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

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Overview

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

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

Background: OCR-D initiative

a. Funding by the *Deutsche Forschungsgemeinschaft*

   $\rightarrow$ 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

Workflow

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Desiderata

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Many thanks for your attention.