European Summer School 2017

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

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

As, in purchasing and selling your merchandise,
<milestone n="1" unit="TLN line">
</milestone>

<note anchored="yes">
  <lemma tauOrder="J">
    | Mercure Indice
  </lemma>
  |
  <foreign lang="la">
    | Mercimonia
  </foreign>
  ." Mercury was the God of trading and merchandise, and was said to have received his name from the Latin word "
  <foreign lang="la">
    | merc.
  </foreign>
  ." See the trademan's prayer to him in the Fasti of Ovid,
  <bibl n="Mo. Fast. 5.685">
    | B. v., 1. 892
  </bibl>
  .
</note>

"you are desirous to render me propitious to your bargains, and that I should assist you in all things:"

ESU CTS Workshop 2017 - Jochen Tiepmar
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
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
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
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

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

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

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