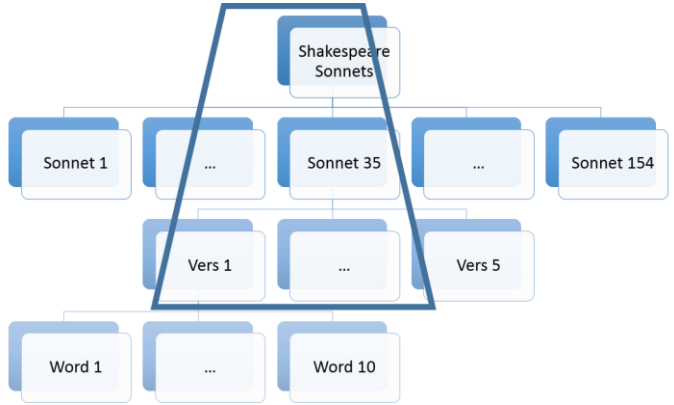


Complex for text spans

urn:cts:german_speeches:Bundespraesident.1990.10.3:1.2.3-1.3.2
urn:cts:german_speeches:Bundespraesident.1990.10.3: 1-2

Search for(first tokens) to find starting URN
foreach (starting URN)
while(result is better)
result = expand to right neighbour

Potentially huge candidate set



Full Text Search

Complex for text spans

urn:cts:demo:Systemofadown.mrjack:1.5	urn:cts:demo:Systemofadown.mrjack:1.6
Hey where you at?	On the side of a freeway in the car

Query: „at? On the side of a freeway in the car“

Search for („at“) to find starting URN
foreach (starting URN) *//-> Every URN for a text passage that contains „at“*
while(result is better)
result = expand to right neighbour

Limit candidates!

Candidate Search for Full Text Search

Limit candidates!

Various approaches: (Document search)

Full Text Search MySQL Fulltext Index & Lucene Fulltext Index

Result of Full Text Search for query *„at? On the side of a freeway in the car“*

Term-Document-matrix.

Documents that contain *„at“, „On”, “the”, “side”, “of”, “a”, “freeway”, “in”, “the”* and *“car“*

Document signatures (e.g. word length)

„2 . 2 3 4 2 1 7 2 3 3“

Combinations

Full Text Search

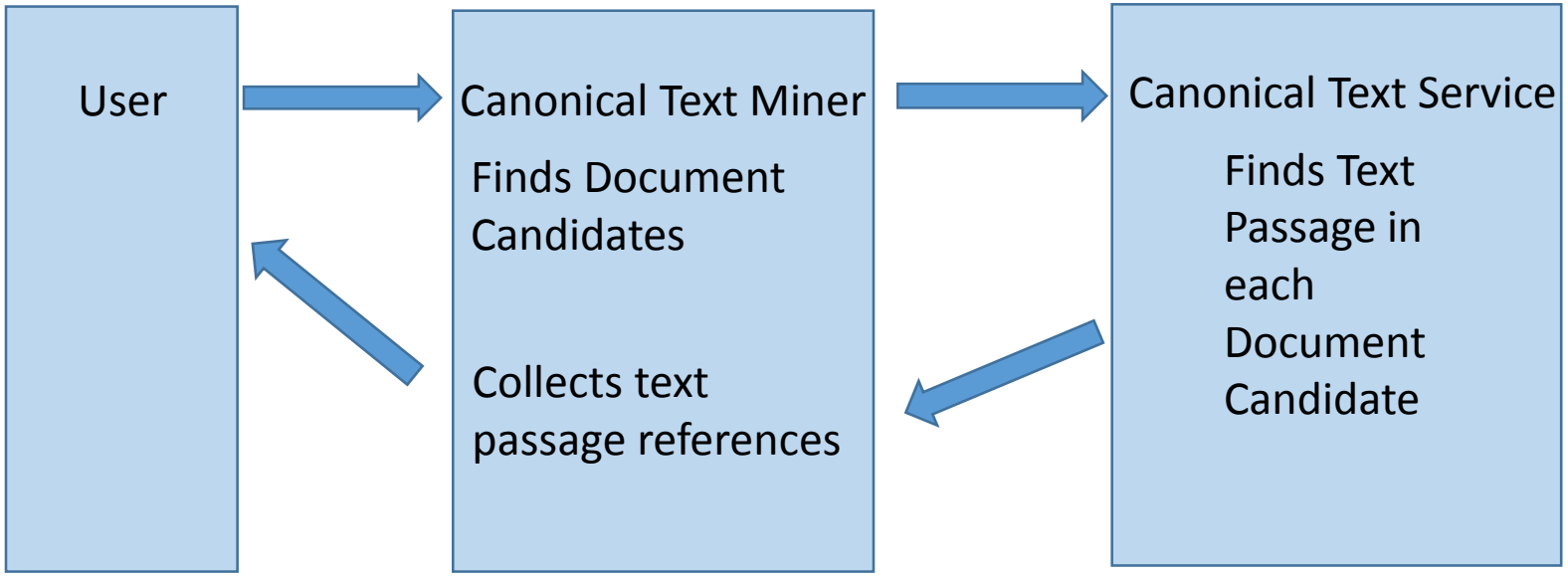
Complex for text spans

urn:cts:demo:Systemofadown.mrjack:1.5	urn:cts:demo:Systemofadown.mrjack:1.6
Hey where you at?	On the side of a freeway in the car

Query: „at? On the side of a freeway in the car“

Search for („at“ in document candidates) to find starting URN
foreach (starting URN)
while(result is better)
result = expand to right neighbour

Full Text Search



Text Reuse

Who cites Whom?

Find very similar text passages

Similarity Analysis

Similarity s for every sentence combination
s > threshold-> citation

Example projects:

Picapica (Martin Potthast)
Etracer (Marco Büchler)
Winowing (Schleimer, Wilkinson, Aiken)

Text Reuse

:1
:1.1
:1.1.1 O Tannenbaum, O Tannenbaum,
:1.1.2 Wie treu sind deine Blätter.
:1.1.3 Du grünst nicht nur zur Sommerzeit,
:1.1.4 Nein auch im Winter wenn es schneit.
:1.1.5 O Tannenbaum, O Tannenbaum,
:1.1.6 Wie grün sind deine Blätter!
:1.2
:1.2.1 O Tannenbaum, O Tannenbaum,
:1.2.2 Du kannst mir sehr gefallen!
:1.2.3 Wie oft hat schon zur Winterszeit
:1.2.4 Ein Baum von dir mich hoch erfreut!
:1.2.5 O Tannenbaum, O Tannenbaum,
:1.2.6 Du kannst mir sehr gefallen!
:1.3
:1.3.1 O Tannenbaum, O Tannenbaum,
:1.3.2 Dein Kleid will mich was lehren:
:1.3.3 Die Hoffnung und Beständigkeit
:1.3.4 Gibt Mut und Kraft zu jeder Zeit!
:1.3.5 O Tannenbaum, O Tannenbaum,
:1.3.6 Dein Kleid will mich was lehren.

:1.1.1 -> 1.1.5, 1.2.1, 1.2.5, 1.3.1, 1.3.5

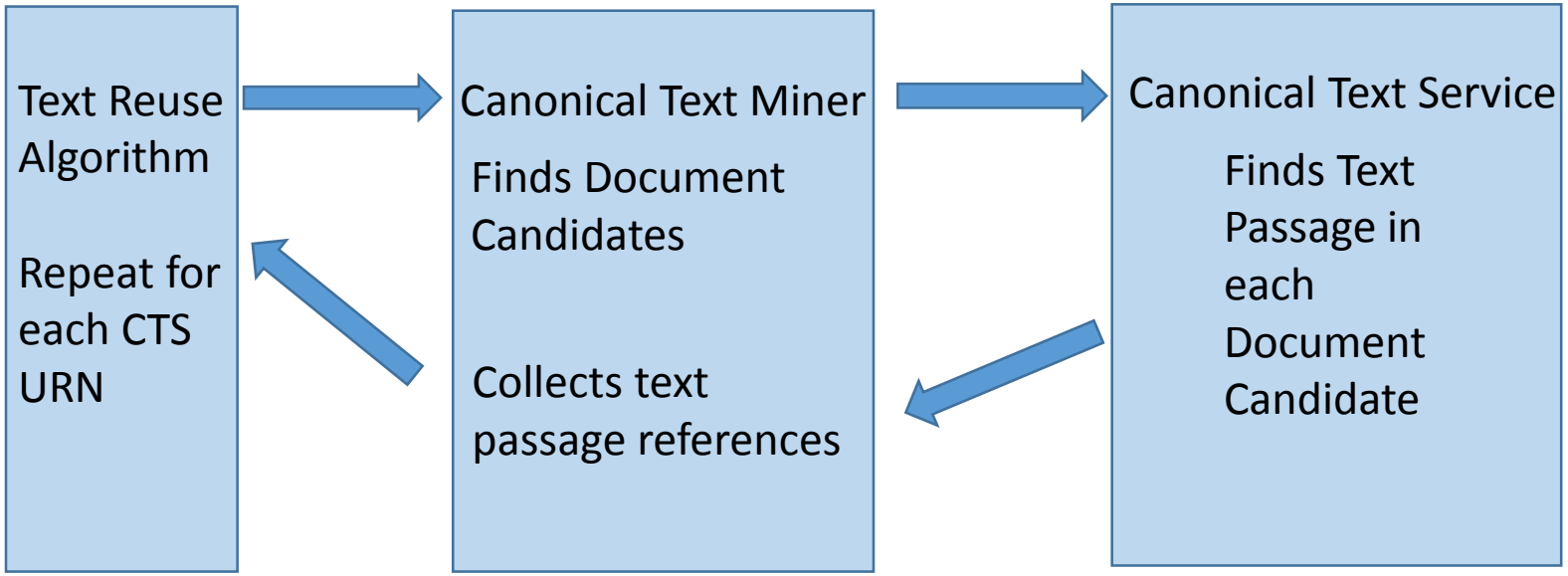
:1.2.2 -> 1.2.6
:1.3.2 -> 1.3.6

:1.1.2 -> 1.1.6(?)

TextReuse = Persistent IDs + similar text passages + Publication date

Here: CTS URNs Text passage search Meta information

Text Reuse


Text Reuse Algorithm
Repeat for each CTS URN
Canonical Text Miner
Finds Document Candidates
Collects text passage references
Canonical Text Service
Finds Text Passage in each Document Candidate

Text Reuse

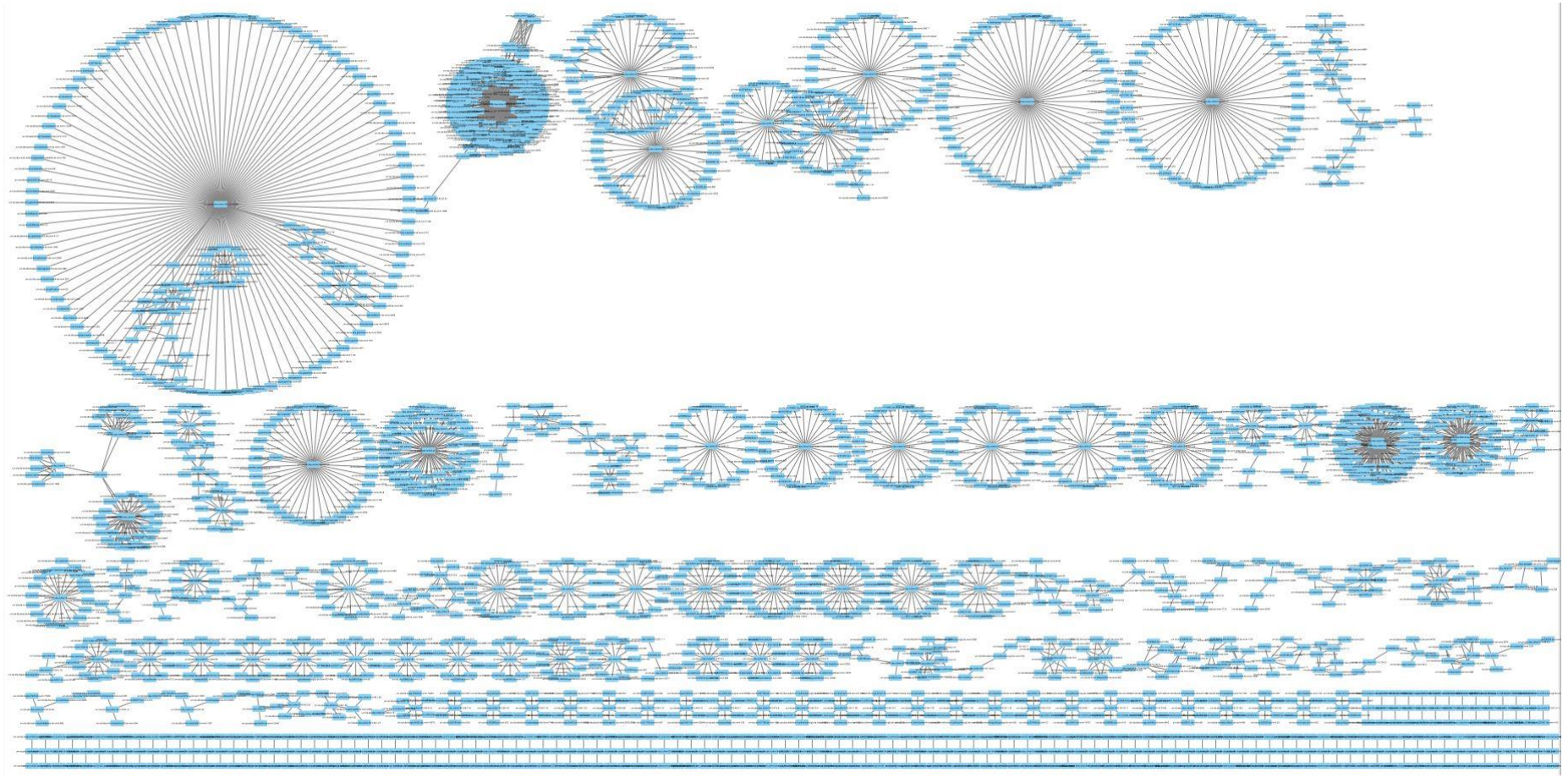
Case Study based on Parallel Bible Corpus and Das Deutsche Textarchiv
http://paralleltext.info/data/
http://www.deutschestextarchiv.de/
Visualisation done using Cytoscape

passage:Am Anfang schuf Gott Himmel und Erde . source:urn:cts:pbc:bible.parallel.deu.luther1545:1.1.1
urn:cts:dta:weise.ertznarren.de.norm:1352_#_secht ihr herren sagte er am anfang schuf gott himmel(…)
urn:cts:dta:justi.geschichte.de.norm:2062_#_am anfang schuf gott himmel und erde
urn:cts:dta:seyfried.medulla.de.norm:853_#_am anfang schuf gott himmel und erden
urn:cts:dta:hundtradowsky.judenschule01.de.norm:750_#_am anfang schuf gott himmel und
urn:cts:dta:bullinger.haussbuoch.de.norm:13540_#_(…)ersten buchs im anfang schuf gott den himmel
urn:cts:dta:luetkemann.auffmunterung2.de.norm:8421_#_im anfang schuf gott himmel und erden (…)
urn:cts:dta:fontane.kinderjahre.de.norm:1747-1748_#_am anfang schuf gott himmel und erde(…)
urn:cts:dta:fontane.kinderjahre.de.norm:1748_#_im anfang schuf gott himmel und erde
urn:cts:dta:luther.betbuechlein.de.norm:1570_#_am anfang schuf gott himmel und erden genes
2016.10.08 at 12:52:57

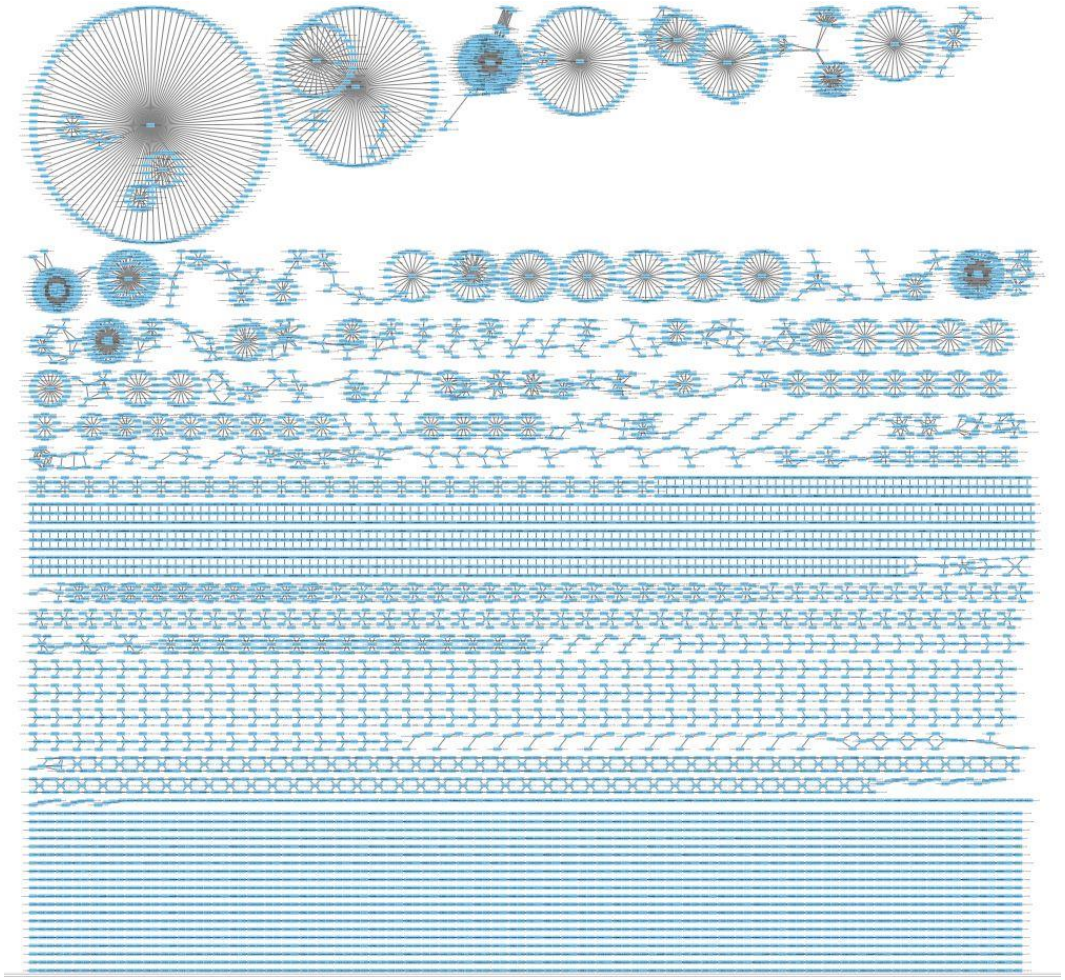
*(…) -> shortened for readability

Bible Citations in Deutsche Textarchiv

For .tsv edge lists and .pdf visualisations see http://cts.informatik.uni-leipzig.de/asusedin.html



Bible Citations in Deutsche Textarchiv

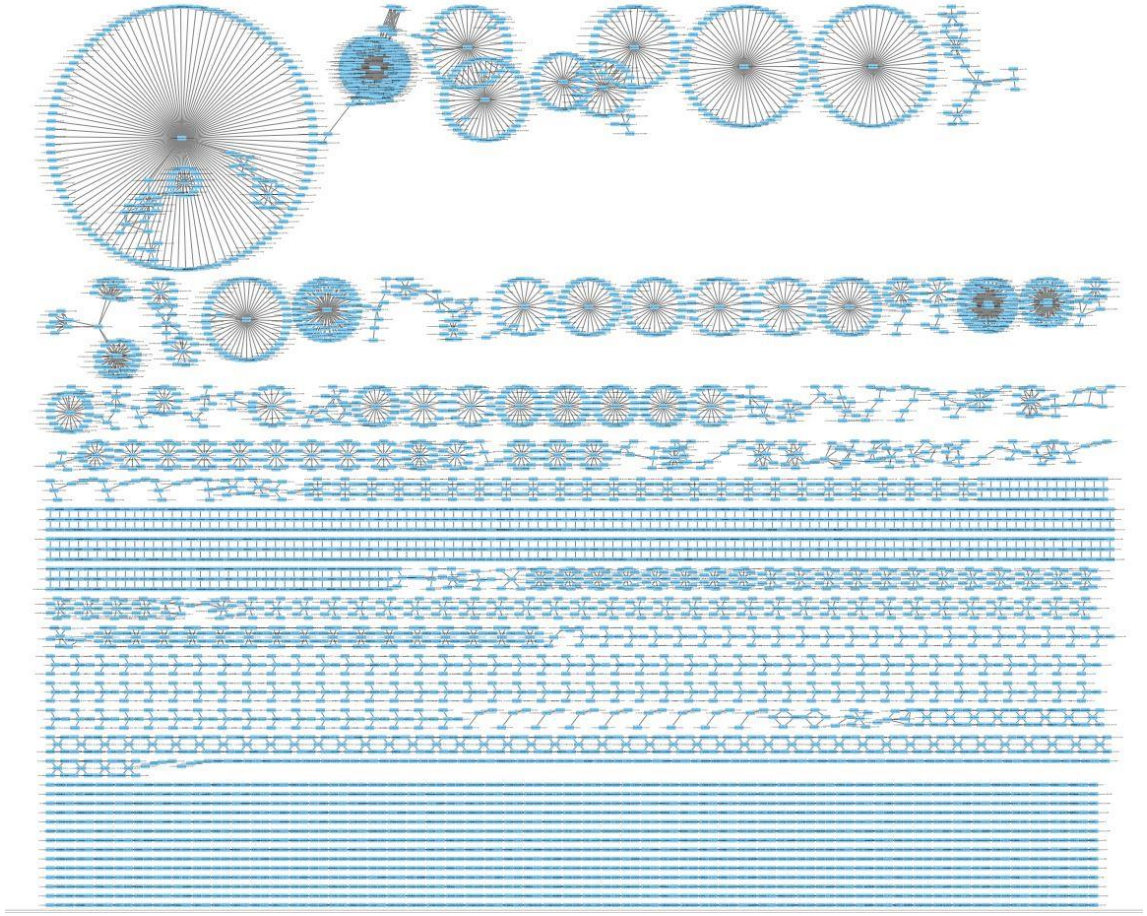


Luther1545

For .tsv edge lists and .pdf visualisations see http://cts.informatik.uni-leipzig.de/asusedin.html

Bible Citations in Deutsche Textarchiv

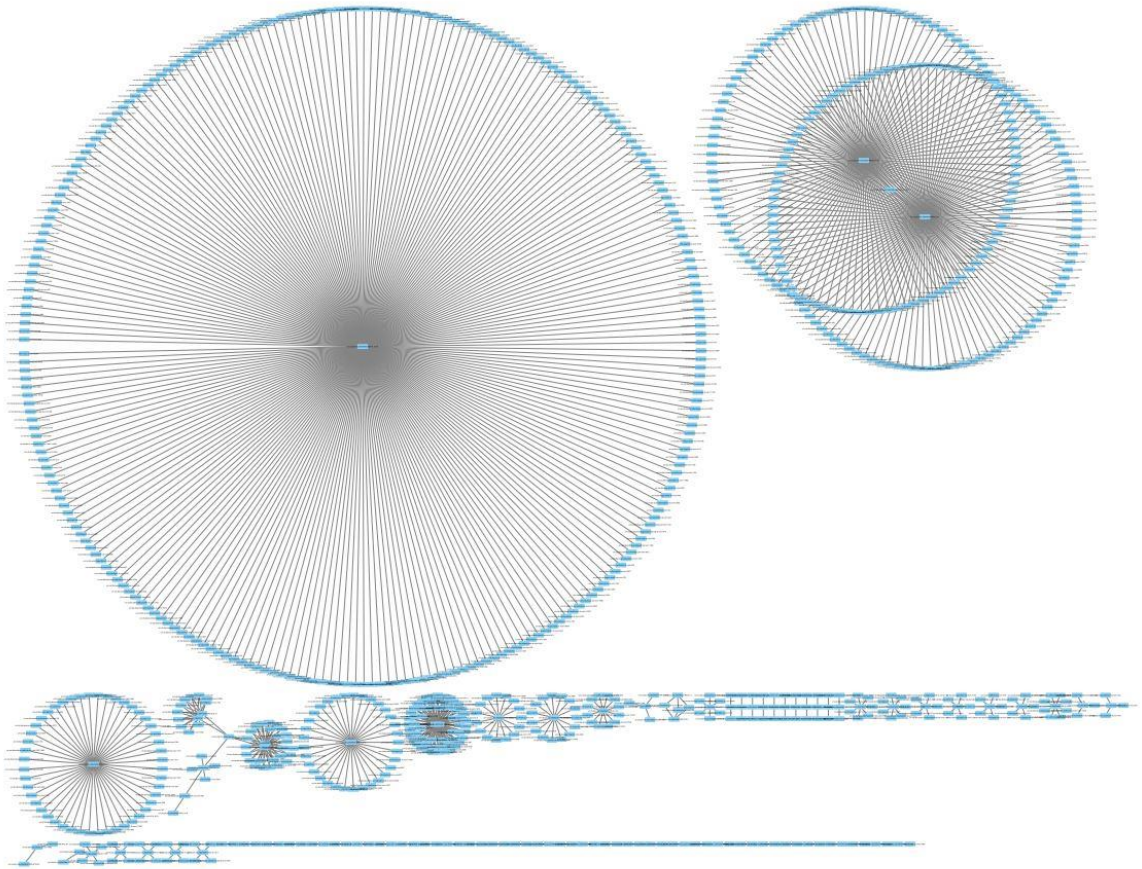
Luther1912



For .tsv edge lists and .pdf visualisations see http://cts.informatik.uni-leipzig.de/asusedin.html

Bible Citations in Deutsche Textarchiv

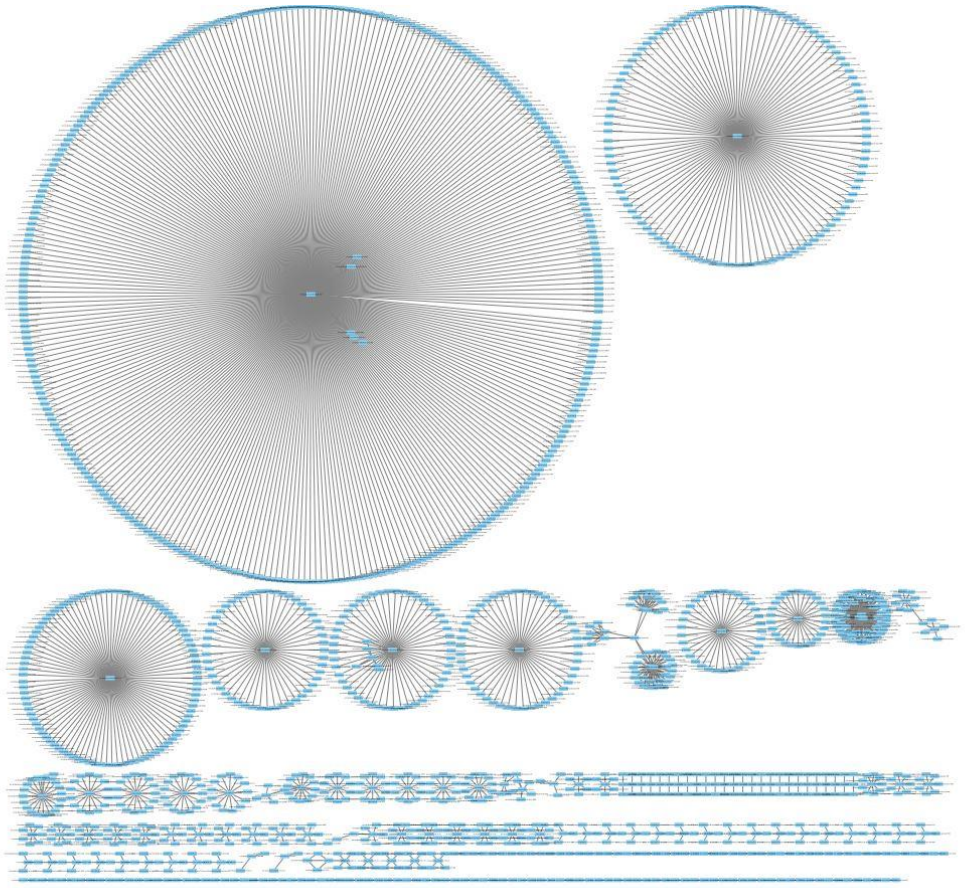
Elberfelder1905



For .tsv edge lists and .pdf visualisations see http://cts.informatik.uni-leipzig.de/asusedin.html

Bible Citations in Deutsche Textarchiv

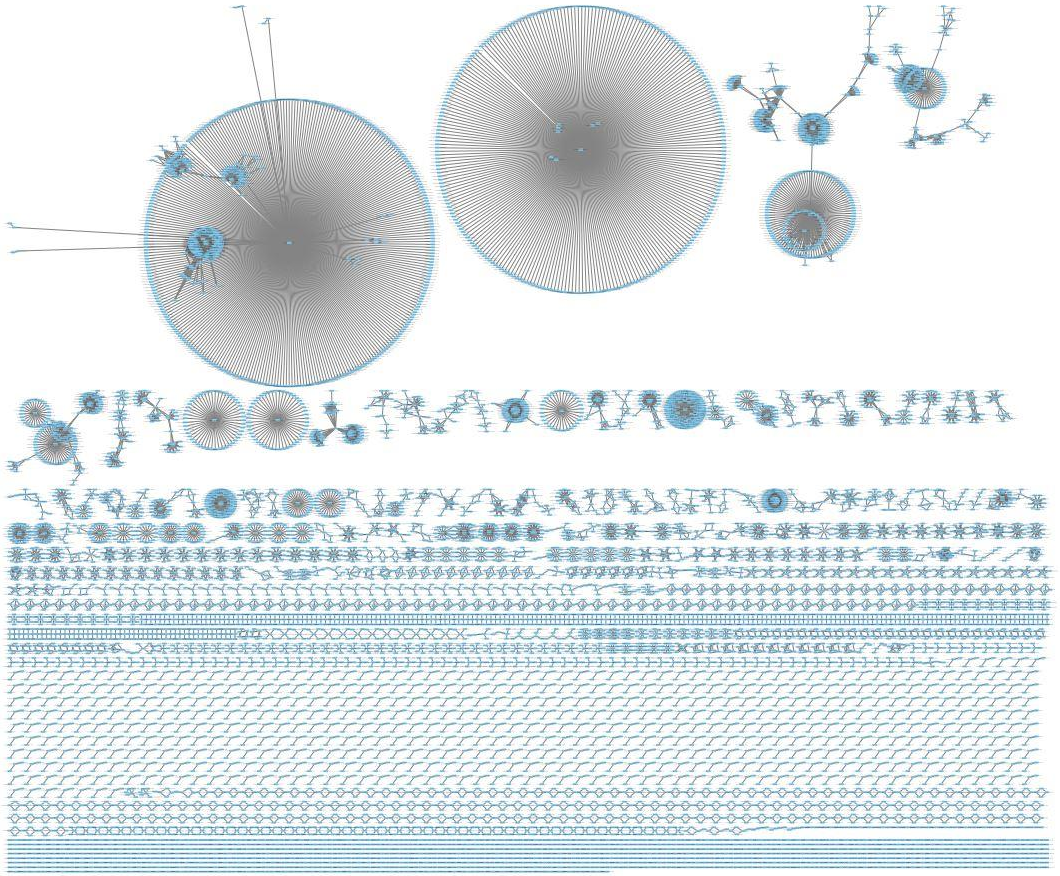
Schlachter



For .tsv edge lists and .pdf visualisations see http://cts.informatik.uni-leipzig.de/asusedin.html

Bible Citations in Deutsche Textarchiv

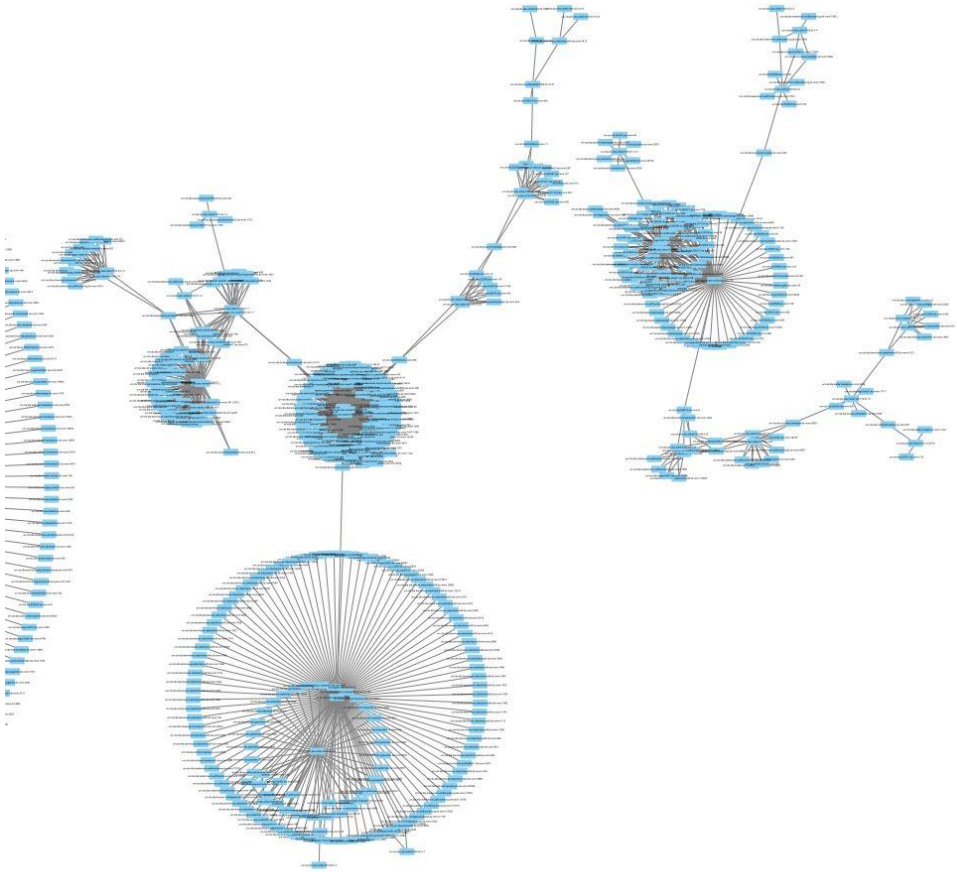
Combined



For .tsv edge lists and .pdf visualisations see http://cts.informatik.uni-leipzig.de/asusedin.html

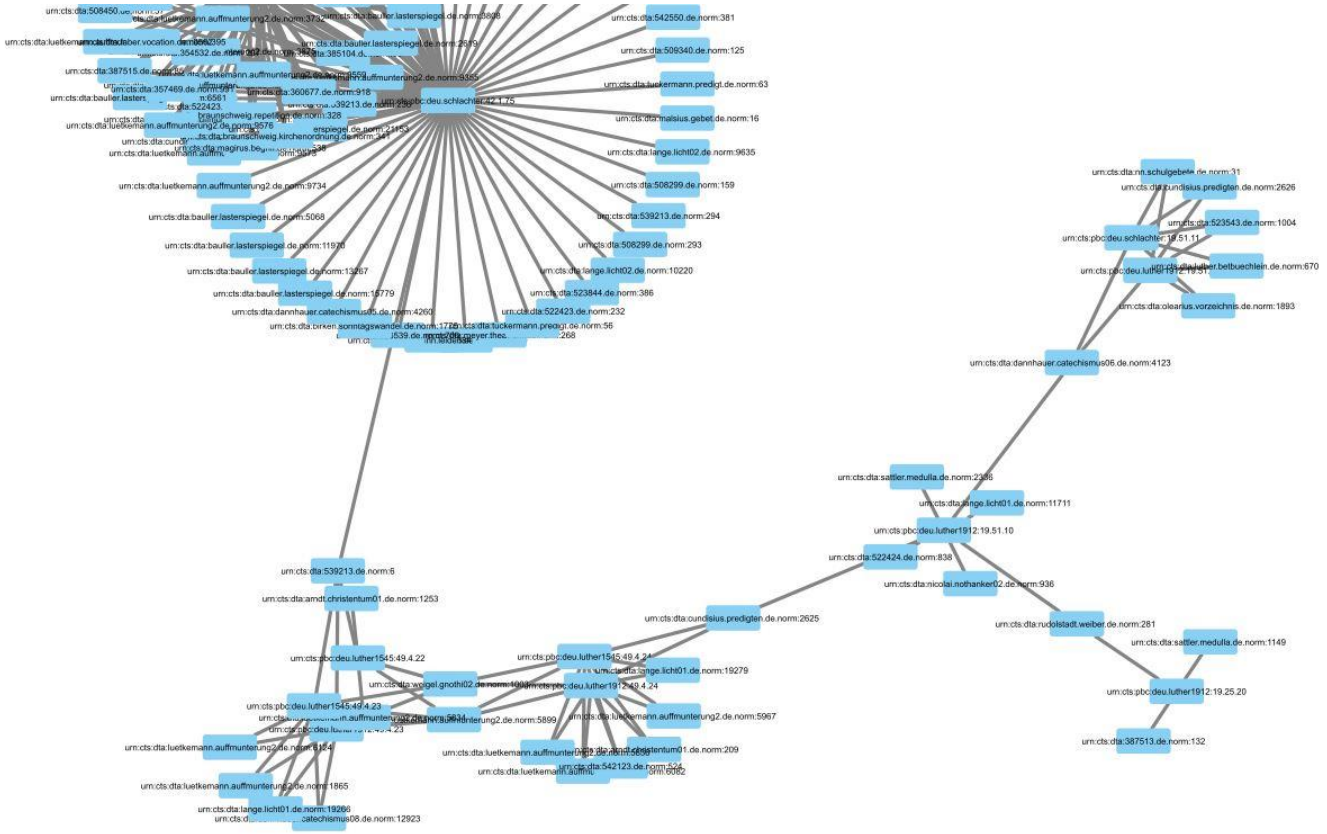
Bible Citations in Deutsche Textarchiv

Combined Zoom 1

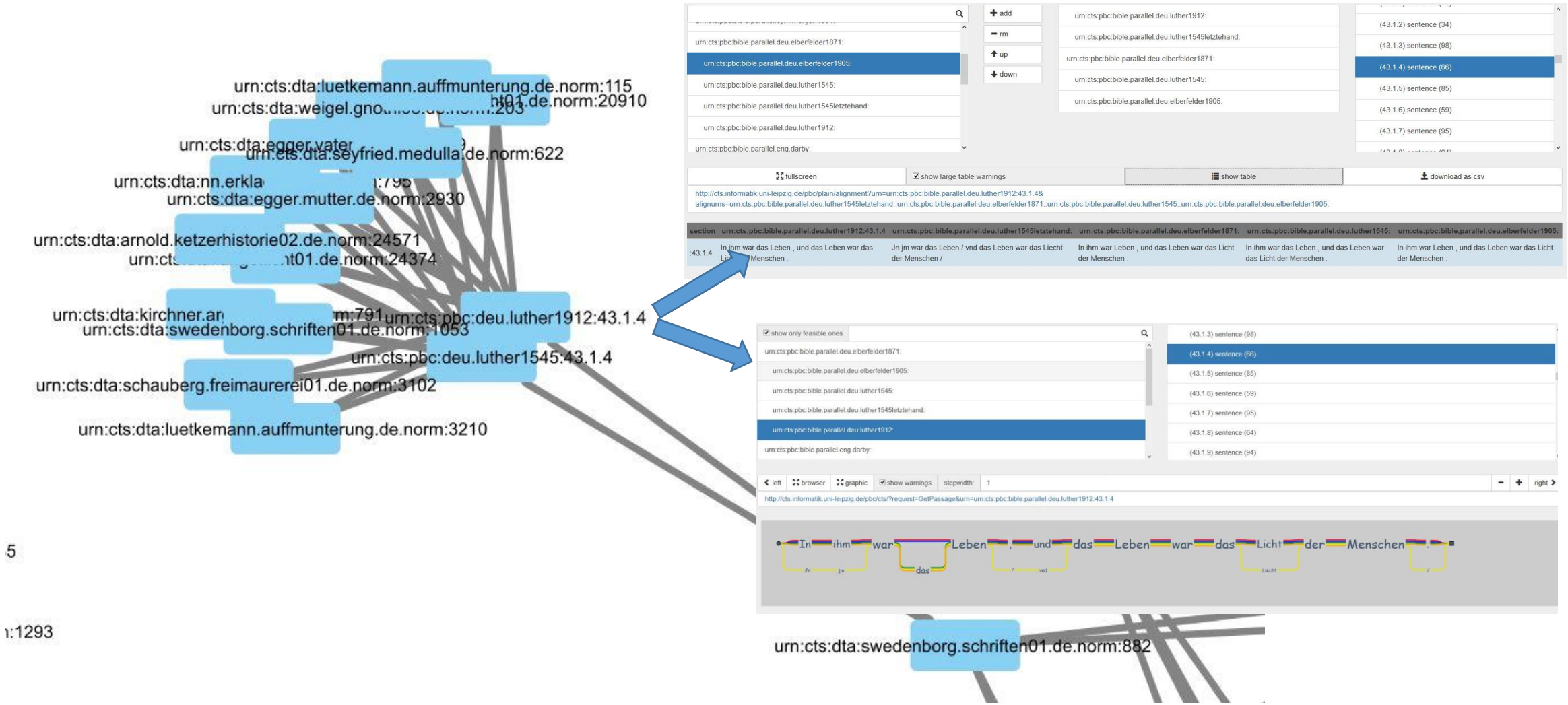


Bible Citations in Deutsche Textarchiv

Combined Zoom 2



Bible Citations in Deutsche Textarchiv



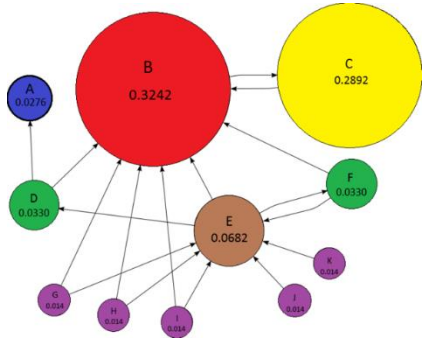
Text Reuse

Pagerank

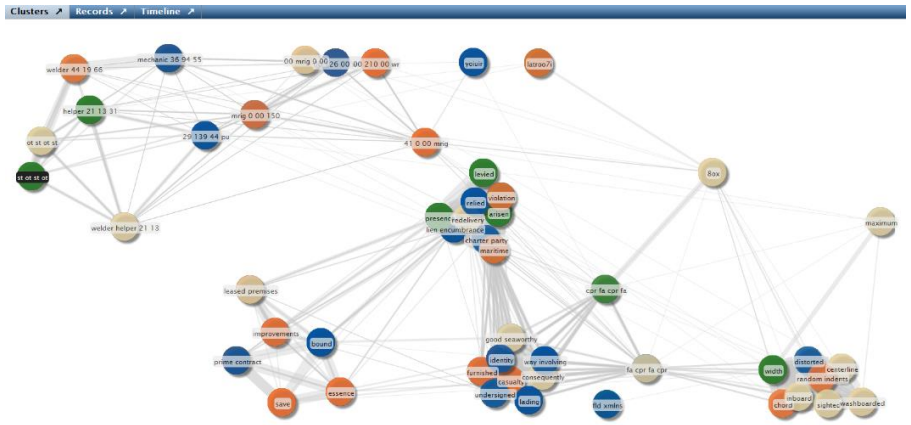
- Find often cited passages
- Find rarely cited passages

Clustering

- Disciplinary networks
- self referencing hubs



https://de.wikipedia.org/wiki/PageRank



http://orcatec.com/wp-content/uploads/2013/09/cluster2.png

Contact

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

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