Layout Detection and Table Recognition

Recent Challenges in Digitizing Historical Documents and Handwritten Tabular Data

About us

Constantin Lehenmeier

IT EMPLOYEE / UNIVERSITY LIBRARY OF REGENSBURG

PHD STUDENT / CHAIR OF MEDIA INFORMATICS / UNIVERSITY OF REGENSBURG

Junior Prof. Dr. Manuel Burghardt

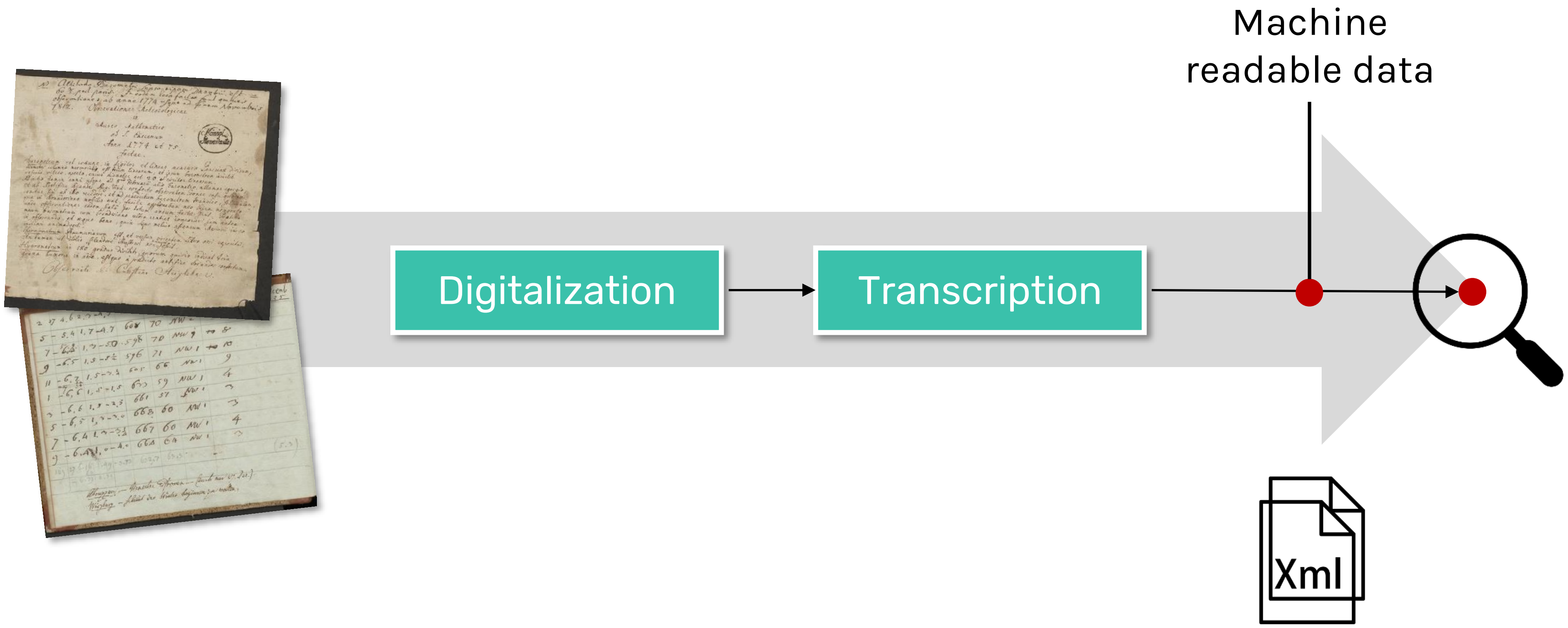
CHAIR HOLDER COMPUTATIONAL HUMANITIES / UNIVERSITY OF LEIPZIG

Bernadette Mischka

PHD STUDENT / CHAIR OF EUROPEAN HISTORY/ UNIVERSITY OF REGENSBURG

Libraries and the digitzed humanities

The digitization of cultural artifacts is part of the “digitized humanities”



Digitalization

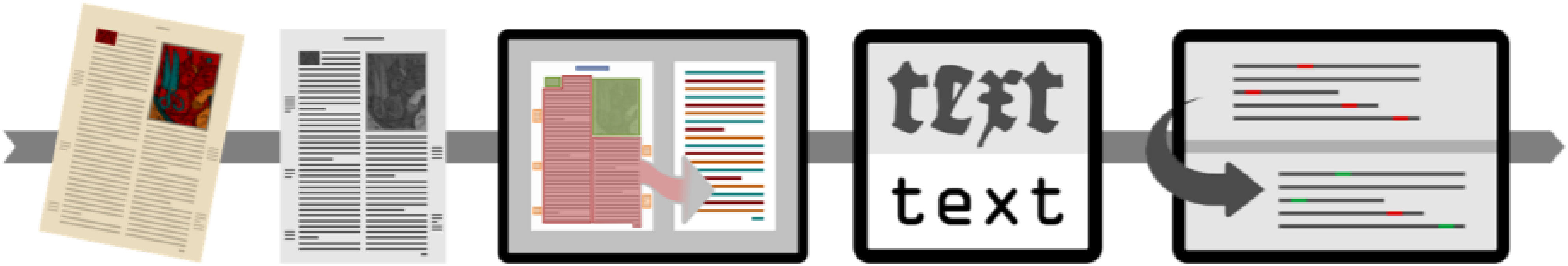
Transcription

Machine readable data

… there are billions of documents

Automated document recognition

OCR is used to automatically convert visible language to a searchable digital format



SCAN PREPROCESSING LAYOUT ANALYSIS TEXT RECOGNITION POSTPROCESSING

Figure 1: The main steps of the OCR process

State-of-the-art

CER under 1% on modern printed books without specific training

CER under 2% on early printed books with specific training

Character Error Rate (CER) is the edit distance between two sequences

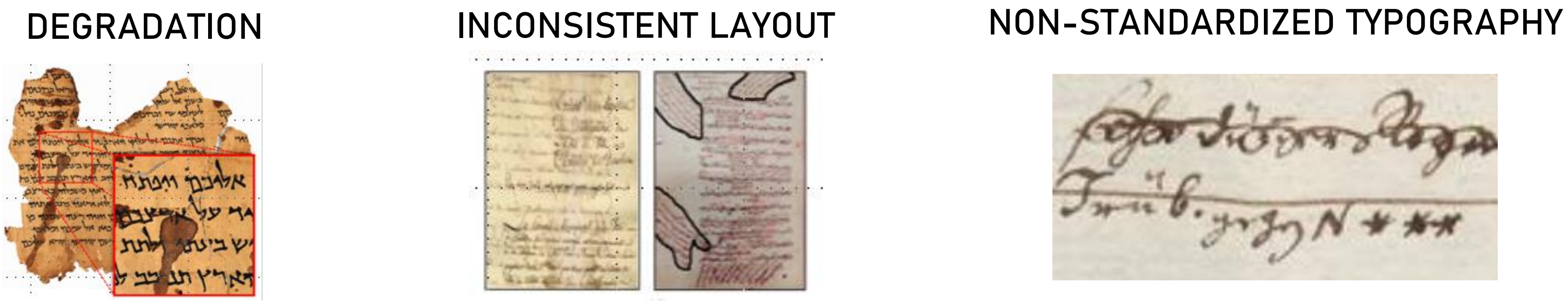
State-of-the-art

CER under 1% on modern printed books without specific training

CER under 2% on early printed books with specific training

Character Error Rate (CER) is the edit distance between two sequences

❗ Historical documents pose special challenges


DEGRADATION

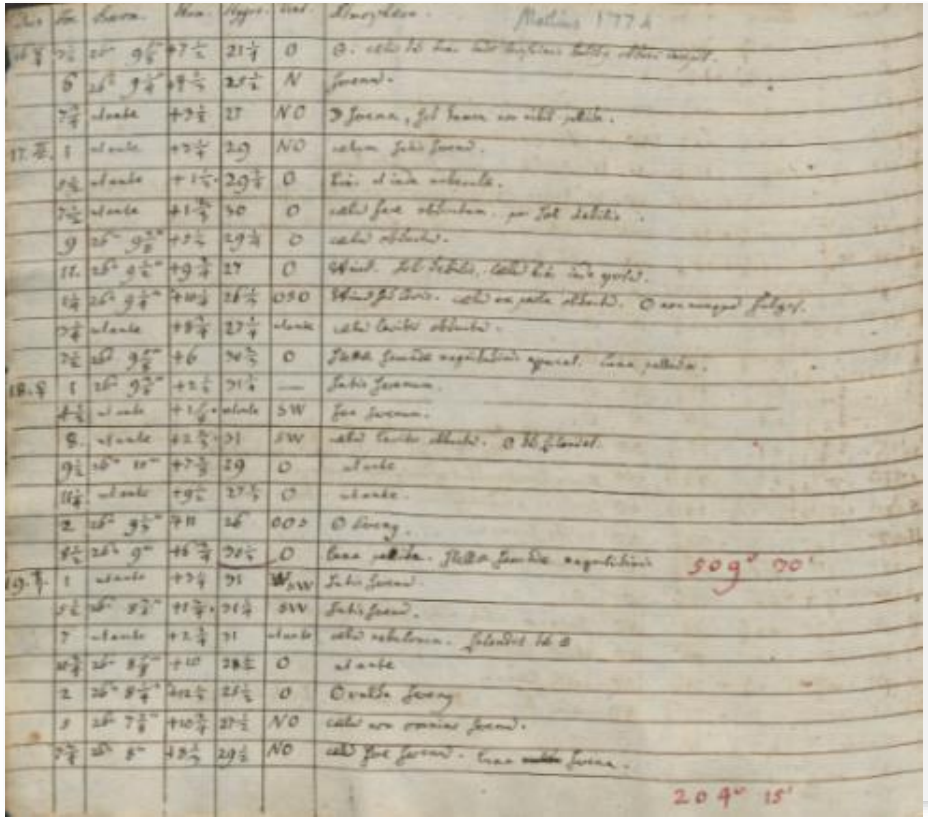
INCONSISTENT LAYOUT

NON-STANDARDIZED TYPOGRAPHY

Figure 2: Main challenges of the automatic recognition of historical documents

Observationes meteorologicae

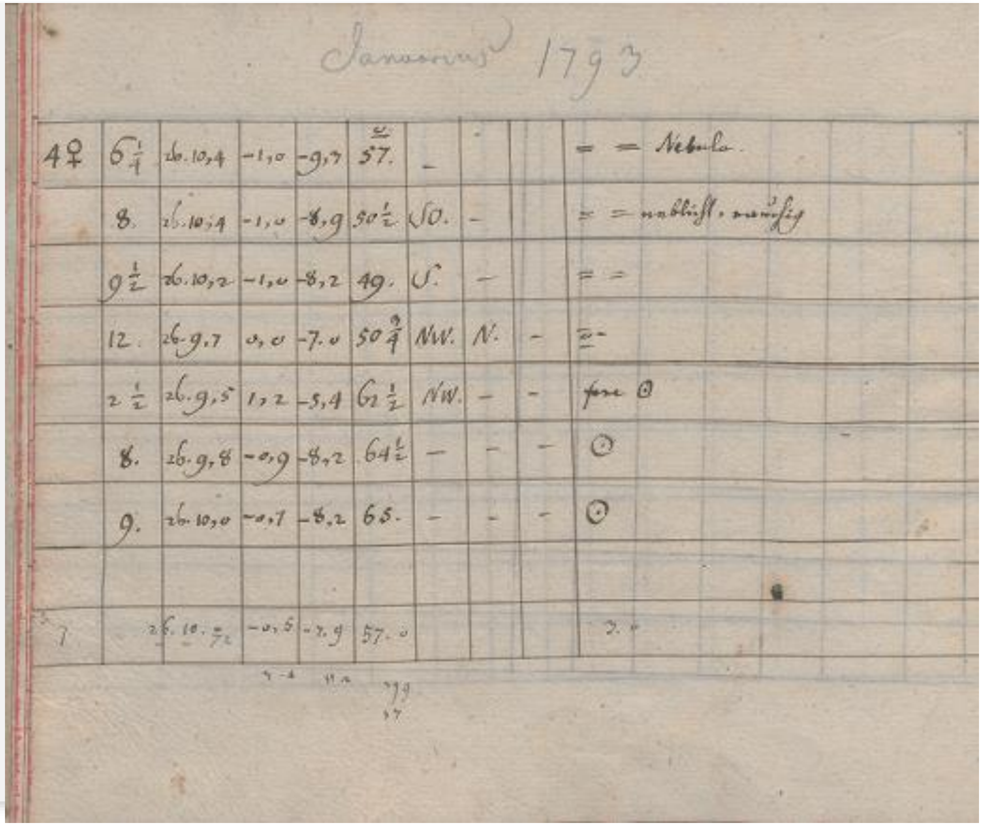
Continious weather reports for over 53 years. The university library owns 55 volumes for the years 1774 – 1827.



Martius 1774

Dies	Hor.	Baro.	Ther.	Hygr.	Vent.	Atmosphaera.
16 ☿	7½	26" 9⅝	+7½	21¼	O	☉. [illegible]
	6	26" 9¾	+9½	25½	N	Serenum.
	7¾	ut ante	+3½	27	NO	☽ serena, [illegible]
17. ♃	1	ut ante	+7¾	29	NO	[illegible]
	1½	ut ante	+1½	29¾	O	[illegible]
	7½	ut ante	+1¾	30	O	[illegible]
	9	26" 9⅞"	+3½	29¾	O	[illegible]
	11	26" 9½"	+9¾	27	O	[illegible]
	1¼	26" 9¼"	+10½	26½	OSO	[illegible]
	7¾	ut ante	+8¾	27¼	ut ante	[illegible]
	7½	26" 9⅝"	+6	34⅔	O	[illegible] Luna pallida.
18. ♀	1	26" 9⅞"	+2½	31¼	—	[illegible]
	4½	ut ante	+1½	ut ante	SW	[illegible]
	8	ut ante	+2⅔	31	SW	[illegible]
	9½	26" 10"	+7⅔	29	O	ut ante
	11¼	ut ante	+9½	27½	O	ut ante.
	2	26" 9½"	+11	26	OOS	☉ [illegible]
	8½	26" 9"	+6¾	31½	O	Luna pallida. [illegible] 509° 00'
19. ♄	1	ut ante	+3¾	31	WSW	[illegible]
	5½	26" 8⅞"	+1¾	31¼	SW	[illegible]
	7	ut ante	+2¾	31	ut ante	[illegible]
	10¾	26" 8⅞"	+10	28½	O	ut ante
	2	26" 8¼"	+12½	25½	O	[illegible]
	5	26" 7⅞"	+10¾	27½	NO	[illegible]
	7¾	26" 8"	+8⅔	29½	NO	[illegible]

204° 15'



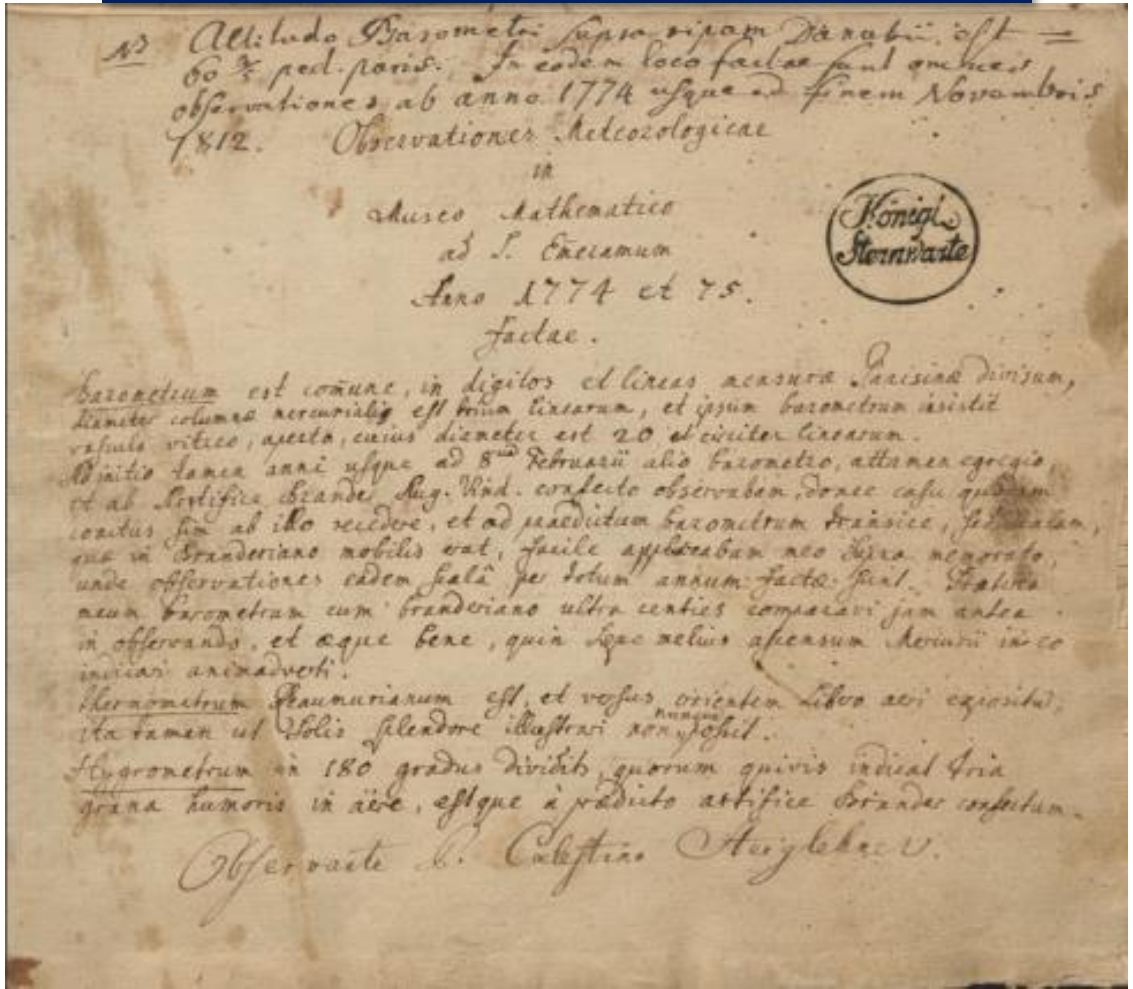
Januarius 1793

4♀	6¼	26.10,4	-1,0	-9,7	57.	—				= = Nebula.
	8.	26.10,4	-1,0	-8,9	50½	SO.	—			= = [illegible]
	9½	26.10,2	-1,0	-8,2	49.	S.	—			= =
	12.	26.9,7	0,0	-7.0	50¾	NW.	N.	—		≡ -
	2½	26.9,5	1,2	-5,4	62½	NW.	—	—		fere ☉
	8.	26.9,8	-0,9	-8,2	64½	—	—	—		☉
	9.	26.10,0	-0,7	-8,2	65.	—	—	—		☉
7		26.10.½	-0,5	-7,9	57.0					3.

799
57

There is something about the weather…

First page of the reports in 1774



N.° Altitudo Barometri supra ripam Danubii est = 60 ⅔ ped. paris. In eodem loco factae sunt omnes observationes ab anno 1774 usque ad finem Novembris 1812. Observationes Meteorologicae

in

Museo Mathematico

ad S. Emeramum

Anno 1774 et 75.

factae.

Königl. Sternwarte

Barometrum est comune, in digitos et lineas mensura Parisina divisum, diameter columnae mercurialis est trium linearum, et ipsum barometrum insistit vasculo vitreo, aperto, cuius diameter est 20 et circiter linearum.

Initio tamen anni usque ad 8um Februarii alio barometro, attamen egregio, et ab Artifice Brander Aug. Vind. confecto observabam, donec casu quodam coactus sum ab illo recedere, et ad praedictum barometrum transire, scalam, quae in Branderiano mobilis erat, facile applicabam meo supra memorato, unde observationes eadem scalâ per totum annum factae sunt. Praeterea meum barometrum cum Branderiano ultra centies comparavi jam antea in observando, et aeque bene, quin saepe melius ascensum Mercurii in eo indicari animadverti.

Thermometrum Reaumurianum est, et versus orientem libero aeri expositum, ita tamen ut Solis splendore illustrari nunquam non possit.

Hygrometrum in 180 gradus divisum, quorum quivis indicat tria grana humoris in aëre, estque à praedicto artifice Brander confectum.

Observante P. Coelestino Steiglehner.

Reconstruction of past climate conditions

Created by Max Hancock
from Noun Project

Historical climate impact research

Created by hidayat
from Noun Project

Scientific history of the climate

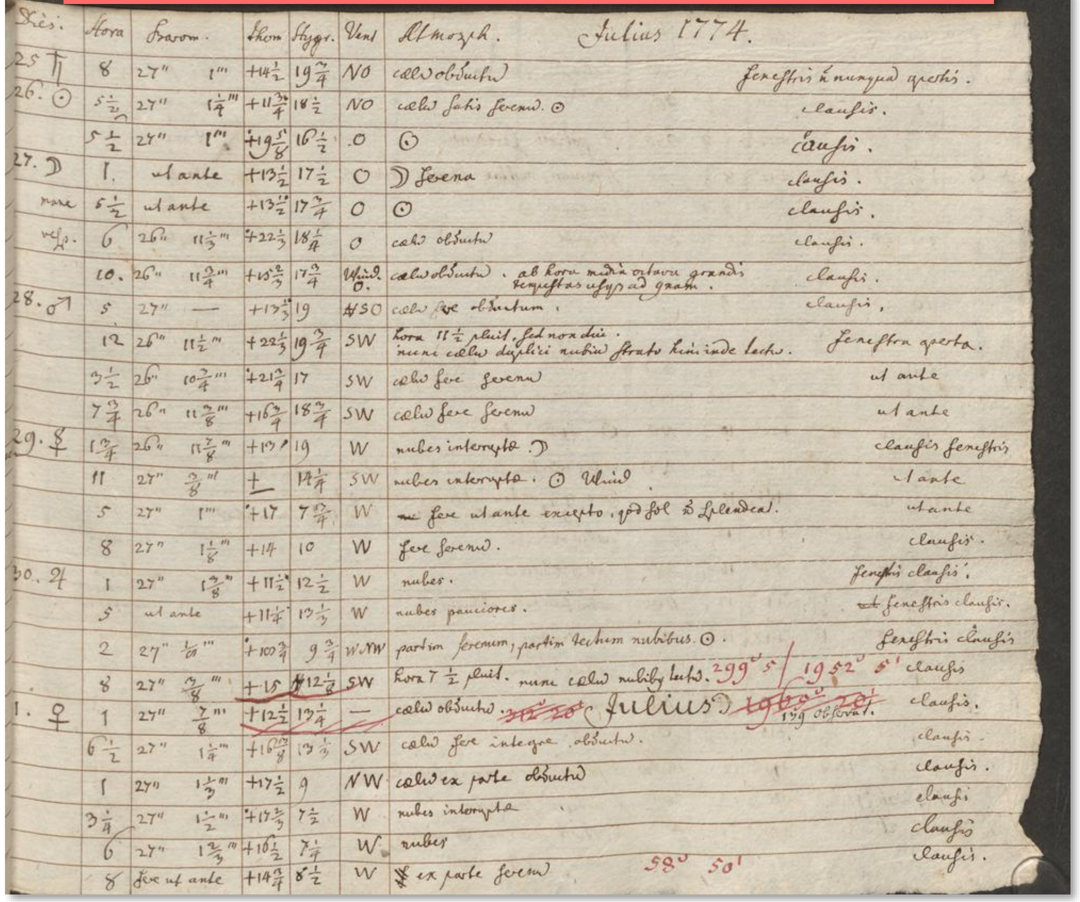
Created by Maxo
from Noun Project

Asses futues climate fluctuations

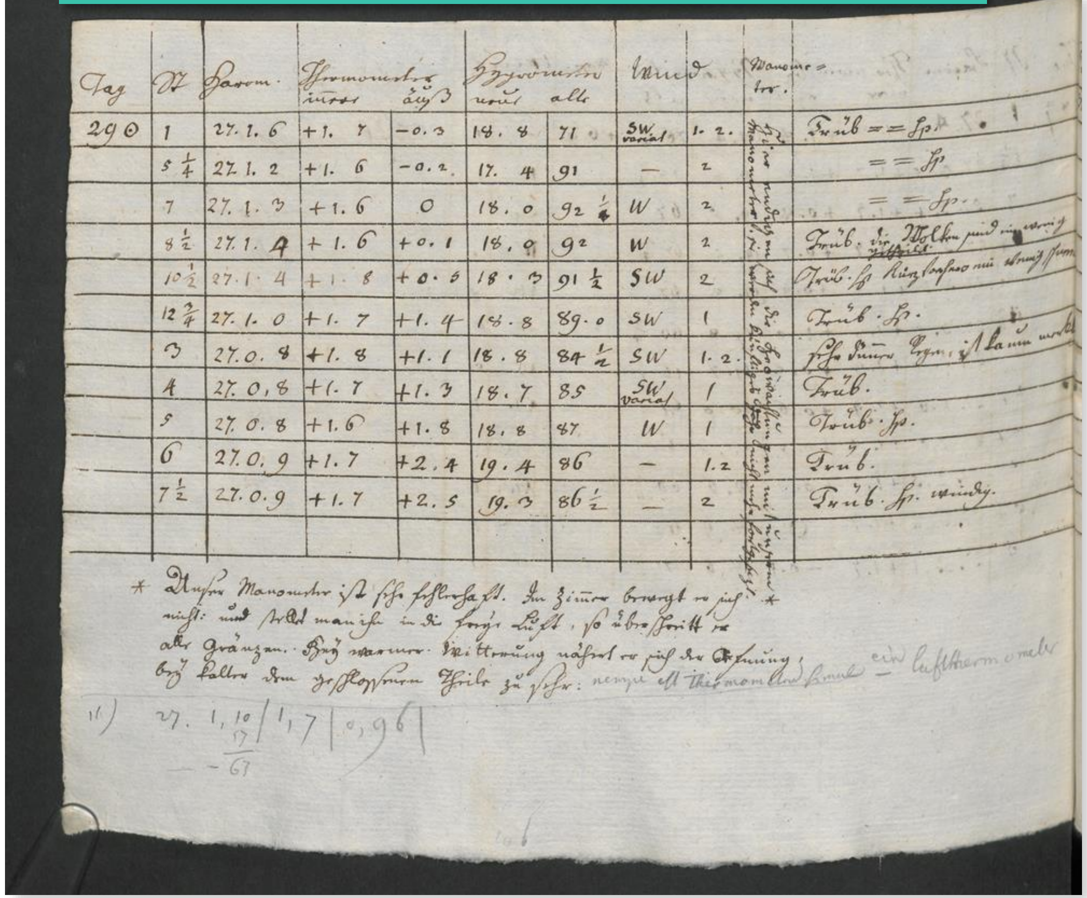
Created by [illegible]
from the Noun Project

Problems

Tables + **Handwritten text**



Dies:	Hora	Barom.		Thom	Hygr.	Vent	Atmosph. Julius 1774.	
25 ♄	8	27"	1'''	+14½	19¾	NO	cœlum obductum	fenestris nunquam apertis.
26. ☉	5½	27"	1¼'''	+11¾	18½	NO	cœlum satis serenum. ☉	clausis.
	5½	27"	1'''	+19⅝	16½	O	☉	clausis.
27. ☽	1.	ut ante		+13½	17½	O	☽ serena	clausis.
mane	5½	ut ante		+13½	17¾	O	☉	clausis.
vesp.	6	26"	11⅓'''	+22⅓	18¼	O	cœlum obductum	clausis.
	10.	26"	11¾'''	+15⅔	17¾	Wind. O.	cœlum obductum. ab hora media octava grandis tempestas usque ad 9nam.	clausis.
28. ♂	5	27"	—	+13⅓	19	~~N~~SO	cœlum fere obductum.	clausis,
	12	26"	11½'''	+22⅓	19¾	SW	hora 11½ pluit, sed non diu. nunc cœlum duplici nubium strato hinc inde tectum.	fenestra aperta.
	3½	26"	10¾'''	+21¾	17	SW	cœlum fere serenum	ut ante
	7¾	26"	11⅞'''	+16¾	18¾	SW	cœlum fere serenum	ut ante
29. ☿	1¾	26"	11⅞'''	+19	19	W	nubes interruptæ. ☽	clausis fenestris
	11	27"	⅜'''	+	14¼	SW	nubes interruptæ. ☉ Wind.	ut ante
	5	27"	1'''	+17	7¾	W	fere ut ante excepto, quod sol non splendeat.	ut ante
	8	27"	1⅛'''	+14	10	W	fere serenum.	clausis.
30. ♃	1	27"	1⅜'''	+11½	12½	W	nubes.	fenestris clausis.
	5	ut ante		+11¼	13⅓	W	nubes pauciores.	~~ut~~ fenestris clausis.
	2	27"	1/10'''	+10¾	9¾	WNW	partim serenum, partim tectum nubibus. ☉.	fenestris clausis
	8	27"	⅜'''	+15	12½	SW	hora 7½ pluit. nunc cœlum nubibus tectum. 299° 5' | 1952° 5'	clausis
1. ♀	1	27"	⅞'''	+12½	13¼	—	cœlum obductum. ~~312° 20'~~ (Julius) ~~1960° 20'~~ 139 observat.	clausis.
	6½	27"	1¼'''	+16⅞	13⅓	SW	cœlum fere integre obductum.	clausis.
	1	27"	1⅓'''	+17½	9	NW	cœlum ex parte obductum	clausis.
	3¾	27"	1½'''	+17⅔	7½	W	nubes interruptæ	clausis
	6	27"	1⅔'''	+16½	7¼	W	nubes	clausis
	8	fere ut ante		+14¾	8½	W	ex parte serenum 58° 50'	clausis.



Tag	St	Barom.	Thermometer innen	äuß	Hygrometer neu	alte	Wind		Manometer.	
29 ☉	1	27.1.6	+1.7	−0.3	18.8	71	SW variat	1.2.	[illegible]	Trüb = = Sp.
	5¼	27.1.2	+1.6	−0.2	17.4	91	—	2		= = Sp
	7	27.1.3	+1.6	0	18.0	92¼	W	2		= = Sp.
	8½	27.1.4	+1.6	+0.1	18.0	92	W	2		Trüb. die Wolken sind ein wenig [illegible]
	10½	27.1.4	+1.8	+0.5	18.3	91½	SW	2		Trüb. Sp. [illegible]
	12¾	27.1.0	+1.7	+1.4	18.8	89.0	SW	1		Trüb. Sp.
	3	27.0.8	+1.8	+1.1	18.8	84½	SW	1.2.		[illegible]
	4	27.0,8	+1.7	+1.3	18.7	85	SW variat	1		Trüb.
	5	27.0.8	+1.6	+1.8	18.8	87	W	1		Trüb. Sp.
	6	27.0.9	+1.7	+2.4	19.4	86	—	1.2		Trüb.
	7½	27.0.9	+1.7	+2.5	19.3	86½	—	2		Trüb. Sp. windy.

* Unser Manometer ist sehr fehlerhaft. Im Zimmer bewegt es sich nicht: und stellt man es in die freye Luft, so überschreitet es alle Gränzen. Bey warmer Witterung nähert es sich der Öffnung; bey kalter dem geschlossenen Theile zu sehr: [illegible] ein Luftthermometer

1) 27. 1,10 17 | 1,7 | 0,96 | − 67

The authors

The reports were written by three scientists

Coelestin Steiglehner

1771 - 1778

Placidus Heinrich

1771 - 1825

Ferdinand von Schmöger

1825 - 1827

Figure 3: The main authors of the observationes meteorologicae

The authors

The reports were written by three scientists + various students

Coelestin Steiglehner

1771 - 1778

Students

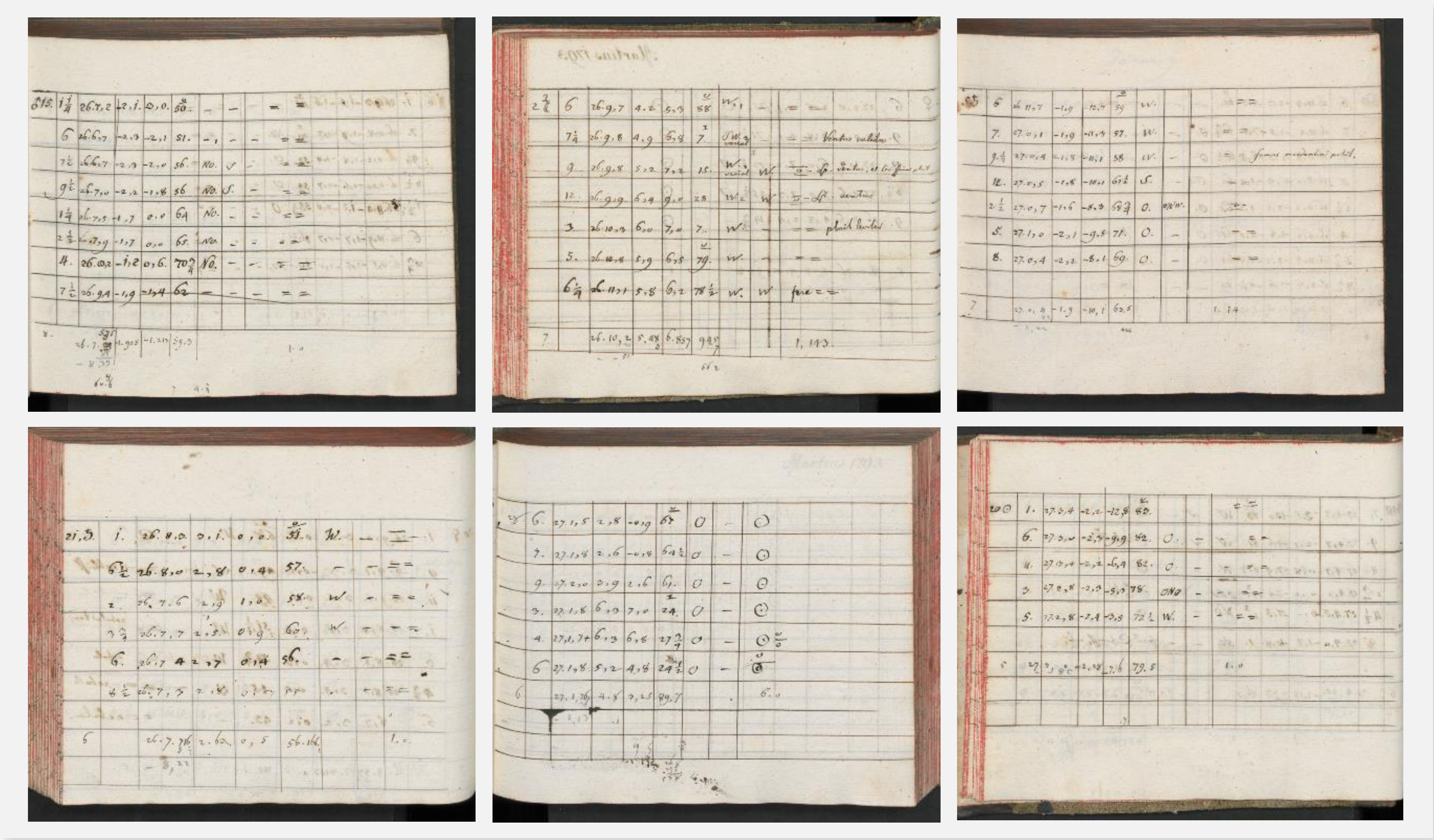
1791 - 1798

Ferdinand von Schmöger

1825 - 1827

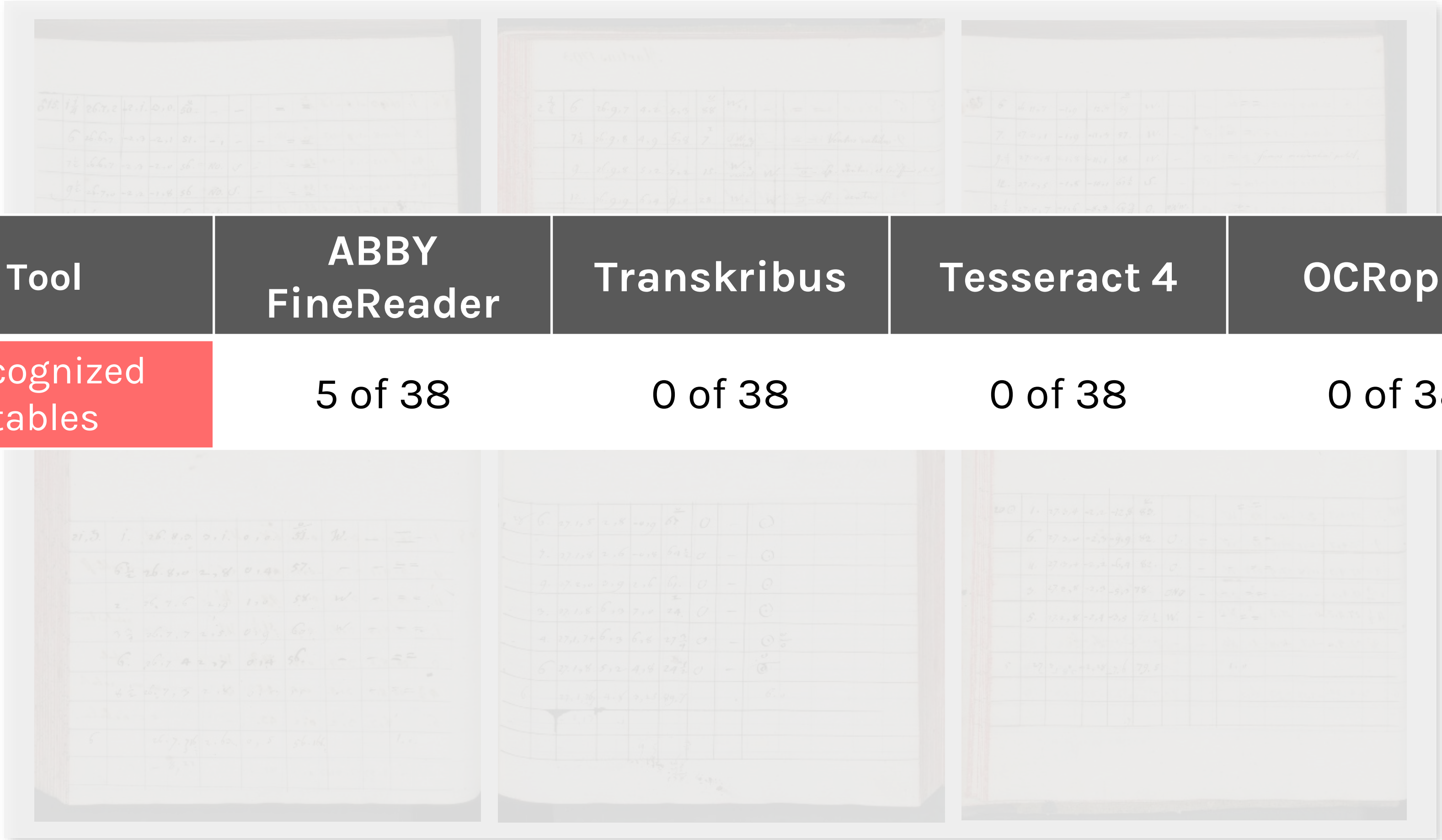
Results of out-of-the-box tools

A test with four existing OCR engines was conducted on 38 pages of one volume (1793)



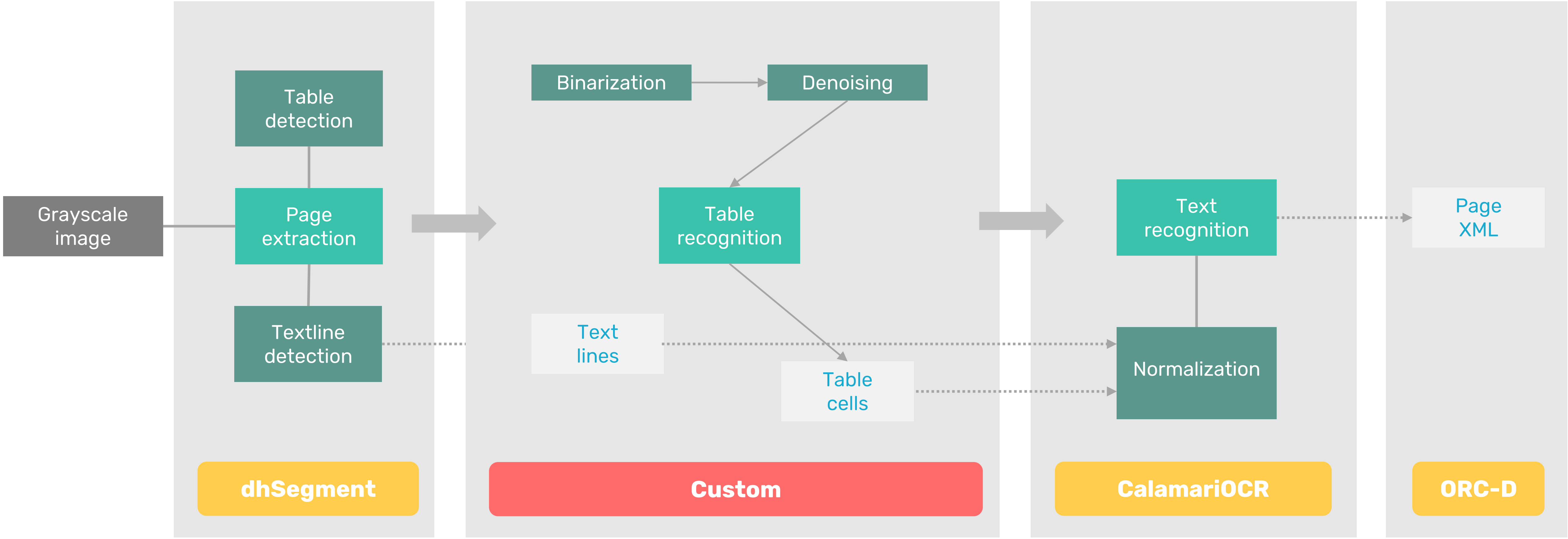
Results of out-of-the-box tools

A test with four existing OCR engines was conducted on 38 pages of one volume

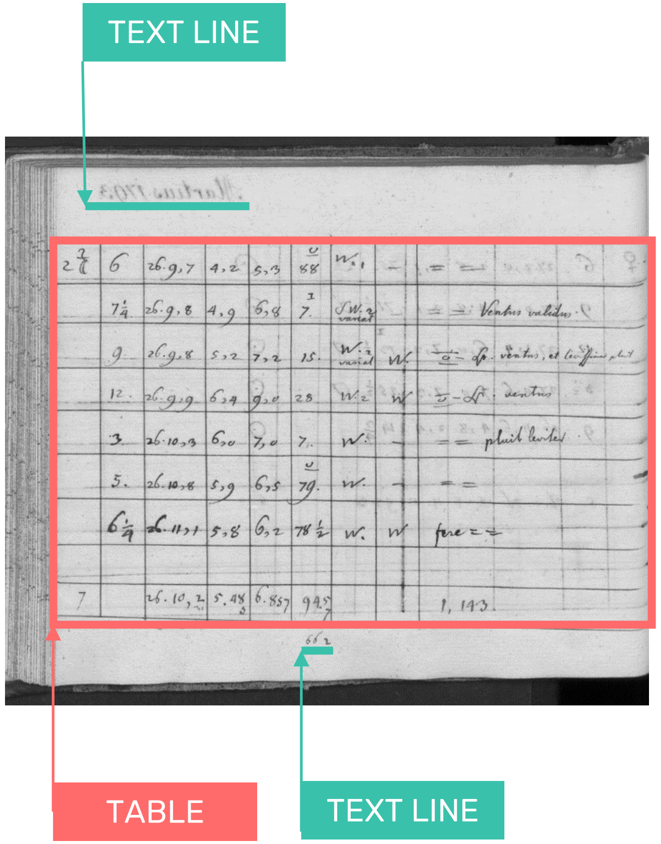


Tool	ABBY FineReader	Transkribus	Tesseract 4	OCRopus
Recognized tables	5 of 38	0 of 38	0 of 38	0 of 38

Current workflow



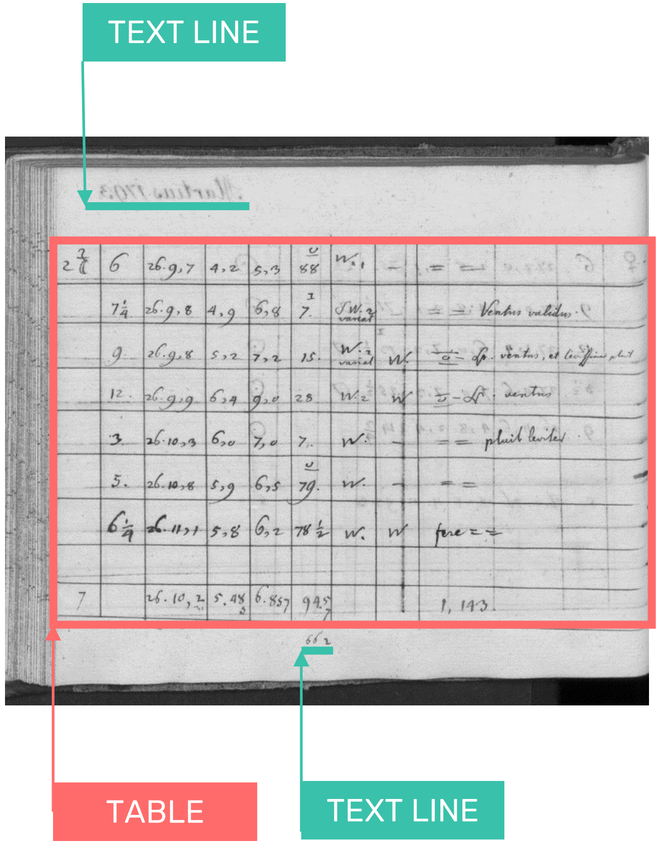
Layout analysis and table recognition



2 ☾	6	26.9,7	4,2	5,3	88	W.1	–	= =
	7¼	26.9,8	4,9	6,8	7.	SW.2 variat	–	= = Ventus validus
	9.	26.9,8	5,2	7,2	15.	W.2 variat	W	– – ventus, et levissime pluit
	12.	26.9,9	6,4	9,0	28	W.2	W	– ventus
	3.	26.10,3	6,0	7,0	7.	W.	–	= = pluit leviter
	5.	26.10,8	5,9	6,5	79.	W.	–	= =
	6¼	26.11,1	5,8	6,2	78½	W.	W	fere = =
7		26.10,2	5.48	6.857	94.5			1, 143.

66 2

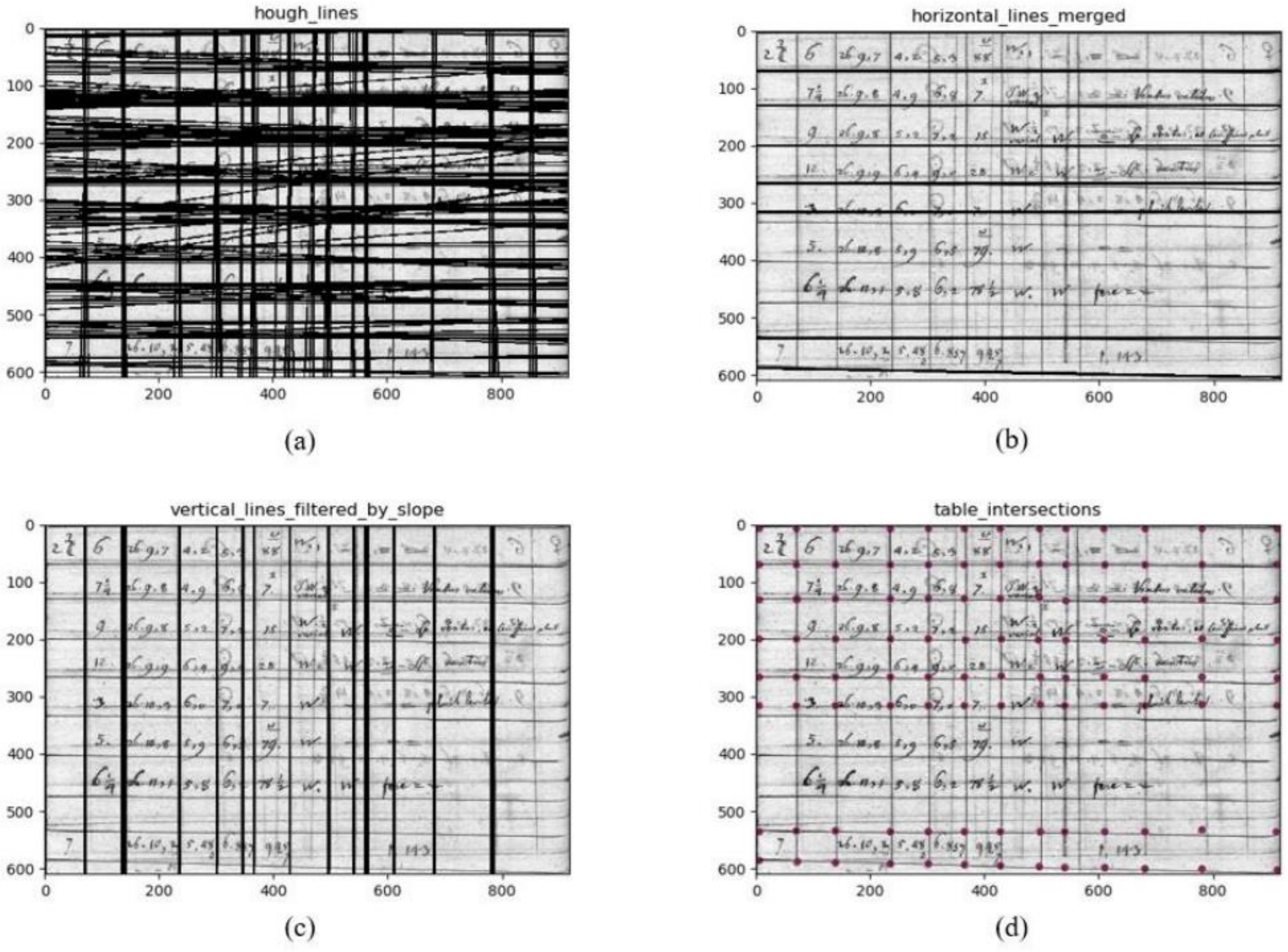
Layout analysis and table recognition



TEXT LINE

TABLE

TEXT LINE



(a)

(b)

(c)

(d)

Evaluation of the table recognition

Training set	40 pages
Evaluation set	38 pages
Table detection	100%
Table recognition	87%

Table detection results were evaluated by using the Jaccard index

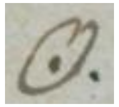
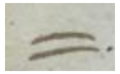
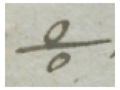
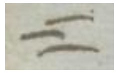
Table recognition results were evaluated by converting the structure to HTML and computing the BLEU metric

Text recognition

Model	C, Mp(2x2), C, Mp(2x2), LSTM(200)
Training set	32 pages (~1.800 text lines)
Evaluation set	4 pages
CER	25,86%

Recent challenges

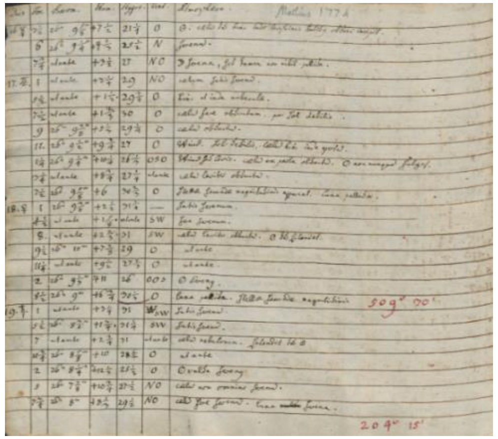
HISTORIC METEOROLOGICAL SYMBOLS

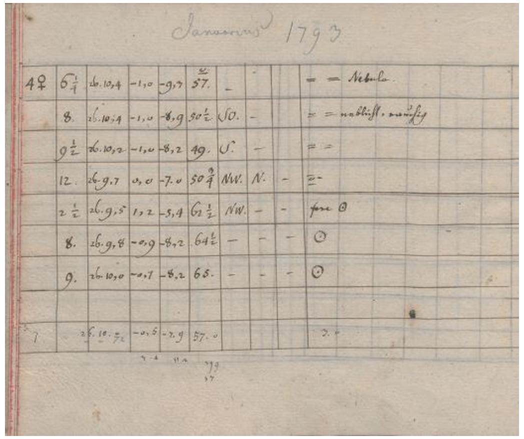
HISTORIC	SEMANTIC	UNICODE	VISUAL
	Coelum totum serenum et fine nube est.	⊙	⊙
	Nubes et coeruleum aeque diuisum.	≋	=
	Nubeculae paucae plerum que albicantes.	ϟ	°o
	Nubes et nubeculae minori coeli parte.	ϟ	ϟ

How to map certain symbols?

- not every symbol has a semantic or visuel Unicode pendant

BIG GT OR SMALLER SETS?

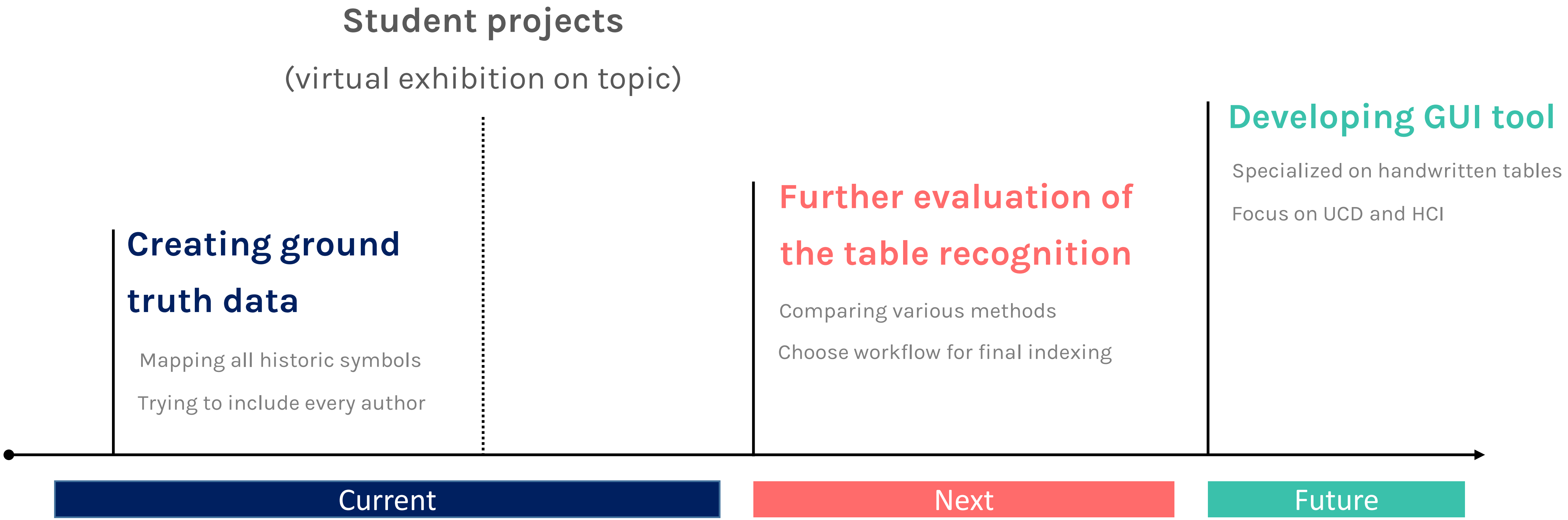




How to generate a sufficient GT data set for the entire collection?

- not every volume covers on year
- layout and writing style changes mainly with autor

Future directions



Student projects

(virtual exhibition on topic)

Creating ground truth data

Mapping all historic symbols

Trying to include every author

Further evaluation of the table recognition

Comparing various methods

Choose workflow for final indexing

Developing GUI tool

Specialized on handwritten tables

Focus on UCD and HCI

Current

Next

Future

Conclusion

OCR engines do not perform optimally out-of-the-box

Layout segmentation can not be done completely automatically

No real guidelines exist when creating ground truth data

Thank you for
your attention!

References

Glaser, R., Hagedorn, H.: Klimageschichte - Antworten auf die Verfinderlichkeit von Wetter, Witterung und Klima? In: Naturwissenschaften 81, pp. 97-107 (1994)

Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A., Puppe, F.: OCR4all – An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings, (2019)

Figures

Figure #	Creator
1	Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A., Puppe, F.: OCR4all – An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings, (2019)
2	Dhali, M., Wit, JW. & Schomaker, L.: BiNet: Degraded-Manuscript Binarization in Diverse Document Textures and Layouts using Deep Encoder-Decoder Networks, (2019)
	Binmakhashen, G. & Mhamoud, S.: Document Layout Analysis: A Comprehensive Survey. ACM Comput. Surv. 52(6), (2020)
3	https://de.wikipedia.org/wiki/Coelestin_II._Steiglehner#/media/Datei:Coelestin_II._Steiglehner.JPG https://de.wikipedia.org/wiki/Placidus_Heinrich#/media/Datei:Placidus_Heinrich.jpg https://rzbvm050.uni-regensburg.de/meteorologie/schmoeger.htm