Canonical Text Service

Jochen Tiepmar
ScaDS, ASV, Uni Leipzig
Overview CTS

Canonical Text Services (CTS)

- protocol for a webbased citable text service
- Unique Identifiers (Unique Resource Name, 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

URL: http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-201773
Text as web resource

Included in project specific website

No -obvious- possibility for automated processing

Crawling can be problematic (request limit, advertisement, Javascript,...)
Erster Aufzug.

Erste Scene.

(Eine Terrasse vor dem Palast.)

(Bernardo und Francisco, zween Schildwachen, treten auf.)

Bernardo. Wer da?


Bernardo. Lang lebe der König!
Hamlet by Wieland

<div class="chapter" id="chap01">
<h3>ERSTER AKT</h3>
<p class="center">ERSTE SZENE</p>
<br />
<i>Helsingör. Eine Terrasse vor dem Schlosse</i></p>
<br />
<p><i>Francisco auf dem Posten, Bernardo tritt auf.</i></p>
<p>BERNARDO</p>
<p>Wer da?</p>
<p>FRANCISCO</p>
<p>Nein,
<em>mir</em> antwortet; steht und gebt Euch kund!</p>
<p>BERNARDO</p>
<p>Lang lebe der König!</p>
SCENE I. Elsinore. A platform before the castle.

FRANCISCO at his post. Enter to him BERNARDO

Who's there? Nay, answer me: stand, and unfold yourself.

Long live the king!
Enter Barnardo and Francisco, two sentinels.

BARNARDO Who's there?

Who's there?
Variety

No commonly used text format & access

No distinction between document structure (chapter), meta information (author) and text

Files are incompatible to other project workflows
ἦ οἵη προλιποῦσα δόμους καὶ πατρίδα γαῖναν

ἤλυθεν ἐς Θήβας μετ' ἀρήιον Ἀμφιτρύων

Ἀλκμήνη, θυγάτηρ λαοσσόου Ἠλεκτρύωνος·

ἥ ῥα γυναικῶν φῦλον ἐκαίνυτο θηλυτεράων

εἴδεΐ τε μεγέθει τε· νόον γε μὲν οὐ τις ἔριζε

τάων, ἃς θνηταὶ θνητοῖς τέκον εὐνηθεῖσαι.
Or like her who left home and country and came to Thebes, following warlike Amphitryon, even Alcmena, the daughter of Electryon, gatherer of the people. She surpassed the tribe of womankind in beauty and in height; and in wisdom none vied with her of those whom mortal women bore of union with mortal men. Her face and her dark eyes wafted such charm as comes from golden Aphrodite. And she so honored her husband in her heart as none of womankind did before her. In truth he had slain her noble father violently when he was angry about oxen; so he left his own country and came to Thebes and was suppliant to the shield-carrying men of Cadmus. There he dwelt with his modest wife without the joys of love, nor might he go in unto the neat-ankled daughter of Electryon until he had avenged the death of his wife’s great-hearted brothers and utterly burned with blazing fire the villages of the heroes, the Taphians and Teleboans; for this thing was laid upon him, and the gods were witnesses to it. And he feared their anger, and hastened to perform the great task to which Zeus had bound him. With him went the horse-driving Boeotians, breathing above their shields, and the Locrians who fight hand to hand, and the gallant Phocians eager for war and battle. And the noble son of Alcaeus led them, rejoicing in his host.

But the father of men and gods was forming another scheme in his heart, to beget one to defend against destruction gods and men who eat bread. So he arose from Olympus by night pondering guile in the deep of his heart, and yearned for the love of the well-girded woman. Quickly he came to Typhaonium, and from there again wise Zeus went on and trod the highest peak of Phicium: A mountain peak near Thebes which took its name from the Sphinx, called in <bibl n="Hes. Th. 326">Hes. Th. 326</bibl> &lt;foreign lang="greek">fi=c</foreign>&lt;/note&gt; there he sat and planned marvellous things in his heart. So in one night Zeus shared the bed and love of the neat-ankled daughter of Electryon and fulfilled his desire; and in the same night Amphitryon, gatherer of the people, the glorious hero, came to his house when he had ended his great task.
Sonnet 1

Vers 1

Word 1

No more be grieved at that which thou hast done,

Roses have thorns, and thorns have roses true,

And sweetest flowers do sometimes sting with rue,

And in the sunniest days do storms do come.
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:35

URL: http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-201773
CTS Requests

http://ctstest.informatik.uni-leipzig.de/perseus/cts/?request=GetPassage&urn=urn:cts:latinLit:phi1014.phi001.lat1:1.pr@magis-1.2.1@est

Server, Endpoint, Access Point
CTS Requests

http://ctstest.informatik.uni-leipzig.de/perseus/cts/?request=GetPassage&urn=urn:cts:latinLit:phi1014.phi001.lat1:1.pr@magis-1.2.1@est

Request

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

http://ctstest.informatik.uni-leipzig.de/perseus/cts/?request=GetPassage&urn=urn:cts:latinLit:phi1014.phi001.lat1:1.pr@magis-1.2.1@est

Parameters for request
  URN, (citation)level,…
O Tannenbaum, O Tannenbaum,
Wie treu sind deine Blätter.
Du grünst nicht nur zur Sommerzeit,
Nein auch im Winter wenn es schneit.
O Tannenbaum, O Tannenbaum,
Wie grün sind deine Blätter!

O Tannenbaum, O Tannenbaum,
Du kannst mir sehr gefallen!
Wie oft hat schon zur Winterszeit
Ein Baum von dir mich hoch erfreut!
O Tannenbaum, O Tannenbaum,
Du kannst mir sehr gefallen!

Dein Kleid will mich was lehren:
Die Hoffnung und Beständigkeit
Gibt Mut und Kraft zu jeder Zeit!
O Tannenbaum, O Tannenbaum,
Dein Kleid will mich was lehren.
O Tannenbaum, O Tannenbaum, (...) Wie grün sind deine Blätter! O Tannenbaum, O Tannenbaum, (...) Ein Baum von dir mich hoch erfreut!
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
Helen.
I will do, Mother! I know the substance of these epieiches, I would have long been at the side. As it was, my father's usual good fortune did not allow me to fear for him, nor to be overly anxious. Now that I have the knowledge, I will spare no pains to learn the whole truth in this matter.

Chorus
Go, then, my son. For prosperity yields advantage even for him who learns of it. (Euripides, on the side, and the Nurse into the house.)

Chorus
For with longing heart, as I hear, Despoina, this battle-price, now like some meanful bed, never restless eyes' longing that they might be without tears, but remaining a well remembered fear for her husband's travels is constantly affecting her anxious, widowed marriage-bed and in her youth anticipates misfortune.

Chorus
For just as one may well follow after boding-wrath and pacing over the wide deep before the tireless south-wind, or the north, or the great bull of the sea, many their back the hair of Cynoche, now wheat in. But some god stayeth him unerring from the house of Hades.

(To Despoina.)

Chorus
On, sale, my son. For prosperity yields advantage even for him who learns of it. (Euripides, on the side, and the Nurse into the house.)

Chorus

Copyright 2014 Leipzig University // Martin Reckziegel

2014 Leipzig University // Martin Reckziegel
CTS Cloning

URNs specify @n-Value of <div>s

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

-> Content of one CTS can be cloned

URNs can be used to filter data

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>
  </div1>
</passage>
CTS Cloning

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

Find comparable text parts in different documents

-> Comparison of documents/editions

-> Input for translation memory tools
Alignment

Naiv: Sentence tokenisation
- Sentences can be splitted/joined

Named Entity Recognition
- Statistical methods

Pre-calculation, Cross Comparison
- DTA: 3.5 GB * 5136 documents > 17 TB

Error propability increases with each text unit
Alignment

http://ctstest.informatik.uni-leipzig.de/Alignment/?cts=pbc&urn=urn:cts:pbc:deu.luther1545:1.1.4-1.1.7

Visualisierung: Stefan Jaenickes TRAViz

Und Gott sah das Licht, daß es gut war. Und Gott schied das Licht von der Finsternis, und nannte das Licht Tag, und die Finsternis Nacht. Und Gott sprach: Es werde eine Ausdehnung inmitten der Wasser, und sie schiede zwischen den Wassern.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>O Tannenbaum, O Tannenbaum,</td>
<td>O Christmas tree, O Christmas tree !</td>
<td>O abies, o abies,</td>
</tr>
<tr>
<td>1.2</td>
<td>Wie treu sind deine Blätter.</td>
<td>How are thy leaves so verdant !</td>
<td>Fidissima cum fronde,</td>
</tr>
<tr>
<td>1.3</td>
<td>Du grünst nicht nur zur Sommerzeit</td>
<td>Not only in the summertime.</td>
<td>Vires et ars artis tempore,</td>
</tr>
<tr>
<td>1.4</td>
<td>Nein auch im Winter wenn es schneit</td>
<td>But even in winter is thy prime.</td>
<td>Vires et ars artis tempore,</td>
</tr>
<tr>
<td>1.5</td>
<td>O Tannenbaum, O Tannenbaum,</td>
<td>O Christmas tree, O Christmas tree,</td>
<td>O abies, o abies,</td>
</tr>
<tr>
<td>1.6</td>
<td>Wie grün sind deine Blätter?</td>
<td>How are thy leaves so verdant !</td>
<td>Fidissima cum fronde !</td>
</tr>
<tr>
<td>1.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2.1</td>
<td>O Tannenbaum, O Tannenbaum,</td>
<td>O Christmas tree, O Christmas tree,</td>
<td>O abies, o abies,</td>
</tr>
<tr>
<td>1.2.2</td>
<td>Du kannst mir sehr gefallen!</td>
<td>Much pleasure doth thou bring me !</td>
<td>Tu mihi vidae places:</td>
</tr>
<tr>
<td>1.2.3</td>
<td>Wie oft hat schon zur Wintersonzeit</td>
<td>For every year the Christmas tree.</td>
<td>Quotiens natalis tempore</td>
</tr>
<tr>
<td>1.2.4</td>
<td>Ein Baum von dir mich hoch erfreut!</td>
<td>Brings to us all both joy and glee.</td>
<td>Dextevirasti me fronde,</td>
</tr>
<tr>
<td>1.2.5</td>
<td>O Tannenbaum, O Tannenbaum,</td>
<td>O Christmas tree, O Christmas tree,</td>
<td>O abies, o abies,</td>
</tr>
<tr>
<td>1.2.6</td>
<td>Du kannst mir sehr gefallen!</td>
<td>Much pleasure doth thou bring me !</td>
<td>Tu mihi vidae places:</td>
</tr>
<tr>
<td>1.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3.1</td>
<td>O Tannenbaum, O Tannenbaum,</td>
<td>O Christmas tree, O Christmas tree,</td>
<td>O abies, o abies,</td>
</tr>
<tr>
<td>1.3.2</td>
<td>Dein Kleid wil mich was leihen.</td>
<td>Thy candies shine out brighty !</td>
<td>Veste tua me doces:</td>
</tr>
<tr>
<td>1.3.3</td>
<td>Die Hoffnung und Beständigkeit</td>
<td>Each bough doth hold its tiny light.</td>
<td>Spes gignunt et constanta</td>
</tr>
<tr>
<td>1.3.4</td>
<td>Gott Mut und Kraft zu jeder Zeit!</td>
<td>That makes each toy to sparkle bright.</td>
<td>Semper vire, solac!</td>
</tr>
<tr>
<td>1.3.5</td>
<td>O Tannenbaum, O Tannenbaum,</td>
<td>O Christmas tree, O Christmas tree,</td>
<td>O abies, o abies,</td>
</tr>
<tr>
<td>1.3.6</td>
<td>Dein Kleid will mich was leihen.</td>
<td>Thy candies shine out brighty !</td>
<td>Veste tua me doces:</td>
</tr>
</tbody>
</table>

Canonical Text Service
Alignment via CTS

Alignment of URNs

urn:cts:demo:shakespeare.sonnets.en.1:3.2
urn:cts:demo:shakespeare.sonnets.de.1:3.2
urn:cts:demo:shakespeare.sonnets.fr.1:3.2

-> Error probability resets for each new chapter

-> fast & real time, no pre-calculation

...paper incoming...
### Alignment via CTS

Canonical Text Service


<table>
<thead>
<tr>
<th>Page</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.4.21</td>
<td>Et la nom de son frère était Jubaïl. Il fut père de tous ceux qui jouent de la harpe et du chalumeau.</td>
</tr>
<tr>
<td>1.4.22</td>
<td>Et Tolêla, elle aussi, enfanta Tubal-Cain, qui forgeait toute sorte d'instruments tranchants d'alait et d'acier, et le seur de Tubal-Cain fut Naema.</td>
</tr>
<tr>
<td>1.4.23</td>
<td>Et Lémech dit à ses femmes : Ada et Tolêla, écoutez ma voix, femmes de Lémech, prêtez l'oreille à ma parole : Oui ! J'ai bu un hommage pour ma blessure, et un jeune homme pour ma maîtrise.</td>
</tr>
<tr>
<td>1.4.24</td>
<td>Car si Cain est vengé sept fois, Lémech le sera soixante-dix-sept fois.</td>
</tr>
<tr>
<td>1.4.25</td>
<td>Et Adam connut encore sa femme, et elle enfanta un fils qu'elle appela Seth (remplaçant) ; car Dieu dit : Il me donnera un autre fils au lieu d'Abel, car Cain a bu.</td>
</tr>
<tr>
<td>1.4.26</td>
<td>Et un fils naquit aussi à Seth, et il l'appela Enos. Alors on commença à invoquer le nom de l'Ehelül.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Page</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>Voici le livre de la proéminence d'Adam. Au jour où Dieu créa l'homme, il le fit à la ressemblance de Dieu.</td>
</tr>
<tr>
<td>1.5.1</td>
<td>Il les créa mâle et femelle, et les bénit, et leur donna le nom d'Homme, au jour qu'ils furent créés.</td>
</tr>
<tr>
<td>1.5.2</td>
<td>Or, Adam vécut cent trente ans, et endangea un fils à sa ressemblance, selon son image, et il lui donna le nom de Seth.</td>
</tr>
<tr>
<td>1.5.3</td>
<td>Et les jours d'Adam, après qu'il eut engendré Seth, furent de huit cent ans, et il engendra des fils et des filles.</td>
</tr>
<tr>
<td>1.5.4</td>
<td>Tout le temps qu'Adam vécut, fut donc de neuf cents trente ans ; puis il mourut.</td>
</tr>
<tr>
<td>1.5.5</td>
<td>Et Seth vécut cent cinq ans, et endenga Enos.</td>
</tr>
</tbody>
</table>

URL: http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-201773
Topic Models

Find abstract topics/thematic structure in text corpus

Easiest Algorithm:
Latent Dirichlet Allocation (LDA)

Latent:
Documents share prepared set of topics
word/document ratio differs
(90% Evolution, 10% Disease)

( David M. Blei: Introduction to Probabilistic Topic Models
Topic Models

What if the topics are unknown? 
Topics can be generated with 
probability for common occurrence

<table>
<thead>
<tr>
<th>labor</th>
<th>jury</th>
</tr>
</thead>
<tbody>
<tr>
<td>workers</td>
<td>trial</td>
</tr>
<tr>
<td>employees</td>
<td>crime</td>
</tr>
<tr>
<td>union</td>
<td>defendant</td>
</tr>
<tr>
<td>employer</td>
<td>sentencing</td>
</tr>
<tr>
<td>work</td>
<td>judges</td>
</tr>
<tr>
<td>job</td>
<td>punishment</td>
</tr>
<tr>
<td>bargaining</td>
<td>evidence</td>
</tr>
<tr>
<td>unions</td>
<td>sentence</td>
</tr>
<tr>
<td>collective</td>
<td>offense</td>
</tr>
</tbody>
</table>
Topic Models

- Manage and order unknown text corpus
- Help explore unknown text collections
  Search engine that lets you “zoom in or out” of topics
  Reveal different or further aspects of one topic
- Find semantic relations between texts
  Use “topic-fingerprint” of the user to propose documents

<table>
<thead>
<tr>
<th>Genetics</th>
<th>Evolution</th>
<th>Disease</th>
<th>Computers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Evolution</td>
<td>Disease</td>
<td>Computer</td>
</tr>
<tr>
<td>Genome</td>
<td>Species</td>
<td>Host</td>
<td>Model</td>
</tr>
<tr>
<td>DNA</td>
<td>Organism</td>
<td>Bacteria</td>
<td>Information</td>
</tr>
</tbody>
</table>
Topic Models

- **Data:** TED Subtitle Corpus, 51770 documents, 105 languages, 1938 english documents, big variety of topics

- **Tokens for Topic 25**
  
  25 -> brain, and, neurons, brains, memory, consciousness, autism, cells, mind, sleep

- **Texts about topic 25**
  
  "Re-engineering the brain", "Parkinson's, depression and the switch that might turn them off", "A prosthetic eye to treat blindness", "How your brain tells you where you are", "The mysterious workings of the adolescent brain", "A monkey that controls a robot with its thoughts. No, really.", "Your brain is more than a bag of chemicals", "A mouse. A laser beam. A manipulated memory", "3 clues to understanding your brain", "A neural portrait of the human mind", "Growing evidence of brain plasticity", "How we read each other's minds", "I am my connectome"

- **Topics for text "A monkey that controls a robot with its thoughts. No, really."**
  
  4 -> and, light, video, now, game, through, show, different, also, look
  14 -> and, technology, computer, new, use, machine, now, system, design, build
  19 -> now, thing, little, got, something, lot, look, kind, doing, say
  25 -> brain, and, neurons, brains, memory, consciousness, autism, cells, mind, sleep
Topics generated from TED Subtitles

0  car, cars, fly, miles, road, drive, around, driving, only, light
1  and, percent, there, than, now, today, two, year, change, look
2  music, sound, play, and, applause, song, now, hear, voice, thank
3  and, africa, india, countries, there, children, country, poor, many, African
4  and, light, video, now, game, through, show, different, also, look
5  and, city, building, cities, new, space, design, built, buildings, project
6  laughter, applause, now, three, two, yeah, four, good, five, okay
7  and, cells, body, stem, cell, tissue, cancer, skin, blood, legs
8  and, feel, life, human, self, social, good, being, happiness, experience
9  she, her, women, and, men, woman, girls, girl, young, sex
10 his, she, him, there, did, went, back, her, after, didn
11 data, internet, information, media, online, web, and, google, new, now
12 and, species, dna, life, human, animals, humans, evolution, bacteria, different
13 and, there, now, than, even, many, question, different, example, much
14 and, technology, computer, new, use, machine, now, system, design, build
15 food, and, eat, bees, plants, plant, feed, farmers, eating, bread
16 and, there, ocean, water, sea, fish, down, ice, animals, species
17 and, god, his, story, say, man, love, book, him, says
18 now, thing, little, got, something, lot, look, kind, doing, say
19 school, kids, and, children, students, education, child, she, learning, teachers
20 and, energy, water, oil, use, need, carbon, climate, power, gas
21 life, day, even, every, need, thank, applause, many, than, most
22 and, money, business, dollars, companies, company, need, market, work, value
23 and, art, design, work, made, kind, laughter, film, sort, new
24 brain, and, neurons, brains, memory, consciousness, autism, cells, mind, sleep
25 and, power, country, political, china, chinese, democracy, government, american, states
26 war, violence, security, there, police, military, killed, afghanistan, conflict, peace
27 and, universe, there, earth, space, planet, life, stars, sun
28 and, health, cancer, disease, patients, care, medical, patient, drug, treatment
Topics generated from arab TED Subtitles
Topics generated from engl. Perseus
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.

URL: http://hdw.eweb4.com/out/1369880.html
Jochen Tiepmar
ScaDS BigData Competence Centre
Ritterstrasse 9-13, 2.OG
04109 Leipzig
jtiepmar@informatik.uni-leipzig.de
This work was funded by the German Federal Ministry of Education and Research within the project Competence Center for Scalable Data Services and Solutions (ScaDS) Dresden/Leipzig (BMBF 01IS14014B)