# 🚦 Ground Truth Roadmap 🚗

Matthias Boenig (BBAW)

2021

Kick-off workshop: July 29–30

# Personal

- since 2010 research assistant at the berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN
  - from 2010 to 2017, at the DTA
  - involved in text production, supervising the transcription production and conversion/transformation of texts according to DTABf.
- since 2015 member of the OCR-D DFG-funded Initiative for Optical Character Recognition Development
  - initiated and supervised the call for proposals for project phase 2 (module project phase) together with my former colleague Kay-Michael Würzner
  - as this time, setting up the Ground Truth department together with my OCR-D colleagues, my former colleagues and some members of modul projects

# Tasks

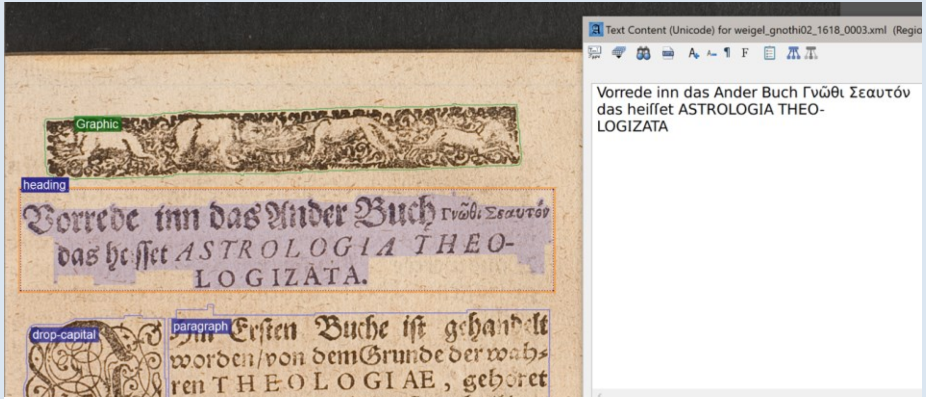
I. Coordinating Ground Truth

II. Implementation

III. Dissemination

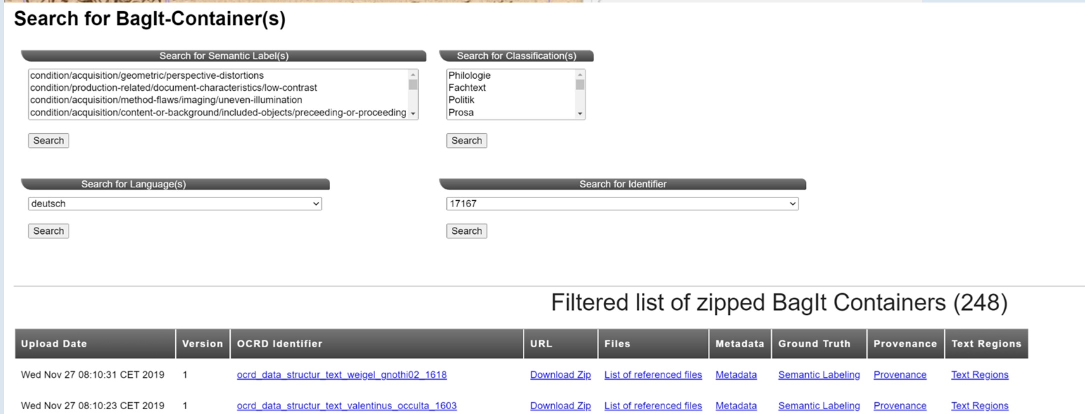
# Task I: Coordinating Ground Truth

- Structure-GT
  - contains regions of the printed page only
- Structure-Text-GT
  - contains both regions and text of the printed page
- GT stored in the GT-Repository



Text Content (Unicode) for weigel_gnothi02_1618_0003.xml (Regio

Vorrede inn das Ander Buch Γνῶθι Σεαυτόν
das heiſſet ASTROLOGIA THEO-
LOGIZATA



## Search for BagIt-Container(s)

Search for Semantic Label(s)

condition/acquisition/geometric/perspective-distortions
condition/production-related/document-characteristics/low-contrast
condition/acquisition/method-flaws/imaging/uneven-illumination
condition/acquisition/content-or-background/included-objects/preceeding-or-proceeding

Search for Classification(s)

Philologie
Fachtext
Politik
Prosa

Search for Language(s): deutsch

Search for Identifier: 17167

Filtered list of zipped BagIt Containers (248)

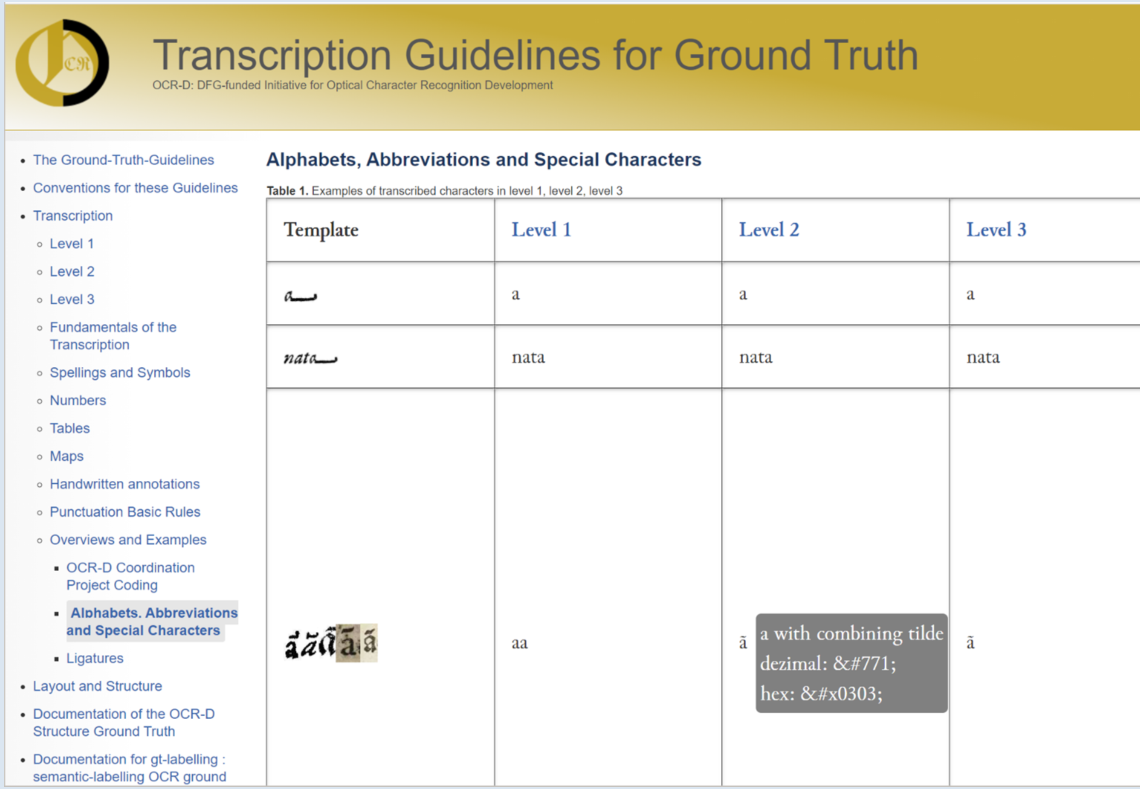
| Upload Date | Version | OCRD Identifier | URL | Files | Metadata | Ground Truth | Provenance | Text Regions |
|---|---|---|---|---|---|---|---|---|
| Wed Nov 27 08:10:31 CET 2019 | 1 | ocrd_data_structur_text_weigel_gnothi02_1618 | Download Zip | List of referenced files | Metadata | Semantic Labeling | Provenance | Text Regions |
| Wed Nov 27 08:10:23 CET 2019 | 1 | ocrd_data_structur_text_valentinus_occulta_1603 | Download Zip | List of referenced files | Metadata | Semantic Labeling | Provenance | Text Regions |

# Vision and Goal for Task I

1. **Improving** the existing ground truth
2. **20,000 pages** of structural GT
3. Ground Truth **Call**
4. **Relaunching** the GT-Repository
5. **Curating** GT

# Task II: Implementation

- Developing transcription rules for both characters and structures based on the PAGE format
- Setting up guidelines for the structure GT
- Classification of metadata for the GT
- Documentation of the PAGE format



## Transcription Guidelines for Ground Truth

OCR-D: DFG-funded Initiative for Optical Character Recognition Development

### Alphabets, Abbreviations and Special Characters

**Table 1.** Examples of transcribed characters in level 1, level 2, level 3

| Template | Level 1 | Level 2 | Level 3 |
|---|---|---|---|
| | a | a | a |
| nata | nata | nata | nata |
| | aa | ã a with combining tilde<br>dezimal: &#771;<br>hex: &#x0303; | ã |

# Vision and Goal for Task II

- **Establishing** standards and specifications
- Developing guidelines with human and **machine-readable rules**
- **Maintenance** and development of standards and specifications

# Task III: Dissemination

- Restructuring the
  “DFG Guidelines: Digitization”

DFG-Praxisregeln

„Digitalisierung“

Deutsche Forschungsgemeinschaft
Kennedyallee 40 · 53175 Bonn · Postanschrift: 53170 Bonn
Telefon: + 49 228 885-1 · Telefax: + 49 228 885-2777 · postmaster@dfg.de · www.dfg.de

DFG

# Contact and Information

Contact:

Matthias Boenig

✉ boenig@bbaw.de

Information:

- Official website: The Ground-Truth-Guidelines (ocr-d.de)
- Preprint-Guidelines: gt-guidelines (en), gt-guidelines (de)
- Data-Site and further Informations: OCR-D data