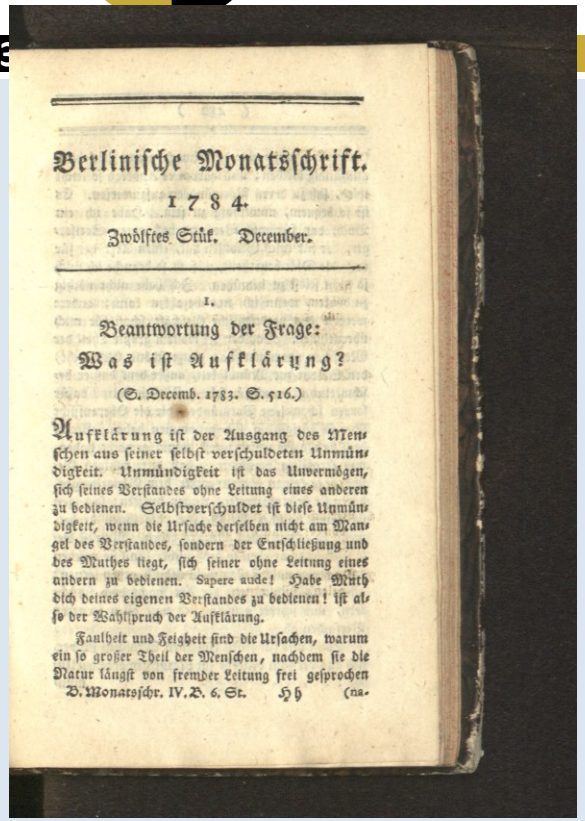
Massendigitalisierung mit OCR-D

2022

8. Bibliothekskongress

Was ist OCR-D?

- Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR)
- Ziel: Volltexttransformation der VDs
  - Phase I (2015–2018): Analyse von Bedarfen und Funktionsmodell
  - Phase II (2018–2020): acht Modulprojekte
  - Phase III (2021–2024): vier Implementierungs- und drei Modulprojekte, produktiven Einsatz vorbereiten
- Frei verfügbare Tools
- Individuelle Workflows für OCR möglich



Berlinische Monatsschrift.

1784.

Zwölftes Stük. December.

I.

Beantwortung der Frage:

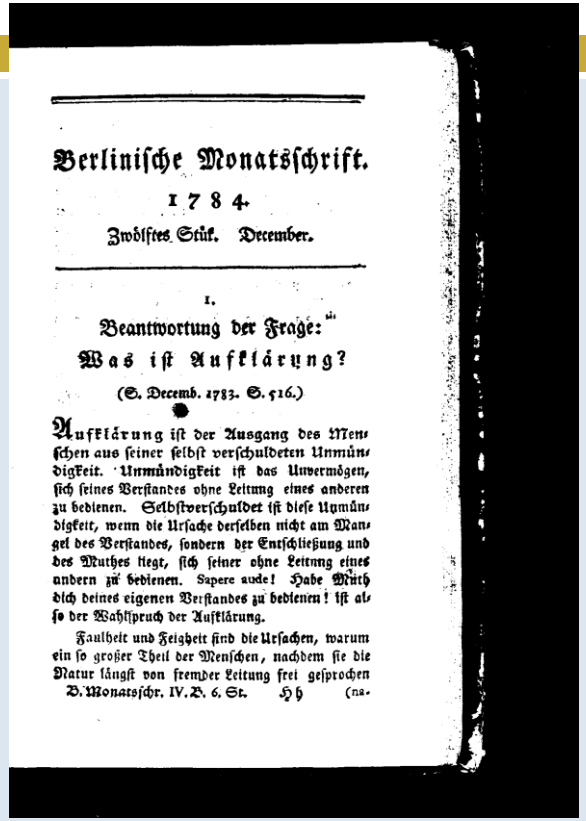
Was ist Aufklärung?

(S. Decemb. 1783. S. 516.)

Aufklärung ist der Ausgang des Menschen aus seiner selbst verschuldeten Unmündigkeit. Unmündigkeit ist das Unvermögen, sich seines Verstandes ohne Leitung eines anderen zu bedienen. Selbstverschuldet ist diese Unmündigkeit, wenn die Ursache derselben nicht am Mangel des Verstandes, sondern der Entschließung und des Muthes liegt, sich seiner ohne Leitung eines andern zu bedienen. Sapere aude! Habe Muth dich deines eigenen Verstandes zu bedienen! ist also der Wahlspruch der Aufklärung.

Faulheit und Feigheit sind die Ursachen, warum ein so großer Theil der Menschen, nachdem sie die Natur längst von fremder Leitung frei gesprochen

B. Monatsschr. IV. B. 6. St. Hh (na-



Berlinische Monatsschrift.

1784.

Zwölftes Stük. December.

I.

Beantwortung der Frage:

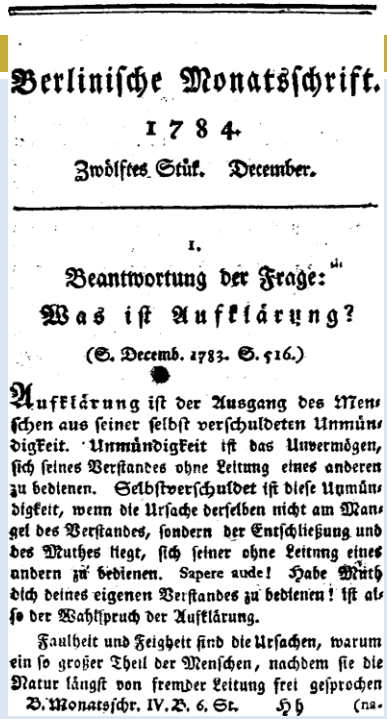
Was ist Aufklärung?

(S. Decemb. 1783. S. 516.)

Aufklärung ist der Ausgang des Menschen aus seiner selbst verschuldeten Unmündigkeit. Unmündigkeit ist das Unvermögen, sich seines Verstandes ohne Leitung eines anderen zu bedienen. Selbstverschuldet ist diese Unmündigkeit, wenn die Ursache derselben nicht am Mangel des Verstandes, sondern der Entschließung und des Muthes liegt, sich seiner ohne Leitung eines andern zu bedienen. Sapere aude! Habe Muth dich deines eigenen Verstandes zu bedienen! ist also der Wahlspruch der Aufklärung.

Faulheit und Feigheit sind die Ursachen, warum ein so großer Theil der Menschen, nachdem sie die Natur längst von fremder Leitung frei gesprochen

B. Monatsschr. IV. B. 6. St. Hh (na-



Berlinische Monatsschrift.

1784.

Zwölftes Stük. December.

I.

Beantwortung der Frage:

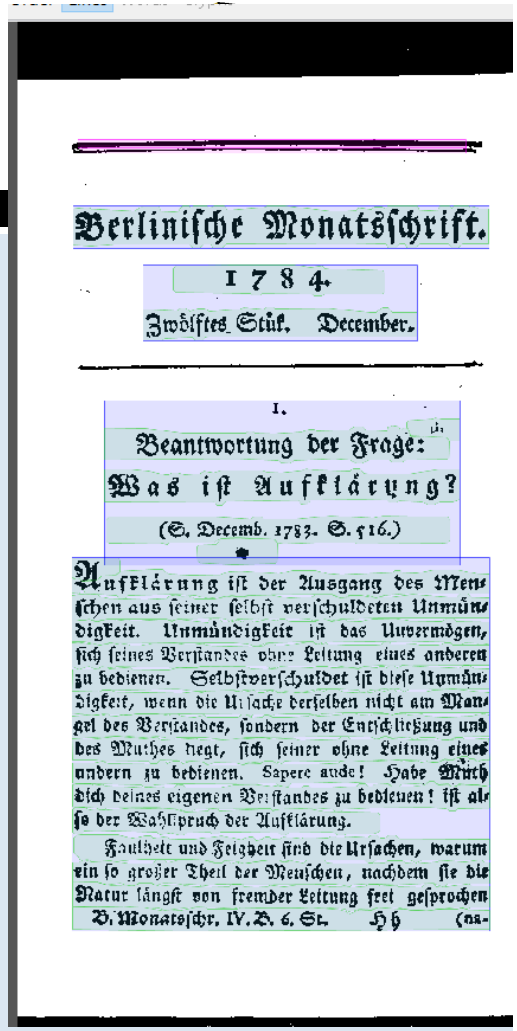
Was ist Aufklärung?

(S. Decemb. 1783. S. 516.)

Aufklärung ist der Ausgang des Menschen aus seiner selbst verschuldeten Unmündigkeit. Unmündigkeit ist das Unvermögen, sich seines Verstandes ohne Leitung eines anderen zu bedienen. Selbstverschuldet ist diese Unmündigkeit, wenn die Ursache derselben nicht am Mangel des Verstandes, sondern der Entschließung und des Muthes liegt, sich seiner ohne Leitung eines andern zu bedienen. Sapere aude! Habe Muth dich deines eigenen Verstandes zu bedienen! ist also der Wahlspruch der Aufklärung.

Faulheit und Feigheit sind die Ursachen, warum ein so großer Theil der Menschen, nachdem sie die Natur längst von fremder Leitung frei gesprochen

B. Monatsschr. IV. B. 6. St. Hh (na-



Berlinische Monatsschrift.

1784.

Zwölftes Stük. December.

I.

Beantwortung der Frage:

Was ist Aufklärung?

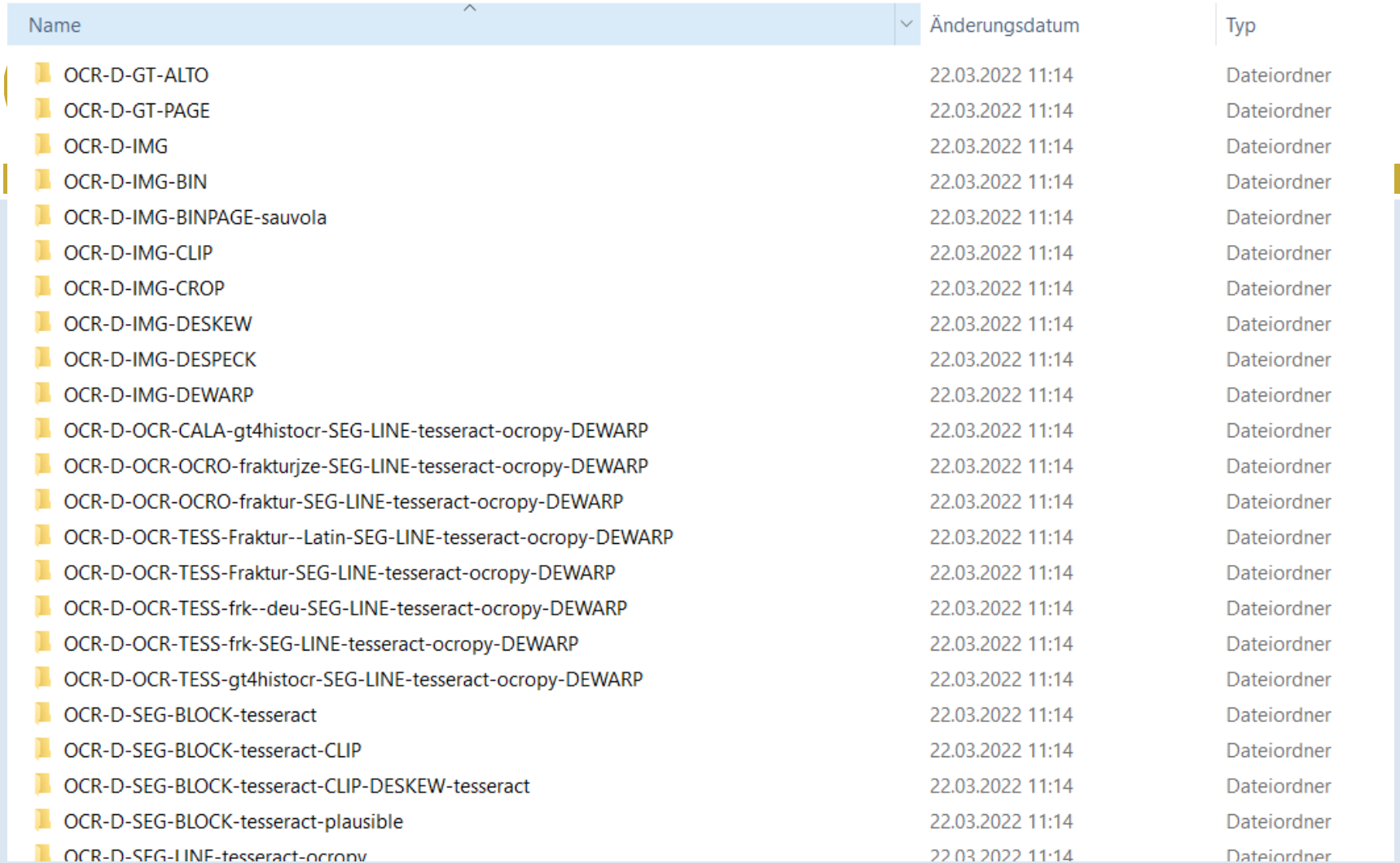
(S. Decemb. 1783. S. 516.)

Aufklärung ist der Ausgang des Menschen aus seiner selbst verschuldeten Unmündigkeit. Unmündigkeit ist das Unvermögen, sich seines Verstandes ohne Leitung eines anderen zu bedienen. Selbstverschuldet ist diese Unmündigkeit, wenn die Ursache derselben nicht am Mangel des Verstandes, sondern der Entschließung und des Muthes liegt, sich seiner ohne Leitung eines andern zu bedienen. Sapere aude! Habe Muth dich deines eigenen Verstandes zu bedienen! ist also der Wahlspruch der Aufklärung.

Faulheit und Feigheit sind die Ursachen, warum ein so großer Theil der Menschen, nachdem sie die Natur längst von fremder Leitung frei gesprochen

B. Monatsschr. IV. B. 6. St. Hh (na-

```
    <pc:TextEquiv>
        <pc:Unicode>fich feines Verstandes ohne Leitung eines anderen</pc:Unicode>
    </pc:TextEquiv>
    <pc:TextStyle fontFamily="Arial" fontSize="7.5"/>
</pc:TextLine>
<pc:TextLine id="tl_12" primaryLanguage="German" custom="readingOrder {index:4;} te
    <pc:Coords points="111,1271 921,1271 921,1311 111,1311"/>
    <pc:Baseline points="111,1304 921,1304"/>
    <pc:Word id="w_w1aab1b3b2b7b9ab1" language="German" custom="readingOrder {index
        <pc:Coords points="111,1279 146,1279 146,1309 111,1309"/>
        <pc:TextEquiv>
            <pc:Unicode>zu</pc:Unicode>
        </pc:TextEquiv>
        <pc:TextStyle fontFamily="Arial" fontSize="7.5"/>
    </pc:Word>
    <pc:Word id="word_1478541388499_834" language="German" custom="readingOrder {in
        <pc:Coords points="290,1302 161,1302 161,1273 290,1273"/>
        <pc:TextEquiv>
            <pc:Unicode>bedienen</pc:Unicode>
        </pc:TextEquiv>
        <pc:TextStyle fontFamily="Arial" fontSize="7.5"/>
    </pc:Word>
    <pc:Word id="word_1478541388497_833" language="German" custom="readingOrder {in
        <pc:Coords points="299,1302 290,1302 290,1273 299,1273"/>
        <pc:TextEquiv>
            <pc:Unicode>.</pc:Unicode>
        </pc:TextEquiv>
        <pc:TextStyle fontFamily="Arial" fontSize="7.5"/>
    </pc:Word>
    <pc:Word id="w_w1aab1b3b2b7b9ac27" language="German" custom="readingOrder {inde
        <pc:Coords points="338,1272 648,1272 648,1310 338,1310"/>
        <pc:TextEquiv>
            <pc:Unicode>Selbftverfchuldet</pc:Unicode>
```



Name	Änderungsdatum	Typ
OCR-D-GT-ALTO	22.03.2022 11:14	Dateiordner
OCR-D-GT-PAGE	22.03.2022 11:14	Dateiordner
OCR-D-IMG	22.03.2022 11:14	Dateiordner
OCR-D-IMG-BIN	22.03.2022 11:14	Dateiordner
OCR-D-IMG-BINPAGE-sauvola	22.03.2022 11:14	Dateiordner
OCR-D-IMG-CLIP	22.03.2022 11:14	Dateiordner
OCR-D-IMG-CROP	22.03.2022 11:14	Dateiordner
OCR-D-IMG-DESKEW	22.03.2022 11:14	Dateiordner
OCR-D-IMG-DESPECK	22.03.2022 11:14	Dateiordner
OCR-D-IMG-DEWARP	22.03.2022 11:14	Dateiordner
OCR-D-OCR-CALA-gt4histocr-SEG-LINE-tesseract-ocropy-DEWARP	22.03.2022 11:14	Dateiordner
OCR-D-OCR-OCRO-frakturjze-SEG-LINE-tesseract-ocropy-DEWARP	22.03.2022 11:14	Dateiordner
OCR-D-OCR-OCRO-fraktur-SEG-LINE-tesseract-ocropy-DEWARP	22.03.2022 11:14	Dateiordner
OCR-D-OCR-TESS-Fraktur--Latin-SEG-LINE-tesseract-ocropy-DEWARP	22.03.2022 11:14	Dateiordner
OCR-D-OCR-TESS-Fraktur-SEG-LINE-tesseract-ocropy-DEWARP	22.03.2022 11:14	Dateiordner
OCR-D-OCR-TESS-frk--deu-SEG-LINE-tesseract-ocropy-DEWARP	22.03.2022 11:14	Dateiordner
OCR-D-OCR-TESS-frk-SEG-LINE-tesseract-ocropy-DEWARP	22.03.2022 11:14	Dateiordner
OCR-D-OCR-TESS-gt4histocr-SEG-LINE-tesseract-ocropy-DEWARP	22.03.2022 11:14	Dateiordner
OCR-D-SEG-BLOCK-tesseract	22.03.2022 11:14	Dateiordner
OCR-D-SEG-BLOCK-tesseract-CLIP	22.03.2022 11:14	Dateiordner
OCR-D-SEG-BLOCK-tesseract-CLIP-DESKEW-tesseract	22.03.2022 11:14	Dateiordner
OCR-D-SEG-BLOCK-tesseract-plausible	22.03.2022 11:14	Dateiordner
OCR-D-SEG-LINE-tesseract-ocropy	22.03.2022 11:14	Dateiordner

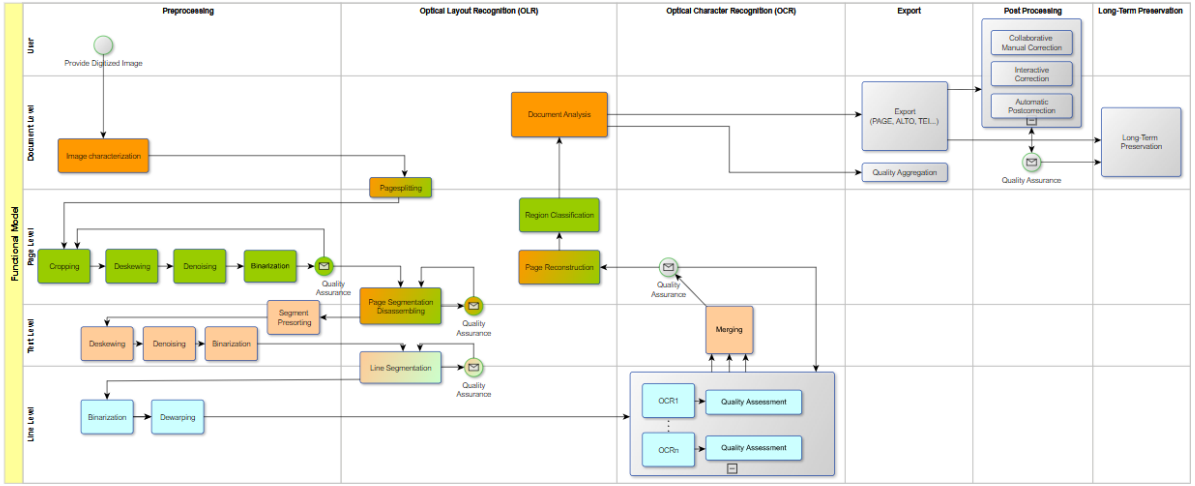
Unterstützte OCR-Schritte

- Bildoptimierung
- Schriftartenerkennung
- Binarisierung
- Zuschneiden
- Rauschunterdrückung
- Deskewing
- Dewarping
- (Re-)Segmentierung
- OCR-Nachkorrektur
- Evaluation (Layout/OCR)

ocr-d.de/en/workflows

Workflows

There are several steps necessary to get the fulltext of a scanned print. The whole OCR process is shown in the following figure:



The following instructions describe all steps of an OCR workflow. Depending on your particular print (or rather images), not all of those steps might be necessary to obtain good results. Whether a step is required or optional is indicated in the description of each step. This guide provides an overview of the available OCR-D processors and their required parameters. For more complex workflows and recommendations see the OCR-D-Website-Wiki. Feel free to add your own experiences and recommendations in the Wiki! We will regularly amend this guide with valuable

OCR-D nutzen

- Setup Guide: https://ocr-d.de/en/setup
  - https://github.com/OCR-D/ocrd_all - erster Anlaufpunkt für die Installation
- Voraussetzungen: aktuell Linux 18.04 und Python 3.6/3.7 (oder Docker)
- Kommandozeilentool
- ausprobieren/Code herunterladen/mitentwickeln: https://github.com/OCR-D

Implementierungs- und Modulprojekte

- Implementierungsprojekte
  - Integration von Kitodo und OCR-D zur produktiven Massendigitalisierung
  - OPERANDI: OCR-D Performance Optimisation and Integration
  - OCR4all libraries –Volltexterkennung historischer Sammlungen
  - ODEM: OCR-D Erweiterung für Massendigitalisierung
- Modulprojekte
  - Workflow für werkspezifisches Training auf Basis generischer Modelle mit OCR-D sowie Ground Truth Aufwertung
  - Font Group Recognition for Improved OCR
  - OLA-HD Service -Ein generischer Dienst für die Langzeitarchivierung historischer Drucke

Offene virtuelle Treffen

- OCR-D Forum
- TechCall
- GT-Call
- Termine und Links: https://ocr-d.de/de/community

Kontakt und Dokumentation

- E-Mail: hinrichsen@hab.de (Lena Hinrichsen)
- Chat: https://gitter.im/OCR-D/Lobby
- Website: https://ocr-d.de/de/
- Tutorials und weitere Ressourcen:

https://github.com/OCR-D/ocrd-website/wiki

www.ocr-d.de