Testbericht OCR-D-Teststellung an Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)

Matthias Boenig (BBAW OCR-D) boenig@bbaw.de, Christian Thomas (BBAW) thomas@bbaw.de, Britta Hermann (BBAW Bibliothek) hermann@bbaw.de

Hintergrund

Die Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) ist eine der größten außeruniversitären geisteswissenschaftlichen Forschungseinrichtung in der Region Berlin-Brandenburg. Die 1992 per Staatsvertrag zwischen den Bundesländern Berlin und Brandenburg gegründete Institution folgt in ihrer Tradition und ihrem Anspruch der Kurfürstlich Brandenburgischen Societät der Wissenschaften, die auf Anregung von Gottfried Wilhelm Leibniz 1700 gegründet wurde, und der Preußischen Akademie der Wissenschaften. In der über 300 Jahre währenden Geschichte dieser Einrichtung wurde eine beachtliche Anzahl von wissenschaftlichen Fragestellungen im Rahmen von Projekten bearbeitet, deren Ergebnisse in vielfältiger Art und Weise publiziert wurden. Diese Akademieschriften spiegeln das facettenreiche wissenschaftliche Themenspektrum dieser Institution und ihrer Wissenschaftler wider und sind in der retrospektiven Betrachtung für Interessierte und Forschende ein ergiebiger Quellenfundus.

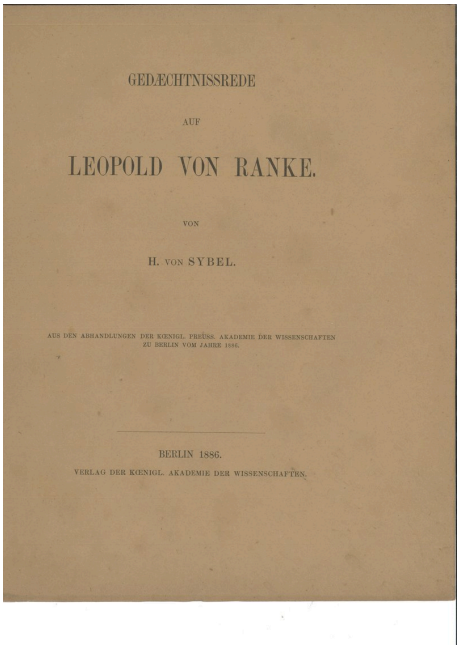
Aus diesem Grund wurden diese Schriften um die Jahrtausendwende im Rahmen eines von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekts bilddigitalisiert. Eine Volltextdigitalisierung mit Hilfe von OCR konnte zum damaligen Zeitpunkt nicht realisiert werden, da die technischen Voraussetzungen dafür noch nicht geschaffen waren. Zwanzig Jahre später kann nun im Rahmen der OCR-D-Teststellung an der BBAW anhand ausgewählter Beispiele mit der Realisierung der Volltextdigitalisierung begonnen werden.

Übersicht der Akademieschriften an der BBAW: http://bibliothek.bbaw.de/bibliothek-digital/digitalequellen/schriften

Vorlagen

Testkorpus

Das Testkorpus umfasst 13 Gedächtnisreden auf Mitglieder der Akademie. Diese Reden wurden sowohl als Sonderdruck als auch in den jeweiligen Jahresberichten bzw. Abhandlungen der Königlichen Preußischen Akademie der Wissenschaften zu Berlin veröffentlicht.


GEDÆCHTNISSREDE

AUF

LEOPOLD VON RANKE.

VON

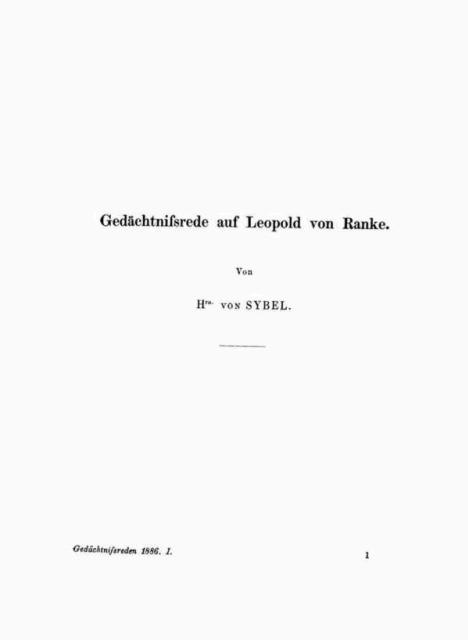
H. VON SYBEL.

AUS DEN ABHANDLUNGEN DER KÖNIGL. PREUSS. AKADEMIE DER WISSENSCHAFTEN ZU BERLIN VOM JAHRE 1886.

BERLIN 1886.

VERLAG DER KÖNIGL. AKADEMIE DER WISSENSCHAFTEN.

Sonderdruck


Gedächtnifsrede auf Leopold von Ranke.

Von

Hrn. von SYBEL.

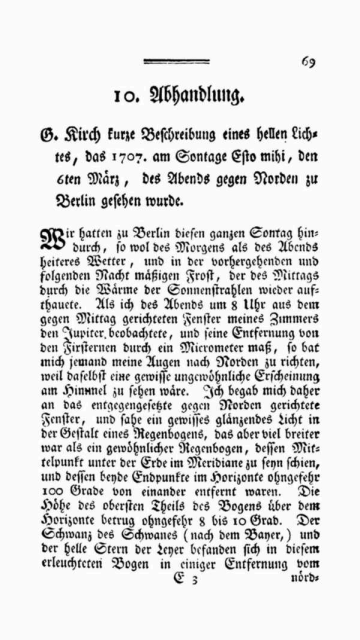
Gedächtnifsreden 1886. I. 1

Veröffentlichung in den Abhandlungen der Königlichen Preußischen Akademie der Wissenschaften zu Berlin

Typographie

Die Gedächtnisreden sind in einzelne Absätze und kapitelähnliche Teile gegliedert. Wie in wissenschaftlichen Publikationen üblich, finden sich Fußnoten und Hervorhebungen im Text, z.B. bei Personennamen. Da die betrachteten Texte zwischen 1886 und 1911 publiziert wurden, handelt es sich um industriell hergestellte Drucke mit einem gleichförmigen Druckbild.

Als Brotschrift wurden für diese Drucke, wie überhaupt für die Mehrheit der Akademieschriften der BBAW von 1700 bis 1900, Antiqua-Typen verwendet. Durch den wissenschaftlichen Austausch und die Abfassung von Publikationen in den damaligen Wissenschaftssprachen Latein und Französisch war Antiqua in dieser Zeit weitaus gebräuchlicher als Fraktur. Fraktur und weitere gebrochene Schriften finden sich bei den Ende des 18. Jahrhunderts herausgegebenen deutschen Publikationen.



69

10. Abhandlung.

G. Kirch kurze Beschreibung eines hellen Lichtes, das 1707. am Sontage Esto mihi, den 6ten März, des Abends gegen Norden zu Berlin gesehen wurde.

Wir hatten zu Berlin diesen ganzen Sontag hindurch, so wol des Morgens als des Abends heiteres Wetter, und in der vorhergehenden und folgenden Nacht mäßigen Frost, der des Mittags durch die Wärme der Sonnenstrahlen wieder aufthauete. Als ich des Abends um 8 Uhr aus dem gegen Mittag gerichteten Fenster meines Zimmers den Jupiter beobachtete, und seine Entfernung von den Fixsternen durch ein Micrometer maß, so bat mich jemand meine Augen nach Norden zu richten, weil daselbst eine gewisse ungewöhnliche Erscheinung am Himmel zu sehen wäre. Ich begab mich daher an das entgegengesetzte gegen Norden gerichtete Fenster, und sahe ein gewisses glänzendes Licht in der Gestalt eines Regenbogens, das aber viel breiter war als ein gewöhnlicher Regenbogen, dessen Mittelpunkt unter der Erde im Meridiane zu seyn schien, und dessen beyde Endpunkte im Horizonte ohngefehr 100 Grade von einander entfernt waren. Die Höhe des obersten Theils des Bogens über dem Horizonte betrug ohngefehr 8 bis 10 Grad. Der Schwanz des Schwanes (nach dem Bayer,) und der helle Stern der Leyer befanden sich in diesem erleuchteten Bogen in einiger Entfernung vom

E 3 nörd-

Kirch, Gottfried: Kurze Beschreibung eines hellen Lichtes, das 1707. am Sontage Esto mihi, den 6ten März, des Abends gegen Norden zu Berlin gesehen wurde. in: Physikalische und medicinische Abhandlungen der Königlichen Academie der Wissenschaften zu Berlin. Erster Band (Gotha: Ettinger, 1781) http://bibliothek.bbaw.de/bbaw/bibliothek-digital/digitalequellen/schriften/anzeige/index_-html?band=04-phys/1&seite:int=85

Ab 1793 wird in der *Sammlung der deutschen Abhandlungen, welche in der Königlichen Akademie der Wissenschaften zu Berlin vorgelesen worden*, Antiqua auch für deutschsprachige Texte verwendet. In den darauffolgenden Jahren werden in der Mehrheit nur noch diese Typen verwendet und nicht mehr die gebrochenen Schriften.

Zwar konzentriert sich das OCR-D Projekt in der Umsetzung der Volltextdigitalisierung auf Bestände, die zu einem großen Prozentsatz in gebrochenen Schriften (z.B. Fraktur) gesetzt wurden; doch sollte mit diesem Test deren Potential für diesen speziellen Bestand abgeschätzt werden.

Physikalische
und
Medicinische
Abhandlungen
der
Königlichen Academie der Wissenschaften
in Berlin.

Aus dem Lateinischen und Französischen übersetzt
von
Johann Ludwig Conrad Mümler,
Doctor der Arzneygelahrtheit und Landphysicus des
Wolfenbüttelschen Districts.

Vierter Band mit Kupfern.

Gotha,
bey Carl Wilhelm Ettinger.
1786.

SAMMLUNG
DER
DEUTSCHEN
ABHANDLUNGEN,
WELCHE IN DER
KÖNIGLICHEN AKADEMIE
DER
WISSENSCHAFTEN ZU BERLIN
VORGELESEN WORDEN
IN DEN JAHREN 1788. UND 1789.

Preis 1. Thaler.

BERLIN,
gedruckt bey GEORGE DECKER, Königlichem Hofbuchdrucker.
1793.

Physikalische und medicinische Abhandlungen der Königlichen Academie der Wissenschaften zu Berlin Erster Band (Gotha: Ettinger, 1781)

http://bibliothek.bbaw.de/bbaw/bibliothek-digital/digitalequellen/schriften/anzeige?band=04-phys/1

Sammlung der deutschen Abhandlungen, welche in der Königlichen Akademie der Wissenschaften zu Berlin vorgelesen worden in den Jahren 1788-89 (Berlin: Decker, 1793)

http://bibliothek.bbaw.de/bbaw/bibliothek-digital/digitalequellen/schriften/anzeige?band=06-samml/17881789

Digitalisate

Für den Test wurden von den Sonderdrucken Digitalisate mit 300 dpi in Farbe angefertigt. Dazu wurden die Drucke aufgeschnitten. Die bereits vorhandenen, ca. 20 Jahre alten Digitalisate (überwiegend Bitonal, 300 dpi) wurden in diesem Test nicht berücksichtigt.

Leitfragen der Teststellung

Der Test sollte Antworten auf folgende Fragen geben:

1. Ist mit der OCR-D-Software eine Volltextdigitalisierung möglich?
2. Wie hoch ist die Texterkennungs-Qualität und Strukturerkennungs-Qualität?
3. Können die erkannten Vorlagen auch als TEI-Dokumente angeboten werden?

Antworten

Frage 1

Zunächst wurde die Software auf einem Testserver installiert. Die dafür notwendigen Schritte sind umfangreich dokumentiert. Die Installation des Software-Pakets `[ocrd_all](https://github.com/OCR-D/ocrd_all)` erfolgte ohne Fehler. Da die Software im Testzeitraum weiterentwickelt wurde, war es notwendig, diese in Abständen zu aktualisieren. Dafür wurde die eingerichtete Hotline genutzt, die mit kompetenter Hilfe zur Seite stand.

Das Korpus wurde mit folgenden **Workflow-Schritten** bearbeitet.

```
ocrd process \
  "olena-binarize -I OCR-D-IMG -O OCR-D-IMG-BIN-XML,OCR-D-IMG-BIN -p '{\"impl\": \"sauvola-ms-split\"}'" \
  "anybaseocr-crop -I OCR-D-IMG-BIN-XML -O OCR-D-IMG-BIN-CROP-XML" \
  "tesserocr-segment-region -I OCR-D-IMG-BIN-CROP-XML -O OCR-D-IMG-BIN-CROP-REGION-XML" \
  "tesserocr-segment-line -I OCR-D-IMG-BIN-CROP-REGION-XML -O OCR-D-IMG-BIN-CROP-LINE-XML" \
  "ocrd-tesserocr-recognize -I OCR-D-IMG-BIN-CROP-LINE-XML -O OCR-D-OCR-TESSEROCR -p '{\"model\": \"deu+Latin\"}'"
```

Die OCR-D-Software bietet eine Vielzahl von Prozessoren an. So gibt es bspw. für die Binarisierung mehrere Prozessoren mit unterschiedlichen Parametern zur Verfügung. Für die Durchführung des Tests war ein Grundverständnis der einzelnen Arbeitschritte des OCR-Prozesses notwendig, die anhand der schematischen Darstellung und der Beschreibung des OCR-D-Workflows nachvollzogen werden kann. Die im Setup-Guide empfohlene Workflow-Software Taverna wurde nicht installiert.

Fazit: Ja, alle Vorlagen des Testkorpus konnten mit der OCR-D-Software bearbeitet und ein Volltext erstellt werden.

Durch die einheitliche Syntax der Eingabe und Ausgabe sind die Prozessoren, d. h. die Software-Module, die die einzelnen Aufgaben im OCR-Workflow bearbeiten, einfach anzusprechen. Was in den einzelnen Workflow-Schritten geleistet wird, war zum Testzeitpunkt noch nicht dokumentiert. Da die Software im Testmodus und nicht produktiv eingesetzt wurde, kann keine realistische und repräsentative Angabe zur Laufzeit gemacht werden. Der Test schloss eine dokumentierte Evaluation des Workflows sowie die fortlaufende Anpassung der Parameter und Prozessoren ein.

Frage 2

Die Erkennungs-Qualität wurde auf den Ebenen der Texterkennung und der Strukturerkennung überprüft. Die Qualität der Texterkennung wurde anhand einer der getesteten 13 Akademieschriften überprüft, von der zu diesem Zweck eine korrekte Version (Ground Truth) erstellt wurde. Die Prüfung ist auf diese Vorlage begrenzt und erhebt daher keinen Anspruch auf Repräsentativität. Mit dem Test sollten

nur Indizien für die zu erwartende Qualität gewonnen werden. Auf dieser Grundlage sollte zudem eingeschätzt werden, mit welchen Maßnahmen die Qualität noch verbessert werden kann.

Für die Prüfung wurde der dingelhopper-processor genutzt, der ebenfalls in `ocrd_all` verfügbar ist.

```
ocrd-dinglehopper -m mets.xml -I OCR-D-GT-PAGE,OCR-D-OCR-TESS -O OCR-D-OCR-TESS-EVAL
```

Das Testergebnis kann eingesehen werden unter:

- Report Gesamt
- Report Image 5

Viele Fehler sind auf die falsche Erkennung des langen s ("ſ", U+017F) zurückzuführen. Zwar ist der Text in Antiqua gesetzt, dennoch wird das lange s verwendet. Dies ist kein Einzelfall in den Akademieschriften. So ist in wesentlich älteren und fremdsprachigen Publikationen ebenfalls diese Art der Schrifttypographie zu beobachten. Für die Erkennung wurde das Standard-Erkennungs-Modell Deutsch und Latein verwendet. Da in diesem Modell kein langes s vorkommt, konnte die OCR dieses auch nicht korrekt erkennen.

Lösungen: Zur Verbesserungen der Erkennung kann folgendes in der Konfiguration verändert werden:

1. Erweiterung der verwendeten Erkennungsmodelle (Antiqua-, Frakturmodell) `ocrd-tesserocr-recognize -I OCR-D-IMG-BIN-CROP-LINE-XML -O OCR-D-OCR-TESSEROCR -p <(echo '{"model": "frk+deu+Fraktur+Latin"}') -m mets.xml`
2. Training eines speziellen Modells auf Grundlage der vorhandenen Modelle Auf Basis des erstellten Ground-Truth wurde das vorhandene tesseract Deutsch-Modell (`deu`) durch Training erweitert. Damit konnte der Fehler des langen s beseitigt werden. Die bereits hohe Erkennungsqualität konnte dadurch weiter verbessert werden. So betrug die Zeichenfehlerquote auf dem Image 5 0.0059 mit der Beseitigung der Fehler des langen s konnte diese um die Hälfte auf 0.00236 gesenkt werden. Die Wortfehlerquote auf Image 5 betrug zuvor 0,0032 und konnte auf 0,00142 ebenfalls gesenkt werden.

Die Prüfung der Strukturerkennung erfolgte manuell. Für die Segmentierung wurde der `tesserocr-segment-region`-Prozessor verwendet. Dabei wurde festgestellt, dass die Absätze des Textes nicht vollständig erkannt wurden und stattdessen eine Untersegmentierung vorliegt. Der Test mit anderen Segmentierungs-Prozessoren ergab, dass diese in einigen Bereichen übersegmentieren. Nach Abwägung der Vor- und Nachteile der verschiedenen Segmentierungen wurde eine Untersegmentierung vorgezogen. Für diese Entscheidung sprach, dass dieser Fehler leichter, schneller und ohne größeren Aufwand zu korrigieren ist. Die vorgesehene Korrektur sollte im digitalen Volltext, Format TEI, vorgenommen werden.

Frage 3

Ziel der Volltextdigitalisierung der Akademieschriften ist es, gemäß TEI- bzw. DTABf-Richtlinien kodierte, strukturierte Forschungsdaten zu schaffen. Die erzeugten Texte sollen als Grundlage u.a. für Editionen, für wissenschaftlich nutzbare Sprachkorpora und für Untersuchungen im Umfeld der Digital Humanities dienen können. Dazu war zu prüfen, ob OCR-D das TEI-Format als Ausgabe anbietet.

Das primäre Ausgabeformat der OCR-D-Software ist PAGE XML. Dieses Format kann mit dem PAGE-Viewer betrachtet und mit Werkzeugen wie Aletheia bearbeitet werden. Da das Abspeichern in TEI/DTABf direkt aus der OCR-D-Software heraus nicht möglich ist, erfolgte die Konvertierung/Transformation in das TEI/DTABf erst nachdem die OCR-D-Software den Volltext erzeugt hatte. OCR-D bietet zu diesem Zweck Konvertierungsprogramme u.a. von PAGE-XML nach TEI (page2tei) an.

So wurde zuerst der XSLT-Transformation Prozessor saxon (HE-Version = Home-Edition-Version) auf dem Rechner installiert. Das benötigte Transformationsprogramm ist unter github.com/tboenig/page2tei zu finden. Mit dem folgendem Kommando wurde die Transformation durchgeführt:

```
java -jar saxon-he-10.0.jar -xsl:page2tei-0.xsl -s:mets.xml PAGEXML=OCR-D-OCR-TESSEROCR PAGEprogram=OCR-D -o:

[gewünschter Name der TEI-Datei].xml
```

Die erstellte TEI-Datei enthält Informationen zum digitalen Faksimile (Digitalisat; siehe dazu TEI-Dokumentation: 11.1 Digital Facsimiles) und die Textinformationen (siehe ebd., 11.2.2 Embedded Transcription). Vorhandene Metadateninformationen, die sich entweder in der METS-Datei oder im Katalogisat stammen könnten, werden in TEI-Header nicht übernommen. Aus diesem Grund besteht der TEI-Header aus einem Skelett, das zu ergänzen ist.

Beispiel für ein Ergebnis der TEI-Konvertierung

```
<?xml version="1.0" encoding="UTF-8"?>

<TEI xmlns="http://www.tei-c.org/ns/1.0">

    <teiHeader>

        <fileDesc>

            <titleStmt/>

            <publicationStmt>

                <publisher>OCR-D</publisher>

            </publicationStmt>

            <seriesStmt/>

            <sourceDesc/>

        </fileDesc>

    </teiHeader>

    <facsimile>

        <surface ulx="0" uly="0" lrx="5100" lry="7016" xml:id="facs_">
```

```
            <graphic url="OCR-D-OCR-TESSEROCR_bbaw_0001.tif" width="5100px" height="7016px"/>
            <zone points="1577,874 3099,874 3099,1029 1577,1029" rendition="TextRegion" xml:id="facs__region0000">
              <zone points="1577,874 3099,874 3099,1029 1577,1029" rendition="Line" xml:id="facs__region0000_line0000"/
>
            </zone>
        </surface>
    </facsimile>

    <text>
        <body>
            <div>
                <pb facs="#facs_00" n="" xml:id="img_00"/>
                <p facs="#facs__region0000">
                    <lb facs="#facs__region0000_line0000"/>GEDÄCHTNISREDE</p>
            </div>
        </body>
    </text>
</TEI>
```

Fazit der Transformation Die Transformation in TEI/DTABf erfolgt problemlos. Jedoch die Überführung von METS/MODS-Metadaten in den TEI-Header wird mit diesem Konvertierungsprogramm nicht realisiert. Aus diesem Grund wurden Metadaten manuell für das Textkorpus auf Grundlage der Katalogisate erfasst. Eine automatische Übernahme dieser Metadaten ist aus zeitökonomischen und personellen Gründen zu empfehlen. In einem nächsten Testlauf wird dies versucht praktisch umzusetzen und weiterhin evaluiert.

Zusammenfassung

Der Testlauf zeigt, dass die Volltexttransformation der Akademieschriften mit der OCR-D-Software geleistet werden kann. Im Rahmen dieser Transformation sind jedoch Phasen zur Optimierung des Workflows, zur Reprozessierung und zum OCR-Training mit einzuplanen.

Für das OCR-Training ist ausreichender Ground Truth zu erzeugen. Dieser kann paralell zur Volltextdigitalisierung erstellt werden. Die OCR sollte an den *einfachen* Vorlagen begonnen werden. Gerade im Hinblick auf die OCR der mathematischen und chemischen Texte ist ein spezialisiertes Training wahrscheinlich unumgänglich, während die diskursiven, vorranging in Fließtext verfassten Schriften bereits mit wenigen Anpassungen in ausgezeichneter Qualtität und mit standardkonformer TEI/DTABf-Strukturierung erfasst werden können.

Die Ergebnisse sind innerhalb der Qualitätssicherungsplattform des *Deutschen Textarchivs*, DTAQ, einzusehen:

Diels, Hermann: Gedächtnisrede auf Eduard Zeller. Berlin, 1908.

Dümmler, Ernst: Gedächtnissrede auf Paul Scheffer-Boichorst. Berlin, 1902.

Fischer, Emil: Gedächtnisrede auf Jacobus Henricus van't Hoff. Berlin, 1911.

Hirschfeld, Otto: Gedächtnisrede auf Theodor Mommsen. Berlin, 1904.

Köhler, Ulrich: Gedächtnissrede auf Ernst Curtius. Berlin, 1897.

Pischel, Richard: Gedächtnisrede auf Albrecht Weber. Berlin, 1903.

Rubens, Heinrich: Gedächtnisrede auf Friedrich Kohlrausch. Berlin, 1910.

Schmidt, Erich: Gedächtnissrede auf Karl Weinhold. Berlin, 1902.

Schulze, Wilhelm: Gedächtnisrede auf Heinrich Zimmer. Berlin, 1911.

Sybel, Heinrich von: Gedächtnissrede auf Leopold von Ranke. Berlin, 1886.

Van't Hoff, Jakobus Heinrich: Gedächtnisrede auf Hans Heinrich Landolt. Berlin, 1911.

Wattenbach, Wilhelm: Gedächtnissrede auf Georg Waitz. Berlin, 1886.

Wilamowitz-Moellendorff, Ulrich von: Gedächtnisrede auf Adolf Kirchhoff. Berlin, 1908.