Compilation of a Large Ground-Truth Data Set Using Transkribus

Matthias Boenig & Kay-Michael Würzner
{boenig|wuerzner}@bbaw.de

Transkribus User Conference
Vienna, 2nd November 2017

Overview

Goal: Compilation of a large, **homogeneous** Ground Truth (GT) data set

- Various **heterogeneous** sources
- Annotation on the **textual** and/or **structural** level

Background: OCR-D initiative

a. Funding by the *Deutsche Forschungsgemeinschaft*

   → Improvement of OCR-D tools for **historical printings** (i.e. VD 16, 17, 18)

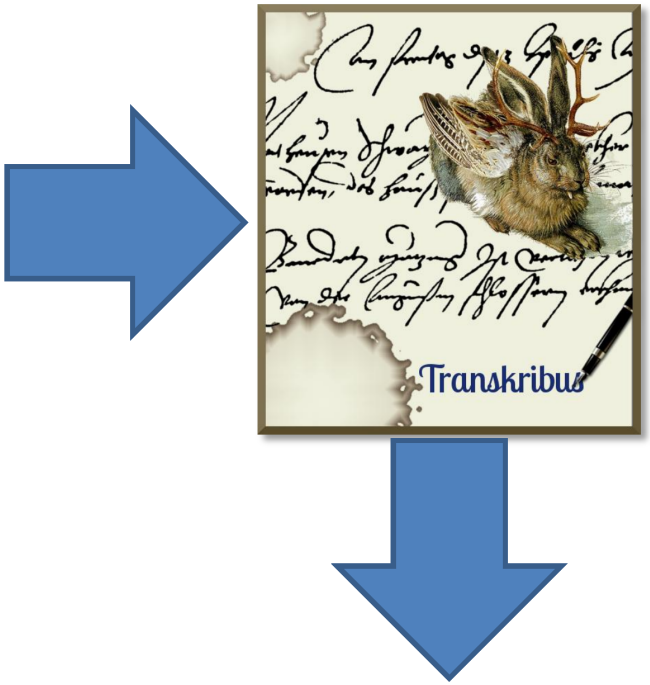
b. Coordination project

   - Identify to-dos, desiderata and improvement options
   - Development of a call for proposals
   - Merge (sub-)project results into a **productive workflow**

Procedure: Annotation with `Transkribus`

1. Import images and existing text and/or structural information
2. **Harmonization** and completion within `Transkribus`

Overview



- Various GT sources
- Containing either text or structural annotations in differing quality
- By now, ≈ 130 documents with ≈ 500 pages
- A lot more to come!

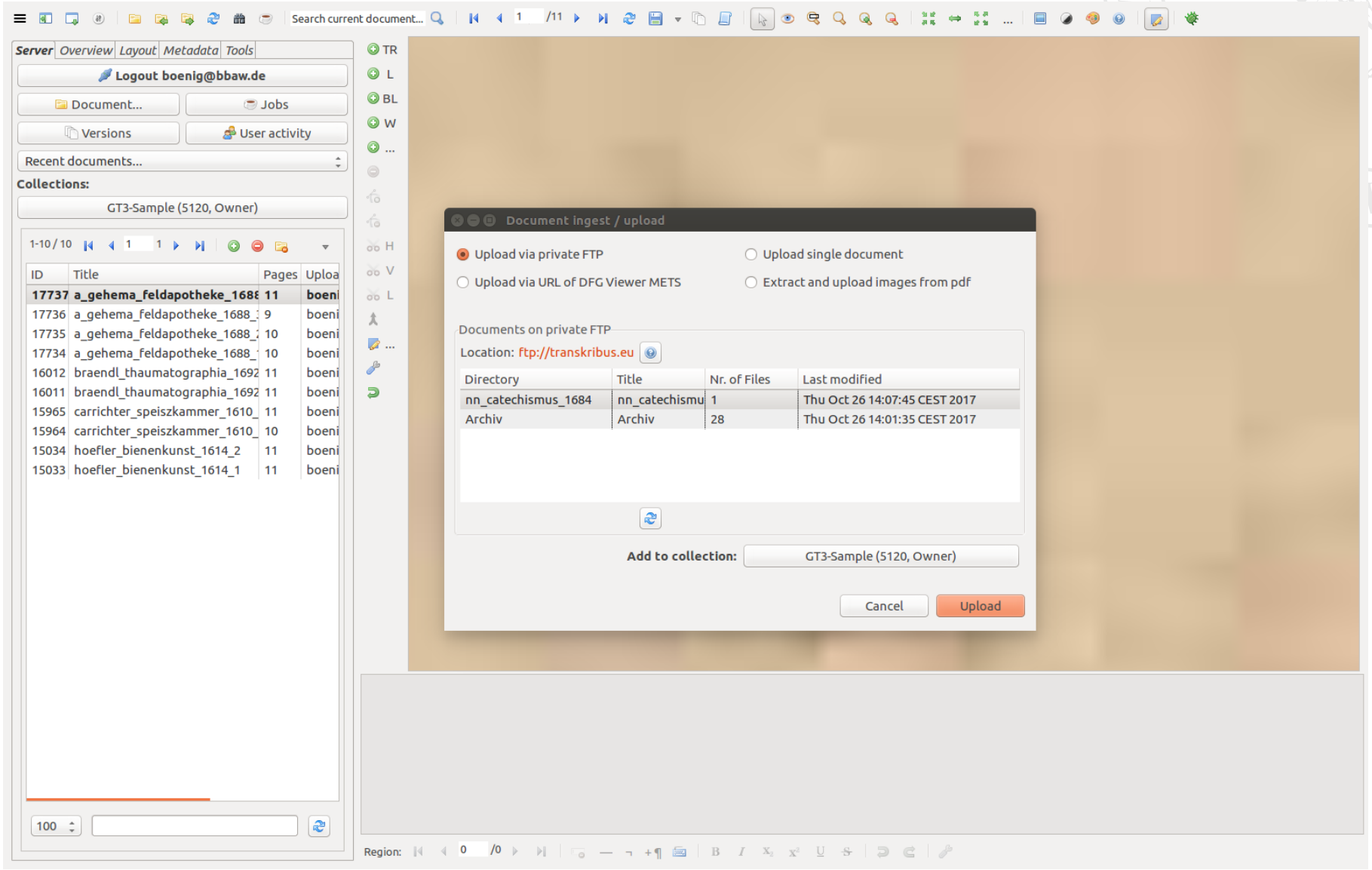
Workflows

Existing text	Existing structure

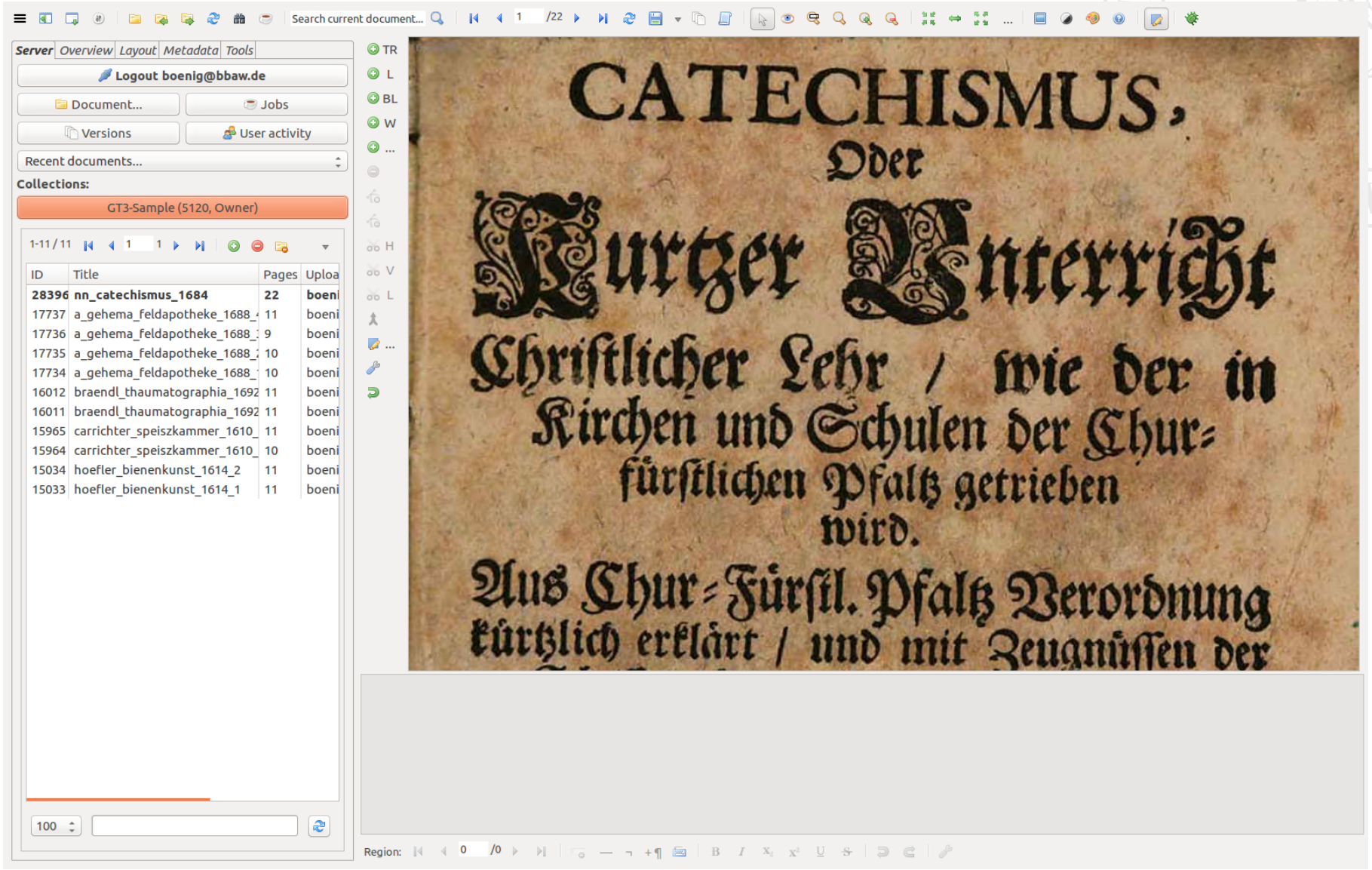
Workflows

Existing text	Existing structure
Import images	

Workflows


Search current document...
1 /11
Server
Overview
Layout
Metadata
Tools
Logout boenig@bbaw.de
Document...
Jobs
Versions
User activity
Recent documents...
Collections:
GT3-Sample (5120, Owner)
1-10 / 10
ID
Title
Pages
Uploa
17737 a_gehema_feldapotheke_1688 11 boeni
17736 a_gehema_feldapotheke_1688_ 9 boeni
17735 a_gehema_feldapotheke_1688_ 10 boeni
17734 a_gehema_feldapotheke_1688_ 10 boeni
16012 braendl_thaumatographia_1692 11 boeni
16011 braendl_thaumatographia_1692 11 boeni
15965 carrichter_speiszkammer_1610_ 11 boeni
15964 carrichter_speiszkammer_1610_ 10 boeni
15034 hoefler_bienenkunst_1614_2 11 boeni
15033 hoefler_bienenkunst_1614_1 11 boeni
100
TR
L
BL
W
H
V
Document ingest / upload
Upload via private FTP
Upload single document
Upload via URL of DFG Viewer METS
Extract and upload images from pdf
Documents on private FTP
Location: ftp://transkribus.eu
Directory
Title
Nr. of Files
Last modified
nn_catechismus_1684
nn_catechismu
1
Thu Oct 26 14:07:45 CEST 2017
Archiv
Archiv
28
Thu Oct 26 14:01:35 CEST 2017
Add to collection:
GT3-Sample (5120, Owner)
Cancel
Upload
Region:
0 /0

Workflows

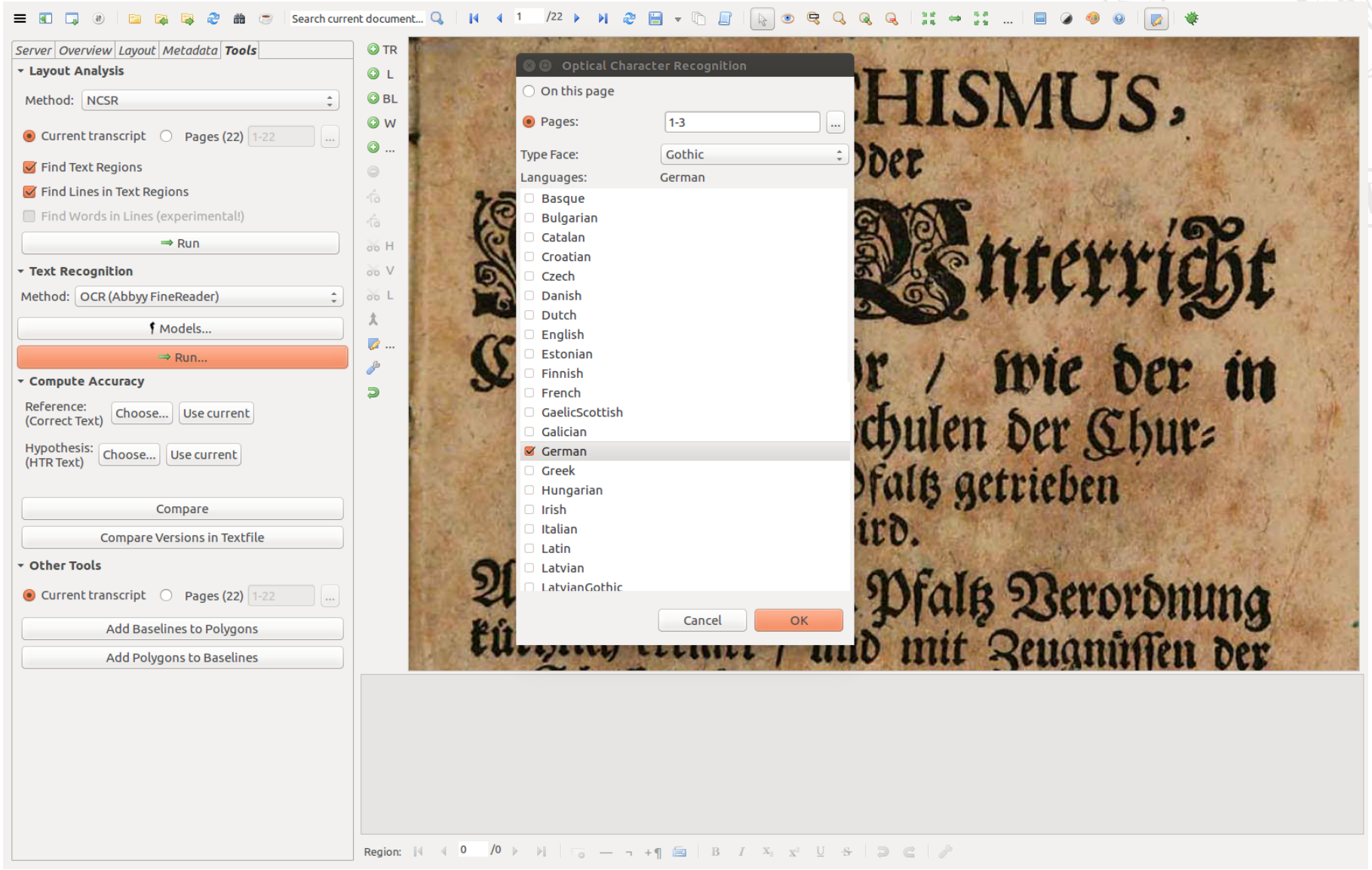

Search current document...
1 /22
Server Overview Layout Metadata Tools
Logout boenig@bbaw.de
Document... Jobs
Versions User activity
Recent documents...
Collections:
GT3-Sample (5120, Owner)
1-11 / 11
ID Title Pages Uploa
28396 nn_catechismus_1684 22 boeni
17737 a_gehema_feldapotheke_1688_ 11 boeni
17736 a_gehema_feldapotheke_1688_ 9 boeni
17735 a_gehema_feldapotheke_1688_ 10 boeni
17734 a_gehema_feldapotheke_1688_ 10 boeni
16012 braendl_thaumatographia_1692 11 boeni
16011 braendl_thaumatographia_1692 11 boeni
15965 carrichter_speiszkammer_1610_ 11 boeni
15964 carrichter_speiszkammer_1610_ 10 boeni
15034 hoefler_bienenkunst_1614_2 11 boeni
15033 hoefler_bienenkunst_1614_1 11 boeni
100
TR
L
BL
W
H
V
L
CATECHISMUS,
Oder
Kurtzer Unterricht
Christlicher Lehr / wie der in
Kirchen und Schulen der Chur=
fürstlichen Pfaltz getrieben
wird.
Aus Chur=Fürstl. Pfaltz Verordnung
kürtzlich erklärt / und mit Zeugnüssen der
Region: 0 /0

Workflows

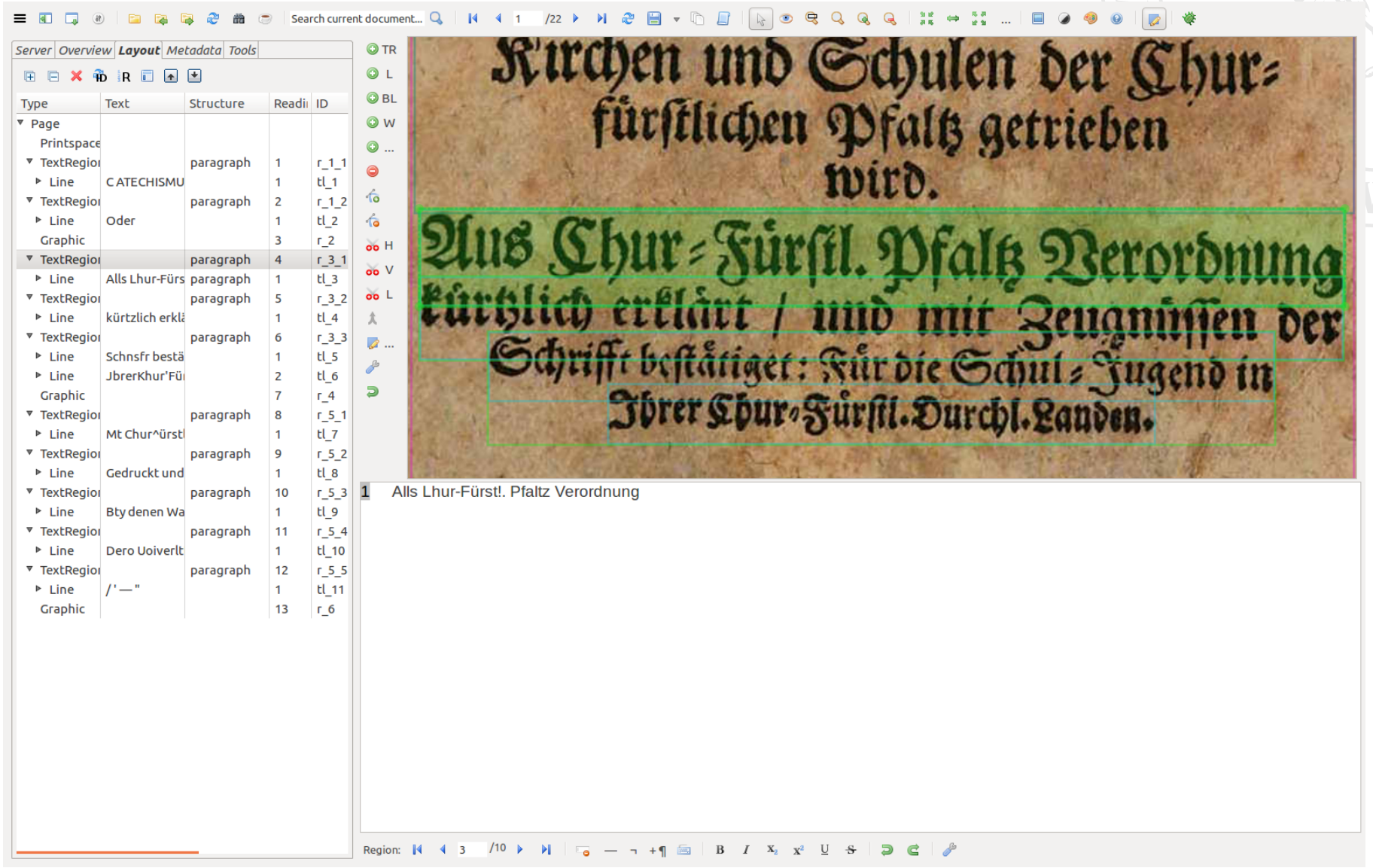
Existing text	Existing structure
Import images	
Run FineReader for initial layout version	Import `Page XML`

-
-
-
-

Workflows


Search current document...
Server
Overview
Layout
Metadata
Tools
Layout Analysis
Method: NCSR
Current transcript
Pages (22) 1-22
Find Text Regions
Find Lines in Text Regions
Find Words in Lines (experimental!)
Run
Text Recognition
Method: OCR (Abbyy FineReader)
Models...
Run...
Compute Accuracy
Reference: (Correct Text)
Choose...
Use current
Hypothesis: (HTR Text)
Choose...
Use current
Compare
Compare Versions in Textfile
Other Tools
Current transcript
Pages (22) 1-22
Add Baselines to Polygons
Add Polygons to Baselines
Optical Character Recognition
On this page
Pages: 1-3
Type Face: Gothic
Languages: German
Basque
Bulgarian
Catalan
Croatian
Czech
Danish
Dutch
English
Estonian
Finnish
French
GaelicScottish
Galician
German
Greek
Hungarian
Irish
Italian
Latin
Latvian
LatvianGothic
Cancel
OK
Region:

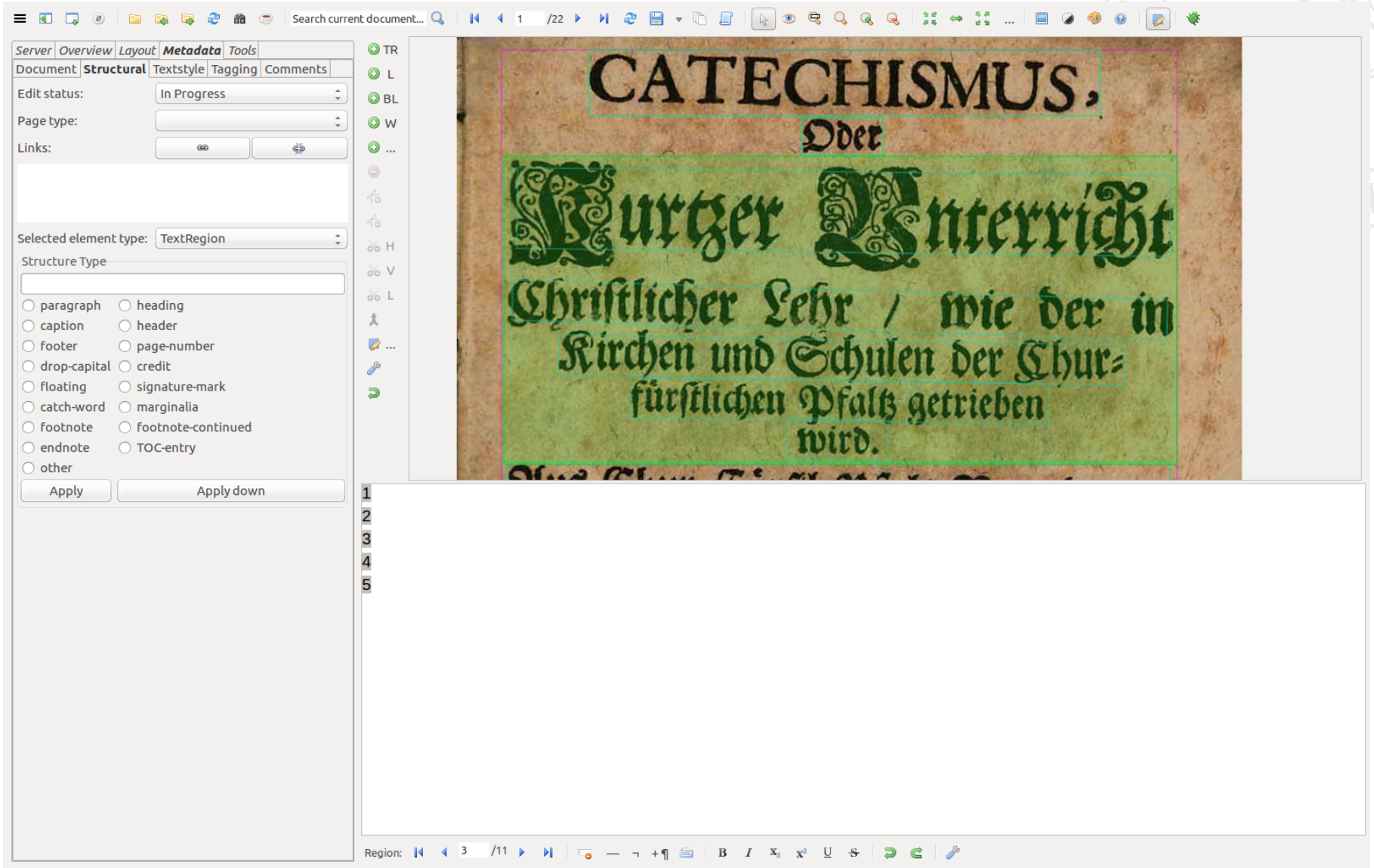
Workflows


Search current document...
1 /22
Server Overview Layout Metadata Tools
Type Text Structure Readi ID
Page
Printspace
TextRegion paragraph 1 r_1_1
Line CATECHISMU 1 tl_1
TextRegion paragraph 2 r_1_2
Line Oder 1 tl_2
Graphic 3 r_2
TextRegion paragraph 4 r_3_1
Line Alls Lhur-Fürs paragraph 1 tl_3
TextRegion paragraph 5 r_3_2
Line kürtzlich erklä 1 tl_4
TextRegion paragraph 6 r_3_3
Line Schnsfr bestä 1 tl_5
Line JbrerKhur'Fü 2 tl_6
Graphic 7 r_4
TextRegion paragraph 8 r_5_1
Line Mt Chur^ürstl 1 tl_7
TextRegion paragraph 9 r_5_2
Line Gedruckt und 1 tl_8
TextRegion paragraph 10 r_5_3
Line Bty denen Wa 1 tl_9
TextRegion paragraph 11 r_5_4
Line Dero Uoiverlt 1 tl_10
TextRegion paragraph 12 r_5_5
Line /'—" 1 tl_11
Graphic 13 r_6
TR
L
BL
W
H
V
L
Kirchen und Schulen der Chur-
fürstlichen Pfaltz getrieben
wird.
Aus Chur-Fürstl. Pfaltz Verordnung
1 Alls Lhur-Fürst!. Pfaltz Verordnung
Region: 3 /10

Workflows

Existing text	Existing structure
Import images	
Run FineReader for initial layout version	Import `Page XML`
Manually correct layout	Run external OCR for initial text version

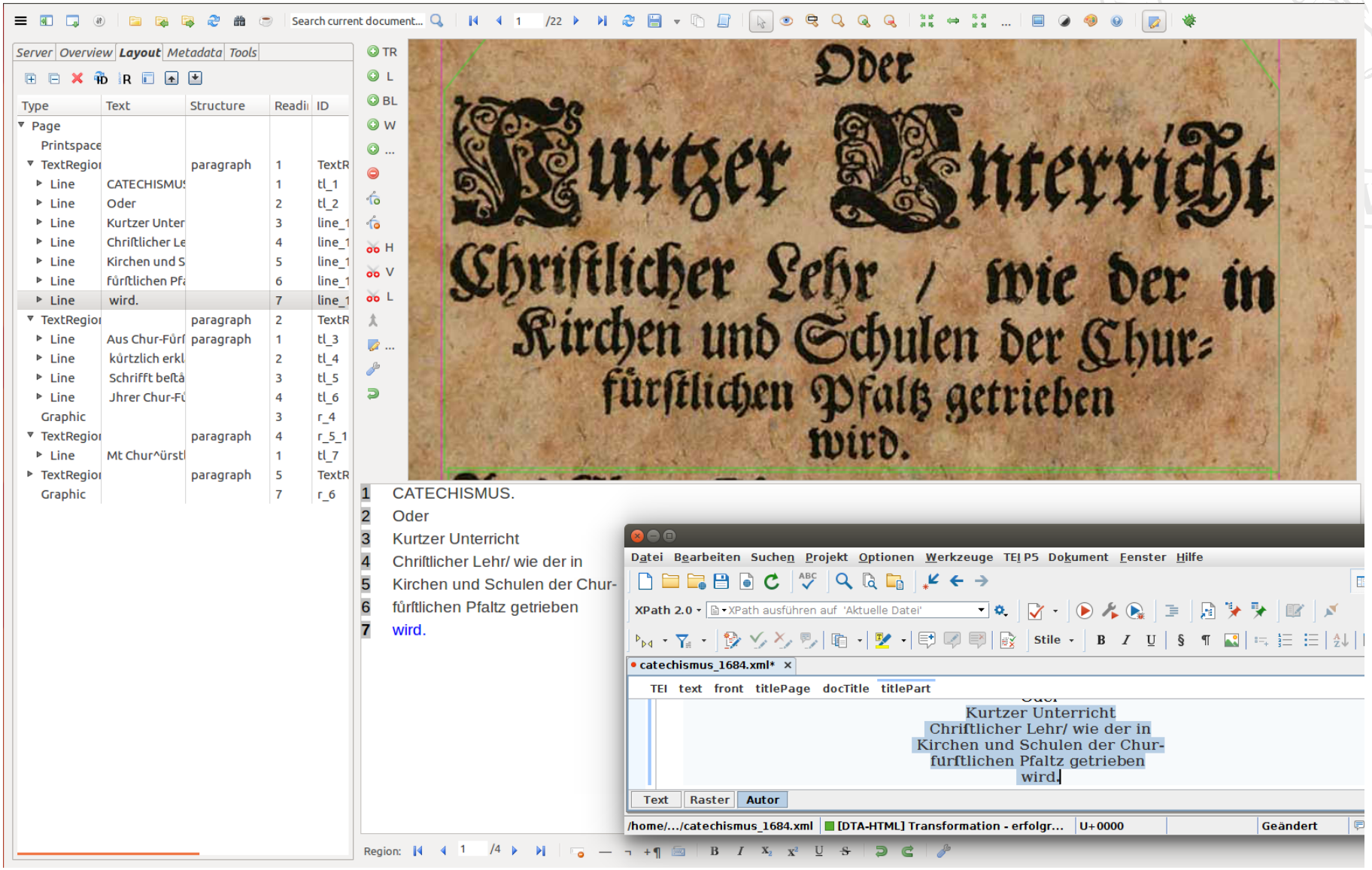
Workflows


Search current document...
1 /22
Server Overview Layout Metadata Tools
Document Structural Textstyle Tagging Comments
Edit status: In Progress
Page type:
Links:
Selected element type: TextRegion
Structure Type
paragraph heading
caption header
footer page-number
drop-capital credit
floating signature-mark
catch-word marginalia
footnote footnote-continued
endnote TOC-entry
other
Apply
Apply down
TR
L
BL
W
H
V
L
CATECHISMUS,
Oder
Kurtzer Unterricht
Christlicher Lehr / wie der in
Kirchen und Schulen der Chur=
fürstlichen Pfaltz getrieben
wird.
1
2
3
4
5
Region: 3 /11

Workflows

Existing text	**Existing structure**
Import images	
Run FineReader for initial layout version	Import `Page XML`
Manually correct layout	Run external OCR for initial text version
Copy and paste text region by region	

Workflows


Server Overview Layout Metadata Tools
Type Text Structure Readi ID
Page
Printspace
TextRegio paragraph 1 TextR
Line CATECHISMU 1 tl_1
Line Oder 2 tl_2
Line Kurtzer Unter 3 line_1
Line Chriſtlicher Le 4 line_1
Line Kirchen und S 5 line_1
Line fürſtlichen Pfa 6 line_1
Line wird. 7 line_1
TextRegio paragraph 2 TextR
Line Aus Chur-Fürſ paragraph 1 tl_3
Line kürtzlich erkl 2 tl_4
Line Schrifft beſtä 3 tl_5
Line Jhrer Chur-Fü 4 tl_6
Graphic 3 r_4
TextRegio paragraph 4 r_5_1
Line Mt Chur^ürstl 1 tl_7
TextRegio paragraph 5 TextR
Graphic 7 r_6
Oder
Kurtzer Unterricht
Chriſtlicher Lehr / wie der in
Kirchen und Schulen der Chur-
fürſtlichen Pfaltz getrieben
wird.
1 CATECHISMUS.
2 Oder
3 Kurtzer Unterricht
4 Chriſtlicher Lehr/ wie der in
5 Kirchen und Schulen der Chur-
6 fürſtlichen Pfaltz getrieben
7 wird.
Datei Bearbeiten Suchen Projekt Optionen Werkzeuge TEI P5 Dokument Fenster Hilfe
XPath 2.0
catechismus_1684.xml*
TEI text front titlePage docTitle titlePart
Kurtzer Unterricht
Chriſtlicher Lehr/ wie der in
Kirchen und Schulen der Chur-
fürſtlichen Pfaltz getrieben
wird.
Text Raster Autor
/home/.../catechismus_1684.xml
[DTA-HTML] Transformation - erfolgr...
U+0000
Geändert
Region:

Workflows

Existing text	**Existing structure**
Import images	
Run FineReader for initial layout version	Import `Page XML`
Manually correct layout	Run external OCR for initial text version
Copy and paste text region by region	
	Manually correct text

Workflows

Existing text	**Existing structure**
Import images	
Run FineReader for initial layout version	Import `Page XML`
Manually correct layout	Run external OCR for initial text version
Copy and paste text region by region	
	Manually correct text

- Somewhat naïve approach
- External Page XML creation or
- Intermediate export and (re-)import as alternative options
- **Not very comfortable**

Desiderata

- `Transkribus` is a wonderful tool!
  - Support for **polygonal** regions
  - **Multiple** OCR options
  - **Collaborative** working environment with basic version control
  - TEI export
- For **GT creation**, we would welcome
  - OCR application on **specific regions** also for FineReader
  - Dedicated **text import** functionalities (e.g. on paragraph level)
  - METS import which accounts for **existing structural annotations** and linked ALTO
  - **Automatic support** during manual post correction
  - TEI import

Collaboration

- OCR-D GT Guidelines
  - Documentation of existing OCR-D GT
  - Instructions for GT creation
    - Already used within the OCR-D project
    - Perspectively also used in a **broader context** (community use)
  - Automatic **validation** of GT data
  - (Semi-)automatic **conversion** of existing GT data sets
  - Plans for setting up a **GT repository** for print publications and handwritten documents
- Availability

  View: `http://kaskade.dwds.de/~matthias/ocr-d/`

  Sources: `https://github.com/OCR-D/`

Collaboration

- Transkribus User Documentation : A proposal
  1. Step: Change the documentation format from Wiki to DITA
     - XML-based documentation format
     - **Topic-oriented** internal and "external" structure (i.e. presentation)
     - Various automatically generated **presentation modes**
  2. Step: Build and organize a documentation **source repository** (e.g. on github)
  3. Step: Involve the user community into the documentation process
     - Non-developer view point
     - Recipes for frequent tasks

Many thanks for your attention.