Amir Moghaddass Esfehani
Campus Library

Freie Universität Berlin

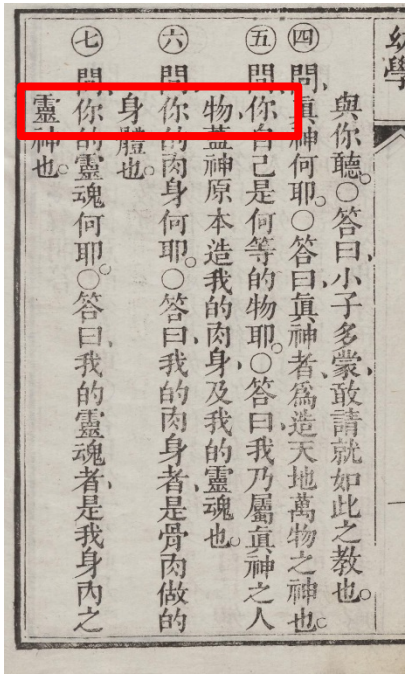
# No text – no mining.
# And what about dirty OCR?

ADHO DH2019 Workshop
"Towards Multilingualism In Digital Humanities: Achievements, Failures And Good Practices In DH Projects With Non-latin Scripts"

# No text – no mining.
# And what about dirty OCR?

- OCR
- Metadata
- Tools & APIs

# Dirty OCR: Layout & Text



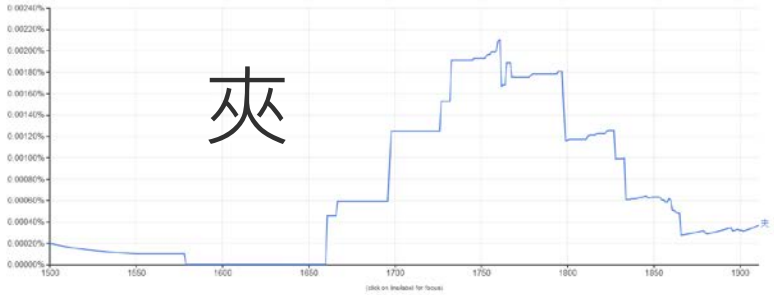

㊅
@
問
靈你'身栎物肩
" 1 ， ？ \^ , 41安5 ~
10 ，
與I

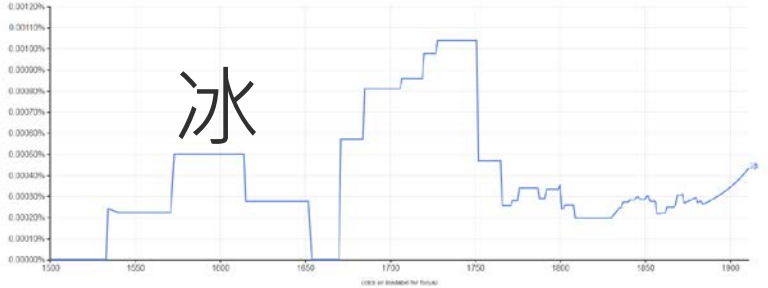
Precision = 0,448
Recall = 0,371
F-measure = 0,272
Error rate = 58,3%

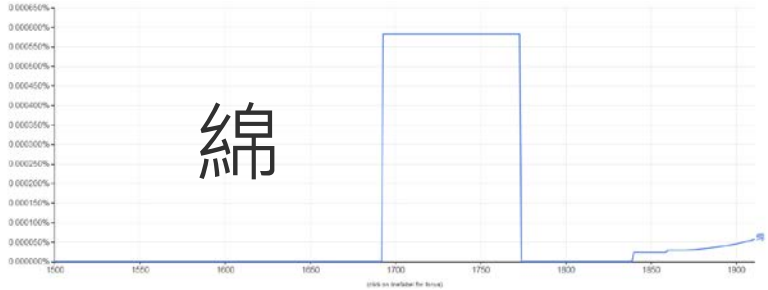
0
原
耶。
荷本狗' 0

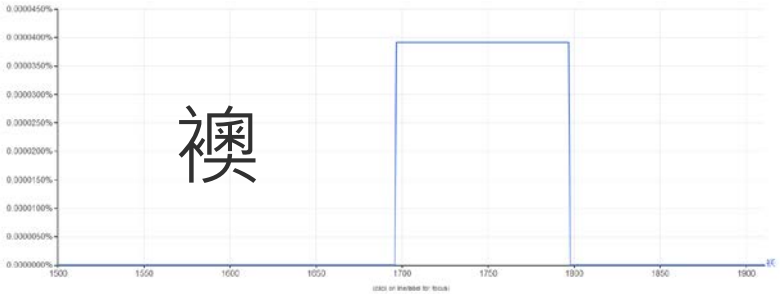
耶。造等答白，
0我的'曰，小
愁的物眞子
答爲
承
是
魂
也。
地如
~萬此
^ # ^
|+之教
之神也。
人

# „Discovery“ Of A Chinese Ice Age



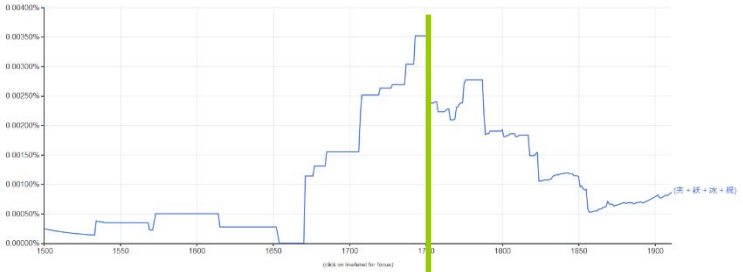



# „Discovery“ Of A Chinese Ice Age

來 + 永 + 秀 + 澳

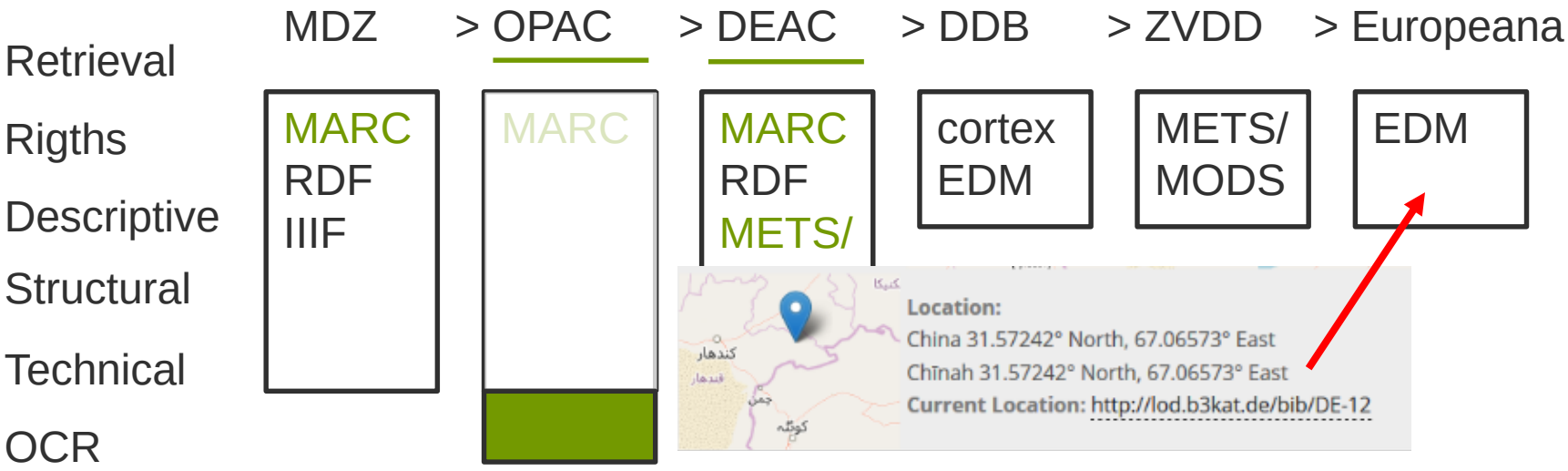
夾 + 冰 + 綿 + 襖




1750

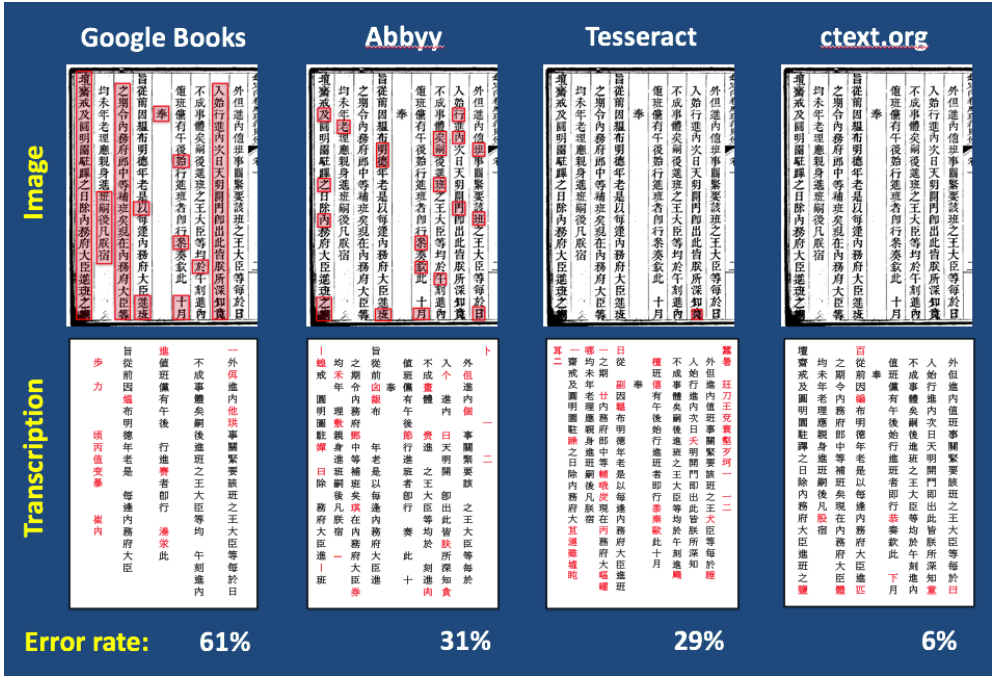
# Metadata

嚴如熤 Yan Ruyi: 苗防備覽: [22卷] Miao fang bei lan [22 juan]. 紹義堂, Daoguang 23 [China, 1843]. https://nbn-resolving.org/urn:nbn:de:bvb:12-bsb11123105-5




non-latin script

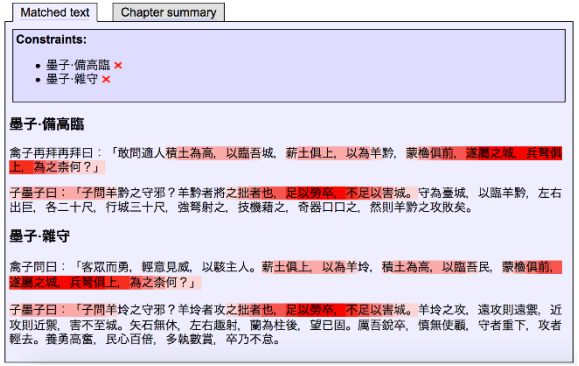
# Chinese Text Project: OCR

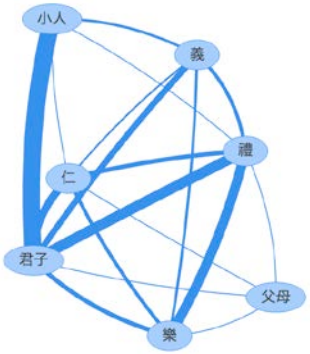


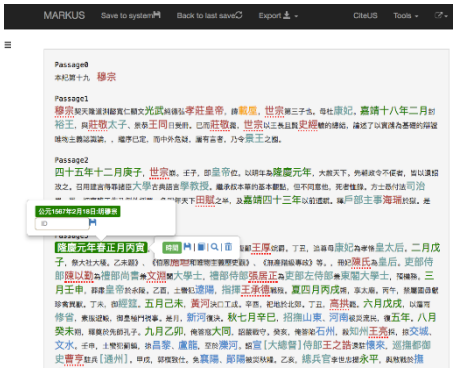

Sturgeon, Donald (2018): Large-scale Optical Character Recognition of Pre-modern Chinese Texts. *International Journal of Buddhist Thought and Culture* (2), p. 11-44.

https://digitalsinology.org/zh/wiki/File:Ctext-ocr.png

# Chinese Text Project / MARKUS: Tools & APIs




Text reuse

Regex

NER (MARKUS)

**Thank you!**

AMIR MOGHADDASS ESFEHANI
Campus Library
Freie Universität Berlin

amir.moghaddass@fu-berlin.de