FID Nahost
NORDAFRIKA UND ISLAMSTUDIEN

The Specialised Information Service Middle East

Digitization - OCR/HTR – what comes next?

The conversion of non-Latin scripts into reliable data sets

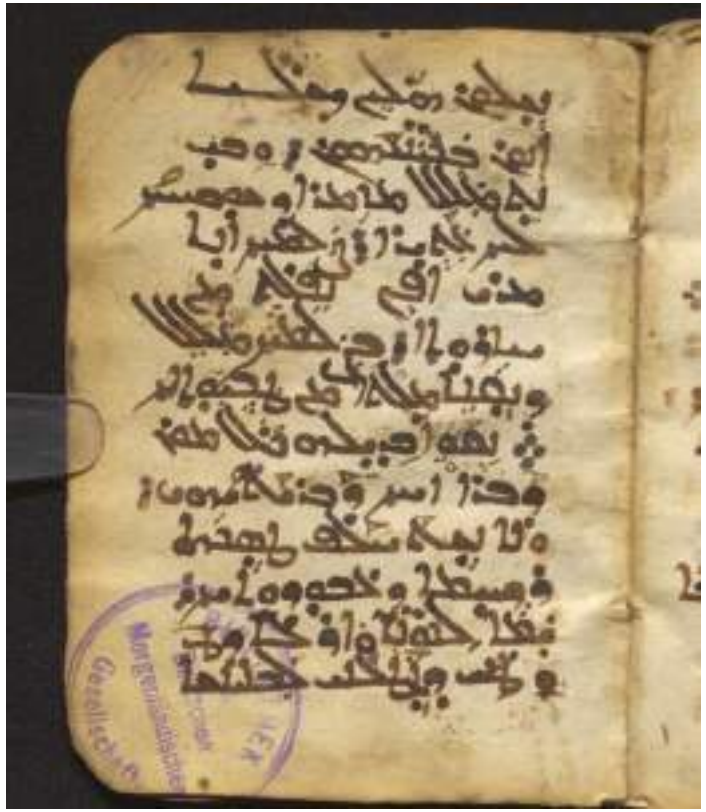
Dr. Volker Adam, 22.05.2025, Paris

UNIVERSITÄTS- UND LANDESBIBLIOTHEK
SACHSEN-ANHALT

DFG Deutsche Forschungsgemeinschaft

Digitised Cultural Heritage

Middle East libraries have digitized valuable, copyright-free sources with public funds (manuscripts, periodicals, monographs ...)







Digitised Cultural Heritage

They offer the digitized material including high-quality metadata for use (III viewer, PDF download)



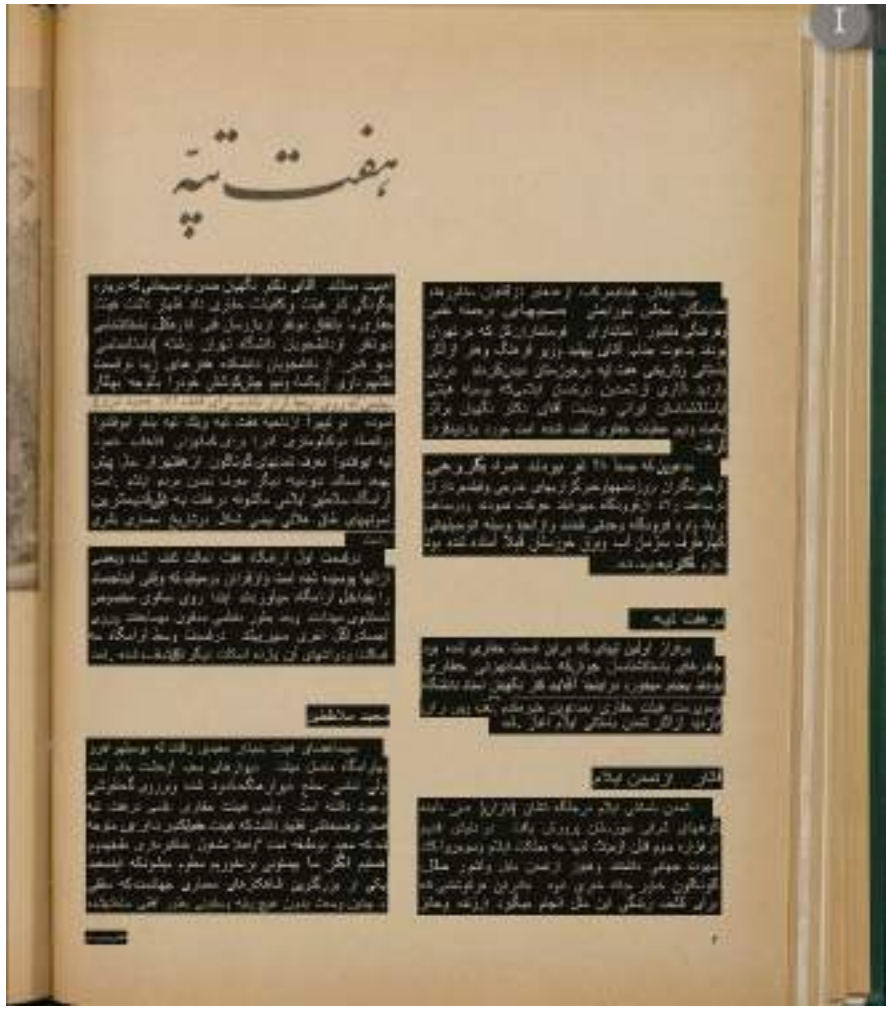
Metadata


```
<mets:mets xmlns:doc="http://www.lyncode.com/xoai" xmlns:a="https:/zvdd" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns"
xmlns:b="https://bibliothek.uni-halle.de" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:mods="http://www.loc.gov/mods/v3" xmlns:dv="http://dfg-
viewer.de/" xmlns:mets="http://www.loc.gov/METS/" xsi:schemaLocation="http://www.loc.gov/mods/v3
http://www.loc.gov/standards/mods/mods.xsd http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/mets.xsd">
  <mets:metsHdr xmlns:ulb="https://bibliothek.uni-halle.de" CREATEDATE="2024-10-10T13:15:37" LASTMODDATE="2024-10-11T05:12:37Z">
    <mets:agent OTHERTYPE="SOFTWARE" ROLE="CREATOR" TYPE="OTHER">
      <mets:name>Kitodo - kitodo-ugh-2.1.4-SNAPSHOT-v1.11.1-108-g2b516b3 - 03-January-2021</mets:name>
      <mets:note>Kitodo</mets:note>
    </mets:agent>
    <mets:agent TYPE="OTHER" ROLE="OTHER" OTHERTYPE="SOFTWARE">
      <mets:name>DigitalDerivans V1.8.4</mets:name>
      <mets:note>PDF FileGroup for PDF_180036390719282 created at 2024-10-10T15:22:06</mets:note>
    </mets:agent>
  </mets:metsHdr>
  <mets:dmdSec xmlns:ulb="https://bibliothek.uni-halle.de" ID="DMDLOG_0001">
    <mets:mdWrap MDTYPE="MODS">
      <mets:xmlData>
        <mods:mods>
          <mods:identifier type="urn">urn:nbn:de:gbv:3:5-1192015415-180036390719282-17</mods:identifier>
          <mods:relatedItem type="host">
            <mods:recordInfo>
              <mods:recordIdentifier source="gvk-ppn">1800363907</mods:recordIdentifier>
            </mods:recordInfo>
          </mods:relatedItem>
          <mods:recordInfo>
            <mods:recordIdentifier source="gvk-ppn">180036390719282</mods:recordIdentifier>
```

Digitised Cultural Heritage

Some use or even develop techniques to not only provide images, but also to recognize and extract the textual content (HTR / OCR)





Open source tools for full-text recognition

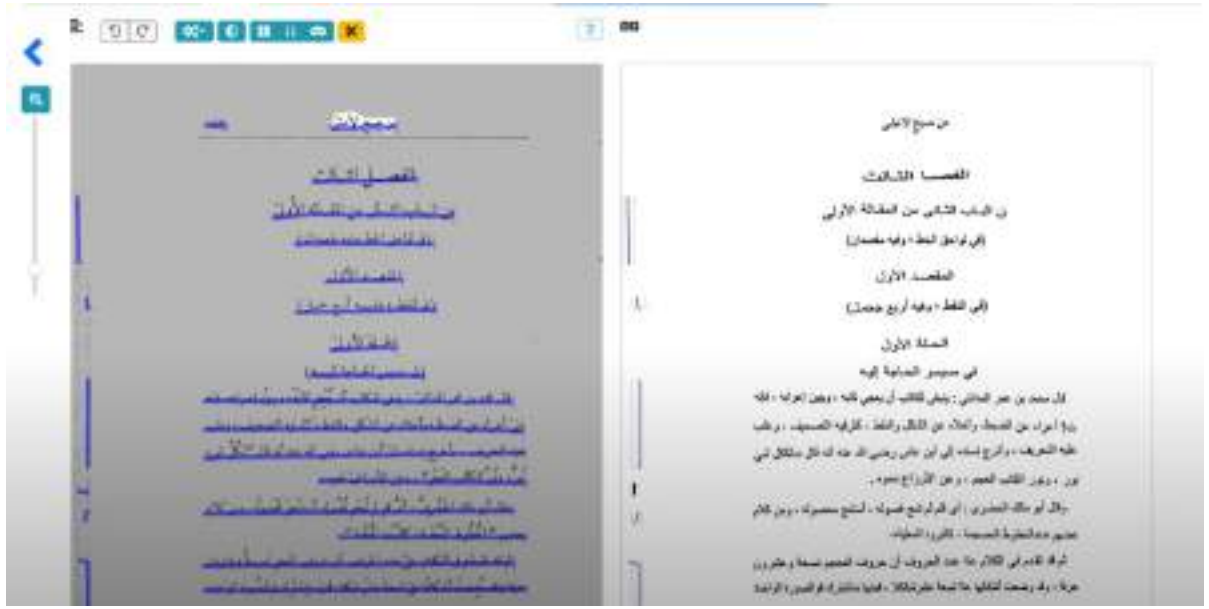
Transkribus:

Provider	Supported Languages / Focus	Key Features	Pricing / Licensing
Transkribus	Handwritten and printed texts with a strong emphasis on historical documents. Supports diverse scripts (including many Latin, Arabic, and regional historical forms) and is especially popular among the digital humanities and academic communities.	AI-powered handwriting recognition (HTR) and printed OCR, advanced transcription editor for correction and formatting retention, custom OCR/HTR model training based on your document types, integrated search capabilities, and access to a growing repository of public models developed by the community and the Transkribus team.	Accounts are issued with free monthly credits (approximately 100 pages for handwritten texts or 600 pages for printed material). Additional usage requires subscription plans, and special support (e.g., grants for students and researchers) is also available.



Open source tools for full-text recognition

eSciptorium:



Provider	Supported Languages / Focus	Key Features	Pricing / Licensing
eScriptorium	Focuses on historical manuscripts and printed texts, making it ideal for archival research and digital humanities. It supports various scripts found in historical documents.	An open-source platform designed for both manual and automated segmentation and OCR. It leverages OCR engines such as Kraken—with optional Tesseract extensions—for robust text recognition. Additionally, it provides transcription editing tools, customizable workflows, and flexible local installation options, making it a strong alternative to commercial solutions like Transkribus.	Free to use (open source) with self-hosted installation; supported and continuously improved by a community of developers.

Open source tools for full-text recognition

Provider	Supported Languages / Focus	Key Features	Pricing / Licensing
LAREX	Primarily designed for historical printed texts and manuscripts. It focuses on the layout analysis and region segmentation of early printed books and archival documents.	A semi-automatic, rule-based tool for segmenting document layouts. It enables fine-grained classification of regions (e.g., text blocks, images, headings, marginalia) with an intuitive manual correction interface. LAREX generates PAGE XML output to integrate seamlessly into broader OCR workflows, notably within the OCR4all ecosystem.	Free to use; open-source and released under the MIT License, available on GitHub.

LAREX:



What comes next?

We integrate the OCR-generated full texts into our catalog search

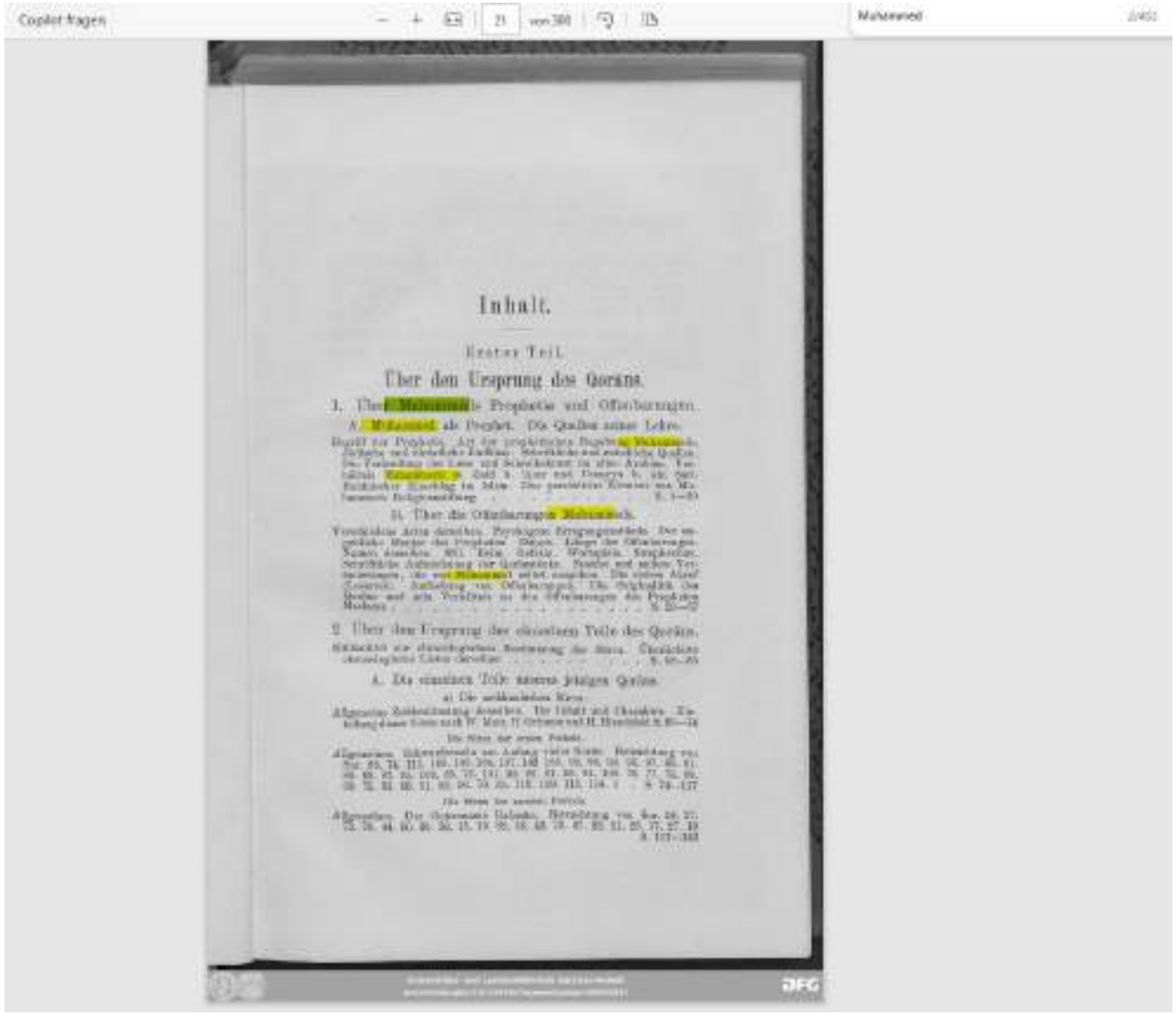




What comes next?

Text Recognition: Creating Searchable PDFs





What if…

…libraries took over all the tasks themselves?

- Open-source OCR engines as a foundation (e.g., Tesseract, Kraken)
- Integration of right-to-left (RTL) support and handling of ligatures, diacritics
- Custom modules to process calligraphic and historical scripts
- Continuous benchmarking and error analysis pipelines
- Structured corpus with metadata management
- Extending eScriptorium as a modular platform with well-documented APIs
- computational linguistics methods, such as dictionary-based error detection and contextual correction
- Seamless integration with existing library systems
- Interdisciplinary teams (IT, linguistics, library science)

From text to corpora



What application scenarios should we develop in future for the use of the text corpora that we make available via our interfaces?

- Which library is already creating text corpora from the digitized data?
- How are the corpora made available (internally or worldwide)?
- For which scientific applications are the recognized texts made available?
- Which tools are used to make this data available in a world of LLM, TDM, LoD, semantic links, etc.?
- What usage rights are granted to the texts and how is this digitally documented and identified?

From text to corpora





Orientalia collection in the Gotha Research Library

What application scenarios should we develop in future for the use of the text corpora that we make available via our interfaces?

- To what extent can commercial providers of digital content and applications be integrated into a cross-regional offering, or are solutions only built for the users of the library itself?
- When data is stored in repositories, which licenses govern its protection, and do AI bots have access to it? How can libraries safeguard their data from unauthorized AI-driven usage?
- What is the relationship between digitized public library collections and IT-strong providers of digital content (digitized periodicals, archives, etc.) or AI application processes?
- Should we voluntarily make our data sets available to AI companies to exploit and generate added value?

From text to corpora

What application scenarios should we develop in future for the use of the text corpora that we make available via our interfaces?

- Do we allow the digital corpora we have created to be mined so that double-dipping subscription models are created in the end? How should we deal with mass queries from crawlers?
- Or should we keep corpora in-house and develop personalized application services for the corpora datasets we create? E.g. offer paid services like N-gram services etc.?
- Or should libraries just continue to offer “beautiful digital images” and forget about the textual content?



Siṃha, Śrīdhara [HerausgeberIn] [Hrsg.]: Adbhuta carcā (Lakhanaū, 1949)



CrossAsia | CrossAsia N-gram Service

Fachinformationsdienst Nahost-, Nordafrika- und Islamstudien

خدمة تقديم المعلومات المختصة بدراسة الشرق الاوسط وشمال افريقيا والدراسات الاسلامية

Orta Doğu, Kuzey Afrika ve İslam Araştırmaları Bilgi Hizmetleri

اطلاعات و خدمات کارشناسی در زمینهٔ مطالعات خاورمیانه و شمال افریقا و پژوهشهای اسلامی

Specialised Information Service Middle East, North Africa and Islamic Studies

Thank you for your attention!

Get in touch with us to share your experiences and projects!

volker.adam@bibliothek.uni-halle.de
maida-farooq.buttar@bibliothek.uni-halle.de

UNIVERSITÄTS- UND LANDESBIBLIOTHEK SACHSEN-ANHALT

DFG Deutsche Forschungsgemeinschaft

www.menalib.de