ScaDS DRESDEN LEIPZIG

ASV Automatische Sprachverarbeitung

Canonical Text Service

Jochen Tiepmar

BigData Competence Center ScaDS
Naural Language Processing
Leipzig University

Federal Ministry of Education and Research

Survey

From 20.06.2015 to 30.08.2015

Anonym, no tracking, skipping allowed

Recall 25.06.2015 : 9

www.urncts.de/survey

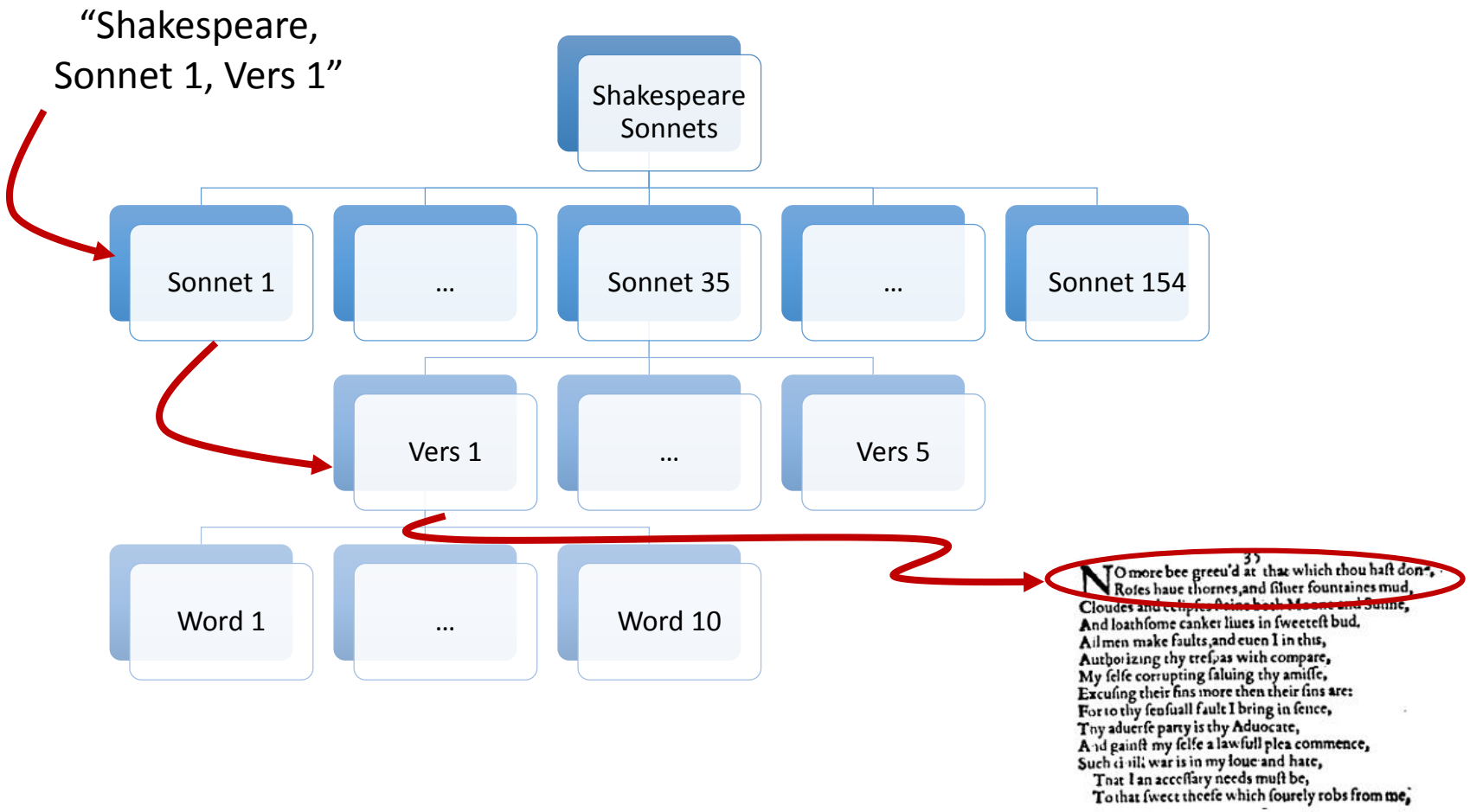
100% know the term terabyte and 71,43% know the term petabyte.

Overview

Canonical Text Services (CTS)

- protocol for a webbased citable text service
- Unique Identifiers(**U**nique **R**esource **N**ame, **URN**) refer to text passages
- 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
- Implementation for Tripelstore and XML-DB not suitable for BW-Project
- Demo webpage: www.urncts.de

Documents Hierarchy



Citation

Document „outer hierarchy“

Shakespeare → Sonnets → english → 1st edition

Text passage „inner hierarchy“

Sonnet 1 → Vers 1

Combined

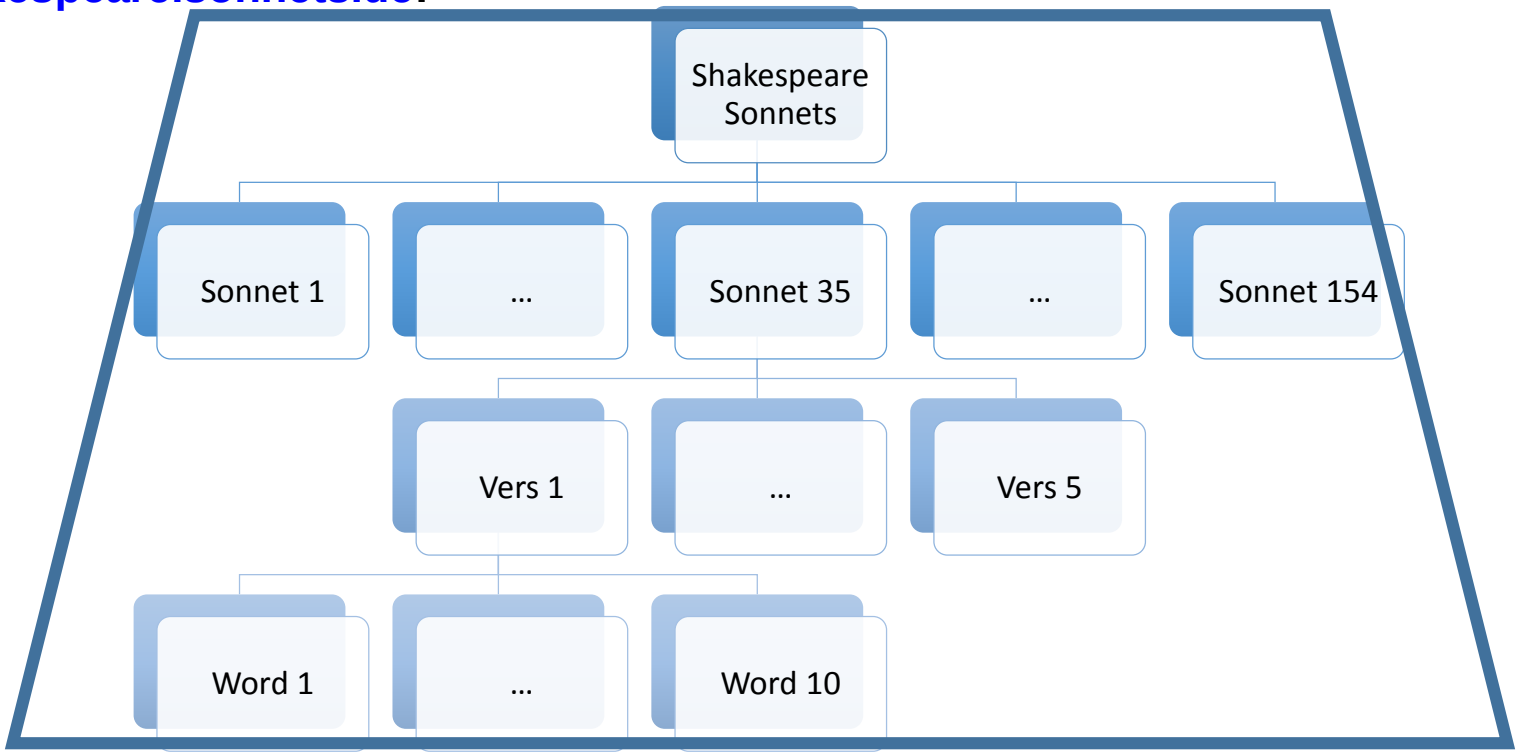
Shakespeare → Sonnets → english → 1st edition → Sonnet 1→ Vers 1

CTS-URN

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

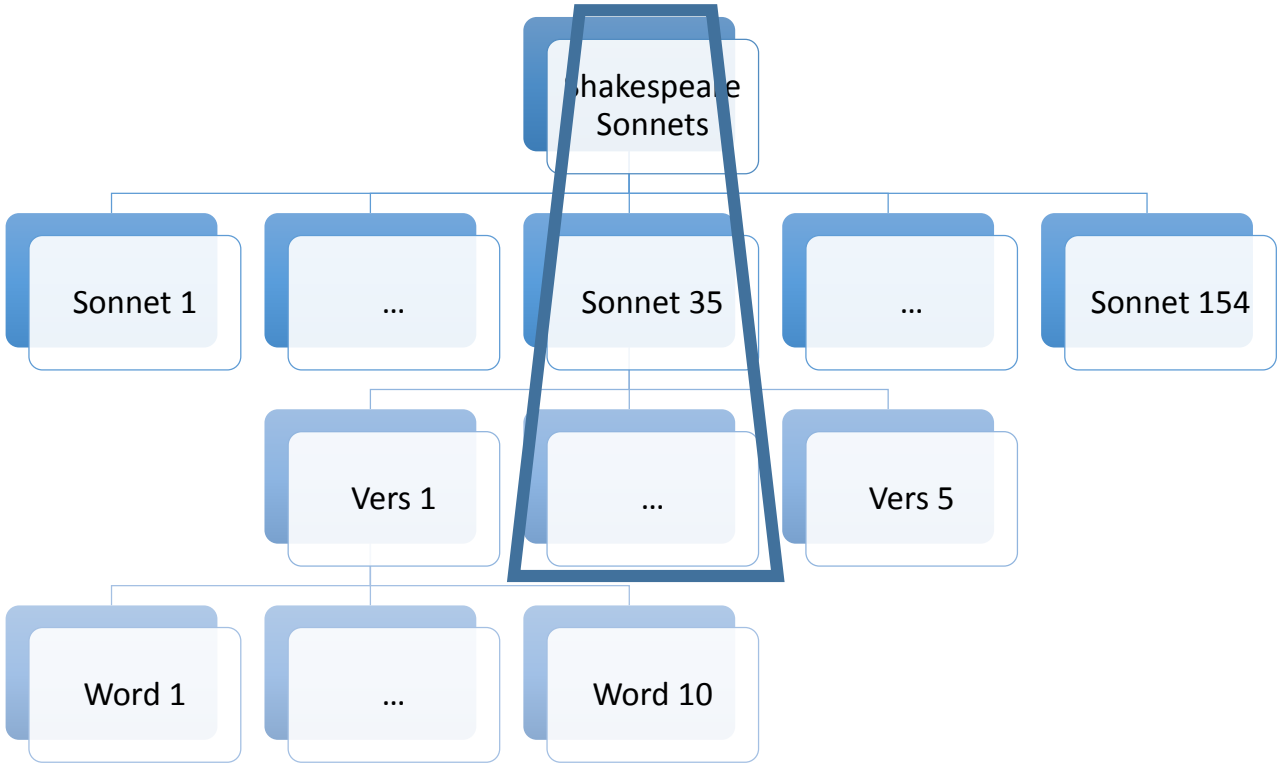
Canonical Citation

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



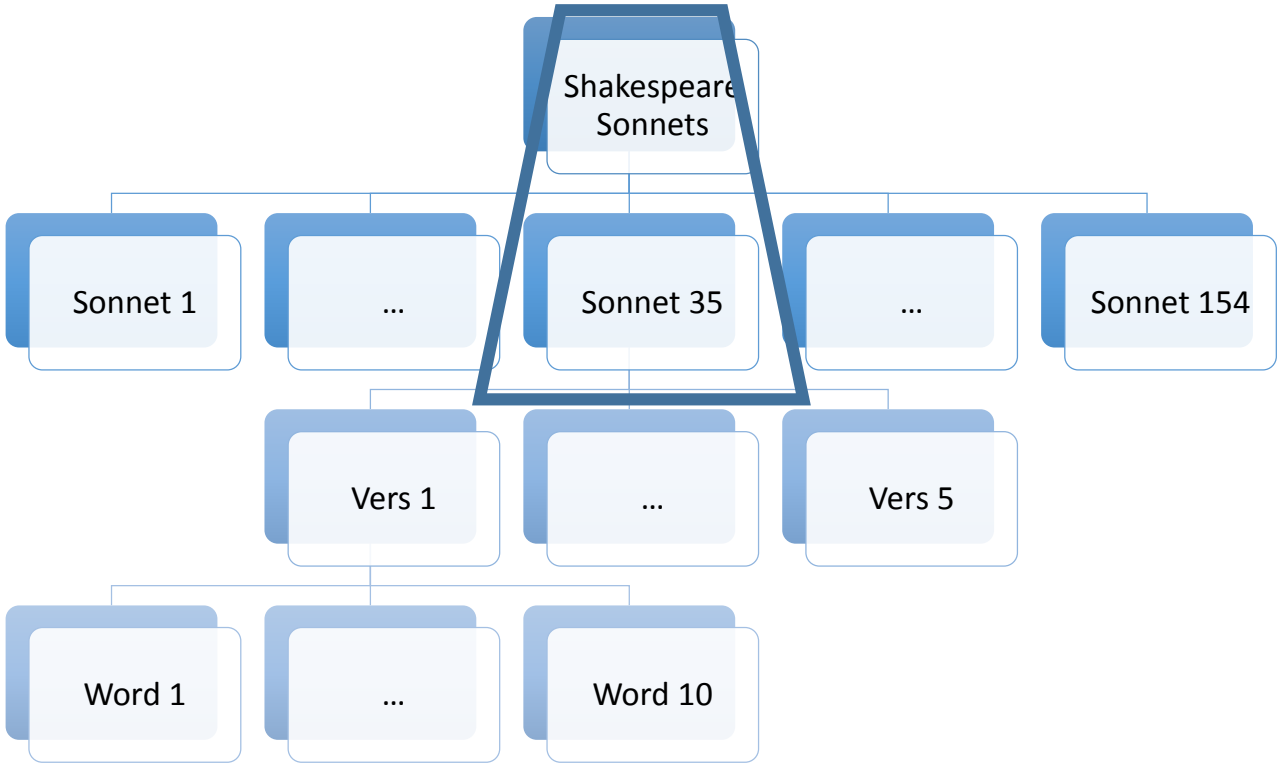
Canonical Citation

urn:cts:demo:shakespeare.sonnets:35.4



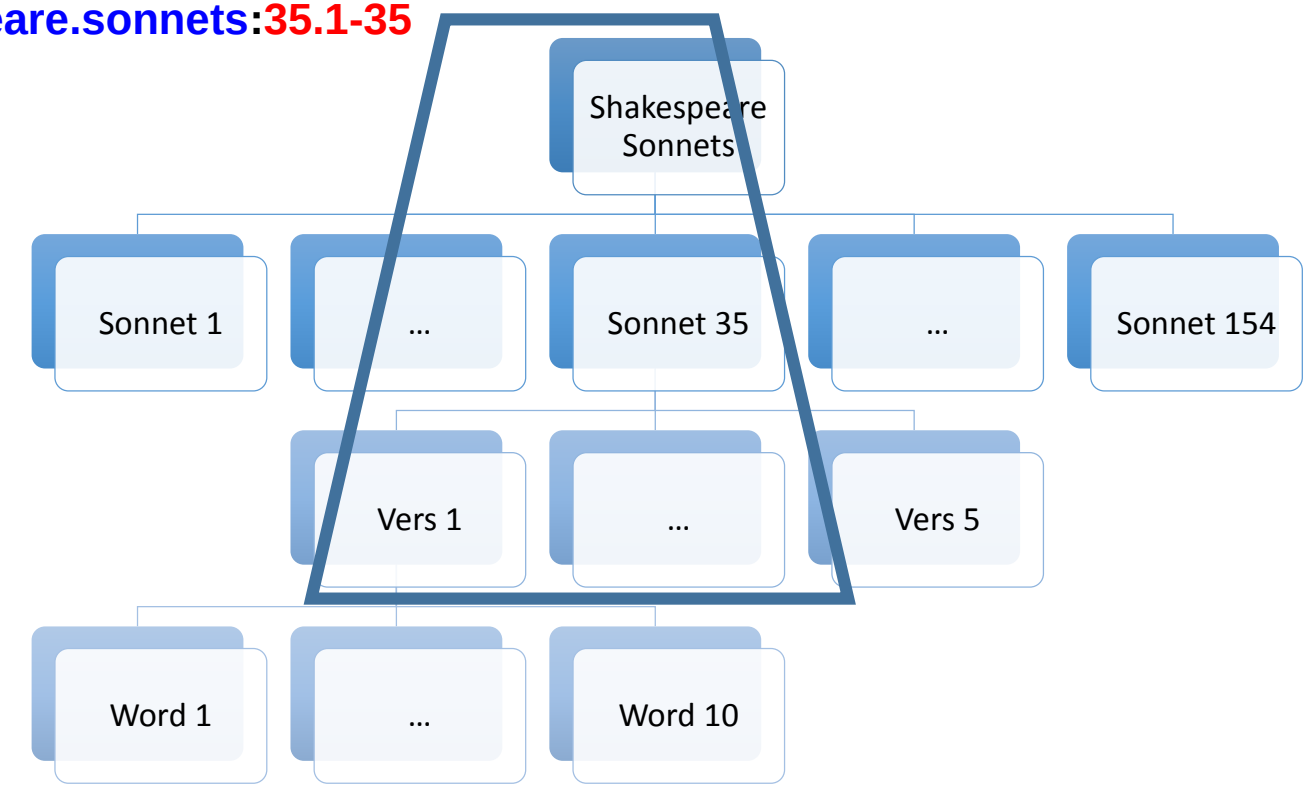
Canonical Citation

urn:cts:demo:shakespeare.sonnets:35



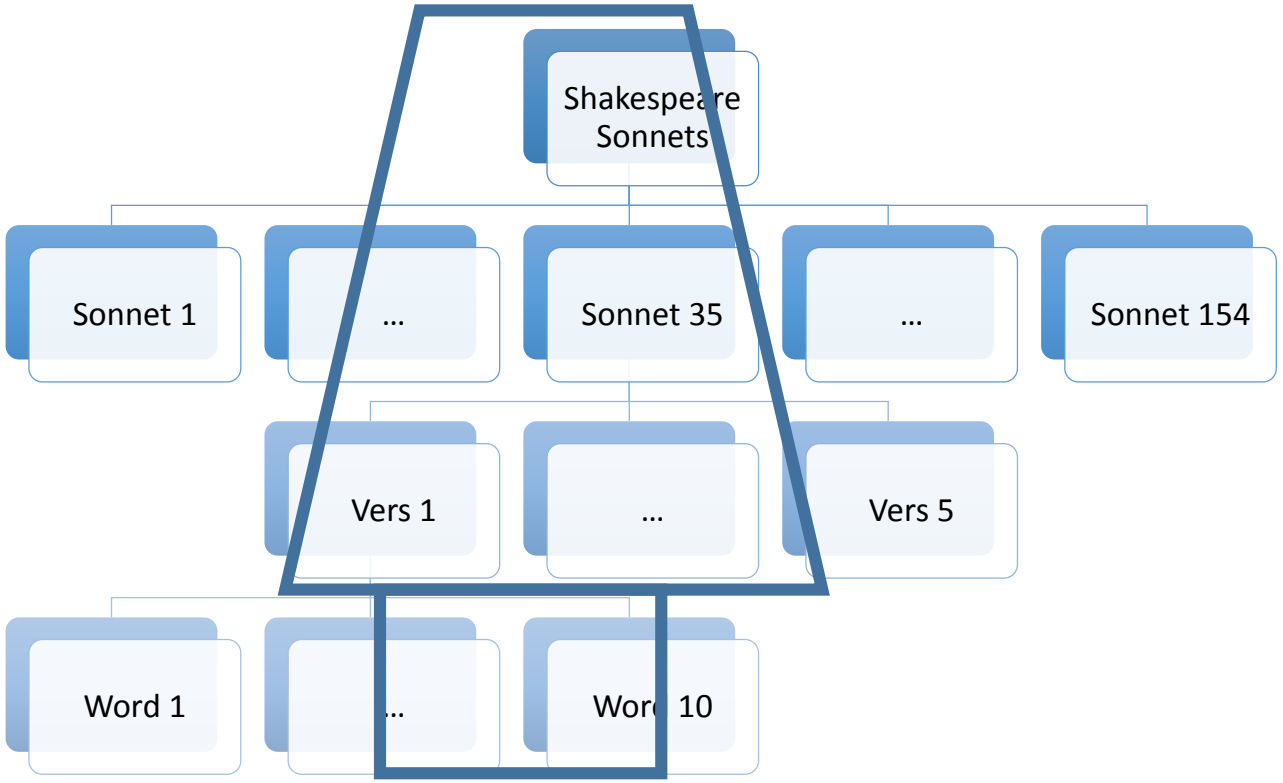
Canonical Citation

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



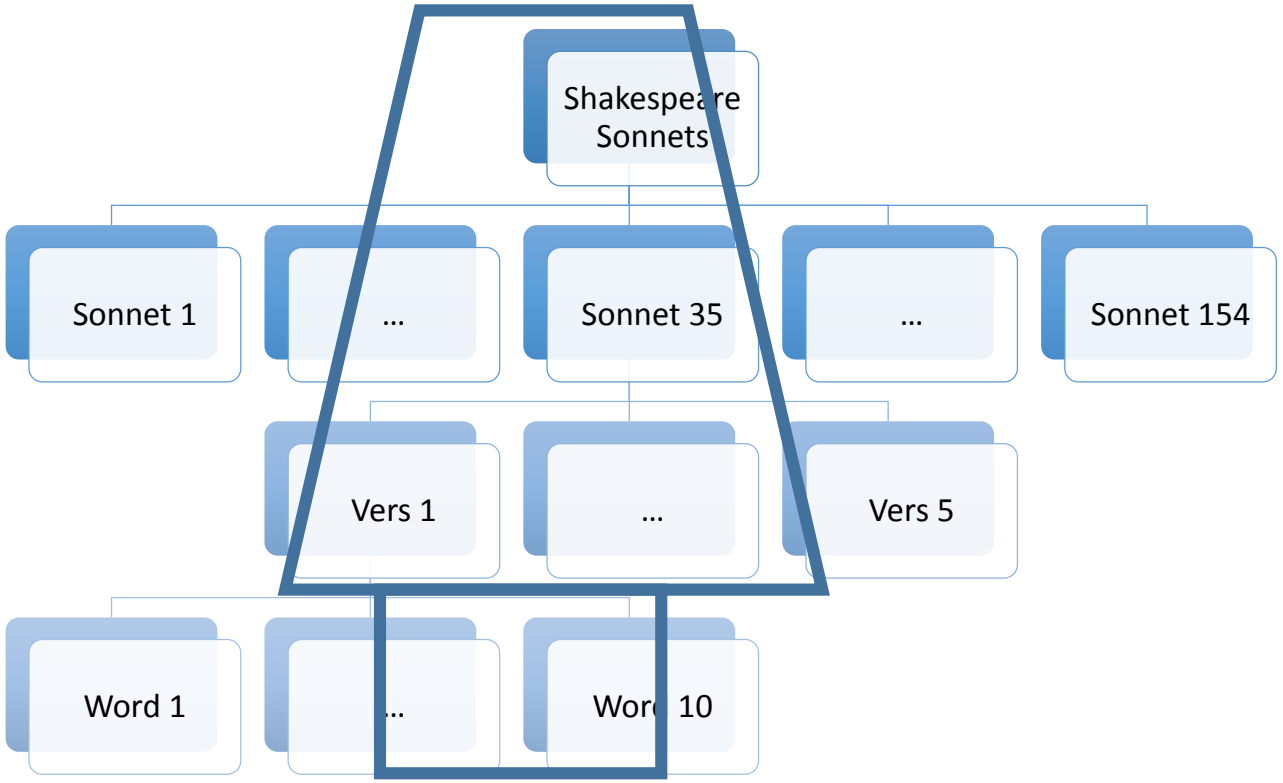
Canonical Citation

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



Canonical Citation

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



Mapping URNs -> Text

: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.

Using CTS to standardize texts

Differentiate text structure from text content and meta information
Refer to generic text parts
Reduce type of text part to label

8/9 think that standardizing documents and access to documents will be (very) important in the next 10 years

8/9 think that referencing documents based on structural text parts (like chapter or sentence) is reasonable. 1 suggests named entities, 1 adds that further standardization and more flexibility is needed

Div-View

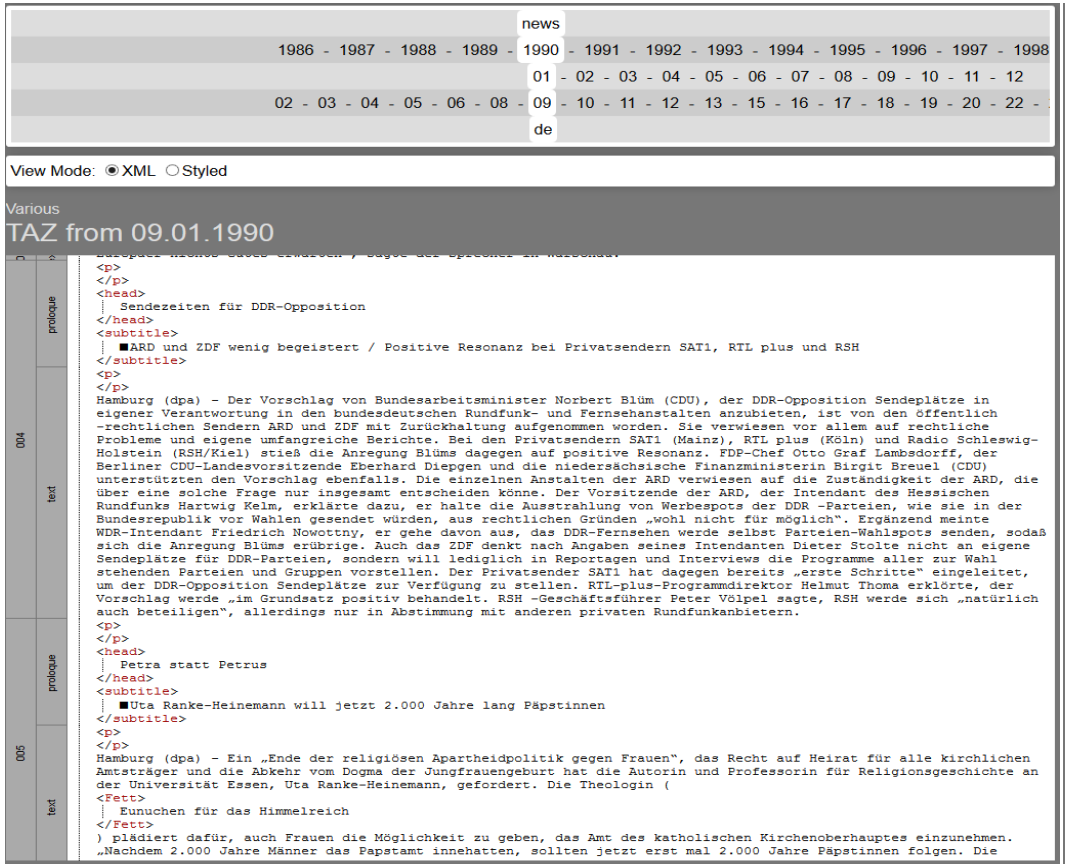
urn:cts:songs:christmas.ohtennenbaum.de.1:1-1.2.4

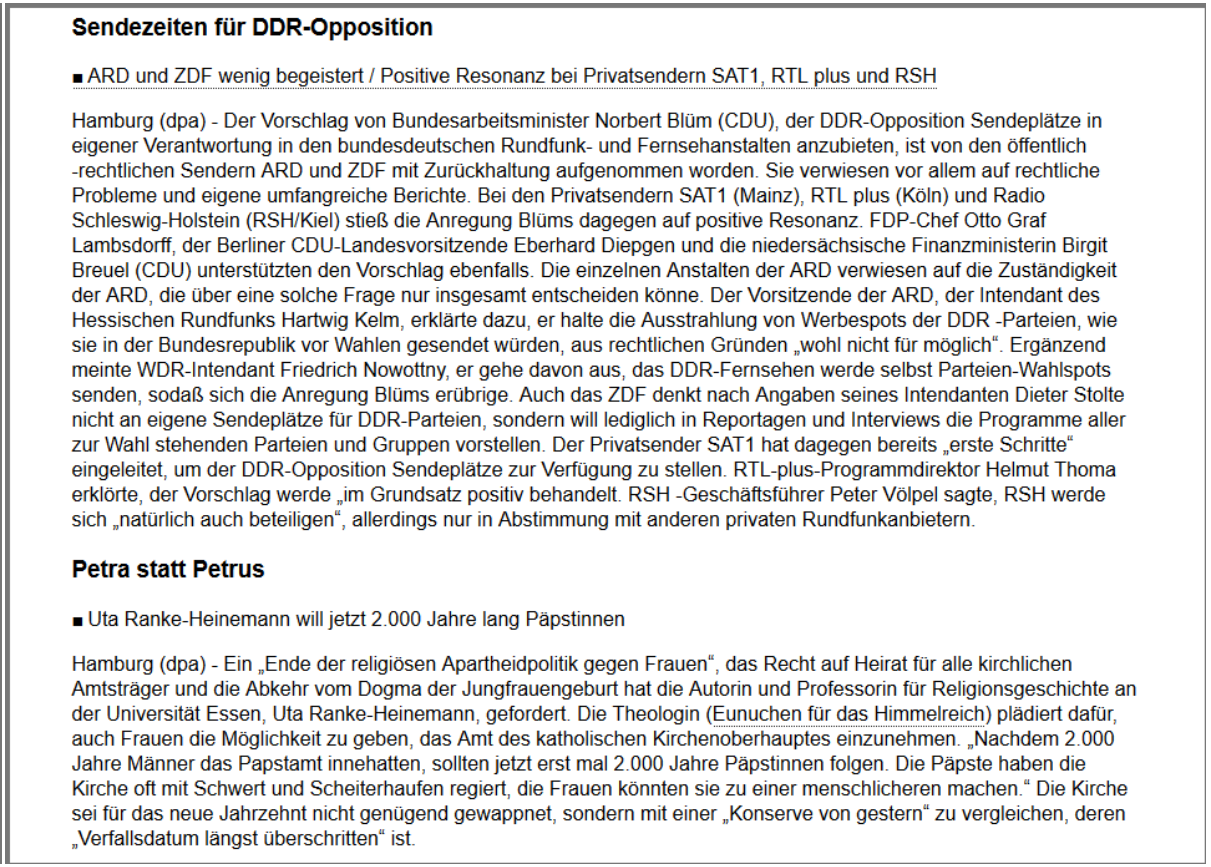
```
<passage>
        O Tannenbaum, O Tannenbaum, (…) Wie grün sind deine Blätter! O Tannenbaum, O Tannenbaum, (…) Ein Baum von dir mich
        hoch erfreut!
</passage>

<passage>
        <div1 n="1" type="song">
                <div2 n="1" type="strophe">
                        <div3 n="1" type="line">O Tannenbaum, O Tannenbaum, </div3>
                        (…)
                        <div3 n="6" type="line">Wie grün sind deine Blätter! </div3>

                </div2>
                <div2 n="2" type="strophe">
                        <div3 n="1" type="line">O Tannenbaum, O Tannenbaum, </div3>
                         (…)
                        <div3 n=„6" type="line">Ein Baum von dir mich hoch erfreut!</div3>
                </div2>
        </div1>
</passage>
```

Generic Reader





Sendezeiten für DDR-Opposition

■ ARD und ZDF wenig begeistert / Positive Resonanz bei Privatsendern SAT1, RTL plus und RSH

Hamburg (dpa) - Der Vorschlag von Bundesarbeitsminister Norbert Blüm (CDU), der DDR-Opposition Sendeplätze in eigener Verantwortung in den bundesdeutschen Rundfunk- und Fernsehanstalten anzubieten, ist von den öffentlich -rechtlichen Sendern ARD und ZDF mit Zurückhaltung aufgenommen worden. Sie verwiesen vor allem auf rechtliche Probleme und eigene umfangreiche Berichte. Bei den Privatsendern SAT1 (Mainz), RTL plus (Köln) und Radio Schleswig-Holstein (RSH/Kiel) stieß die Anregung Blüms dagegen auf positive Resonanz. FDP-Chef Otto Graf Lambsdorff, der Berliner CDU-Landesvorsitzende Eberhard Diepgen und die niedersächsische Finanzministerin Birgit Breuel (CDU) unterstützten den Vorschlag ebenfalls. Die einzelnen Anstalten der ARD verwiesen auf die Zuständigkeit der ARD, die über eine solche Frage nur insgesamt entscheiden könne. Der Vorsitzende der ARD, der Intendant des Hessischen Rundfunks Hartwig Kelm, erklärte dazu, er halte die Ausstrahlung von Werbespots der DDR -Parteien, wie sie in der Bundesrepublik vor Wahlen gesendet würden, aus rechtlichen Gründen „wohl nicht für möglich". Ergänzend meinte WDR-Intendant Friedrich Nowottny, er gehe davon aus, das DDR-Fernsehen werde selbst Parteien-Wahlspots senden, sodaß sich die Anregung Blüms erübrige. Auch das ZDF denkt nach Angaben seines Intendanten Dieter Stolte nicht an eigene Sendeplätze für DDR-Parteien, sondern will lediglich in Reportagen und Interviews die Programme aller zur Wahl stehenden Parteien und Gruppen vorstellen. Der Privatsender SAT1 hat dagegen bereits „erste Schritte" eingeleitet, um der DDR-Opposition Sendeplätze zur Verfügung zu stellen. RTL-plus-Programmdirektor Helmut Thoma erklärte, der Vorschlag werde „im Grundsatz positiv behandelt. RSH -Geschäftsführer Peter Völpel sagte, RSH werde sich „natürlich auch beteiligen", allerdings nur in Abstimmung mit anderen privaten Rundfunkanbietern.

Petra statt Petrus

■ Uta Ranke-Heinemann will jetzt 2.000 Jahre lang Päpstinnen

Hamburg (dpa) - Ein „Ende der religiösen Apartheidpolitik gegen Frauen", das Recht auf Heirat für alle kirchlichen Amtsträger und die Abkehr vom Dogma der Jungfrauengeburt hat die Autorin und Professorin für Religionsgeschichte an der Universität Essen, Uta Ranke-Heinemann, gefordert. Die Theologin (Eunuchen für das Himmelreich) plädiert dafür, auch Frauen die Möglichkeit zu geben, das Amt des katholischen Kirchenoberhauptes einzunehmen. „Nachdem 2.000 Jahre Männer das Papstamt innehatten, sollten jetzt erst mal 2.000 Jahre Päpstinnen folgen. Die Päpste haben die Kirche oft mit Schwert und Scheiterhaufen regiert, die Frauen könnten sie zu einer menschlicheren machen." Die Kirche sei für das neue Jahrzehnt nicht genügend gewappnet, sondern mit einer „Konserve von gestern" zu vergleichen, deren „Verfallsdatum längst überschritten" ist.

2014 Leipzig University // Martin Reckziegel

CTS Cloning

URNs specify @n-Value of s

-> @n-Values can be used to reconstruct URNs

-> Content of one CTS can be cloned

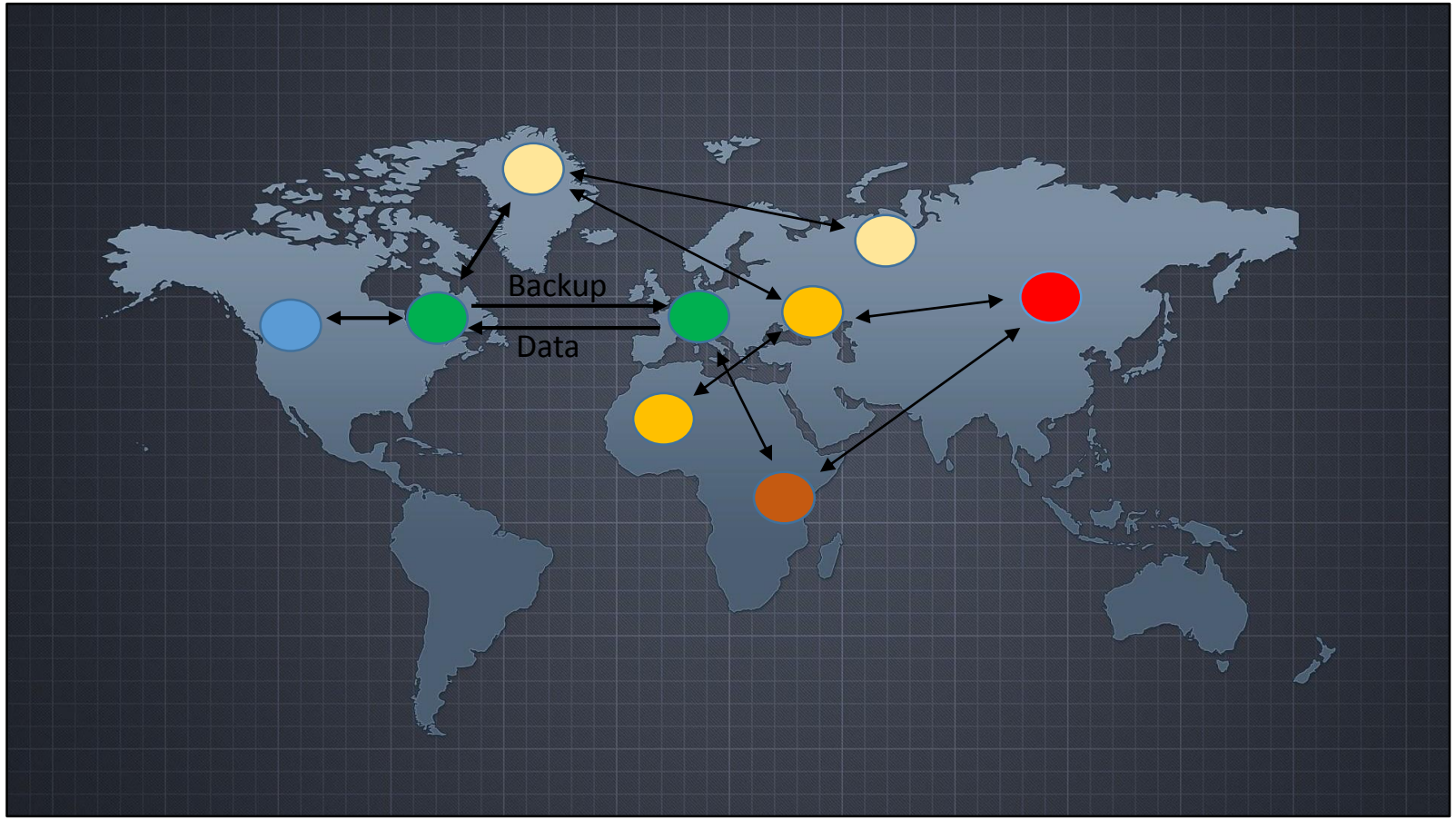
Data can be narrowed down „from left to right“ by URNs

Clone everything from Shakespeare:

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

```
<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>
```

CTS Cloning



7/9 think that a decentralized web of smaller text repositories based on individual researchers or projects is a more realistic scenario than a few central big text repositories containing the digitized documents of multiple researchers or projects

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

Data

Text Collection	Languages	Documents	File size
A german daily newspaper 1986-2012	German	15980	3,2 gb
Deutsches Textarchiv	German	5136	3 gb
PBC	831 Translations	831	1,9 gb
Perseus	Greek, Latin	2569	304 mb
Law	German	12698	226 mb
German Shakespeare works	German	188	21 mb

Alignment

(…)

Text Reuse Analysis

Which text part is a citation of what text part?

Pre calculation necessary

-> calculate similiarity between sentence and all other sentences

-> high similiarity = citation candidate

-> cross comparison, misses need to be calculated

Result: text reuse graph

Text Reuse Analysis per CTS

URNs as IDs for text parts

Fulltext search (WIP) as similiarity search

Unique IDs + fulltext search => Text Reuse Analysis?

To be continued(…)

CTS – Text Miner (CTSTM)

CTS Text Mining Framework

Broad and comprehensive framework for text analysis

Done:

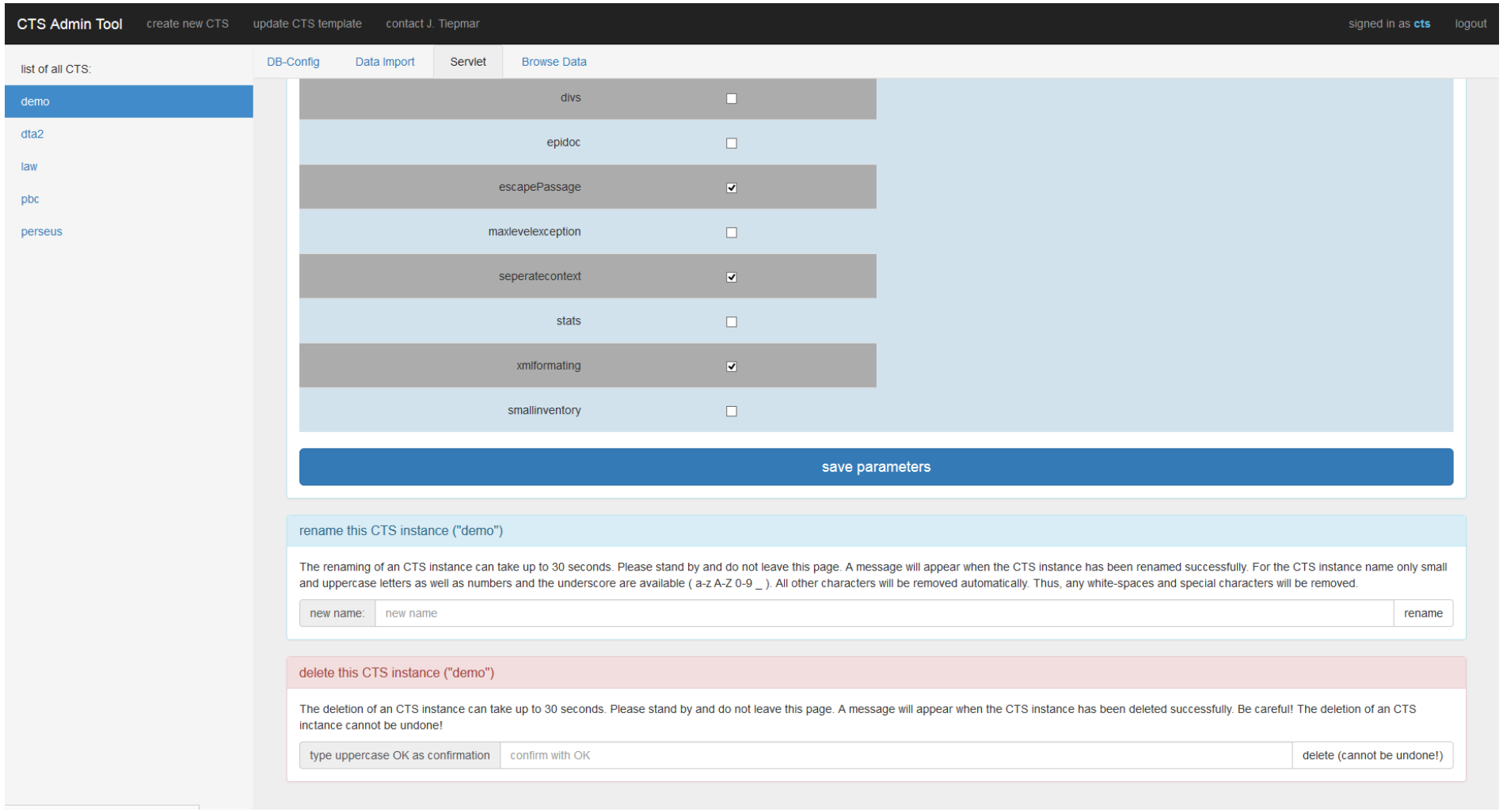
Term-Document Matrix

Token/Types per Document/Corpus

Document- and Termbased Pruning + lists of Stopwords

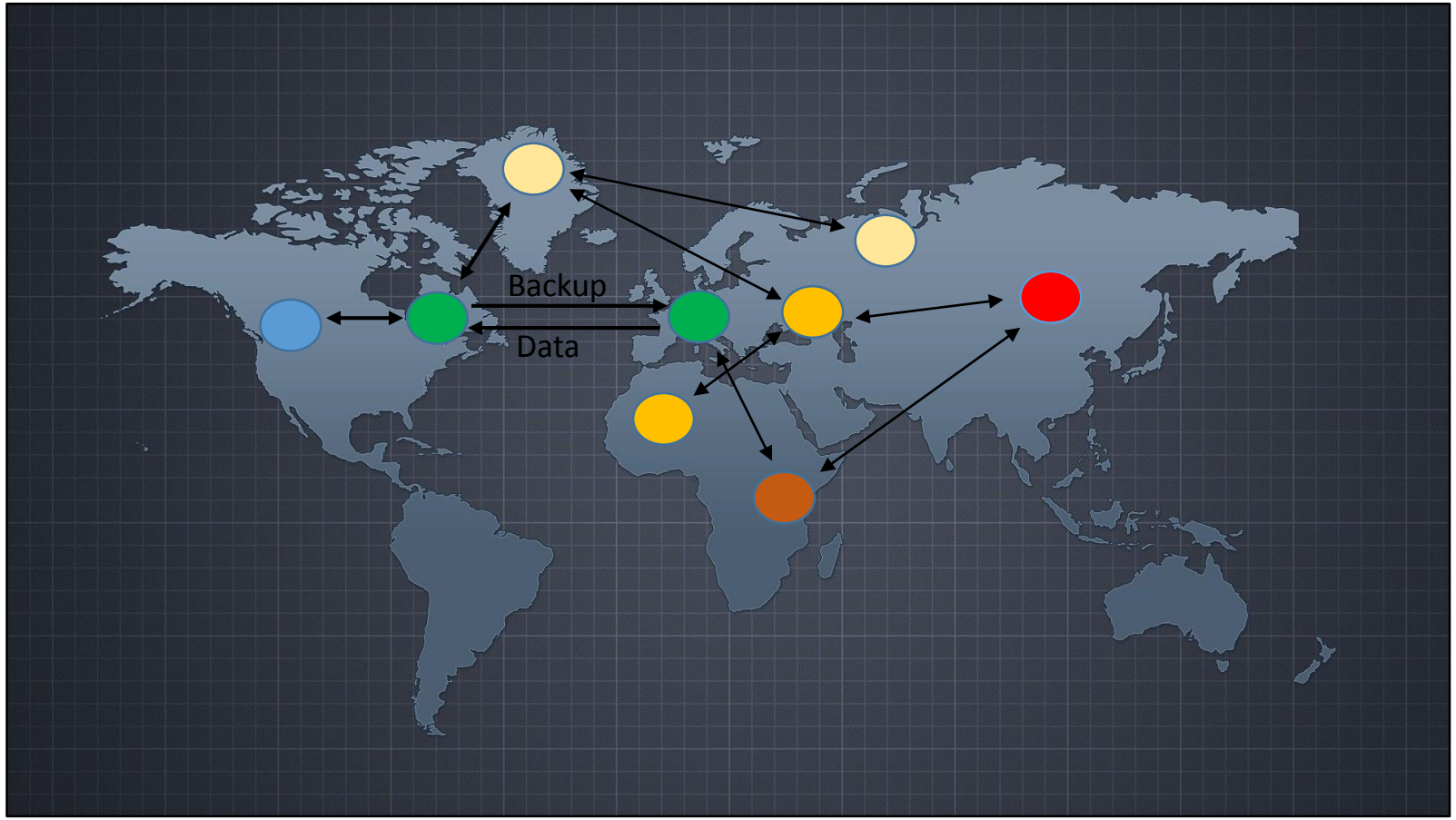
Tokensequence /(Kookurenz)

CTS Admin Tool



Implemented by Sascha Ludwig

Big Picture



global
decentralised
community organised
community backup‘ed
open access
standardized
persistent citable
easy to install
text repository
for browsing, searching
and analysis of text resources.

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