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)

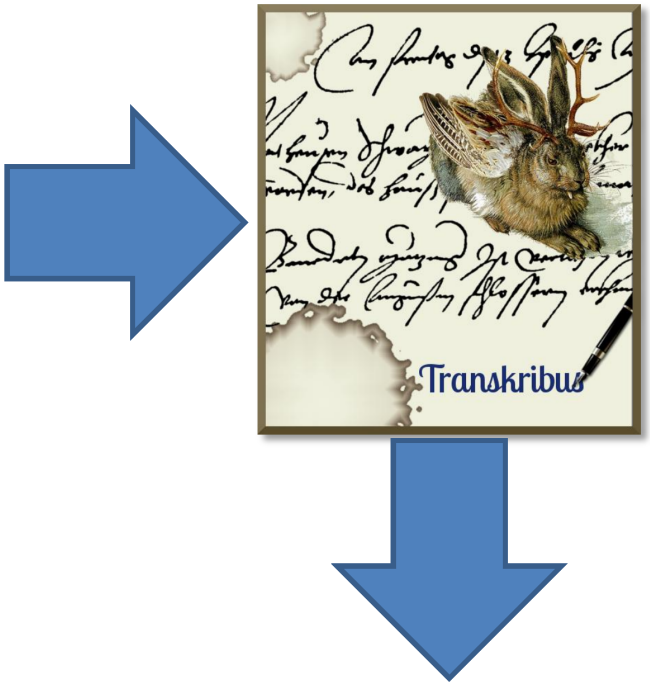
b. Coordination project

   - 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

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

-
-
-
-

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

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 line by line	

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 line by line	
	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 line by line	
	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 (cf. 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 transcription guidelines
  -
- Transkribus documentation in DITA format
  -

Many thanks for your attention.