ETH zürich

Volltext für Wissenschaft und Lehre auf e-rara.ch
Erfahrungen und Probleme mit Fraktur-OCR

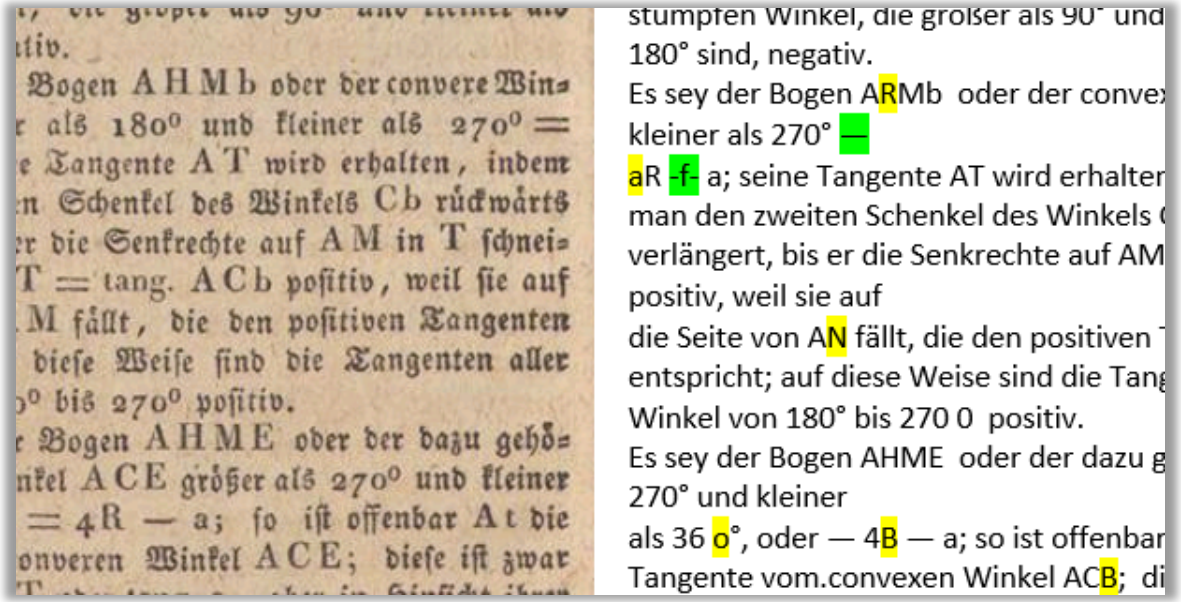
7. Bibliothekskongress Leipzig

18. März 2019

Oliver Ammann

Agenda

- Die Plattform e-rara.ch
- Ausgangslage «e-rara.ch: Volltext» (2016-2017)
- Erweiterung Volltexterkennung: Erfahrungen, Probleme, Ergebnisse
- Fazit
- Perspektive

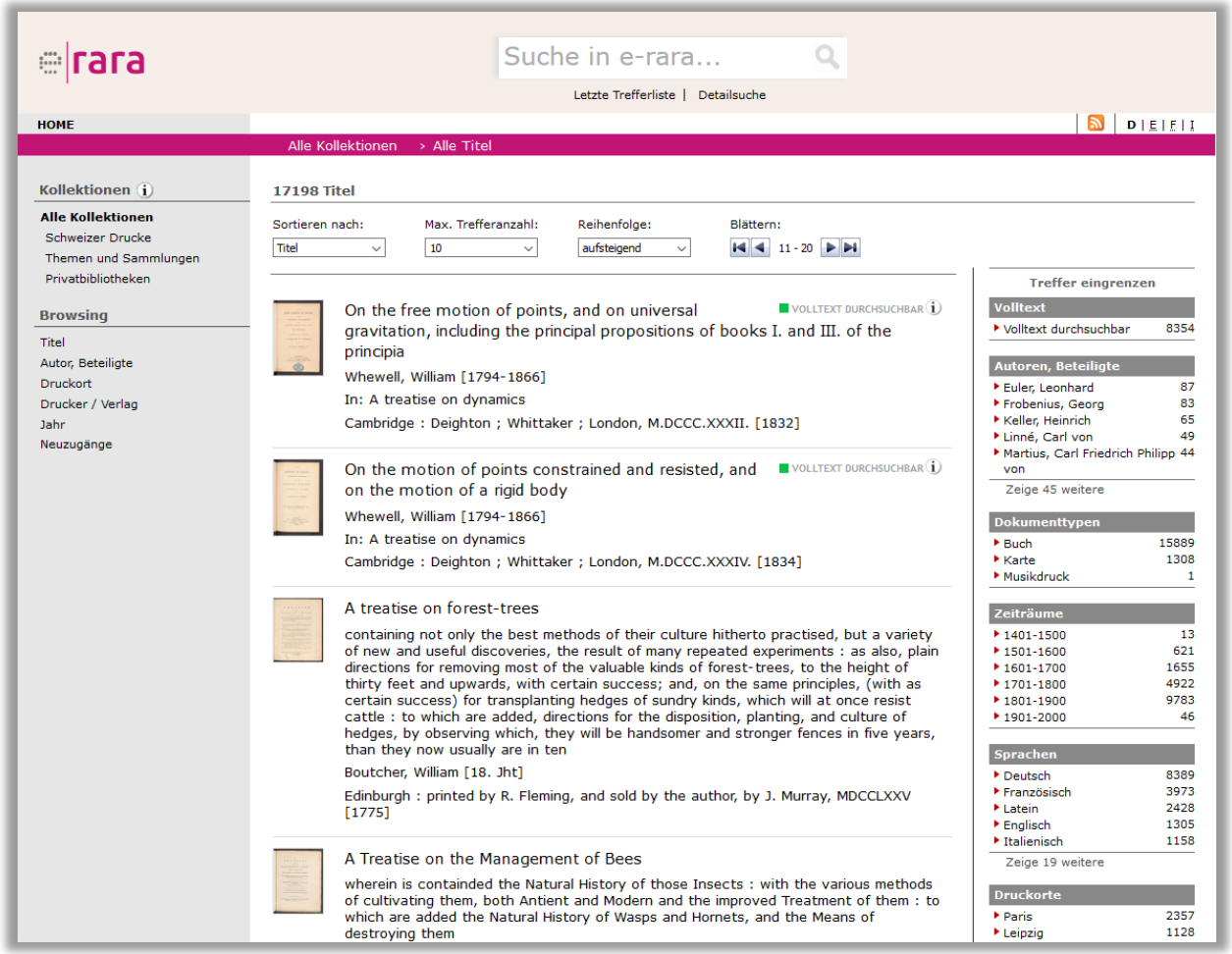


tiv.
Bogen AHMb oder der convexe Win-
r als 180° und kleiner als 270° =
e Tangente AT wird erhalten, indem
n Schenkel des Winkels Cb rückwärts
r die Senkrechte auf AM in T schnei-
T = tang. ACb positiv, weil sie auf
M fällt, die den positiven Tangenten
diese Weise sind die Tangenten aller
0° bis 270° positiv.
Bogen AHME oder der dazu gehö-
nkel ACE größer als 270° und kleiner
= 4R — a; so ist offenbar At die
onveren Winkel ACE; diese ist zwar

stumpfen Winkel, die grösser als 90° und
180° sind, negativ.
Es sey der Bogen ARMb oder der convex
kleiner als 270° —
aR -f- a; seine Tangente AT wird erhalten
man den zweiten Schenkel des Winkels
verlängert, bis er die Senkrechte auf AM
positiv, weil sie auf
die Seite von AN fällt, die den positiven
entspricht; auf diese Weise sind die Tang
Winkel von 180° bis 270 0 positiv.
Es sey der Bogen AHME oder der dazu g
270° und kleiner
als 36 o°, oder — 4B — a; so ist offenbar
Tangente vom.convexen Winkel ACB; di

Die Plattform e-rara.ch

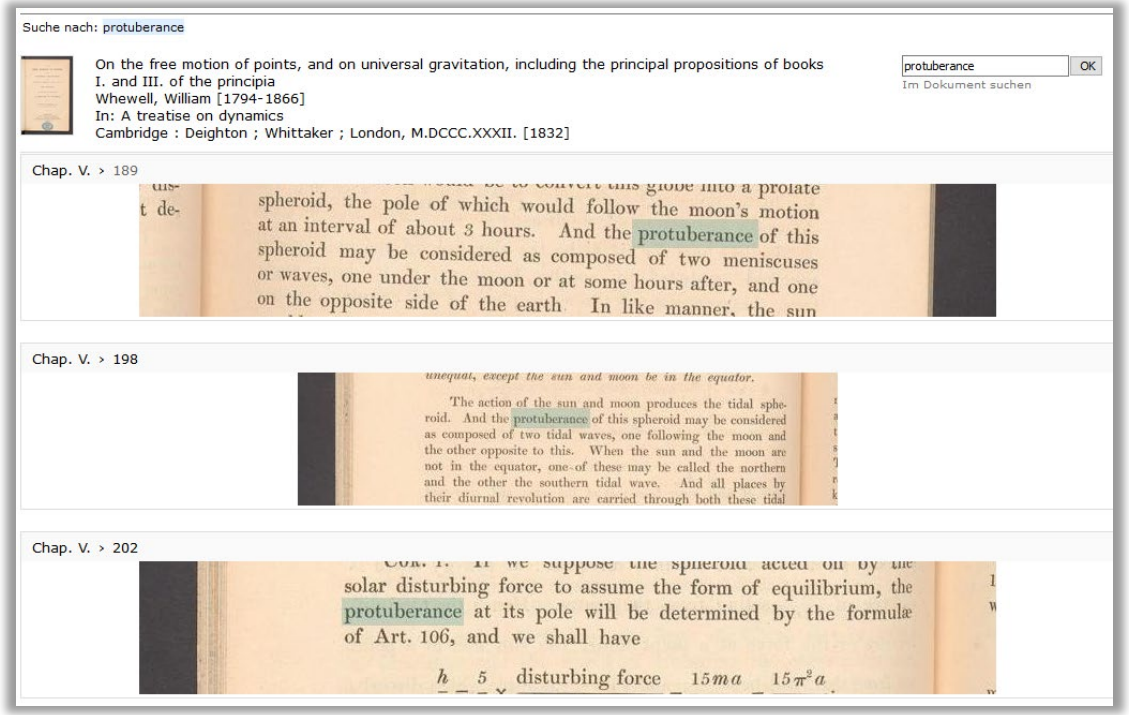
- Plattform für digitalisierte Drucke aus Schweizer Bibliotheken
- 2007 als Projekt gestartet
- Seit 2010 online
- 5 Träger, 19 Institutionen
- 71.000 Titel online, davon 17.000 Titel der ETH-Bibliothek
- Weiterentwicklungen
  - Kollektion Privatbibliotheken
  - Volltext



Digitalisate der ETH-Bibliothek Zürich auf e-rara.ch [Screenshot]

Ausgangslage

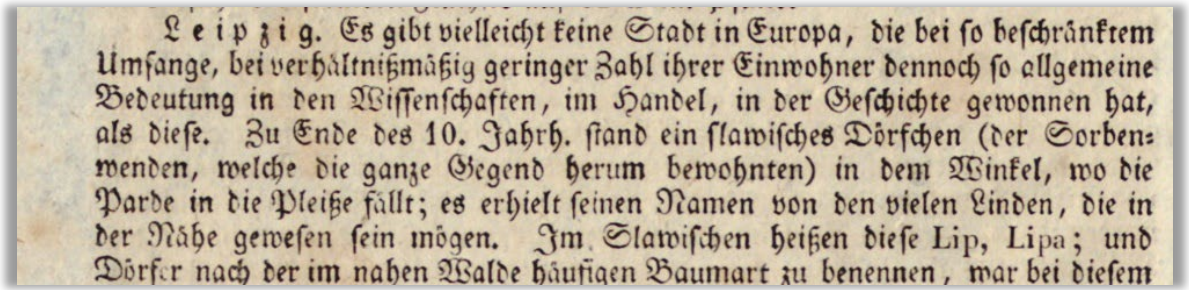
- «e-rara.ch: Volltext» (2016 – 2017)
- OCR für 3.700 Titel ab Druckjahr 1830
- nur Antiqua-Schriften
- ABBYY FineReader
- Volltextsuche online und in PDFs
- Treffer in Snippets hervorgehoben



Suchbegriff hervorgehoben in Snippets. [Screenshot e-rara.ch]

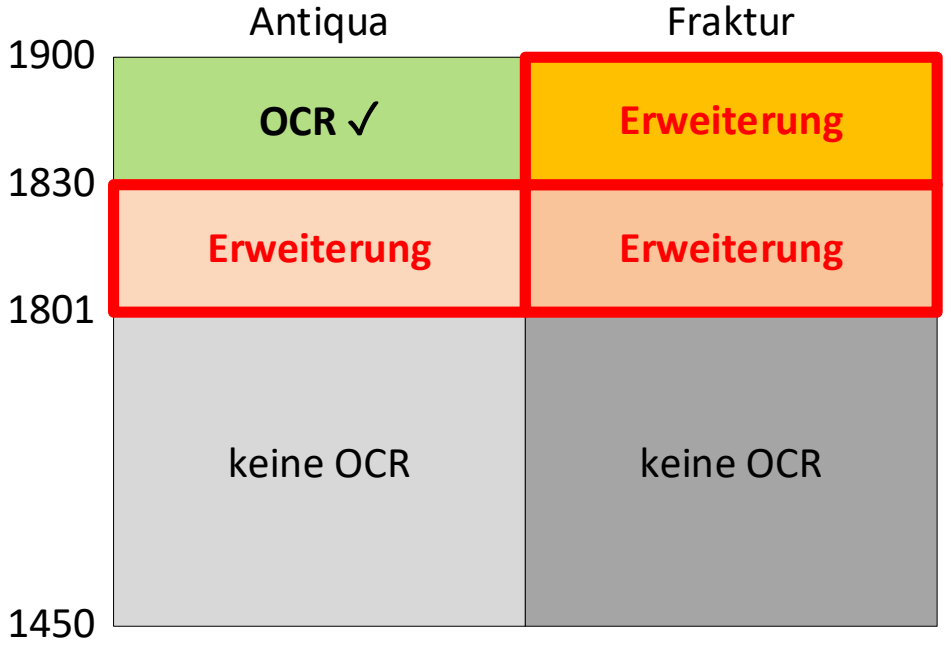
Übersicht Erweiterung Volltexterkennung (2018)

- Volltext für 6.000 Titel (~ 3 Mio. Seiten)
- ab Druckjahr **1801**
- Antiqua- und **Frakturschrift**
- ABBYY FineReader



Leipzig. Es gibt vielleicht keine Stadt in Europa, die bei so beschränktem Umfange, bei verhältnißmäßig geringer Zahl ihrer Einwohner dennoch so allgemeine Bedeutung in den Wissenschaften, im Handel, in der Geschichte gewonnen hat, als diese. Zu Ende des 10. Jahrh. stand ein slawisches Dörfchen (der Sorbenwenden, welche die ganze Gegend herum bewohnten) in dem Winkel, wo die Parde in die Pleiße fällt; es erhielt seinen Namen von den vielen Linden, die in der Nähe gewesen sein mögen. Im Slawischen heißen diese Lip, Lipa; und Dörfer nach der im nahen Walde häufigen Baumart zu benennen, war bei diesem

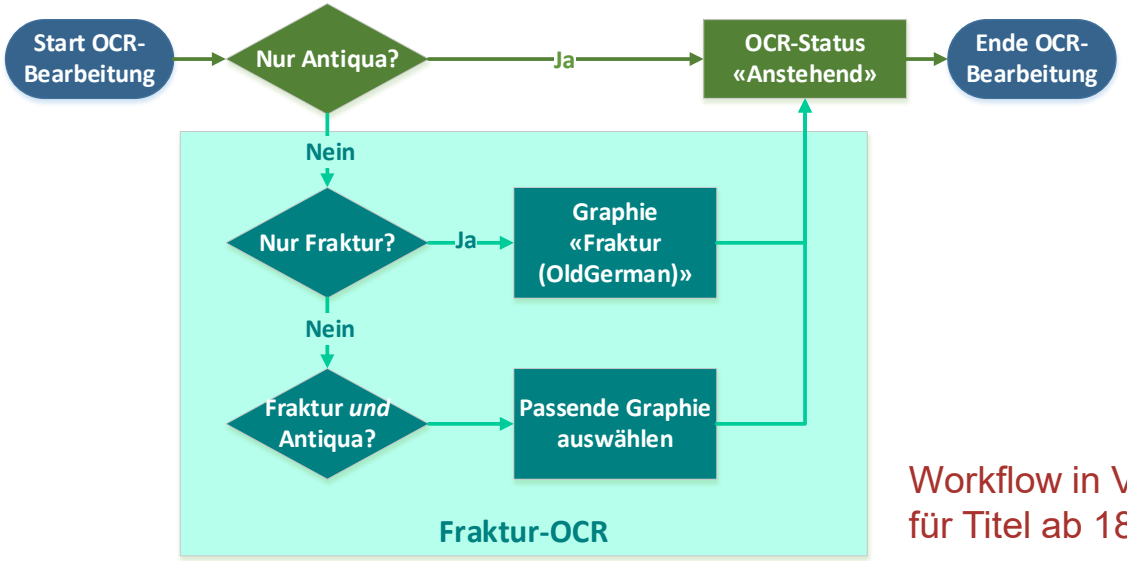
Eintrag «Leipzig» in Frakturschrift aus: *Allgemeine deutsche Real-Encyklopädie [...]*, Reutlingen 1830-1831, Seite 531. ETH-Bibliothek Zürich, Rar 3421, http://doi.org/10.3931/e-rara-17880.



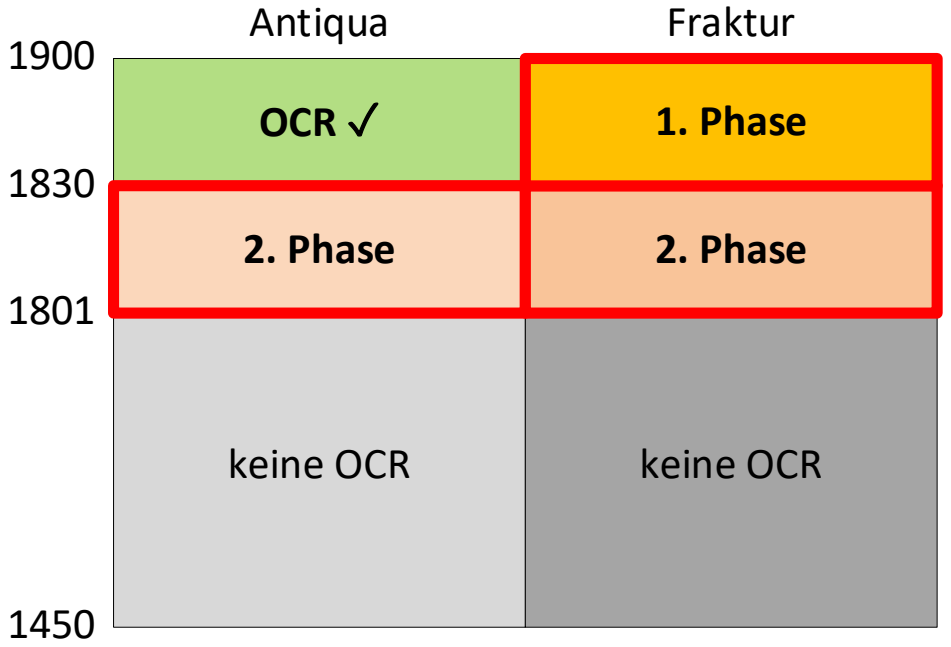
OCR-Bearbeitung der Teilbestände auf e-rara.ch

Umsetzung & Workflow

- Drei Phasen
  1. 3.000 Titel in Frakturschrift nach 1830
  2. 3.000 Titel nach 1801 und vor 1830: Sichtung und Feststellung der Graphie
  3. Workflow für neu hinzukommende Titel



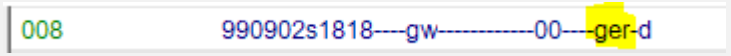
Workflow in VLM für Titel ab 1801



OCR-Bearbeitung der Teilbestände auf e-rara.ch: Vorgehen

Metadaten und Graphie*

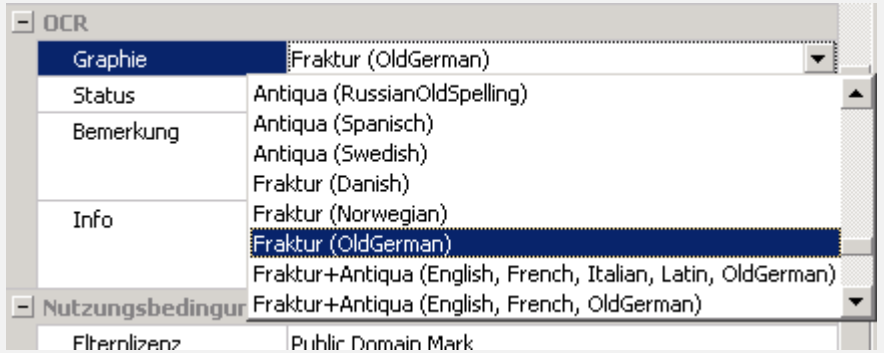
Katalog, MARC-Felder 008, 041: **Sprachcode**



- keine Graphie-Metadaten

?

VLM: **Graphie-Einstellung**
(Kombination aus Graphie und Sprache)



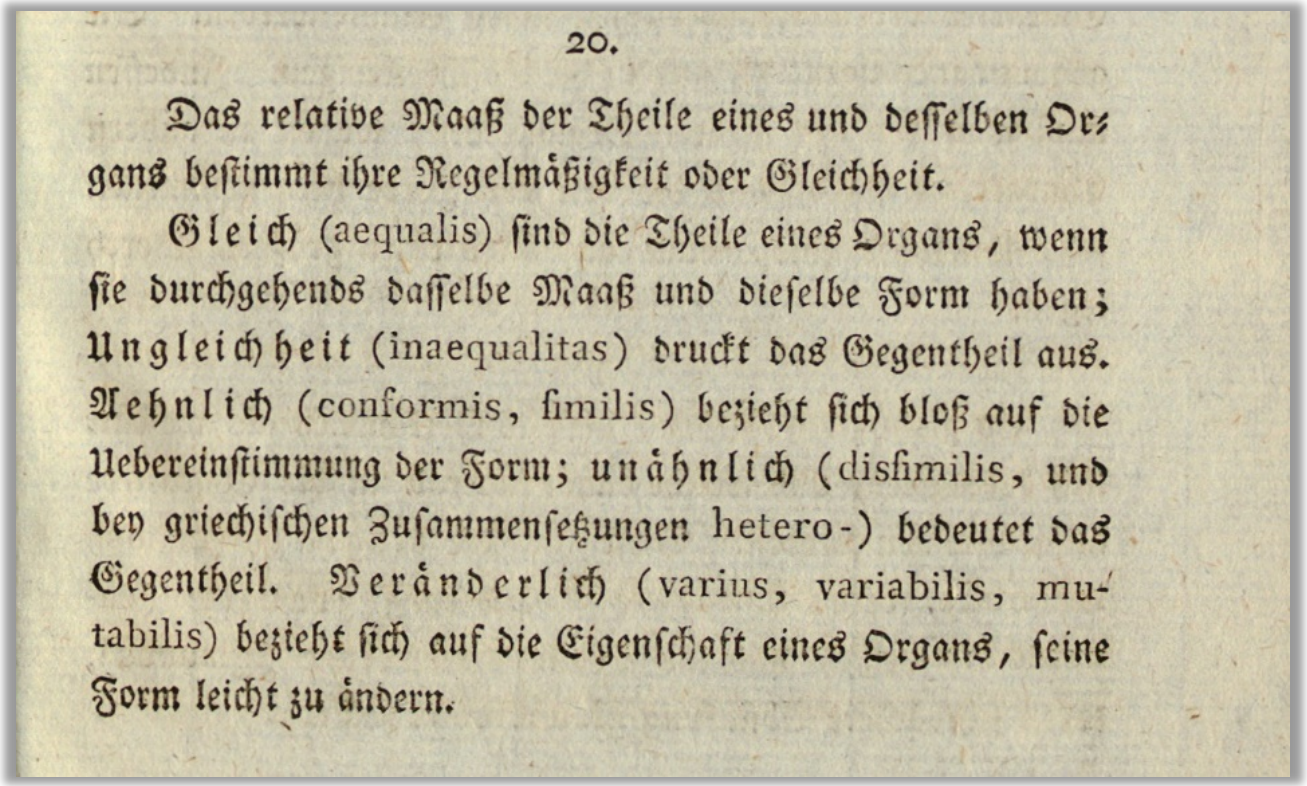
- manuelle Auswahl der passenden Graphie

---

* Einstellung für Schriftart und Sprache bei der OCR-Bearbeitung mit ABBYY FineReader und in Visual Library Manager

Erwartete Probleme mit OCR für Fraktur+Antiqua 1801+

- keine Graphie-Metadaten
- schlechte Druckqualität
- (unterschiedliche) Frakturschriften
- Antiqua-Fraktur-Mischtexte



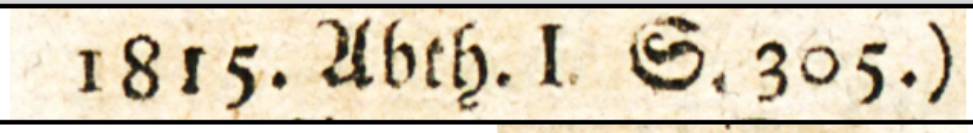
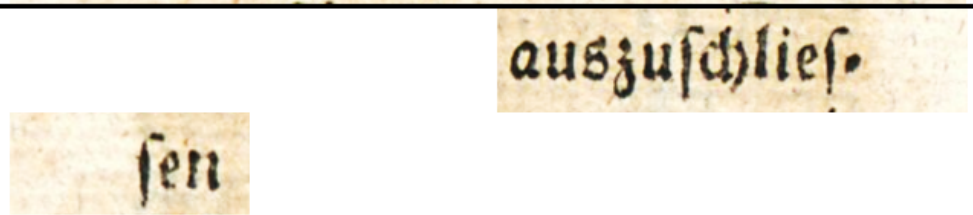
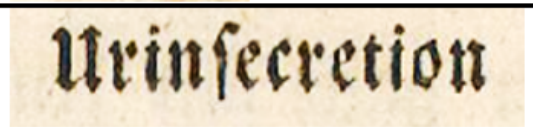
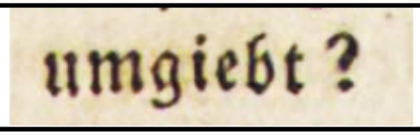
20.

Das relative Maaß der Theile eines und deſſelben Organs beſtimmt ihre Regelmäßigkeit oder Gleichheit.

Gleich (aequalis) ſind die Theile eines Organs, wenn ſie durchgehends daſſelbe Maaß und dieſelbe Form haben; Ungleichheit (inaequalitas) druckt das Gegentheil aus. Aehnlich (conformis, ſimilis) bezieht ſich bloß auf die Uebereinſtimmung der Form; unähnlich (diſſimilis, und bey griechiſchen Zuſammenſetzungen hetero-) bedeutet das Gegentheil. Veränderlich (varius, variabilis, mutabilis) bezieht ſich auf die Eigenſchaft eines Organs, ſeine Form leicht zu ändern.

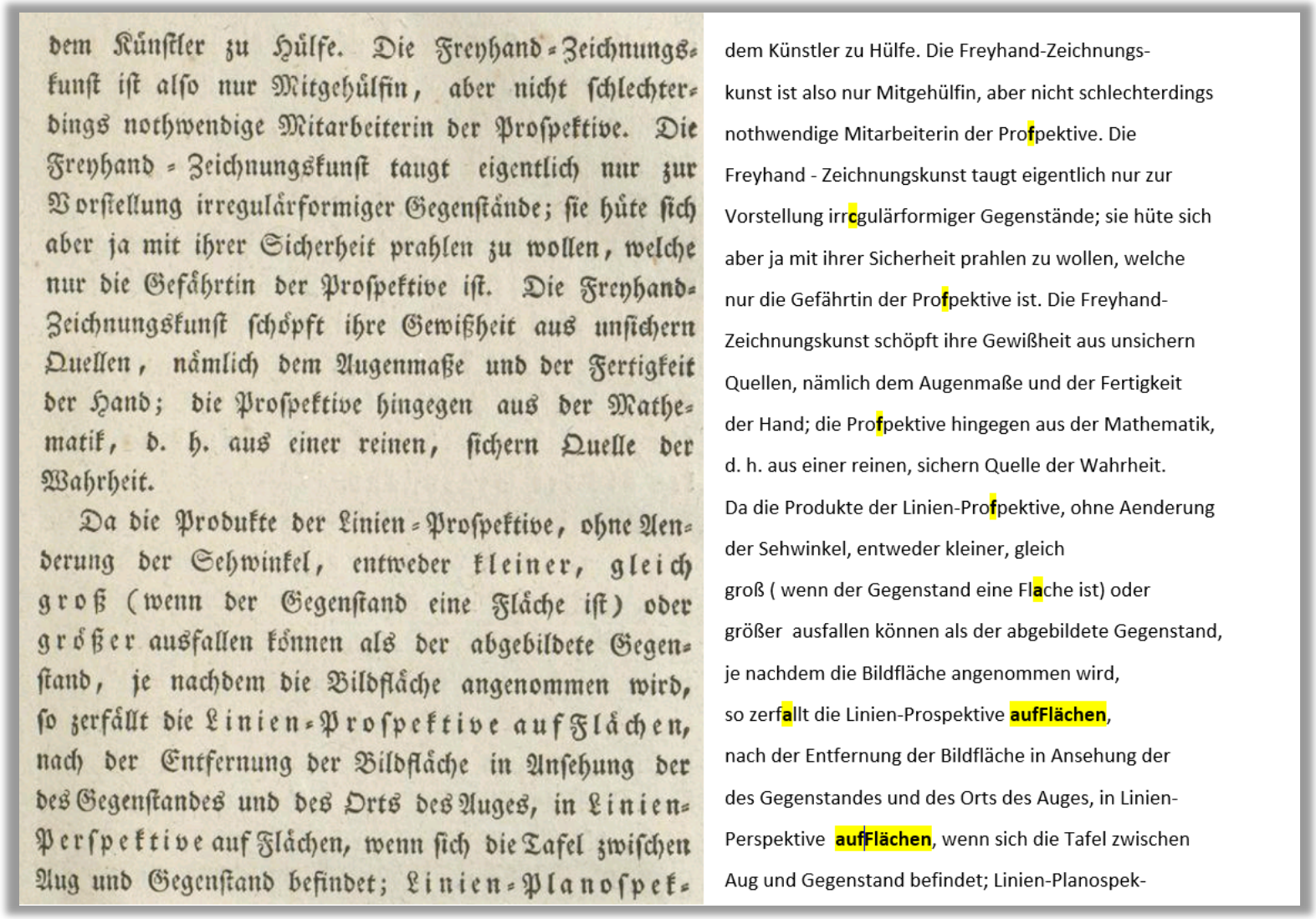
Mischtext in: *A. P. de Candolle's und K. Sprengel's Grundzüge der wissenschaftlichen Pflanzenkunde [...]*. Leipzig 1820, Seite 15. ETH-Bibliothek Zürich, Rar 3722, http://doi.org/10.3931/e-rara-18699.

Ergebnisse I

Fehlertyp	Beispiel	OCR-Ergebnis
Ziffern		i8rz. Abth. l S. zc>;.)
Worttrennung		auszuschlies. sen
c ↔ e		Urinseeretion
u ↔ n		nmgiebt?

Auswahl typischer Fehler bei Fraktur-OCR auf e-rara.ch

Ergebnisse II



dem Künſtler zu Hülfe. Die Freyhand-Zeichnungs-
kunſt iſt alſo nur Mitgehülfin, aber nicht ſchlechter-
dings nothwendige Mitarbeiterin der Proſpektive. Die
Freyhand-Zeichnungskunſt taugt eigentlich nur zur
Vorſtellung irregulärformiger Gegenſtände; ſie hüte ſich
aber ja mit ihrer Sicherheit prahlen zu wollen, welche
nur die Gefährtin der Proſpektive iſt. Die Freyhand-
Zeichnungskunſt ſchöpft ihre Gewißheit aus unſichern
Quellen, nämlich dem Augenmaße und der Fertigkeit
der Hand; die Proſpektive hingegen aus der Mathe-
matik, d. h. aus einer reinen, ſichern Quelle der
Wahrheit.

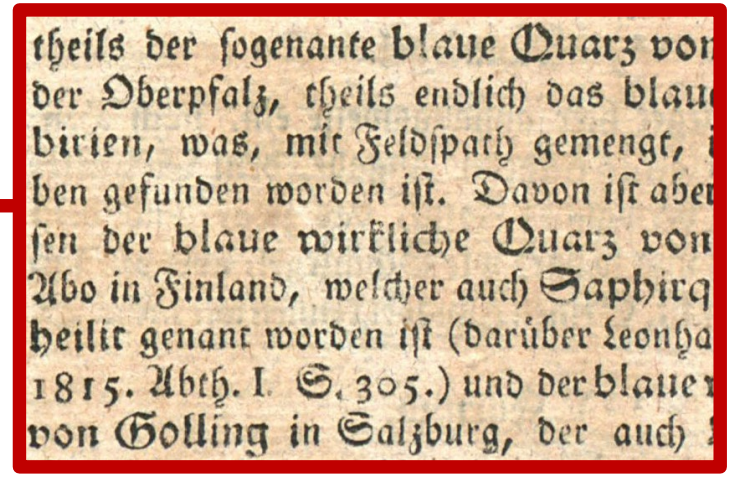
Da die Produkte der Linien-Proſpektive, ohne Aen-
derung der Sehwinkel, entweder kleiner, gleich
groß (wenn der Gegenſtand eine Fläche iſt) oder
größer ausfallen können als der abgebildete Gegen-
ſtand, je nachdem die Bildfläche angenommen wird,
ſo zerfällt die Linien-Proſpektive auf Flächen,
nach der Entfernung der Bildfläche in Anſehung der
des Gegenſtandes und des Orts des Auges, in Linien-
Perſpektive auf Flächen, wenn ſich die Tafel zwiſchen
Aug und Gegenſtand befindet; Linien-Planoſpek-

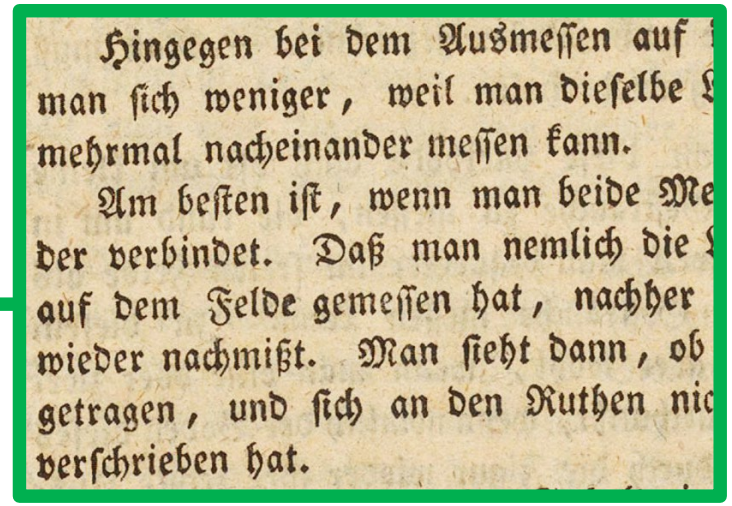
dem Künstler zu Hülfe. Die Freyhand-Zeichnungs-
kunst ist also nur Mitgehülfin, aber nicht schlechterdings
nothwendige Mitarbeiterin der Profpektive. Die
Freyhand - Zeichnungskunst taugt eigentlich nur zur
Vorstellung irrcgulärformiger Gegenstände; sie hüte sich
aber ja mit ihrer Sicherheit prahlen zu wollen, welche
nur die Gefährtin der Profpektive ist. Die Freyhand-
Zeichnungskunst schöpft ihre Gewißheit aus unsichern
Quellen, nämlich dem Augenmaße und der Fertigkeit
der Hand; die Profpektive hingegen aus der Mathematik,
d. h. aus einer reinen, sichern Quelle der Wahrheit.
Da die Produkte der Linien-Profpektive, ohne Aenderung
der Sehwinkel, entweder kleiner, gleich
groß (wenn der Gegenstand eine Flache ist) oder
größer ausfallen können als der abgebildete Gegenstand,
je nachdem die Bildfläche angenommen wird,
so zerfallt die Linien-Prospektive **aufFlächen**,
nach der Entfernung der Bildfläche in Ansehung der
des Gegenstandes und des Orts des Auges, in Linien-
Perspektive **auf**Flächen, wenn sich die Tafel zwischen
Aug und Gegenstand befindet; Linien-Planospek-

Johann Friedrich Ladomus: *Umfang und Eintheilung der Prospektive*. Königsberg 1804, Seite 19. Zentralbibliothek Zürich, NE 1859,3, http://doi.org/10.3931/e-rara-52262

Ergebnisse III: Frakturschrift

Testseite	WG	ZG
https://www.e-rara.ch/zut/content/pageview/11419715 (1817)	86,6	97,2
https://www.e-rara.ch/zut/content/pageview/4223294 (1826)	93,9	98,7
https://www.e-rara.ch/sikjm/content/pageview/2993807 (1802)	96,1	99,1
https://www.e-rara.ch/zut/content/pageview/19329070 (1819)	96,2	99,1
https://www.e-rara.ch/zuz/content/pageview/14339618 (1804)	93,9	99,2
https://www.e-rara.ch/zut/content/pageview/4597935 (1802)	97,5	99,2
https://www.e-rara.ch/zuz/content/pageview/8784910 (1826)	96,1	99,3
https://www.e-rara.ch/zuz/content/pageview/13231297 (1823)	96,5	99,5
https://www.e-rara.ch/zut/content/pageview/16335222 (1818)	98,9	99,6
Mittelwert 9 Testseiten	***95,1***	***99,0***

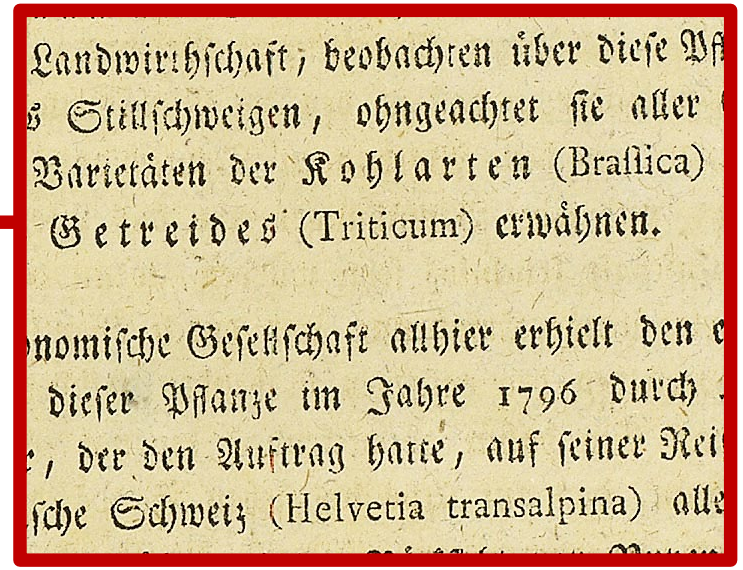

theils der ſogenante blaue Quarz von
der Oberpfalz, theils endlich das blau
birien, was, mit Feldſpath gemengt,
ben gefunden worden iſt. Davon iſt aber
ſen der blaue wirkliche Quarz von
Abo in Finland, welcher auch Saphirq
heilit genant worden iſt (darüber Leonha
1815. Abth. I. S. 305.) und der blaue
von Golling in Salzburg, der auch


Hingegen bei dem Ausmeſſen auf
man ſich weniger, weil man dieſelbe
mehrmal nacheinander meſſen kann.
Am beſten iſt, wenn man beide Me
der verbindet. Daß man nemlich die
auf dem Felde gemeſſen hat, nachher
wieder nachmißt. Man ſieht dann, ob
getragen, und ſich an den Ruthen nic
verſchrieben hat.

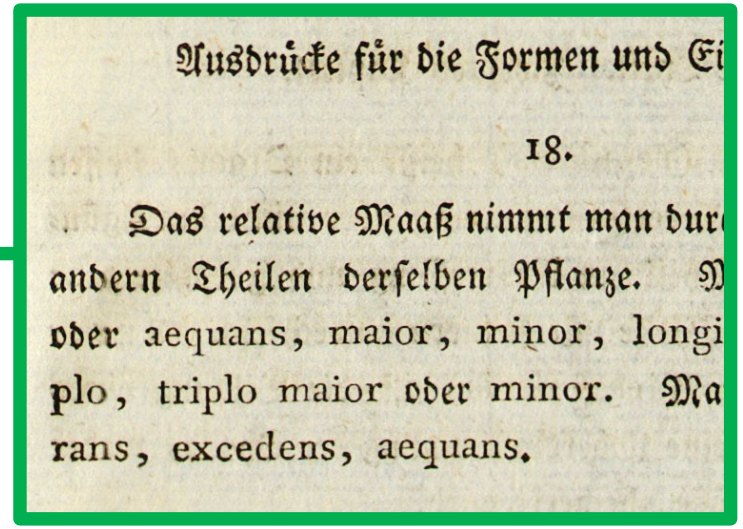
Auszählung Erkennungsfehler in Testseiten aus Testsample «Fraktur 1801-1830»;
WG: Wortgenauigkeit in %, ZG: Zeichengenauigkeit in %. Groß- und Kleinschreibung, Sonderzeichen und Ziffern sind nicht berücksichtigt.

Ergebnisse IV: Mischschrift

Testseite	WG	ZG
https://www.e-rara.ch/zuz/content/pageview/8266791 (1804)	91,8	97,6
https://www.e-rara.ch/zut/content/pageview/5253211 (1810)	90,7	97,8
https://www.e-rara.ch/zut/content/pageview/18577029 (1821)	90,6	97,9
https://www.e-rara.ch/zuz/content/pageview/10777711 (1803)	90,8	98,0
https://www.e-rara.ch/zut/content/pageview/13602392 (1819)	94,6	98,8
https://www.e-rara.ch/zut/content/pageview/12378413 (1824-1825)	94,2	98,9
https://www.e-rara.ch/zuz/content/pageview/8759470 (1805)	95,1	99,0
https://www.e-rara.ch/bau_1/content/pageview/9770818 (1828)	98,1	99,6
https://www.e-rara.ch/zut/content/pageview/5892271 (1820)	99,0	99,7
Mittelwert 9 Testseiten	***93,9***	***98,6***


Landwirthſchaft; beobachten über dieſe Pf
s Stillſchweigen, ohngeachtet ſie aller
Varietäten der Kohlarten (Braſſica)
Getreides (Triticum) erwähnen.

nomiſche Geſellſchaft allhier erhielt den
dieſer Pflanze im Jahre 1796 durch
, der den Auftrag hatte, auf ſeiner Rei
ſche Schweiz (Helvetia transalpina) alle


Ausdrücke für die Formen und Ei

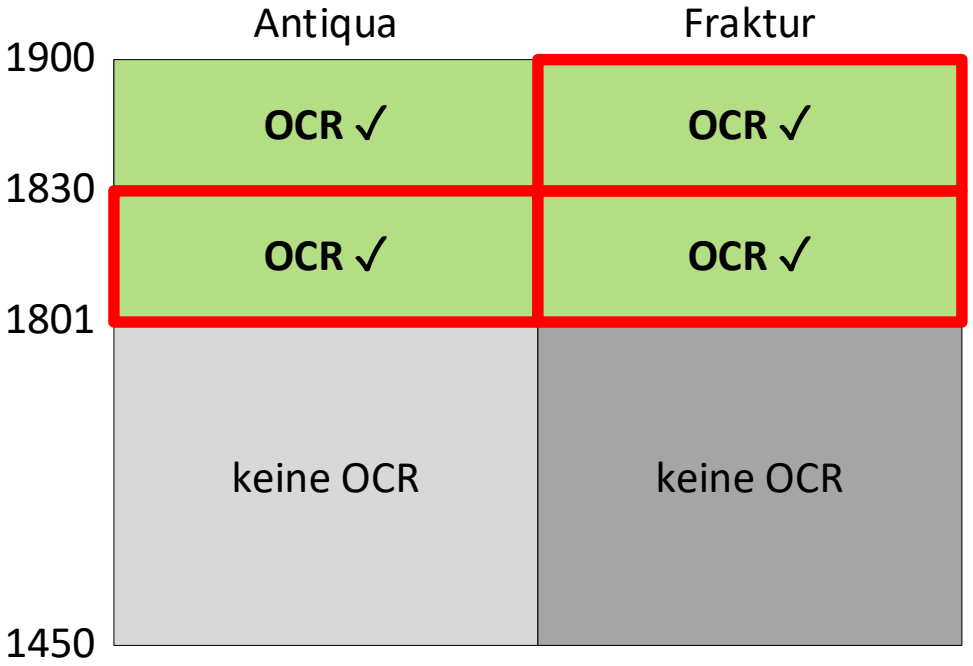
18.

Das relative Maaß nimmt man dur
andern Theilen derſelben Pflanze. N
oder aequans, maior, minor, longi
plo, triplo maior oder minor. Ma
rans, excedens, aequans.

Auszählung Erkennungsfehler in Testseiten aus Testsample «Fraktur+Antiqua 1801-1830»; WG: Wortgenauigkeit in %, ZG: Zeichengenauigkeit in %. Groß- und Kleinschreibung, Sonderzeichen und Ziffern sind nicht berücksichtigt.

Fazit

- 11.700 von 71.000 Titeln mit Volltext
- Relativ aufwendige Feststellung der Graphie
  - ➢ Bessere Ergebnisse rechtfertigen den Aufwand
  - ➢ Kosten für eine kombinierte Lizenz Fraktur+Antiqua rund doppelt so teuer wie einfache Antiqua-Lizenz
- NutzerInnen schätzen die guten Volltexte
- Visits e-rara.ch ↑ 33%



OCR-Status der Teilbestände auf e-rara.ch nach der Erweiterung 2018

Perspektiven: Aktuelle Weiterentwicklungen e-rara.ch

- IIIF-Manifest und neuer Viewer
- JPEG-Download
- Social Media-Share Buttons
- **Volltext-Download (Textdatei)**
- Export Metadaten für Literaturverwaltung im Format .ris
  (verfügbar später im Frühjahr 2019)

Vielen Dank für Ihre Aufmerksamkeit

7. Bibliothekskongress Leipzig | 18. März 2019

ETH Zürich, ETH-Bibliothek, Oliver Ammann, Alte und Seltene Drucke, Rämistrasse 101, 8092 Zürich
Tel. +41 44 632 49 05, oliver.ammann@library.ethz.ch, www.library.ethz.ch