UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

ASV Automatische Sprachverarbeitung

European Summer School 2017

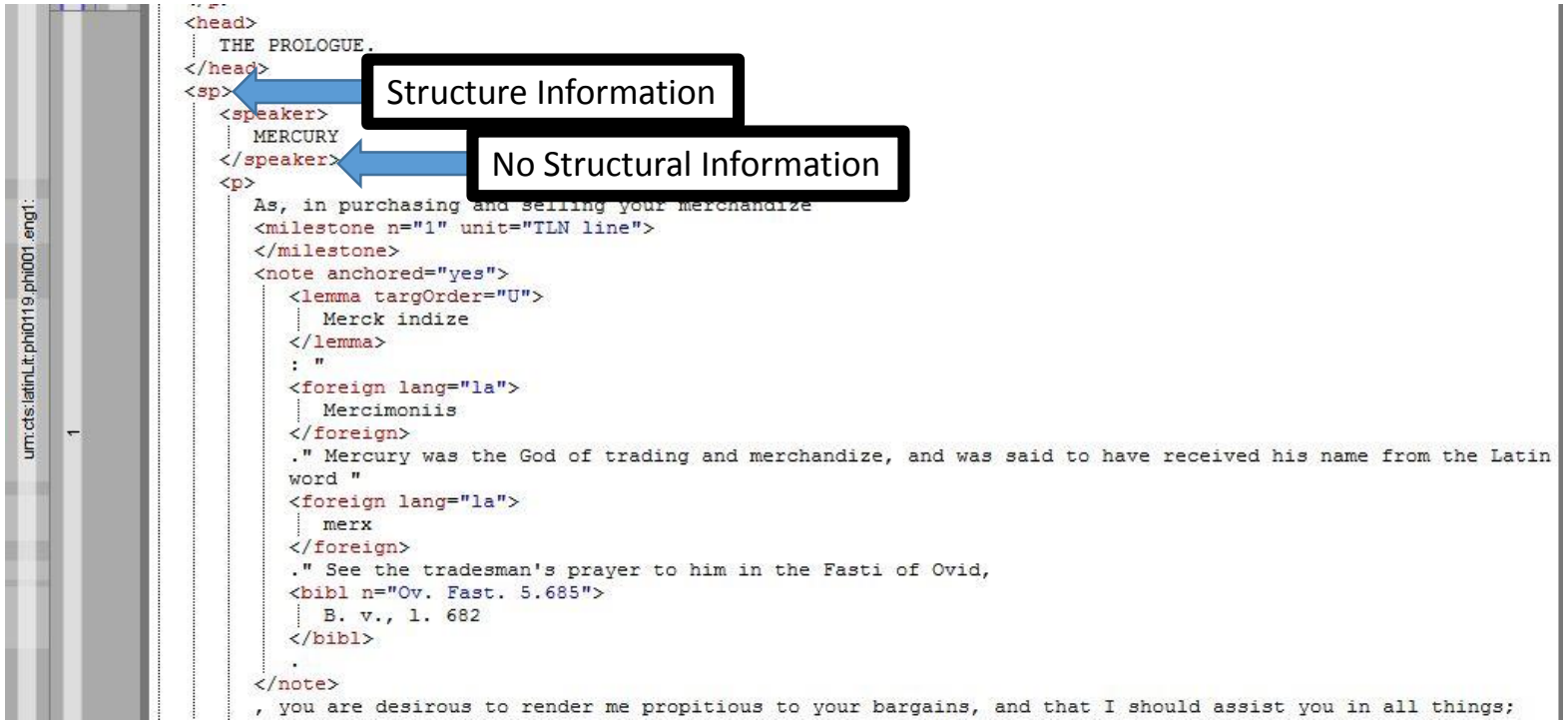
Text Mining with Canonical Text Services

Theory Session 4 – TEI as input for CTS

Federal Ministry of Education and Research

CLARIN-D

Structural Meta 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 seperated 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 = anchor point, “line“ = type

TEI Structure -> CTS URNs

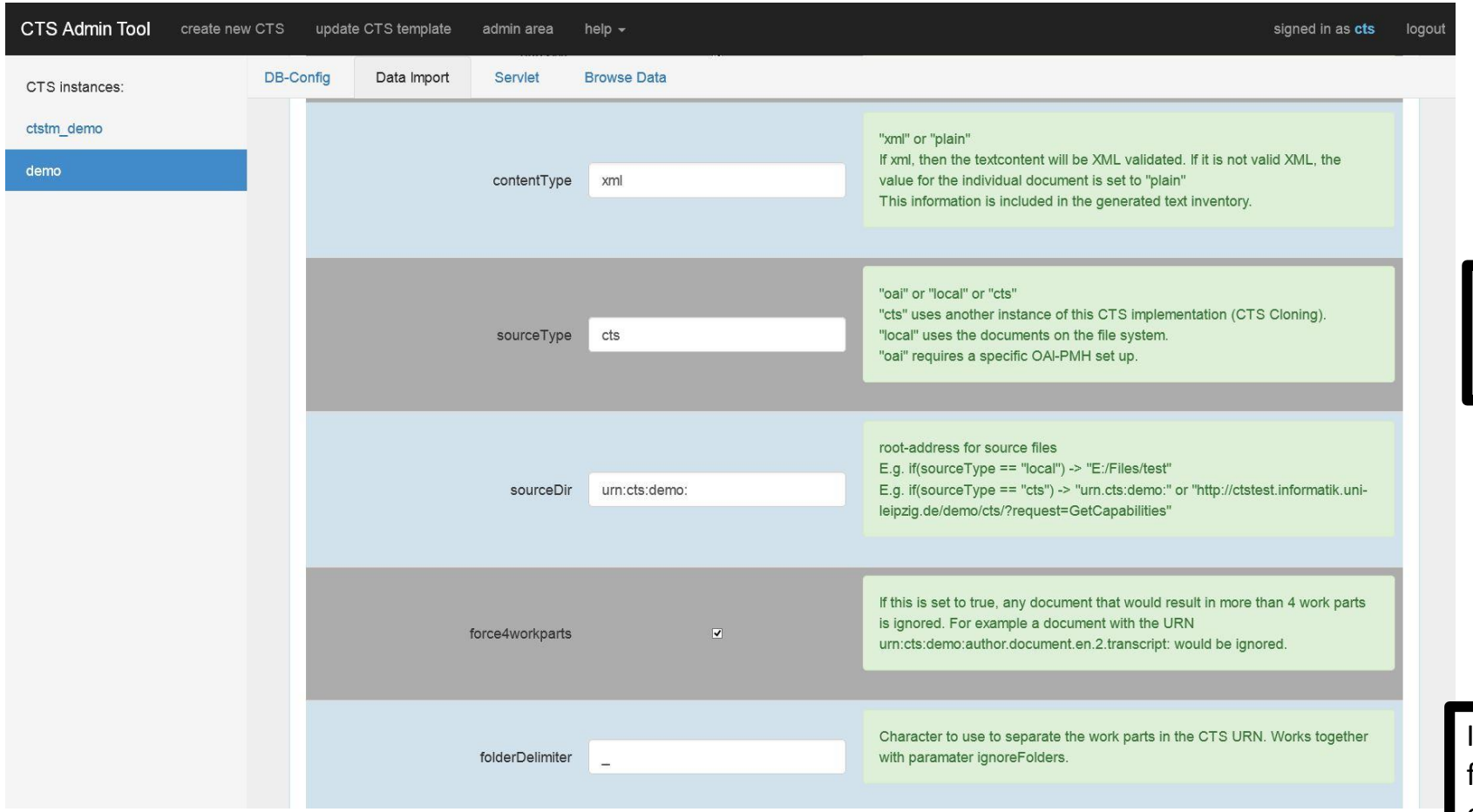
```
– <text>
  – <body>
    – <div1 type="book" n="1">
      – <div2 type="chapter" n="1">
          <div3 type="sentence" xml:lang="cym" n="1">Yn y dechr
        – <div3 type="sentence" xml:lang="cym" n="2">
            A'r ddaear oedd afluniaidd a gwag , a thywyllwch oedd ar
            dyfroedd .
          </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
            .
          </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:pbc:bible.parallel.cym.morgan1804:	edwrap
urn:cts:pbc:bible.parallel.cym.morgan1804:1	book
urn:cts:pbc:bible.parallel.cym.morgan1804:1.1	chapter
urn:cts:pbc:bible.parallel.cym.morgan1804:1.1.1	sentence
urn:cts:pbc:bible.parallel.cym.morgan1804:1.1.2	sentence
urn:cts:pbc:bible.parallel.cym.morgan1804:1.1.3	sentence
urn:cts:pbc:bible.parallel.cym.morgan1804:1.1.4	sentence
urn:cts:pbc:bible.parallel.cym.morgan1804:1.1.5	sentence
urn:cts:pbc:bible.parallel.cym.morgan1804:1.1.6	sentence
urn:cts:pbc:bible.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

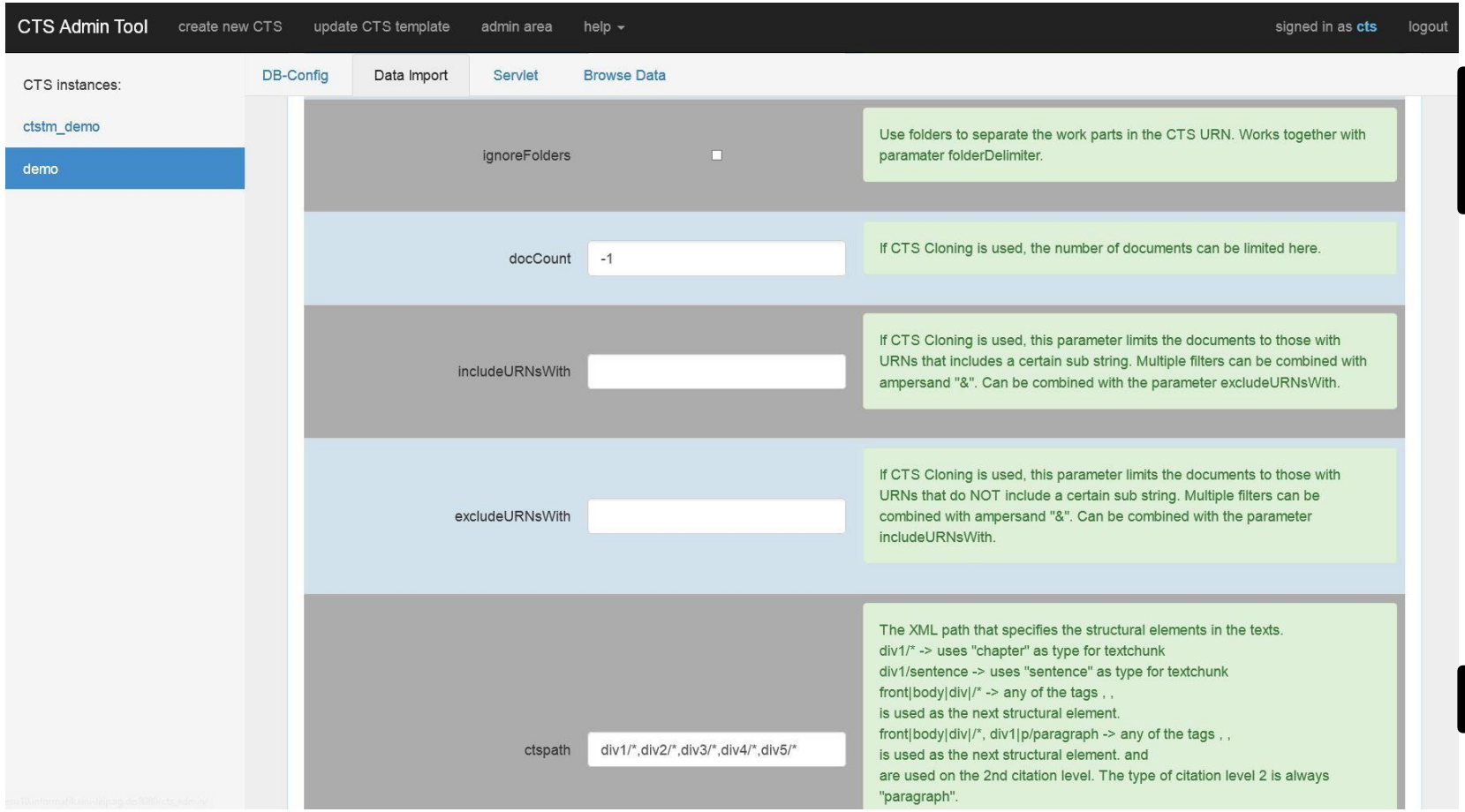
Data Import Configuration



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

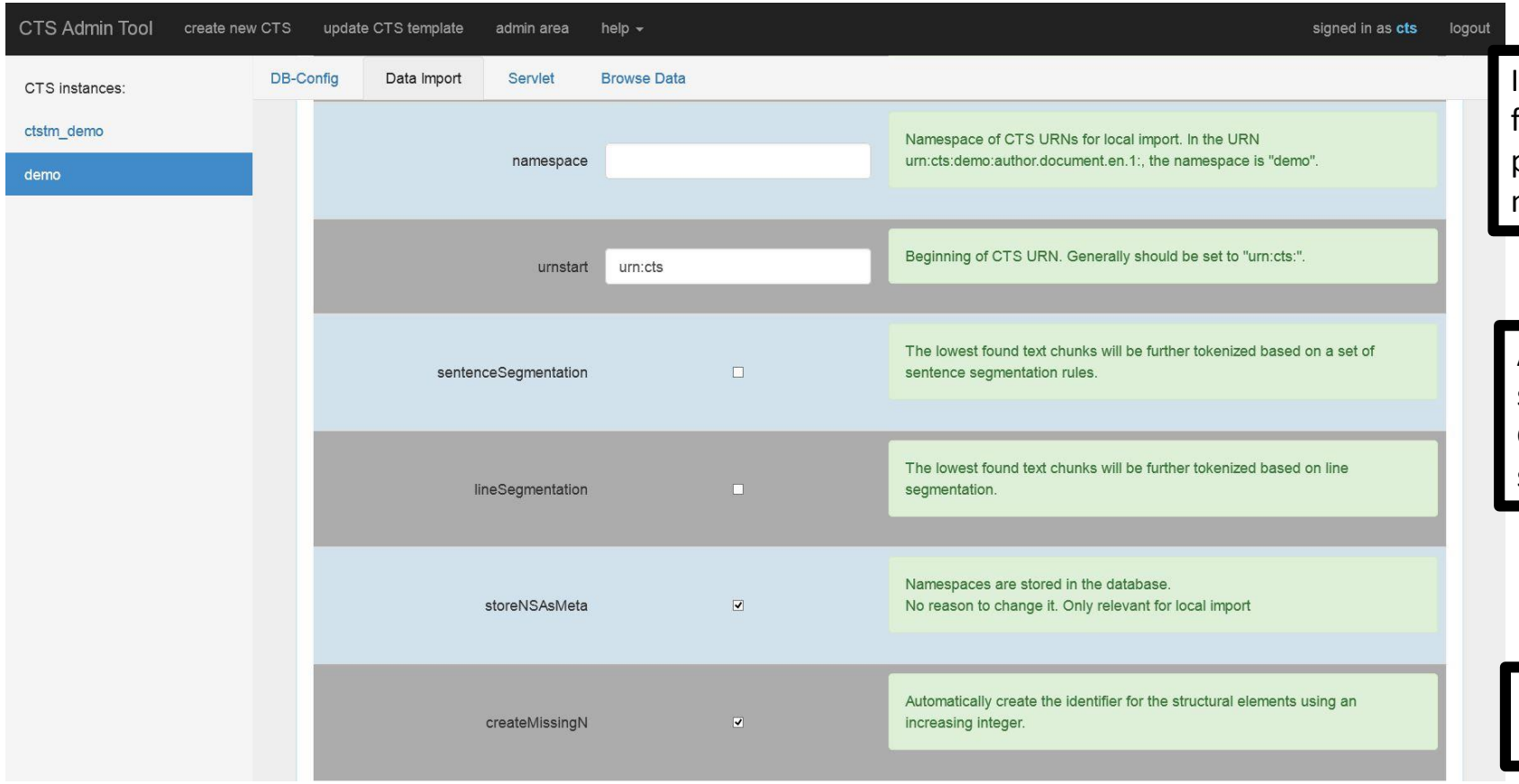
Data Import Configuration



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

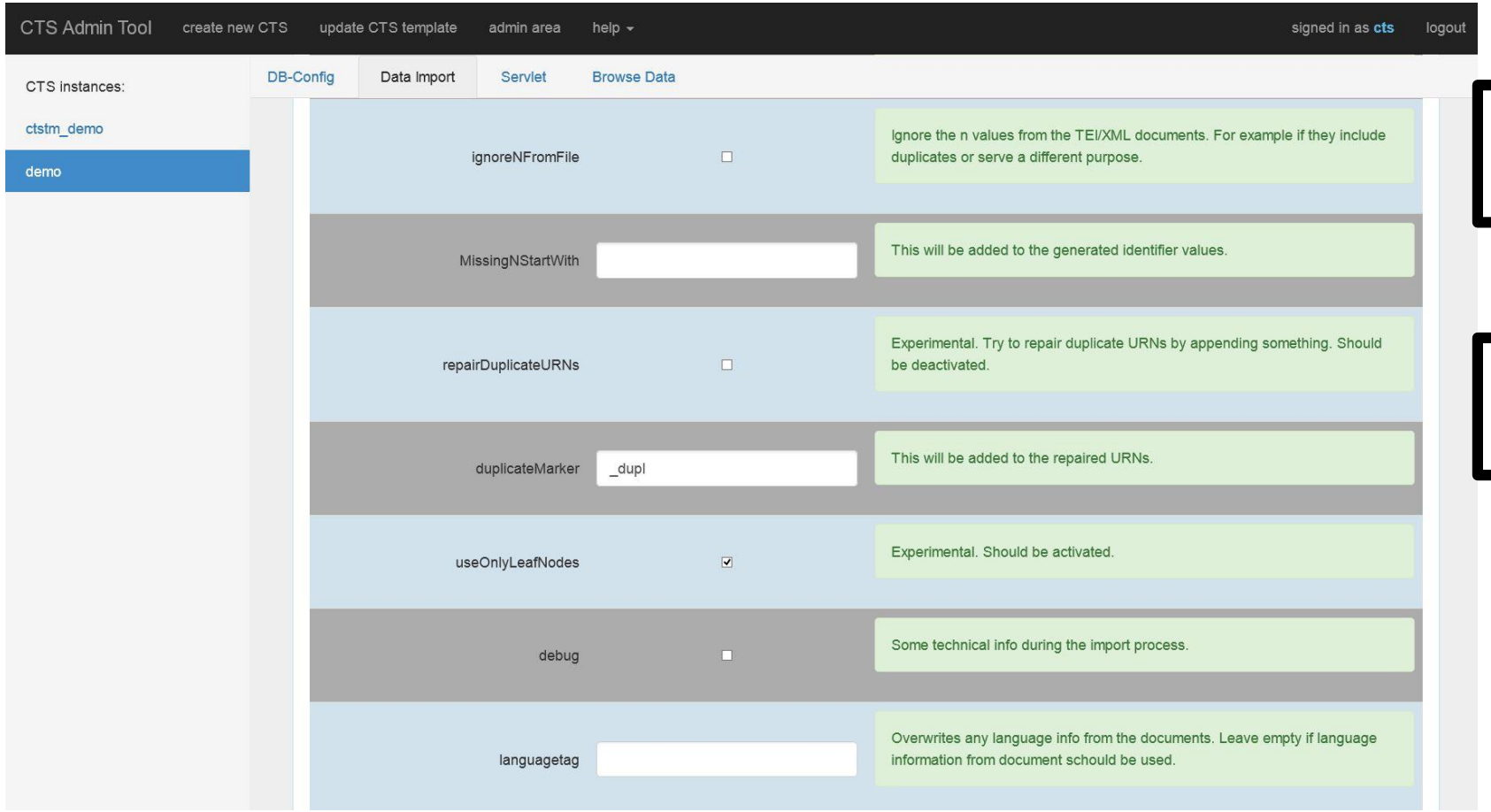


If no namespace is set, the first component of the work part will become the namespace

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