Integration of Kitodo and OCR-D for Productive Mass-Digitisation

OCR-D Phase 3 Kick-Off

Robert Sachunsky

July 30, 2021

Implementation Project Kitodo / OCR-D

- 8 man-years, 2 years, 3 libraries:
    - Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden
    - Universitätsbibliothek der TU Braunschweig
    - Universitätsbibliothek Mannheim
- integrate Kitodo with OCR-D “backend” as distributed system
- extend both OCR-D and Kitodo for robust mass production

Premises

- Kitodo: Workflow Management System for libraries
    - Open-source, community-driven
    - Modules:
        - Kitodo.Production (digitisation workflows)
        - Kitodo.Presentation (DFG viewer etc.)
    - OCR: only via commercial plugins (black box, license costs)
- OCR-D: operative single-workstation command-line prototype
    - no network interfaces for distribution/scaling yet
    - no error recovery and dynamic workflow execution yet
    - no result quality estimation and runtime evaluation yet
    - no assisted/automatic workflow configuration yet

Goals

1. Implement OCR-D as Web-based distributed system
    - controller + processing servers
    - container virtualisation
2. Develop quality based workflow optimisation for OCR-D
    - automatic quality estimation of results
    - dynamic workflows with quality thresholds and switches
    - predefined, manually optimised workflow configurations
3. Implement OCR-D as OCR module in Kitodo.Production
    - manage data and run workflows
    - track and visualise result progress/quality
    - edit and manage workflow configurations
4. Extend Kitodo.Presentation and DFG Viewer
    - user evaluation of results, versioning
    - user prioritisation of OCR tasks (On-Demand OCR)

OPERANDI

OCR-D Performance Optimisation and Integration

PROJECT VISION

We want to create full-text capture that is up to 80 times faster
- up to 60 pages per minute.

THANK YOU!

Contact: Lilja Sautter, sautter@sub.uni-goettingen.de

PROJECT GOALS

- Implementation package for mass digitisation
- Adaptive, parallelised workflows
- Task management and prioritisation
- Asynchronous interprocess communication via API
- Provide a simple-to-use implementation package
- Simple data storage
- Easy transfer to alternative processing environments (cloud)

OCR4all-libraries
Full-Text Transformation of Historical Collections

Anke Hertling
Georg Eckert Institute for International Textbook Research (GEI)

Christian Reul
Centre for Philology and Digitality (ZPD); University of Würzburg

30.07.2021

Project Goals

1) Comfortable application of OCR-D solutions for non-technical users via OCR4all

- Full control without using the CLI
- several OCR-D processors for each step of the workflow
- optimization Usability and User Experience
- Embedding OCR4all in library digitisation workflow

2) Ensure and optimize the quality of the OCR result

- best workflows (processors, settings) for different works / collections / pages types
- Comparing/Evaluating workflows with / without Ground Truth (GT)
- GT Export and Training Configuration
- Defining Training and Evaluations Sets using Tagging

GEORG ECKERT INSTITUTE
for International Textbook Research

zpd

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

- Project Partners

  Dr. Anke Hertling (GEI)
  hertling@gei.de

  Dr. Christian Reul (ZPD Würzburg)
  christian.reul@uni-wuerzburg.de

  Prof. Dr. Marc Erich Latoschik (HCI Würzburg)
  marc.latoschik@uni-wuerzburg.de

- External Cooperation Partners

  UB Heidelberg

  USB Köln

ODEM:
OCR-**D** **E**rweiterung für **M**assendigitalisierung

(**O**CR-**D** **E**xtension for **M**ass digitization)

Uwe Hartwig
Daniel Brenn
Universitäts- und Landesbibliothek Sachsen-Anhalt
uwe.hartwig@bibliothek.uni-halle.de
daniel.brenn@bibliothek.uni-halle.de

UNIVERSITÄTS- UND
LANDESBIBLIOTHEK
SACHSEN - ANHALT

- *vast **knowledge wastelands** stretch before the mind*
  millions of pages carry billions of letters ==>

I. vehicles **harvest standarized** (OAI) to feed OCR-System
simple worker machines with ocr-d-containers scale horizontal
ever-improving ocr pushed to target

II. integrated processes **adapt construction changes fast**
feedback for experimental stuff from large VD data sets

III. runs **automated** - only alerted by serious disturbances
no human supervision at run-time

IV. **control quality** of processing steps and final OCR-melange
groundtruth + quality estimation service + …

==> *perpetually recognition runs **increase knowledge base***

Lightning Talk

OCR-D module project

Workflow for work-specific training based on generic models with OCR-D and upgrading of ground truth data

Jan Kamlah / Thomas Schmidt (Universitätsbibliothek Mannheim) / **29th – 30th July 2021**

Goals of the project

Enable users to perform high-quality text-recognition based on the OCR-D workflow with work-specific training for **Tesseract** and **Calamari**.

1 Develop and implement a fast and reliable workflow for training

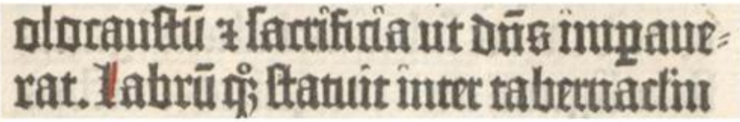
2 Select and improve suitable ground truth data

3 Qualify software tools for the correction and revaluation of ground truth

4 Create a public repository for trained models and ground truth

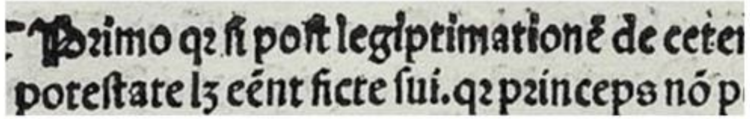
Workflow for work-specific training based on generic models with OCR-D and upgrading of ground truth data

Font Group Recognition



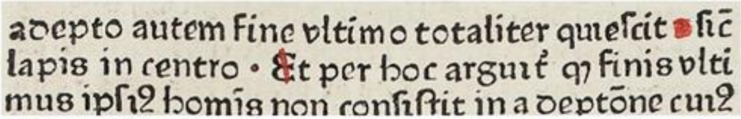
olocauſtū ⁊ ſacrificia ut dñs impaue-
rat. Labrū q̄ȝ ſtatuit inter tabernacłm

Textura



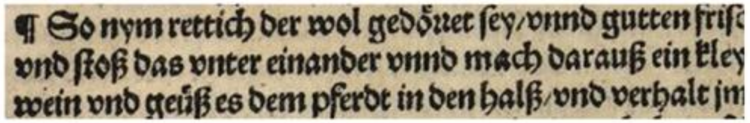
Pꝛimo qꝛ ſi poſt legiptimationē de ceter
poteſtate lȝ eēnt ficte ſui. qꝛ pꝛinceps nō p

Rotunda



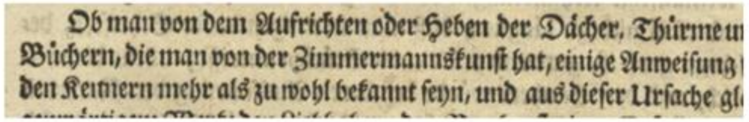
adepto autem fine vltimo totaliter quieſcit ſic
lapis in centro • Et per hoc arguit q̄ finis vlti
mus ipſiꝰ homīs non conſiſtit in a deptōne cuiꝰ

Gotico-Antiqua



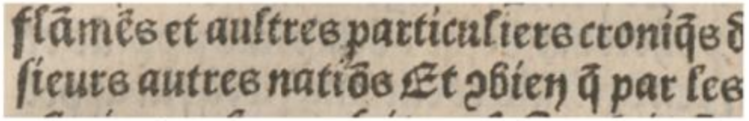
¶ So nym rettich der wol gedőret ſey/vnnd gutten friſ
vnd ſtoß das vnter einander vnnd mach darauß ein kley
wein vnd geüß es dem pferdt in den halß/vnd verhalt jm

Schwabacher



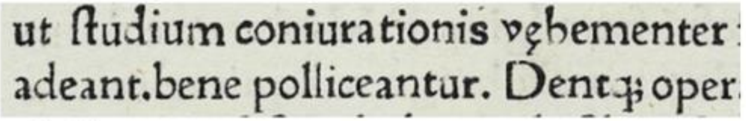
Ob man von dem Aufrichten oder Heben der Dächer, Thürme u
Büchern, die man von der Zimmermannskunſt hat, einige Anweiſung
den Kennern mehr als zu wohl bekannt ſeyn, und aus dieſer Urſache gl

Fraktur



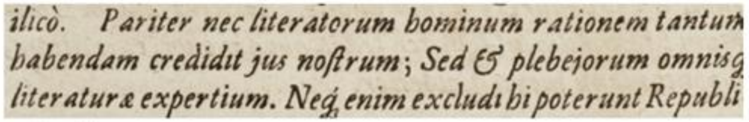
flāmēs et auſtres particuliers croniq̄s d
ſieurs autres natiōs Et ꝯbien q̄ par ſes

Bastarda



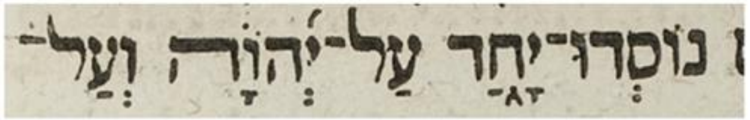
ut ſtudium coniurationis vehementer
adeant.bene polliceantur. Dentq; oper.

Antiqua



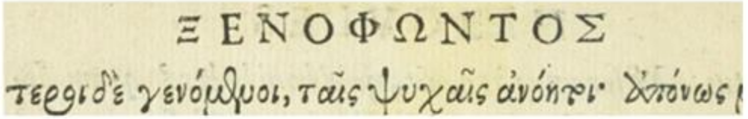
ilicò. Pariter nec literatorum hominum rationem tantum
habendam credidit jus noſtrum; Sed & plebejorum omnisq
literaturæ expertium. Neq; enim excludi hi poterunt Republi

Italic



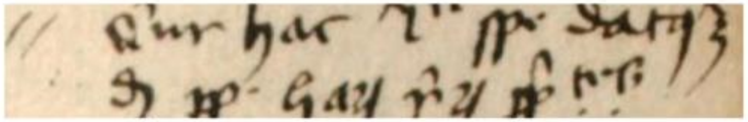
נוֹסְדוּ־יָחַד עַל־יְהוָה וְעַל־

Hebrew



ΞΕΝΟΦΩΝΤΟΣ
περὶ δὲ γενόμενοι, ταῖς ψυχαῖς ἀνόητοι· ἀπόνως

Greek

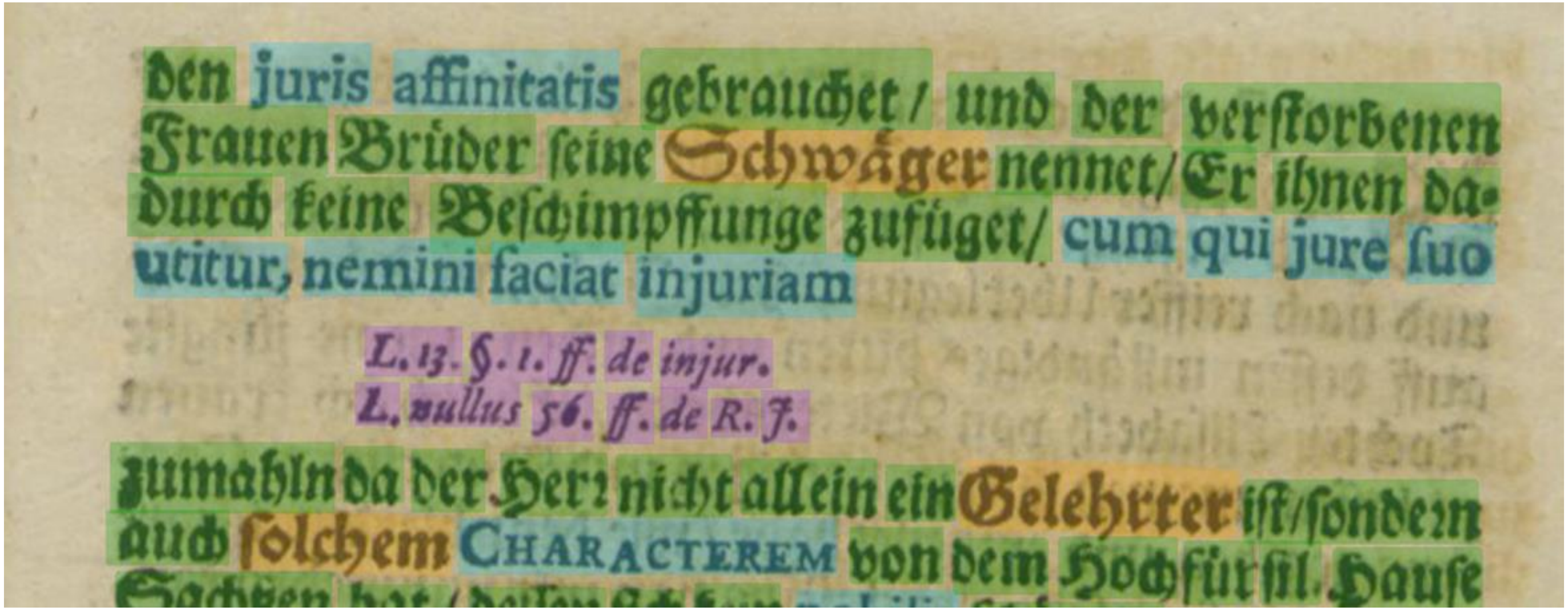


[illegible]

Other (including manuscript)

Objectives of the New Project

- Development of a more fine-granular font group recognition tool
- Generation of font-group-specific OCR training-data
- Training of font-group-specific OCR-models



den juris affinitatis gebrauchet / und der verſtorbenen Frauen Brůder ſeine Schwåger nennet/ Er ihnen dadurch keine Beſchimpffunge zufüget/ cum qui jure ſuo utitur, nemini faciat injuriam

L. 13. §. 1. ff. de injur.
L. nullus 56. ff. de R. J.

zumahln da der Herꝛ nicht allein ein Gelehrter iſt/ ſondern auch ſolchem CHARACTEREM von dem Hochfürſtl. Hauſe

vincent.christlein@fau.de

weichsel@uni-mainz.de

OLA-HD Service

Generic service for the long-term storage of historical prints

Project Vision

We want to have super fast transfer to an archive that offers 100% reliability, searchability and fine-grained referencing.

Thank you!

Contact: Lilja Sautter, sautter@sub.uni-goettingen.de

Project Goals

1) Optimise search and filter
2) Interim storage within OCR workflow (hot storage)
3) API specification
4) Interface linking archive to presentation systems
5) Roles and rights system
6) Service provision
7) Frameworks and best practices for the service
8) Implementation package: integration into OCR workflows

GWDG
Gesellschaft für wissenschaftliche
Datenverarbeitung mbH Göttingen

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

NIEDERSÄCHSISCHE STAATS- UND
UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN
SUB