EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

OCR-BW – Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen

Gefördert durch
Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

16.10.2020

OCR-BW

Projektziele

„Aufbau eines Kompetenzzentrums für Volltexterschließung von handschriftlichen und gedruckten Werken.“

Das Projekt OCR-BW unterstützt Archive, wissenschaftliche Bibliotheken und andere Institutionen in Baden-Württemberg bei der Anwendung von automatischer Texterkennungs- und Transkriptionssoftware.

UB Tübingen:	Transkription und Volltexterschließung von Autographen, Handschriften und Inkunabeln
UB Mannheim:	Volltexterkennung (OCR) von Druckwerken

➢ **https://ocr-bw.bib.uni-mannheim.de/**

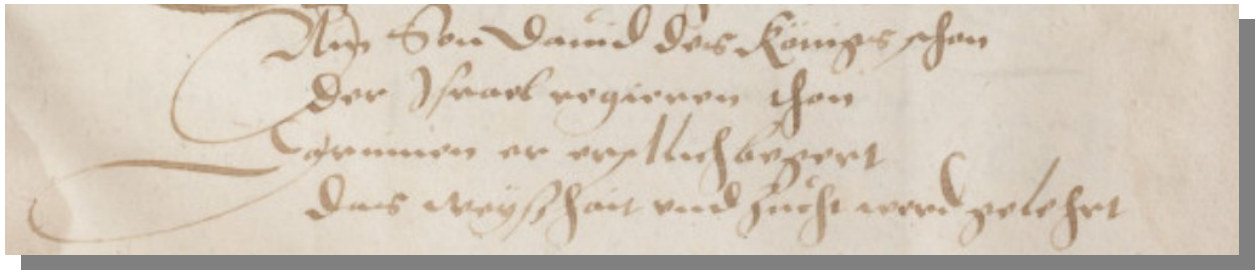
OCR-BW

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

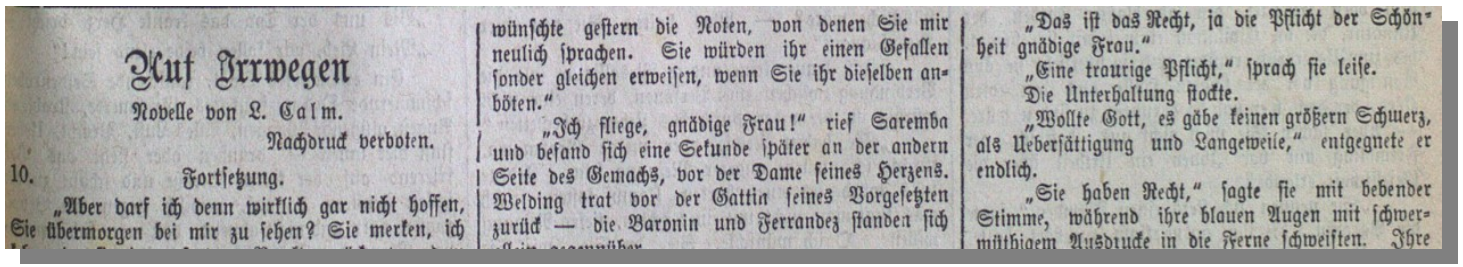
UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

Arbeitsbeispiele für Transkription und Texterkennung

- Handschrift aus dem Bestand der WLB Stuttgart (16. Jhd.)



- Historische Zeitungen (19.–20. Jhd.) für das Stadtarchiv Ladenburg, das MARCHIVUM und die BLB Karlsruhe



Auf Irrwegen

Novelle von L. Calm.

Nachdruck verboten.

10. Fortsetzung.

„Aber darf ich denn wirklich gar nicht hoffen, Sie übermorgen bei mir zu sehen? Sie merken, ich wünschte gestern die Noten, von denen Sie mir neulich sprachen. Sie würden ihr einen Gefallen sonder gleichen erweisen, wenn Sie ihr dieselben anböten."

„Ich fliege, gnädige Frau!" rief Saremba und befand sich eine Sekunde später an der andern Seite des Gemachs, vor der Dame seines Herzens. Welding trat vor der Gattin seines Vorgesetzten zurück — die Baronin und Ferrandez standen sich

„Das ist das Recht, ja die Pflicht der Schönheit gnädige Frau."

„Eine traurige Pflicht," sprach sie leise.

Die Unterhaltung stockte.

„Wollte Gott, es gäbe keinen größern Schmerz, als Uebersättigung und Langeweile," entgegnete er endlich.

„Sie haben Recht," sagte sie mit bebender Stimme, während ihre blauen Augen mit schwermüthigem Ausdrucke in die Ferne schweiften. Ihre

OCR-BW

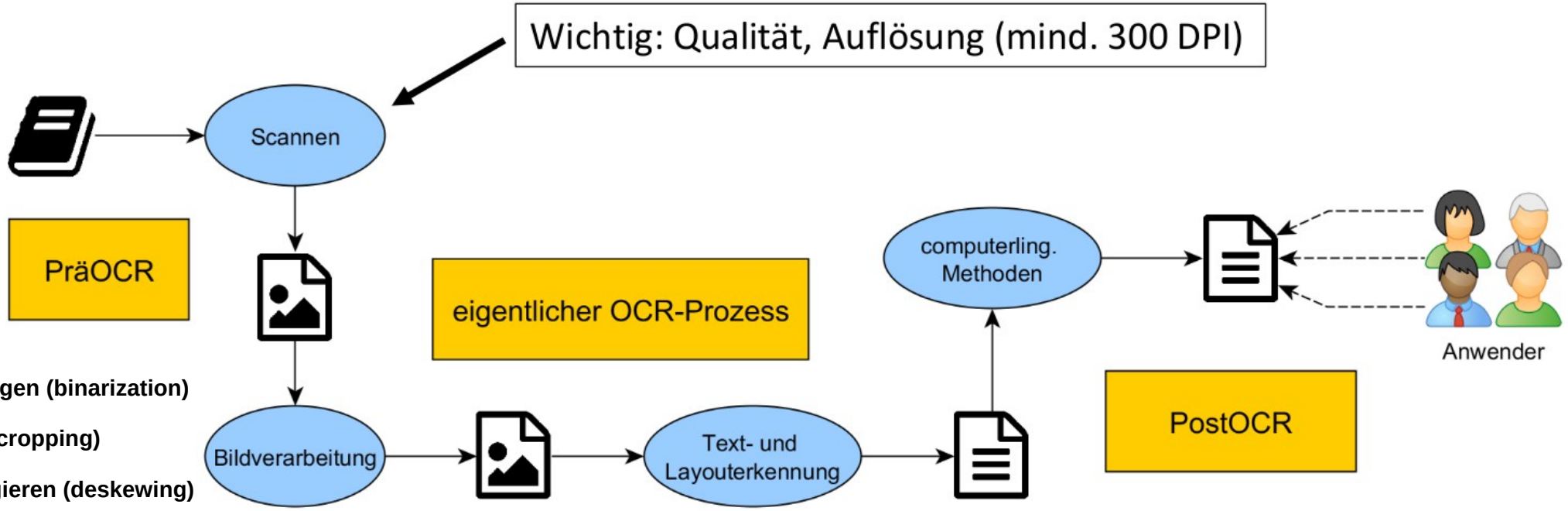
Texterkennung von historischen Drucken
mit OCR-D und Tesseract

Stefan Weil, Jan Kamlah
Universitätsbibliothek Mannheim

16.10.2020

Gefördert durch Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

Vom Werk zum Digitalisat mit Volltext



S/W-Bild erzeugen (binarization)

Beschneiden (cropping)

Neigung korrigieren (deskewing)

Entzerren (dewarping)

Flecken entfernen (despeckling)

aus: Baierer, Zumstein. Verbesserung der OCR in digitalen Sammlungen von Bibliotheken

Tesseract in Kürze https://tesseract-ocr.github.io/

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Open Source
- Komplettlösung „All-in-1“
- Kommandozeilenprogramm (CLI)
- Dient als Kernkomponente in vielen OCR-Applikationen
- Mehr als 100 Sprachen / mehr als 30 Schriften
- Liest Bilder in allen gängigen Formaten (nicht PDF!)
- Erzeugt Text, PDF, hOCR, ALTO, TSV
- Große, weltweite Anwender-Community
- Technologisch aktuell (Texterkennung mit neuronalem Netz)

OCR-D in Kürze https://ocr-d.github.io/de/

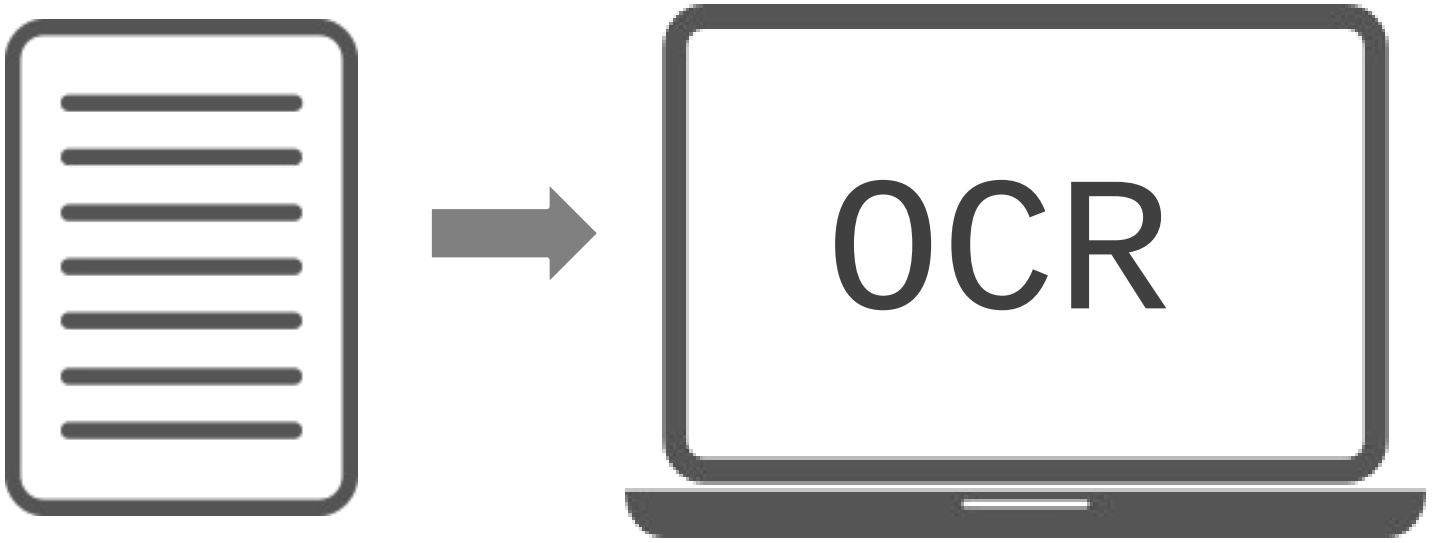
UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Open Source
- Modularer Ansatz – „Prozessoren“ für alle Arbeitsschritte
- Kommandozeilenprogramm (CLI)
- METS als Metadatenformat
- PAGE XML für Seiteninhalte, optional auch Text, PDF, hOCR, ALTO XML, TSV, …
- Definition von Ground-Truth-Erfassungsregeln und Bereitstellung von GT-Korpora
- Langzeitarchivierungsmöglichkeiten
- Fokus auf Drucke des 16.–18. Jahrhunderts (VD16, VD17, VD18)

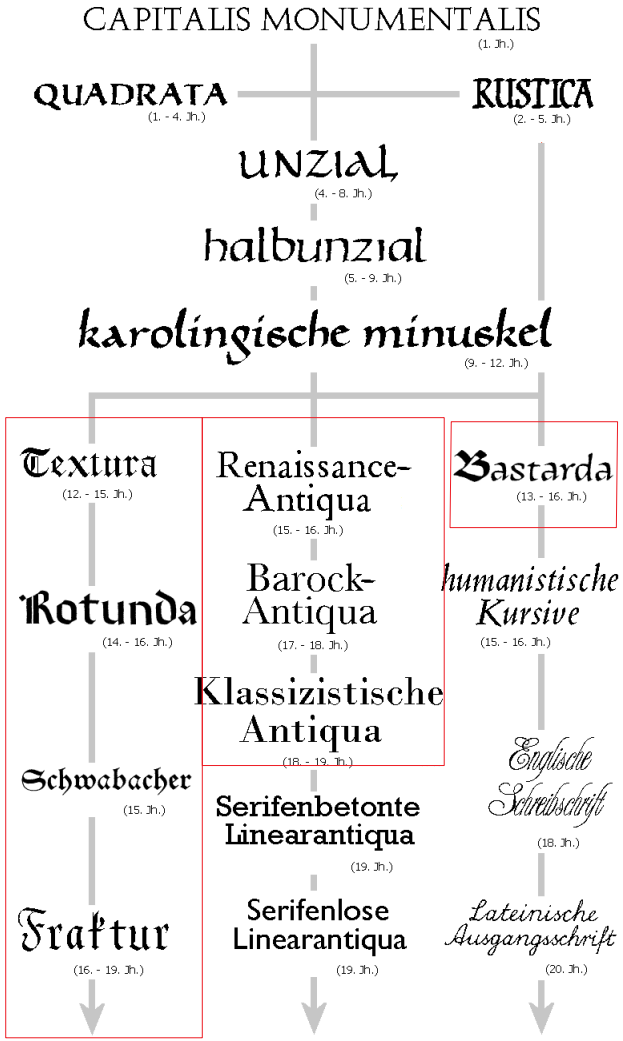
OCR-D
Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR)

Gefördert durch DFG Deutsche Forschungsgemeinschaft

OCR im Einsatz



One model to rule them all?



https://typo-info.de/entwicklung-der-schrift/

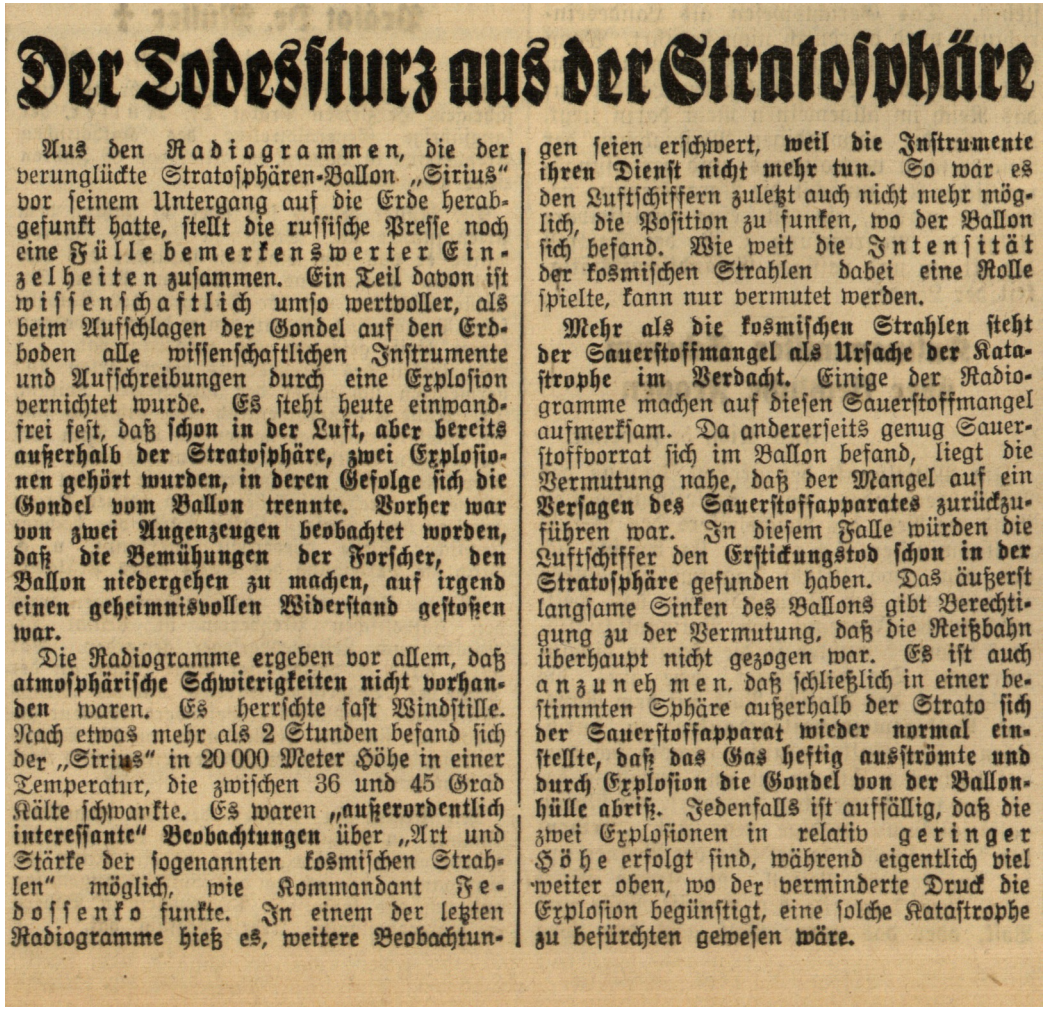
Neue generische Modelle
sind sprachenunabhängig und
decken auch viele Schriftarten ab,
ebenso wie **fett** und *kursiv* gedruckte Variationen.
Erleichtert Volltexterkennung für typische Druckwerke.

Erstes Modell für historische Schriftarten:

GT4HistOCR

https://ub-backup.bib.uni-mannheim.de/~stweil/ocrd-train/data/Fraktur_5000000/tessdata_fast/

Beispiel 1 – BLB Karlsruhe

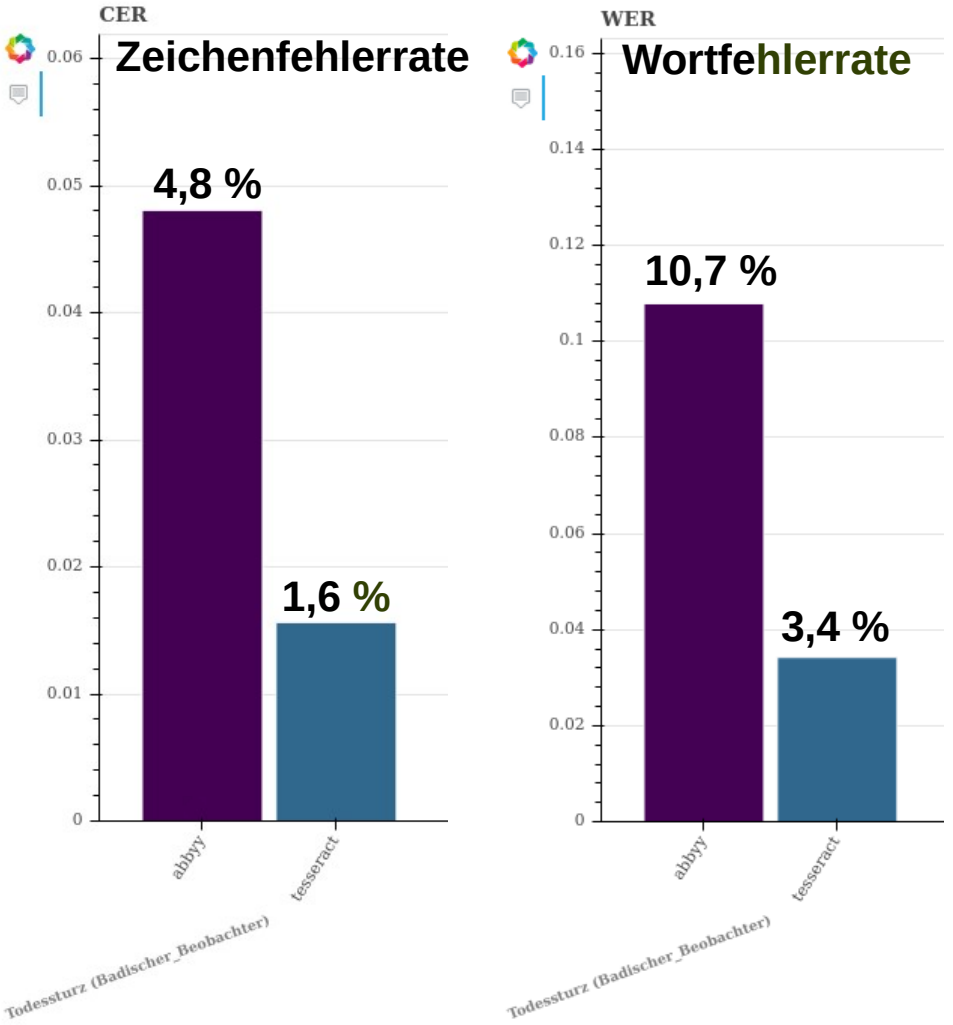


Der Todessturz aus der Stratosphäre

Aus den Radiogrammen, die der verunglückte Stratosphären-Ballon „Sirius" vor seinem Untergang auf die Erde herabgefunkt hatte, stellt die russische Presse noch eine Fülle bemerkenswerter Einzelheiten zusammen. Ein Teil davon ist wissenschaftlich umso wertvoller, als beim Aufschlagen der Gondel auf den Erdboden alle wissenschaftlichen Instrumente und Aufschreibungen durch eine Explosion vernichtet wurde. Es steht heute einwandfrei fest, daß **schon in der Luft, aber bereits außerhalb der Stratosphäre, zwei Explosionen gehört wurden, in deren Gefolge sich die Gondel vom Ballon trennte. Vorher war von zwei Augenzeugen beobachtet worden, daß die Bemühungen der Forscher, den Ballon niedergehen zu machen, auf irgend einen geheimnisvollen Widerstand gestoßen war.**

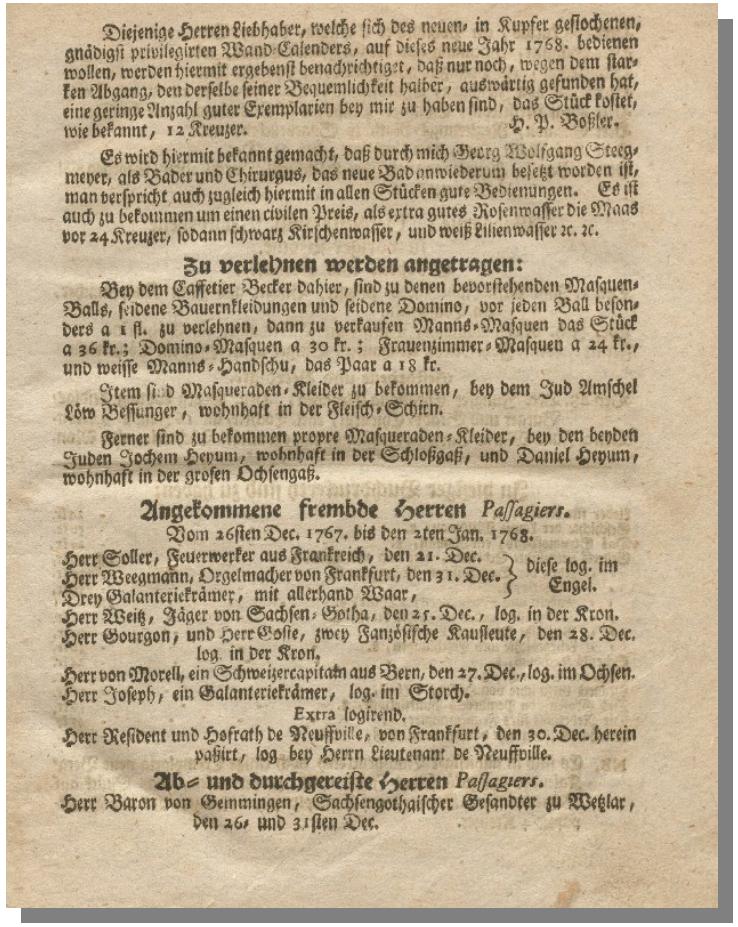
Die Radiogramme ergeben vor allem, daß **atmosphärische Schwierigkeiten nicht vorhanden** waren. Es herrschte fast Windstille. Nach etwas mehr als 2 Stunden befand sich der „Sirius" in 20 000 Meter Höhe in einer Temperatur, die zwischen 36 und 45 Grad Kälte schwankte. Es waren **„außerordentlich interessante" Beobachtungen** über „Art und Stärke der sogenannten kosmischen Strahlen" möglich, wie Kommandant Fedossenko funkte. In einem der letzten Radiogramme hieß es, weitere Beobachtungen seien erschwert, **weil die Instrumente ihren Dienst nicht mehr tun.** So war es den Luftschiffern zuletzt auch nicht mehr möglich, die Position zu funken, wo der Ballon sich befand. Wie weit die Intensität der kosmischen Strahlen dabei eine Rolle spielte, kann nur vermutet werden.

Mehr als die kosmischen Strahlen steht der Sauerstoffmangel als Ursache der Katastrophe im Verdacht. Einige der Radiogramme machen auf diesen Sauerstoffmangel aufmerksam. Da andererseits genug Sauerstoffvorrat sich im Ballon befand, liegt die Vermutung nahe, daß der Mangel auf ein **Versagen des Sauerstoffapparates** zurückzuführen war. In diesem Falle würden die Luftschiffer den **Erstickungstod schon in der Stratosphäre** gefunden haben. Das äußerst langsame Sinken des Ballons gibt Berechtigung zu der Vermutung, daß die Reißbahn überhaupt nicht gezogen war. Es ist auch anzunehmen, daß schließlich in einer bestimmten Sphäre außerhalb der Strato sich **der Sauerstoffapparat wieder normal einstellte, daß das Gas heftig ausströmte und durch Explosion die Gondel von der Ballonhülle abriß.** Jedenfalls ist auffällig, daß die zwei Explosionen in relativ geringer Höhe erfolgt sind, während eigentlich viel weiter oben, wo der verminderte Druck die Explosion begünstigt, eine solche Katastrophe **zu befürchten gewesen wäre.**



Beispiel 2 – ULB Darmstadt (Preprocessing)

CER: 6,84 %



Diejenige Herren Liebhaber, welche sich des neuen in Kupfer gestochenen, gnädigst privilegirten Wand-Calenders, auf dieses neue Jahr 1768. bedienen wollen, werden hiermit ergebenst benachrichtiget, daß nur noch, wegen dem starken Abgang, den derselbe seiner Bequemlichkeit halber, auswärtig gefunden hat, eine geringe Anzahl guter Exemplarien bey mir zu haben sind, das Stück kostet, wie bekannt, 12 Kreuzer. H. P. Boßler.

Es wird hiermit bekannt gemacht, daß durch mich Georg Wolfgang Steegmeyer, als Bader und Chirurgus, das neue Bad erwiederum besetzt worden ist, man verspricht auch zugleich hiermit in allen Stücken gute Bedienungen. Es ist auch zu bekommen um einen civilen Preis, als extra gutes Rosenwasser die Maas vor 24 Kreuzer, sodann schwarz Kirschenwasser, und weiß Lilienwasser ꝛc. ꝛc.

Zu verlehnen werden angetragen:

Bey dem Caffetier Becker dahier, sind zu denen bevorstehenden Masquen-Balls, seidene Bauernkleidungen und seidene Domino, vor jeden Ball besonders a 1 fl. zu verlehnen, dann zu verkaufen Manns-Masquen das Stück a 36 kr.; Domino-Masquen a 30 kr.; Frauenzimmer-Masquen a 24 kr., und weisse Manns-Handschu, das Paar a 18 kr.

Item sind Masqueraden-Kleider zu bekommen, bey dem Jud Amschel Löw Bessunger, wohnhaft in der Fleisch-Schirn.

Ferner sind zu bekommen propre Masqueraden-Kleider, bey den beyden Juden Jochem Heyum, wohnhaft in der Schloßgaß, und Daniel Heyum, wohnhaft in der grosen Ochsengaß.

Angekommene frembde Herren *Passagiers.*

Vom 26sten Dec. 1767. bis den 2ten Jan. 1768.

Herr Soller, Feuerwerker aus Frankreich, den 21. Dec.
Herr Weegmann, Orgelmacher von Frankfurt, den 31. Dec. } diese log. im Engel.
Drey Galanteriekrämer, mit allerhand Waar,
Herr Weitz, Jäger von Sachsen-Gotha, den 25. Dec., log. in der Kron.
Herr Gourgon, und Herr Goste, zwey Janzösische Kaufleute, den 28. Dec. log. in der Kron.
Herr von Morell, ein Schweizercapitain aus Bern, den 27. Dec., log. im Ochsen.
Herr Joseph, ein Galanteriekrämer, log. im Storch.

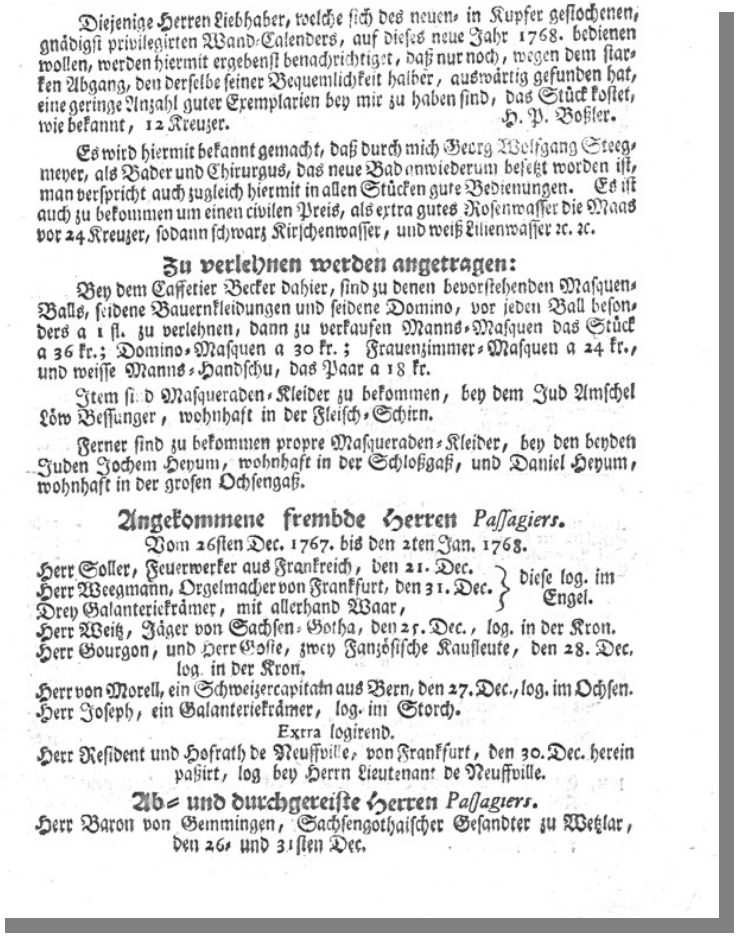
Extra logirend.

Herr Resident und Hofrath de Neuffville, von Frankfurt, den 30. Dec. herein paßirt, log. bey Herrn Lieutenant de Neuffville.

Ab- und durchgereiste Herren *Passagiers.*

Herr Baron von Gemmingen, Sachsengothaischer Gesandter zu Wetzlar, den 26. und 31sten Dec.

CER: 3,76 %



Diejenige Herren Liebhaber, welche sich des neuen in Kupfer gestochenen, gnädigst privilegirten Wand-Calenders, auf dieses neue Jahr 1768. bedienen wollen, werden hiermit ergebenst benachrichtiget, daß nur noch, wegen dem starken Abgang, den derselbe seiner Bequemlichkeit halber, auswärtig gefunden hat, eine geringe Anzahl guter Exemplarien bey mir zu haben sind, das Stück kostet, wie bekannt, 12 Kreuzer. H. P. Boßler.

Es wird hiermit bekannt gemacht, daß durch mich Georg Wolfgang Steegmeyer, als Bader und Chirurgus, das neue Bad erwiederum besetzt worden ist, man verspricht auch zugleich hiermit in allen Stücken gute Bedienungen. Es ist auch zu bekommen um einen civilen Preis, als extra gutes Rosenwasser die Maas vor 24 Kreuzer, sodann schwarz Kirschenwasser, und weiß Lilienwasser ꝛc. ꝛc.

Zu verlehnen werden angetragen:

Bey dem Caffetier Becker dahier, sind zu denen bevorstehenden Masquen-Balls, seidene Bauernkleidungen und seidene Domino, vor jeden Ball besonders a 1 fl. zu verlehnen, dann zu verkaufen Manns-Masquen das Stück a 36 kr.; Domino-Masquen a 30 kr.; Frauenzimmer-Masquen a 24 kr., und weisse Manns-Handschu, das Paar a 18 kr.

Item sind Masqueraden-Kleider zu bekommen, bey dem Jud Amschel Löw Bessunger, wohnhaft in der Fleisch-Schirn.

Ferner sind zu bekommen propre Masqueraden-Kleider, bey den beyden Juden Jochem Heyum, wohnhaft in der Schloßgaß, und Daniel Heyum, wohnhaft in der grosen Ochsengaß.

Angekommene frembde Herren *Passagiers.*

Vom 26sten Dec. 1767. bis den 2ten Jan. 1768.

Herr Soller, Feuerwerker aus Frankreich, den 21. Dec.
Herr Weegmann, Orgelmacher von Frankfurt, den 31. Dec. } diese log. im Engel.
Drey Galanteriekrämer, mit allerhand Waar,
Herr Weitz, Jäger von Sachsen-Gotha, den 25. Dec., log. in der Kron.
Herr Gourgon, und Herr Goste, zwey Janzösische Kaufleute, den 28. Dec. log. in der Kron.
Herr von Morell, ein Schweizercapitain aus Bern, den 27. Dec., log. im Ochsen.
Herr Joseph, ein Galanteriekrämer, log. im Storch.

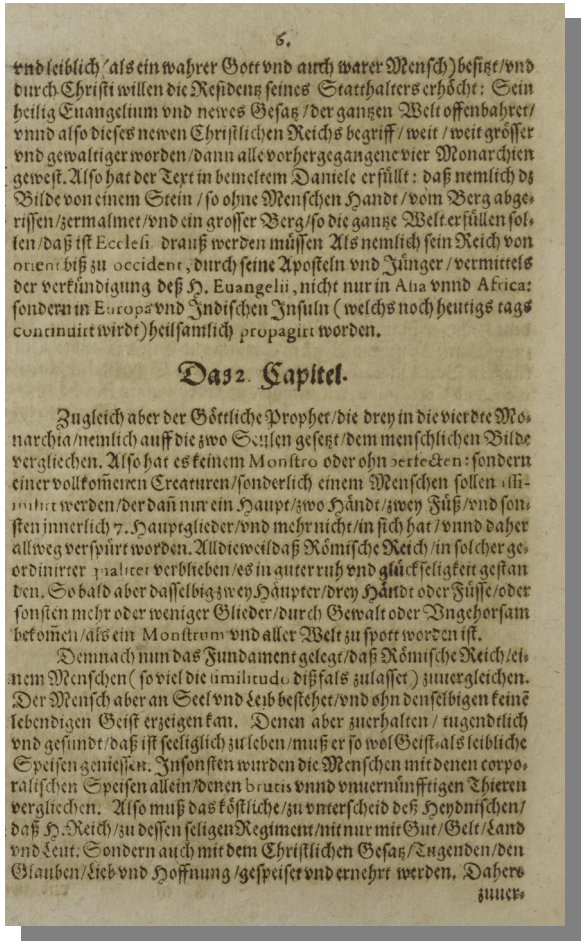
Extra logirend.

Herr Resident und Hofrath de Neuffville, von Frankfurt, den 30. Dec. herein paßirt, log. bey Herrn Lieutenant de Neuffville.

Ab- und durchgereiste Herren *Passagiers.*

Herr Baron von Gemmingen, Sachsengothaischer Gesandter zu Wetzlar, den 26. und 31sten Dec.

Anwendungs-**Beispiel**



6.

vnd leiblich / als ein wahrer Gott vnd auch warer Mensch) besitzt / vnd durch Christi willen die Residentz seines Statthalters erhöcht: Sein heilig Euangelium vnd newes Gesatz / der gantzen Welt offenbahret / vnnd also dieses newen Christlichen Reichs begriff / weit / weit grösser vnd gewaltiger worden / dann alle vorhergegangene vier Monarchien gewest. Also hat der Text in bemeltem Daniele erfüllt: daß nemlich dz Bilde von einem Stein / so ohne Menschen Handt / vom Berg abgerissen / zermalmet / vnd ein grosser Berg / so die gantze Welt erfüllen solten / daß ist Ecclesia drauß werden müssen. Als nemlich sein Reich von oriente biß zu occidente, durch seine Apostoln vnd Jünger / vermittels der verkündigung deß H. Euangelii, nicht nur in Asia vnnd Africa: sondern in Europa vnd Jndischen Jnsuln (welchs noch heutigs tags continuirt wirdt) heilsamlich propagirt worden.

Das 2. Capitel.

Zugleich aber der Göttliche Prophet / die drey in die vier dte Monarchia / nemlich auff die zwo Seulen gesetzt / dem menschlichen Bilde vergliechen. Also hat es keinem Monstro oder ohn gerechten: sondern einer vollkomeneren Creaturen / sonderlich einem Menschen sollen assimilirt werden / der dañ nur ein Haupt / zwo Händt / zwey Füß / vnd sonsten innerlich 7. Hauptglieder / vnd mehr nicht / in sich hat / vnnd daher allweg verspürt worden. Alldieweil daß Römische Reich / in solcher geordinirter qualitet verblieben / es in guter ruh vnd glückseligkeit gestanden. So bald aber dasselbig zwey Häupter / drey Händt oder Füsse / oder sonsten mehr oder weniger Glieder / durch Gewalt oder Vngehorsam bekomen / als ein Monstrum vnd aller Welt zu spott worden ist.

Demnach nun das Fundament gelegt / daß Römische Reich / einem Menschen (so viel die similitudo dißfals zulasset) zuuergleichen. Der Mensch aber an Seel vnd Leib bestehet / vnd ohn denselbigen keine lebendigen Geist erzeigen kan. Denen aber zuerhalten / tugendlich vnd gesundt / daß ist seeliglich zu leben / muß er so wol Geist- als leibliche Speisen genieſſen. Jnsonsten wurden die Menschen mit denen corporalischen Speisen allein / denen brutis vnnd vnuernünfftigen Thieren vergliechen. Also muß das köstliche / zu vnterscheid deß Heydnischen / daß H. Reich / zu dessen seligen Regiment / nit nur mit Gut / Gelt / Land vnd Leut: Sondern auch mit dem Christlichen Gesatz / Tugenden / den Glauben / Lieb vnd Hoffnung / gespeiset vnd ernehrt werden. Dahers

zuuer-

Tesseract Beispiel

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

Installation: **https://tesseract-ocr.github.io/tessdoc/Home.html**

Workflow:

```
# Aufruf-Schema
tesseract -l model input output

# Ausgabe aller zur Verfügung stehender Modelle
# Speicherort der Modelle: Tessdata (Default: /usr/share/tesseract/tessdata/)
tesseract --list-langs

# Einsatz von GT4HistOCR-Modell mit PDF-Ausgabe
# Andere Ausgabemöglichkeiten alto, hocr, tsv, txt (Default: txt)
# Um die Ausgabe auf den Terminal auszugeben: output → stdout
tesseract -l GT4HistOCR input output pdf

# Ausgabe der erweiterten Einstellungsmöglichkeiten
tesseract --help-extra

# Erweiterung des Aufrufs um den PageSegmentationMode (PSM)
# Das Bild ist einspaltig (PSM → 4)
tesseract --psm 4 -l GT4HistOCR input output

# Verarbeitung aller jpg-Dateien in einem Ordner
# Zur Parallelisierung empfohlen: GNUparallel (dabei darauf achten, dass OpenMP nicht aktiv ist)
for i in *.jpg; do echo ${i/.jpg/}; tesseract -l GT4HistOCR $i ${i/.jpg/};done

# Alternative mit „fd-find“
fd -e jpg -x tesseract -l GT4HistOCR {} {.}
```

OCR-D Beispiel

Installation: **https://github.com/OCR-D/ocrd_all**

Workflow für eine Einzelseite ohne bestehende METS-Datei:

```
# Step 0: Create Workspace & METS file
mkdir -p ~/projects/OCR-D/workshop/2020_02_19/
cd ~/projects/OCR-D/workshop/2020_02_19/
# Create workspace including METS file
ocrd workspace init OCRbw_workspace && cd OCRbw_workspace

# Step 1: Download jpg image
mkdir ./OCR-D-IMG && \
wget -O ./OCR-D-IMG/Fraktur_1621.jpg \
https://digi.bib.uni-mannheim.de/~jkamlah/OCRbw-Workshop-2020-02-19/example/ocrd/MAX/Fraktur_1621.jpg

# Step 2: Add image to METS
# Be aware, that the ID and the GROUPID must be identical if the referenced image represents the original image
ocrd workspace add --file-grp OCR-D-IMG --file-id Fraktur_1621 --page-id OCR-D-IMG_0001 \
  --mimetype image/jpeg OCR-D-IMG/Fraktur_1621.jpg

# Install OCR model into Tesseract data path
apt-get install tesseract-ocr-script-frak

# Step 3: Run the workflow (processors)
ocrd process \
  'tesserocr-segment-region -I OCR-D-IMG -O OCR-D-SEG-BLOCK' \
  'tesserocr-segment-line -I OCR-D-SEG-BLOCK -O OCR-D-SEG-LINE' \
  'tesserocr-recognize -I OCR-D-SEG-LINE -O OCR-D-OCR-TESSEROCR -P model frk -P textequiv_level word'
```

OCR-D Beispiel

Installation: **https://github.com/OCR-D/ocrd_all**

Workflow für einen Datensatz mit METS:

```
# Create workspace including METS file
ocrd workspace clone -a METS-URL

# Run the recommended workflow (processors)
ocrd process \
  "cis-ocropy-binarize -I OCR-D-IMG -O OCR-D-BIN" \
  "anybaseocr-crop -I OCR-D-BIN -O OCR-D-CROP" \
  "skimage-binarize -I OCR-D-CROP -O OCR-D-BIN2 -P method li" \
  "skimage-denoise -I OCR-D-BIN2 -O OCR-D-BIN-DENOISE -P level-of-operation page" \
  "tesserocr-deskew -I OCR-D-BIN-DENOISE -O OCR-D-BIN-DENOISE-DESKEW -P operation_level page" \
  "cis-ocropy-segment -I OCR-D-BIN-DENOISE-DESKEW -O OCR-D-SEG-REG -P level-of-operation page" \
  "tesserocr-deskew -I OCR-D-SEG-REG -O OCR-D-SEG-REG-DESKEW" \
  "cis-ocropy-clip -I OCR-D-SEG-REG-DESKEW -O OCR-D-SEG-REG-DESKEW-CLIP" \
  "cis-ocropy-segment -I OCR-D-SEG-REG-DESKEW-CLIP -O OCR-D-SEG-LINE" \
  "cis-ocropy-clip -I OCR-D-SEG-LINE -O OCR-D-SEG-CLIP-LINE -P level-of-operation line" \
  "cis-ocropy-dewarp -I OCR-D-SEG-CLIP-LINE -O OCR-D-SEG-LINE-RESEG-DEWARP" \
  "calamari-recognize -I OCR-D-SEG-LINE-RESEG-DEWARP -O OCR-D-OCR -P checkpoint /path/to/models/\*.ckpt.json"
```

Literatur

- Weil, S. (2019). Training Fraktur. GitHub. https://github.com/tesseract-ocr/tesstrain/wiki
- Weil, S. (2019). Vom Bild zum Text. Automatisierte Texterkennung in historischen Drucken mit der freien Software Tesseract. https://nbn-resolving.org/urn:nbn:de:0290-opus4-163511
- Weil, S., & Zumstein, P. (2016). Mit freier Software Text in Digitalisaten erkennen. https://speakerdeck.com/zuphilip/mit-freier-software-text-in-digitalisaten-erkennen-ocr-praxis-an-der-ub-mannheim

Bildquellen

- Titelseite:
  https://pixabay.com/photos/letter-handwriting-written-ink-447577/
  https://pixabay.com/photos/wash-angle-hook-book-printing-705674/
- Vektorgrafiken:
  https://pixabay.com/de/vectors/flach-design-symbol-icon-www-2126884/
  https://pixabay.com/de/vectors/flach-design-symbol-icon-www-2126884/
  https://pixabay.com/de/vectors/flach-design-symbol-icon-www-2126880/
  https://pixabay.com/de/vectors/werkzeug-schraubenschl%C3%BCssel-3456474/
- Schriftarten: https://typo-info.de/entwicklung-der-schrift/
- GitHub Logos: https://github.com/logos
- Tux: https://de.wikipedia.org/wiki/Linux#/media/Datei:Tux.svg
- Win10: https://commons.wikimedia.org/wiki/File:Windows_10_Logo.svg
- Mac Logo: https://commons.wikimedia.org/wiki/File:Apple_logo_black.svg
- OCR-D Logo: https://www.ocr-d.de/
- DFG-Logo: https://www.dfg.de/service/logo_corporate_design/index.html