

# European Summer School 2017

Text Mining with Canonical Text Services  
Theory Session 4 – TEI as input for CTS

# Structural Meta Information

```
ur:cts:latinLit:phi0119_phi0001_eng1:
1
<head>
  THE PROLOGUE.
</head>
<sp>
  <speaker>
    MERCURY
  </speaker>
<p>
  As, in purchasing and selling your merchandize
  <milestone n="1" unit="TLN line">
  </milestone>
  <note anchored="yes">
    <lemma targOrder="U">
      Merck indize
    </lemma>
    : "
    <foreign lang="la">
      Mercimoniis
    </foreign>
    ." Mercury was the God of trading and merchandize, and was said to have received his name from the Latin
    word "
    <foreign lang="la">
      merx
    </foreign>
    ." See the tradesman's prayer to him in the Fasti of Ovid,
    <bibl n="Ov. Fast. 5.685">
      B. v., 1. 682
    </bibl>
  </note>
  , you are desirous to render me propitious to your bargains, and that I should assist you in all things;
```

Structure Information

No Structural Information

Structure Information is encoded as Meta Information

CTS uses structural markup as anchor for CTS URNs

Computer can not distinguish between structural and non structural markup

-> Needs to be configured

# TEI as Import

- TEI/XML based import process
  - Not technically required, support for other formats can be added
- Filepath is used for work/document component
- Information in <TEIHeader> used for document level meta information
  - <author>, <date>, <license>, <source> | <idno> | <publisher>, <title>
- CTS URNs are built based on configured “structure XML tag path” in <text> element
- Other XML is considered as text
- TEI/XML files are not required after import

# “structure XML tag path”

- div1/\*,div2/paragraph,div3 | p/\*,div4 | div/sentence
  - XML path is traversed according to comma separated elements
  - Type of the structural unit can be configured
- div1/\* -> <div1> = anchor point, @type-value = type
- div2/paragraph-> <div2> = anchor point, “paragraph” = type
- div3 | p/\* -> <div3> or <p> anchor point, @type-value = type
- div4 | div/line -> <div4> or <div> = anchor point, “line” = type

# TEI Structure -> CTS URNs

```
- <text>
- <body>
- <div1 type="book" n="1">
- <div2 type="chapter" n="1">
  <div3 type="sentence" xml:lang="cym" n="1">Yn y dechr
- <div3 type="sentence" xml:lang="cym" n="2">
  A'r ddaear oedd afluniaidd a gwag , a thywyllwch oedd ar
  dyfroedd .
  </div3>
- <div3 type="sentence" xml:lang="cym" n="3">
  A Duw a ddywedodd , Bydded goleuni , a goleuni a fu .
  </div3>
- <div3 type="sentence" xml:lang="cym" n="4">
  A Duw a welodd y goleuni , mai da oedd : a Duw a wahar
  </div3>
- <div3 type="sentence" xml:lang="cym" n="5">
  A Duw a alwodd y goleuni yn Ddydd , a'r tywyllwch a al
  </div3>
- <div3 type="sentence" xml:lang="cym" n="6">
  Duw hefyd a ddywedodd , Bydded ffurfafen yng nghanol y
  .
  </div3>
- <div3 type="sentence" xml:lang="cym" n="7">
  A Duw a wnaeth y ffurfafen , ac a wahanodd rhwng y dyfr
  y bu .
  </div3>
```

```
urn:cts:pbcbible.parallel.cym.morgan1804:      edwrap
urn:cts:pbcbible.parallel.cym.morgan1804:1      book
urn:cts:pbcbible.parallel.cym.morgan1804:1.1    chapter
urn:cts:pbcbible.parallel.cym.morgan1804:1.1.1  sentence
urn:cts:pbcbible.parallel.cym.morgan1804:1.1.2  sentence
urn:cts:pbcbible.parallel.cym.morgan1804:1.1.3  sentence
urn:cts:pbcbible.parallel.cym.morgan1804:1.1.4  sentence
urn:cts:pbcbible.parallel.cym.morgan1804:1.1.5  sentence
urn:cts:pbcbible.parallel.cym.morgan1804:1.1.6  sentence
urn:cts:pbcbible.parallel.cym.morgan1804:1.1.7  sentence
```

# Meta Information

- author, date, license, source | idno | publisher and title are stored for every document
- type, text content and language tag are stored for every CTS URN

# Data Import Configuration

CTS Admin Tool   create new CTS   update CTS template   admin area   help ▾   signed in as **cts**   logout

CTS instances:

- ctstm\_demo
- demo**

DB-Config   **Data Import**   Servlet   Browse Data

contentType	<input type="text" value="xml"/>	"xml" or "plain" If xml, then the textcontent will be XML validated. If it is not valid XML, the value for the individual document is set to "plain" This information is included in the generated text inventory.
sourceType	<input type="text" value="cts"/>	"oai" or "local" or "cts" "cts" uses another instance of this CTS implementation (CTS Cloning). "local" uses the documents on the file system. "oai" requires a specific OAI-PMH set up.
sourceDir	<input type="text" value="urn:cts:demo:"/>	root-address for source files E.g. if(sourceType == "local") -> "E:/Files/test" E.g. if(sourceType == "cts") -> "urn:cts:demo:" or "http://ctstest.informatik.uni-leipzig.de/demo/cts/?request=GetCapabilities"
force4workparts	<input checked="" type="checkbox"/>	If this is set to true, any document that would result in more than 4 work parts is ignored. For example a document with the URN urn:cts:demo:author.document.en.2.transcript: would be ignored.
folderDelimiter	<input type="text" value="-"/>	Character to use to separate the work parts in the CTS URN. Works together with parameter ignoreFolders.

Here you can decide between local import and import from another CTS instance

If you do not want to use the folder structure for the work components of the CTS URNs

# Data Import Configuration

CTS Admin Tool   create new CTS   update CTS template   admin area   help ▾   signed in as cts   logout

CTS instances:   DB-Config   Data Import   Servlet   Browse Data

csttm\_demo

demo

ignoreFolders    Use folders to separate the work parts in the CTS URN. Works together with parameter folderDelimiter.

docCount    If CTS Cloning is used, the number of documents can be limited here.

includeURNsWith    If CTS Cloning is used, this parameter limits the documents to those with URNs that includes a certain sub string. Multiple filters can be combined with ampersand "&". Can be combined with the parameter excludeURNsWith.

excludeURNsWith    If CTS Cloning is used, this parameter limits the documents to those with URNs that do NOT include a certain sub string. Multiple filters can be combined with ampersand "&". Can be combined with the parameter includeURNsWith.

ctspath    The XML path that specifies the structural elements in the texts.  
div1/\* -> uses "chapter" as type for textchunk  
div1/sentence -> uses "sentence" as type for textchunk  
front|body|div|/\* -> any of the tags , ,  
is used as the next structural element.  
front|body|div|/\*, div1|p|paragraph -> any of the tags , ,  
is used as the next structural element. and  
are used on the 2nd citation level. The type of citation level 2 is always "paragraph".

If you do not want to use the folder structure for the work components of the CTS URNs

“structure XML tag path“



# Data Import Configuration

CTS Admin Tool   create new CTS   update CTS template   admin area   help ▾   signed in as **cts**   logout

CTS instances:   **DB-Config**   Data Import   Servlet   Browse Data

ctstm\_demo

**demo**

namespace	<input type="text"/>	Namespace of CTS URNs for local import. In the URN urn:cts:demo:author.document.en.1;, the namespace is "demo".
urnstart	<input type="text" value="urn:cts"/>	Beginning of CTS URN. Generally should be set to "urn:cts:".
sentenceSegmentation	<input type="checkbox"/>	The lowest found text chunks will be further tokenized based on a set of sentence segmentation rules.
lineSegmentation	<input type="checkbox"/>	The lowest found text chunks will be further tokenized based on line segmentation.
storeNSAsMeta	<input checked="" type="checkbox"/>	Namespaces are stored in the database. No reason to change it. Only relevant for local import
createMissingN	<input checked="" type="checkbox"/>	Automatically create the identifier for the structural elements using an increasing integer.

If no namespace is set, the first component of the work part will become the namespace

A basic sentence or line segmentation based on a German sentence segmentation rule set.

If you do not have @n-values in your markup

# Data Import Configuration

CTS Admin Tool   create new CTS   update CTS template   admin area   help ▾   signed in as **cts**   logout

CTS instances:   **DB-Config**   Data Import   Servlet   Browse Data

ctstm\_demo

**demo**

ignoreNFromFile	<input type="checkbox"/>	Ignore the n values from the TEI/XML documents. For example if they include duplicates or serve a different purpose.
MissingNStartWith	<input type="text"/>	This will be added to the generated identifier values.
repairDuplicateURNs	<input type="checkbox"/>	Experimental. Try to repair duplicate URNs by appending something. Should be deactivated.
duplicateMarker	<input type="text" value="_dupl"/>	This will be added to the repaired URNs.
useOnlyLeafNodes	<input checked="" type="checkbox"/>	Experimental. Should be activated.
debug	<input type="checkbox"/>	Some technical info during the import process.
languageTag	<input type="text"/>	Overwrites any language info from the documents. Leave empty if language information from document should be used.

If you have @n-values but do not want to use them for CTS URNs

This can solve some issues with duplicate @n-values but should generally not be used.

# Possible Error Messages

- MySQLConstraintException Duplicate Value for urn
  - @n-value is not unique in its context and would create a duplicate URN
  - Make @n-value unique or activate parameters ignoreNFromFile and createMissingN
- Something with SAXParserException
  - The XML file is somehow invalid. For example there may be a missing closing tag
- Value too big for column type text
  - Text content of a specific static URN is very large. Maybe you forgot to configure the more fine grained XML paths?

# Contact

Jochen Tiepmar

E-Mail: [jtiepmar@informatik.uni-leipzig.de](mailto:jtiepmar@informatik.uni-leipzig.de)

Scalable Data Solutions (ScaDS) Leipzig

Universität Leipzig

Ritterstraße 9-13

04109 Leipzig

