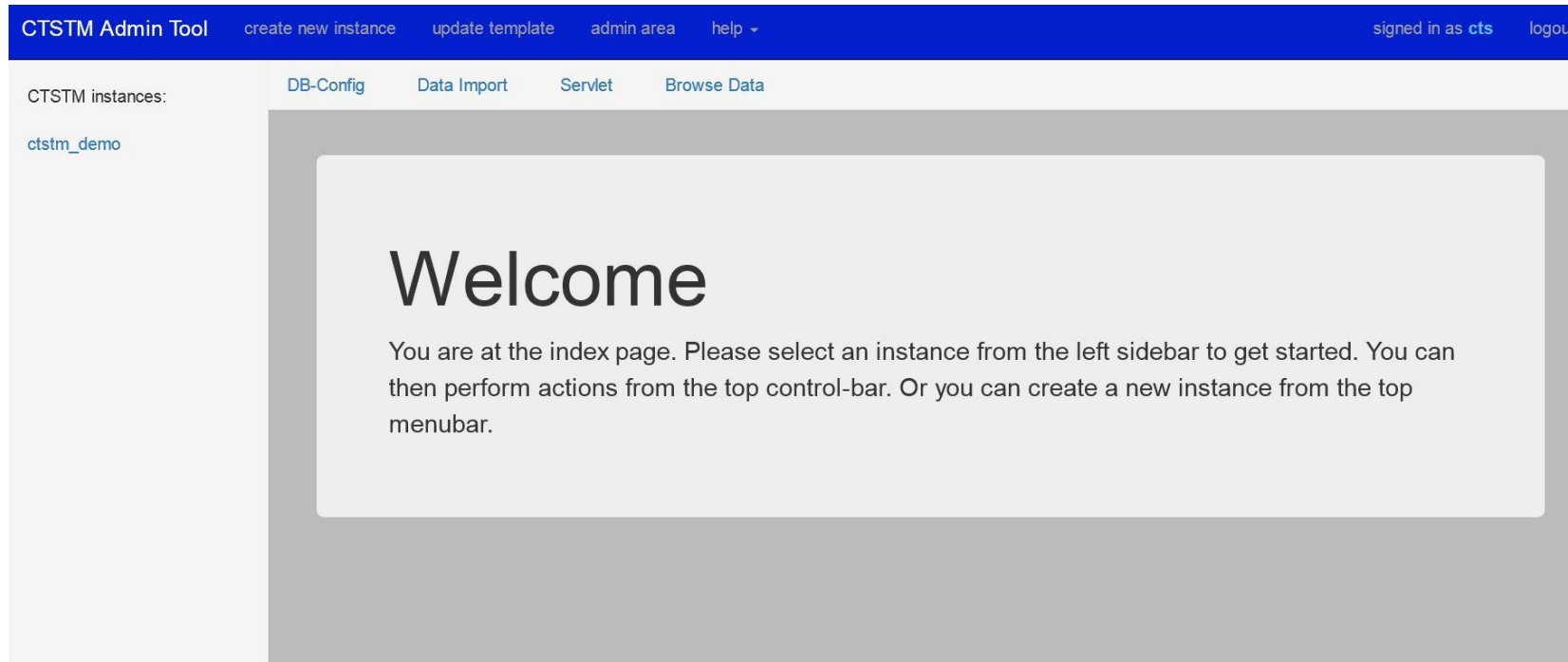


# European Summer School 2017

Text Mining with Canonical Text Services  
Theory Session 6 – Canonical Text Miner

# Overview



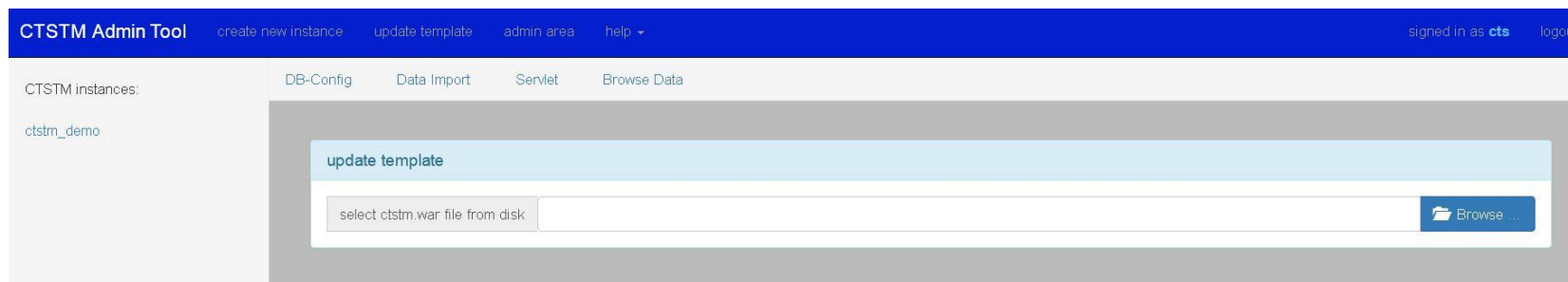
Can be deleted and redeployed without any impact on the CTS instances (In case the admin user is deleted or something else breaks).

Improvised „Re-Skin“ of CTS Admin Tool

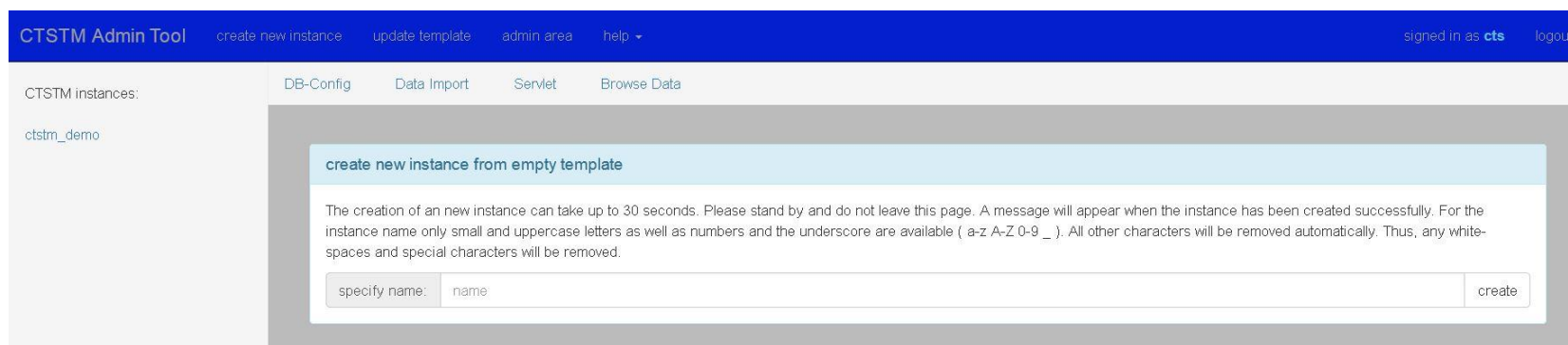
# Overview

The screenshot displays the CTSTM Admin Tool interface. At the top, a blue navigation bar contains the text "CTSTM Admin Tool" and several menu items: "create new instance", "update template", "admin area", and "help". On the right side of this bar, it shows "signed in as cts" and a "logout" link. Below the navigation bar, a secondary bar lists management tools: "DB-Config", "Data Import", "Servlet", and "Browse Data". A left sidebar shows "CTSTM instances:" followed by a link to "ctstm\_demo". The main content area features a large "Welcome" heading and a paragraph of instructions: "You are at the index page. Please select an instance from the left sidebar to get started. You can then perform actions from the top control-bar. Or you can create a new instance from the top menubar." Four callout boxes with arrows point to specific parts of the interface: "Administration Tools" points to the top right navigation bar; "CTSTM Management Tools" points to the secondary management tool bar; "CTSTM Instances" points to the sidebar link; and "Working Area" points to the main content area.

# Create New Instance / Update .war File



The screenshot shows the 'update template' page in the CTSTM Admin Tool. The top navigation bar includes 'CTSTM Admin Tool', 'create new instance', 'update template', 'admin area', and 'help'. The user is signed in as 'cts'. The left sidebar shows 'CTSTM instances:' with 'ctstm\_demo'. The main content area has tabs for 'DB-Config', 'Data Import', 'Servlet', and 'Browse Data'. The 'update template' section contains a text input field with the placeholder 'select ctstm.war file from disk' and a 'Browse ...' button.



The screenshot shows the 'create new instance from empty template' page in the CTSTM Admin Tool. The top navigation bar is identical to the previous screenshot. The left sidebar shows 'CTSTM instances:' with 'ctstm\_demo'. The main content area has tabs for 'DB-Config', 'Data Import', 'Servlet', and 'Browse Data'. The 'create new instance from empty template' section contains a text input field with the placeholder 'specify name: name' and a 'create' button. Below the input field, there is a note: 'The creation of a new instance can take up to 30 seconds. Please stand by and do not leave this page. A message will appear when the instance has been created successfully. For the instance name only small and uppercase letters as well as numbers and the underscore are available ( a-z A-Z 0-9 \_ ). All other characters will be removed automatically. Thus, any white-spaces and special characters will be removed.'

Old .war files are still available in admin area after update

New .war file must be selected in admin area

Updating the template does not update the CTS instances (-> Servlet Menu)

# Admin Area / User Management

The screenshot shows the 'CTSTM Admin Tool' interface. The top navigation bar includes 'create new instance', 'update template', 'admin area', and 'help'. The user is signed in as 'cts'. The main content area is titled 'CTSTM instances:' and shows 'ctstm\_demo'. Below this, there are tabs for 'DB-Config', 'Data Import', 'Servlet', and 'Browse Data'. The 'help menu' section is active, displaying a text area with the following JSON configuration:

```
[
  {
    "text": "Contact",
    "href": "mailto:jtiepmar@informatik.uni-leipzig.de"
  },
  {
    "text": "Another address",
    "href": "http://www.example.org"
  }
]
```

Buttons for 'save' and 'restore default' are visible at the bottom of the configuration area.

Translates to

This close-up shows the 'help' dropdown menu. The 'admin area' and 'help' tabs are visible at the top. The dropdown list contains the following items:

- admin area
- help
- Data Import
- contact
- CTS iterator
- SetCapabilities
- Browse
- Candi

# Admin Area / User Management

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

CTSTM instances:   DB-Config   Data Import   Servlet   Browse Data

ctstm\_demo

save   restore default entries

### user management

Name	Role	New PW	New PW rep.	Save	Del.
cts	Admin ▾	<input type="text"/>	<input type="text"/>		
<input type="text" value="new user"/>	User ▾	<input type="text" value="password"/>	<input type="text" value="repeat password"/>		

### Versioned Servlets

Name	Active	Delete
ctstm.war	<input checked="" type="radio"/>	

save

### database sizes

Database	Size (MB)	Used by CTS	delete
ctstm_ted	635.141		
cts_demo	0.109		

# Database Configuration

The screenshot displays the 'CTSTM Admin Tool' interface. The top navigation bar includes 'create new instance', 'update template', 'admin area', and 'help'. The user is signed in as 'cts'. The main content area is titled 'database parameters of ctstm\_demo' and features a blue banner indicating 'import running for this CTS instance'. Below this is a table with columns for 'Parameter', 'Value', and 'Comment'. The parameters listed are db\_host (localhost), db\_port (3306), db\_pw (9la)Gan3, db\_user (root), mysqlserver (checked), commitbatch (500), and DB\_batchSize (50000). A green callout box provides instructions for the db\_user parameter. At the bottom, there is a yellow warning section for 'drop database of instance ctstm\_demo' with a confirmation form.

CTSTM Admin Tool   create new instance   update template   admin area   help   signed in as cts   logout

CTSTM instances:   DB-Config   Data Import   Servlet   Browse Data

ctstm\_demo

database parameters of ctstm\_demo

import running for this CTS instance

Parameter	Value	Comment
db_host	localhost	
db_port	3306	
db_pw	9la)Gan3	
db_user	root	do not use " for strings parameters for MySQL db_user must have root-access or at least be able to create databases If this is not possible you can create the database manually before running the script
mysqlserver	<input checked="" type="checkbox"/>	
commitbatch	500	
DB_batchSize	50000	

drop database of instance ctstm\_demo

This will drop (delete) the MySQL database holding all the data of this instance. Be careful! This cannot be undone! This will only affect the data of this instance and not of any other instances.

type uppercase OK as confirmation   confirm with OK   drop table (cannot be undone!)

# Data Import

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

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

ctstm\_demo

### data import

This will start the data import. This could take several hours. You can safely leave this page and do something else in the mean time. Please note that only one instance can import data at a given time. A logfile of the last import activities is shown below.

import running for 'ctstm\_demo'

cancel import

### import parameters of ctstm\_demo

import running for this CTS instance

Parameter	Value	Comment
resetDataset	<input checked="" type="checkbox"/>	
requestNew	<input checked="" type="checkbox"/>	
instance_name	ted	
cts	urn.cts.ted:	
cts_configuration	escapePassage=false_epidoc=false_dele	
shutdown	<input type="checkbox"/>	



# Data Import

CTSTM Admin Tool [create new instance](#) [update template](#) [admin area](#) [help](#) signed in as **cts** [logout](#)

CTSTM instances: [DB-Config](#) [Data Import](#) [Servlet](#) [Browse Data](#)

ctstm\_demo

docCount	<input type="text" value="-1"/>	
dateformat	<input type="text" value="yyyy-MM-DD HH.MM:SS"/>	
termdocumentmatrix	<input checked="" type="checkbox"/>	Basic Dataformats. Required for advanced formats
neighbourtable	<input checked="" type="checkbox"/>	
ngram_n	<input type="text" value="3&amp;5"/>	
tokenlengthpassage	<input checked="" type="checkbox"/>	
termmatrix	<input checked="" type="checkbox"/>	
documentpruning	<input checked="" type="checkbox"/>	
termpruning	<input checked="" type="checkbox"/>	
stopwords	<input checked="" type="checkbox"/>	
prunedneighbourtable	<input checked="" type="checkbox"/>	
ngramreduce	<input checked="" type="checkbox"/>	Advanced Dataformats. Require specific basic formats
zipfrank	<input checked="" type="checkbox"/>	
topicmodels	<input checked="" type="checkbox"/>	
topic_word_minweight	<input type="text" value="0.2"/>	
topic_doc_minweight	<input type="text" value="0.2"/>	
ngram	<input checked="" type="checkbox"/>	
fulltextindexmysql	<input type="checkbox"/>	

# Data Import

CTSTM Admin Tool [create new instance](#) [update template](#) [admin area](#) [help](#) signed in as **cts** [logout](#)

CTSTM instances: [DB-Config](#) [Data Import](#) [Servlet](#) [Browse Data](#)

ctstm\_demo

fulltextindexmysql	<input type="checkbox"/>
fulltextindexlucene	<input checked="" type="checkbox"/>
zipfStopwordCount	<input type="text" value="150"/>
modelCount	<input type="text" value="30"/>
iterations	<input type="text" value="1000"/>
numberOfThreads	<input type="text" value="2"/>
threshold_docpruning	<input type="text" value="90"/>
threshold_termpruning	<input type="text" value="1"/>
decimals_statistics	<input type="text" value="65"/>
precision_statistics	<input type="text" value="30"/>
neighbourBatchSize	<input type="text" value="10000"/>
includeURNsWith	<input type="text" value="len:"/>
excludeURNsWith	<input type="text"/>
casesensitive	<input type="checkbox"/>
normalizepassage	<input checked="" type="checkbox"/>

Normalization of the texts

# Servlet Management

The screenshot displays the 'CTSTM Admin Tool' interface. The top navigation bar is blue and contains the following items: 'CTSTM Admin Tool', 'create new instance', 'update template', 'admin area', 'help', 'signed in as cts', and 'logout'. Below the navigation bar, there are tabs for 'DB-Config', 'Data Import', 'Servlet', and 'Browse Data'. The 'Servlet' tab is currently selected. On the left side, there is a sidebar with 'CTSTM instances:' and a list containing 'ctstm\_demo'. The main content area shows three management options for the 'ctstm\_demo' instance:

- rename this instance ("ctstm\_demo")**: A light blue header. The text below states: "The renaming of an instance can take up to 30 seconds. Please stand by and do not leave this page. A message will appear when the instance has been renamed successfully. For the name only small and uppercase letters as well as numbers and the underscore are available ( a-z A-Z 0-9 \_ ). All other characters will be removed automatically. Thus, any white-spaces and special characters will be removed." Below this is a form with a 'new name:' label, a text input field containing 'new name', and a button labeled 'import running for this instance'.
- update this instance ("ctstm\_demo")**: A light blue header. The text below states: "This will update this instance to the newest template file. This could take up to 60 seconds." Below this is a form with a text input field and a button labeled 'import running for this instance'.
- delete this instance ("ctstm\_demo")**: A light red header. The text below states: "The deletion of an instance can take up to 30 seconds. Please stand by and do not leave this page. A message will appear when the instance has been deleted successfully. Be careful! The deletion of an instance cannot be undone!" Below this is a form with a label 'type uppercase OK as confirmation', a text input field containing 'confirm with OK', and a button labeled 'import running for this instance'.

# Servlet Management

The screenshot displays the CTSTM Admin Tool interface. At the top, a blue navigation bar contains the text "CTSTM Admin Tool" and several menu items: "create new instance", "update template", "admin area", and "help". On the right side of this bar, it indicates the user is "signed in as cts" and provides a "logout" link.

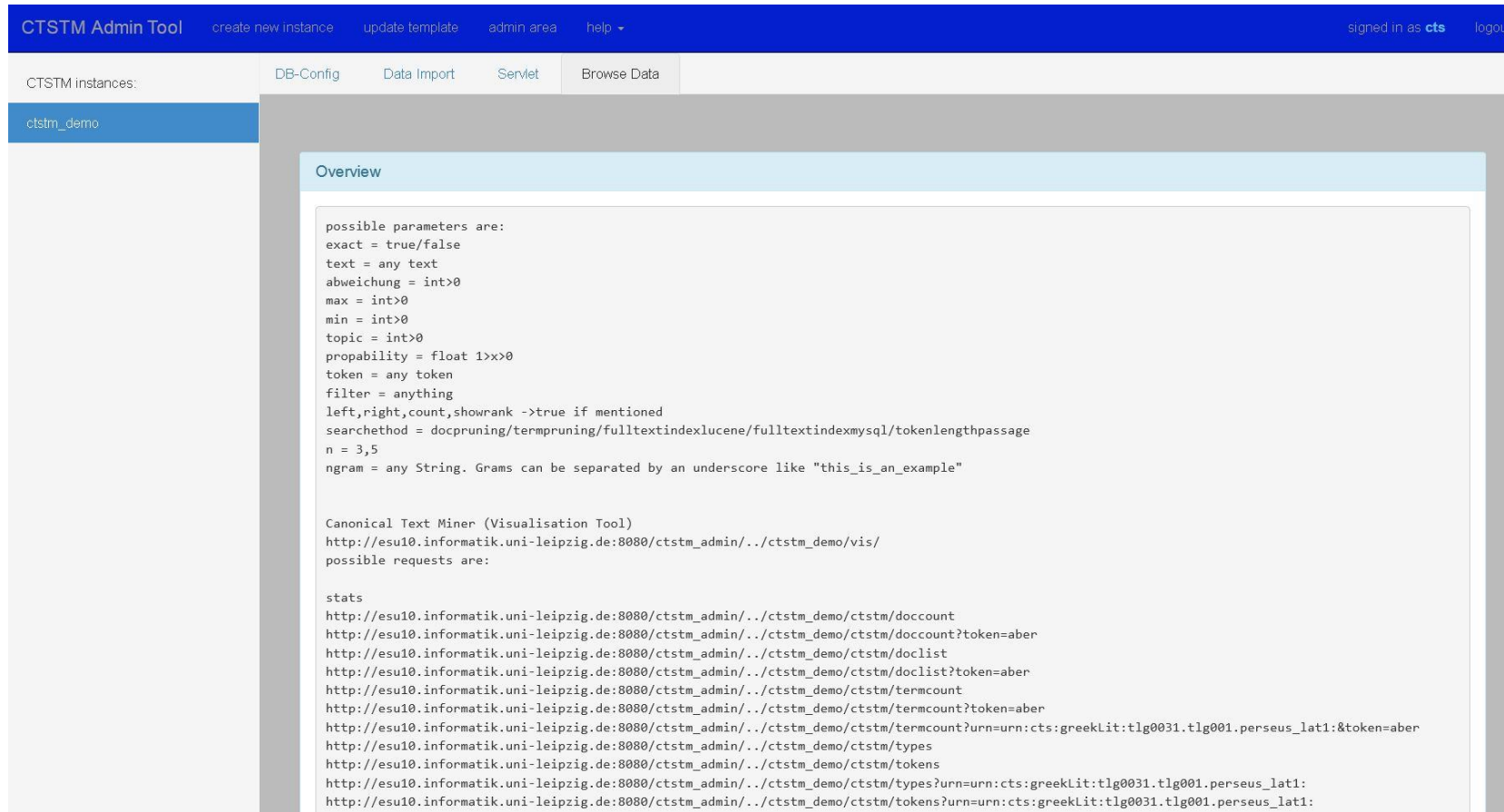
Below the navigation bar, a sidebar on the left lists "CTSTM instances:" with "ctstm\_demo" selected. The main content area has a top navigation bar with "DB-Config", "Data Import", "Servlet", and "Browse Data". The "Servlet" tab is active, showing a message "import running for this instance" in a light blue box.

The main content area is divided into two sections. The first section, titled "delete this instance ('ctstm\_demo')", contains a warning: "The deletion of an instance can take up to 30 seconds. Please stand by and do not leave this page. A message will appear when the instance has been deleted successfully. Be careful! The deletion of an instance cannot be undone!". Below this warning is a confirmation form with a text input field containing "type uppercase OK as confirmation", a "confirm with OK" button, and another "import running for this instance" message.

The second section, titled "servlet parameters of ctstm\_demo", features a blue bar with the text "import running for this CTS instance". Below this is a table with three columns: "Parameter", "Value", and "Comment".

Parameter	Value	Comment
demotoken	aber	
demogram	aber	
demourn	urn:cts:greekLit:tlg0031.tlg001.perseus_k	
demotext	i guess the story ac	
textlength	4629622	
wordlength	31	

# Browse – “API”



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

CTSTM instances:   DB-Config   Data Import   Servlet   Browse Data

ctstm\_demo

### Overview

possible parameters are:

- exact = true/false
- text = any text
- abweichung = int>0
- max = int>0
- min = int>0
- topic = int>0
- propability = float 1>x>0
- token = any token
- filter = anything
- left,right,count,showrank ->true if mentioned
- searchmethod = docpruning/termpruning/fulltextindexlucene/fulltextindexmysql/tokenlengthpassage
- n = 3,5
- ngram = any String. Grams can be separated by an underscore like "this\_is\_an\_example"

Canonical Text Miner (Visualisation Tool)  
[http://esu10.informatik.uni-leipzig.de:8080/ctstm\\_admin/./ctstm\\_demo/vis/](http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/./ctstm_demo/vis/)

possible requests are:

stats

- [http://esu10.informatik.uni-leipzig.de:8080/ctstm\\_admin/./ctstm\\_demo/ctstm/doccount](http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/./ctstm_demo/ctstm/doccount)
- [http://esu10.informatik.uni-leipzig.de:8080/ctstm\\_admin/./ctstm\\_demo/ctstm/doccount?token=aber](http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/./ctstm_demo/ctstm/doccount?token=aber)
- [http://esu10.informatik.uni-leipzig.de:8080/ctstm\\_admin/./ctstm\\_demo/ctstm/doclist](http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/./ctstm_demo/ctstm/doclist)
- [http://esu10.informatik.uni-leipzig.de:8080/ctstm\\_admin/./ctstm\\_demo/ctstm/doclist?token=aber](http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/./ctstm_demo/ctstm/doclist?token=aber)
- [http://esu10.informatik.uni-leipzig.de:8080/ctstm\\_admin/./ctstm\\_demo/ctstm/termcount](http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/./ctstm_demo/ctstm/termcount)
- [http://esu10.informatik.uni-leipzig.de:8080/ctstm\\_admin/./ctstm\\_demo/ctstm/termcount?token=aber](http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/./ctstm_demo/ctstm/termcount?token=aber)
- [http://esu10.informatik.uni-leipzig.de:8080/ctstm\\_admin/./ctstm\\_demo/ctstm/termcount?urn=urn:cts:greekLit:tlg0031.tlg001.perseus\\_lat1:&token=aber](http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/./ctstm_demo/ctstm/termcount?urn=urn:cts:greekLit:tlg0031.tlg001.perseus_lat1:&token=aber)
- [http://esu10.informatik.uni-leipzig.de:8080/ctstm\\_admin/./ctstm\\_demo/ctstm/types](http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/./ctstm_demo/ctstm/types)
- [http://esu10.informatik.uni-leipzig.de:8080/ctstm\\_admin/./ctstm\\_demo/ctstm/tokens](http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/./ctstm_demo/ctstm/tokens)
- [http://esu10.informatik.uni-leipzig.de:8080/ctstm\\_admin/./ctstm\\_demo/ctstm/types?urn=urn:cts:greekLit:tlg0031.tlg001.perseus\\_lat1:](http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/./ctstm_demo/ctstm/types?urn=urn:cts:greekLit:tlg0031.tlg001.perseus_lat1:)
- [http://esu10.informatik.uni-leipzig.de:8080/ctstm\\_admin/./ctstm\\_demo/ctstm/tokens?urn=urn:cts:greekLit:tlg0031.tlg001.perseus\\_lat1:](http://esu10.informatik.uni-leipzig.de:8080/ctstm_admin/./ctstm_demo/ctstm/tokens?urn=urn:cts:greekLit:tlg0031.tlg001.perseus_lat1:)

# 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

