Historical legislation: segmentation, datafication, and automatic metadating

Dr. Annemieke Romein

Annemieke.Romein@Huygens.knaw.nl

Datum: 17 Nov. 2020

Contents

intro | LA: segmentation | Datafication | Automatic metadating | Conclusion

Introduction

- Early modernist (political/ legal/ institutional) focus on political terminology (training in Rotterdam)
- Turned DH about 3 years ago... (although I did statistics)
  - Old-fashioned way of doing paleography… but now *the* go-to person on Transkribus (according to colleagues).
- Postdoc at UGent (2017-2020: Law and Order?);
- Postdoc at KNAW Huygens ING (2020-2023: A Game of Thrones?);
- Researcher-in-Residence at the KB National Library of the Netherlands (2019) to automatically metadate texts.

Project/ program: early modern ordinances

Paalhuis,
Amsterdam
(ca. 1660)

Appearances

PLACAET

De Staten van Hollandt ende West-Vries-landt. Allen den genen die desen sullen sien of hooren lesen/ Salutet, Doen te weten: Dat Wy om goede redenen Ons daer toe moverende/ goetgevonden hebben wel expresselijck te verbieden/ soo als Wy verbieden by desen/ dat niemandt van wat conditie of qualiteyt hy ook soude mogen wesen/ sal vermogen te drucken of doen drucken/ tzy voor hem selfs of voor andere/ eenigerhande Fransche Couranten of Gazettes/ tzy onder de naem van Couranten, Gazettes, Gazette-Raisonnée, Nouvelles-Choisies, Lardons, of wat andere Naem of Titul daer aen soude mogen worden ghegeven; op peene dat die gene/ die bevonden soude mogen worden eenige soodanige Fransche Couranten, Gazettes, Gazette-Raisonnée, Nouvelles-Choisies, Lardons, ofte wat andere Naem of Titul aen die oock soude mogen worden ghegeven/ te hebben gedruckt of doen drucken/ selfs of voor anderen/ soo wel den Maecher als den Drucker/ daer over arbitralijck naer exigentie van saecken sal worden gecorrigeert. Ende op dat niemandt hier van eenige ignorantie soude mogen pretenderen/ soo begeren Wy dat desen alomme gepubliceert ende geaffigeert sal worden daer het behoort ende te ghelschieden gewoonlijck is.

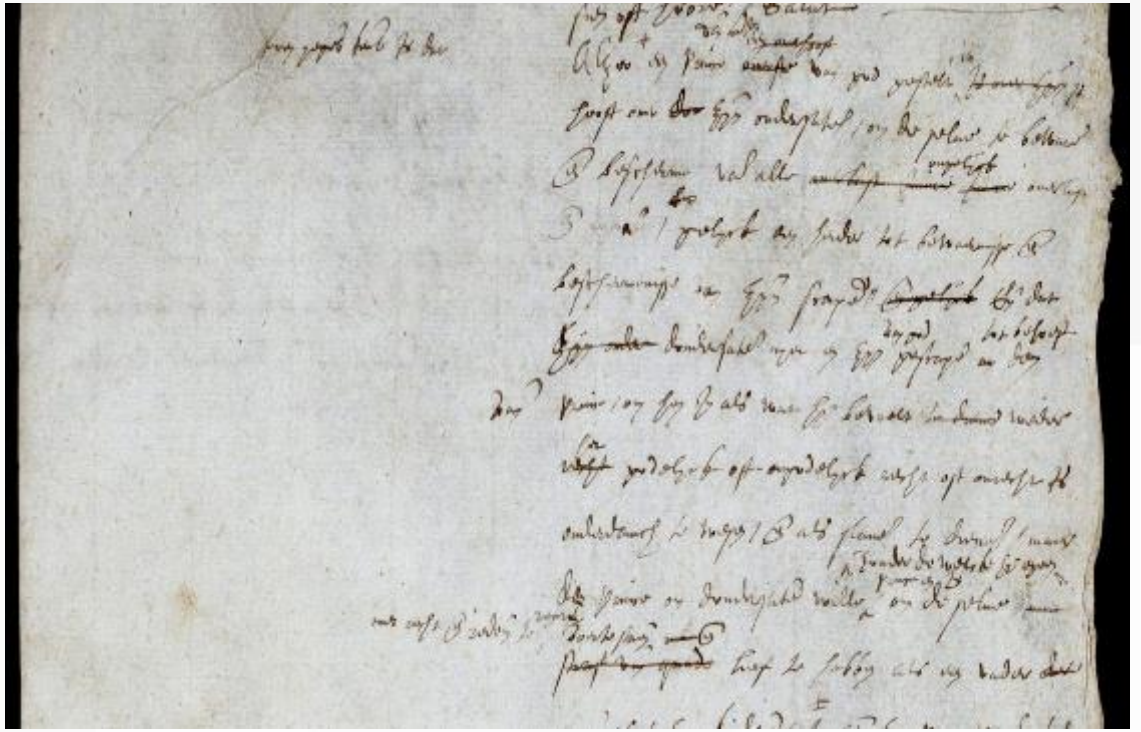
Gedaen in den Hage onder het Tegel van den Lande desen 11 February 1686.

Oorderstondt/

Ter Ordonnantie van de Staten.

Geteeckent/

Simon van Beaumont

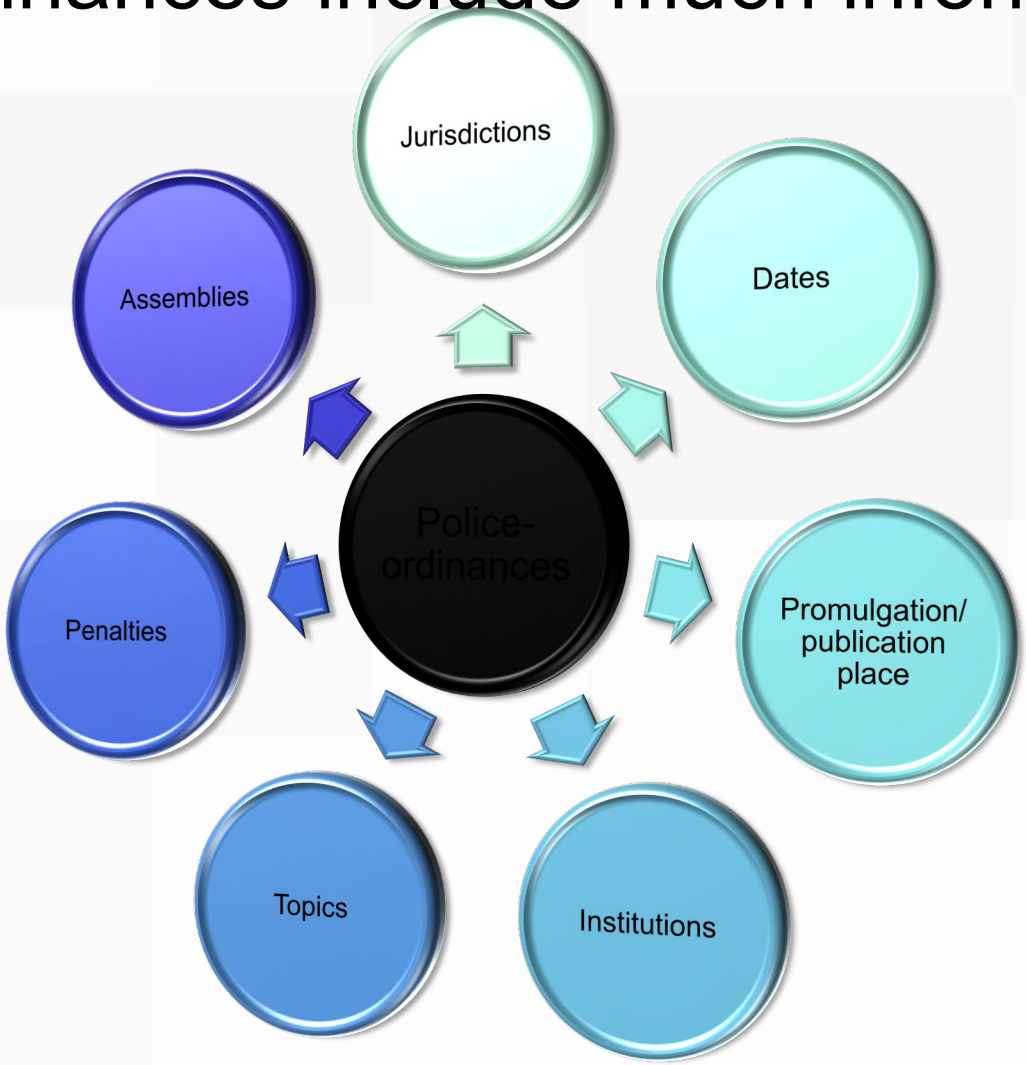


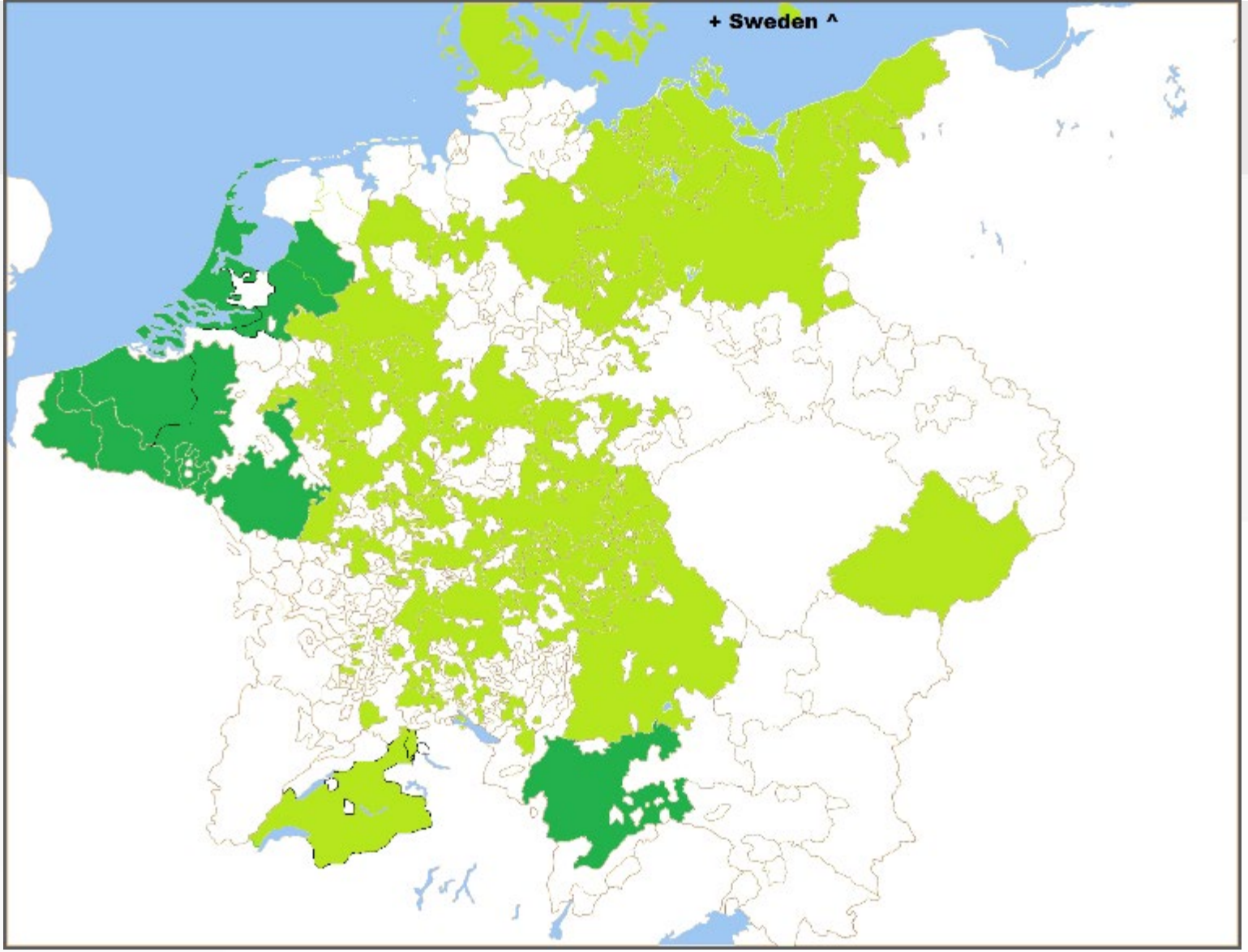
Left: https://library.maastrichtuniversity.nl/collections/databases/digitale-plakkaatzoeker/

Right: Plakkaat van Verlatinghe, pagina 1 (NA)

intro | LA: segmentation | Datafication | Automatic metadating | Conclusion

Police-ordinances include much information





MPIeR *Repertorium and other initiatives (dark green), period 1500-1800*.

The ‘smoking gun’ and the RQ

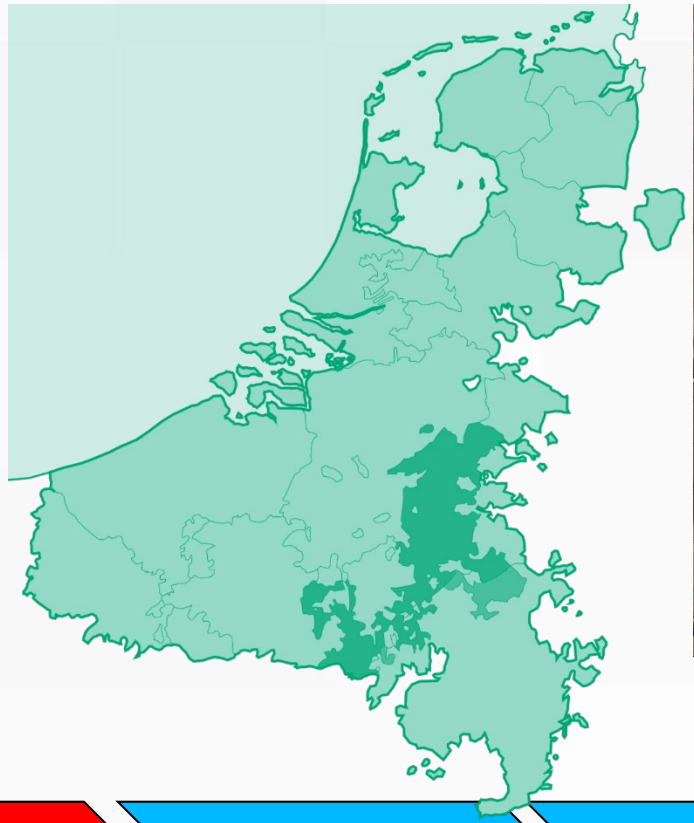
‘Early Modern European states struggled for survival, making it impossible to ‘reinvent the wheel’ each time a problem arose. Hence, it was of tremendous importance to copy, adapt and implement normative rules (aka legislation) that were already proven succesful elsewhere.’

But…

- Where to start?
- How to prove this?
- Data?

intro | LA: segmentation | Datafication | Automatic metadating | Conclusion

Narrowing down the focus



intro | LA: segmentation | Datafication | Automatic metadating | Conclusion

From going narrow to broadening up… DH

‘Entangled Histories’ as a DH-pilot study… to:

- Work on automatic segmentation
- Working on the readability of early modern print (Dutch Gothic/ Dutch Romantype/ French Romantype)
- Automatic metadating early modern laws

Team ‘Entangled Histories’

Core team (6 months)

… Annemieke Romein (PI)

… Sara Veldhoen (Scientific Programmer)

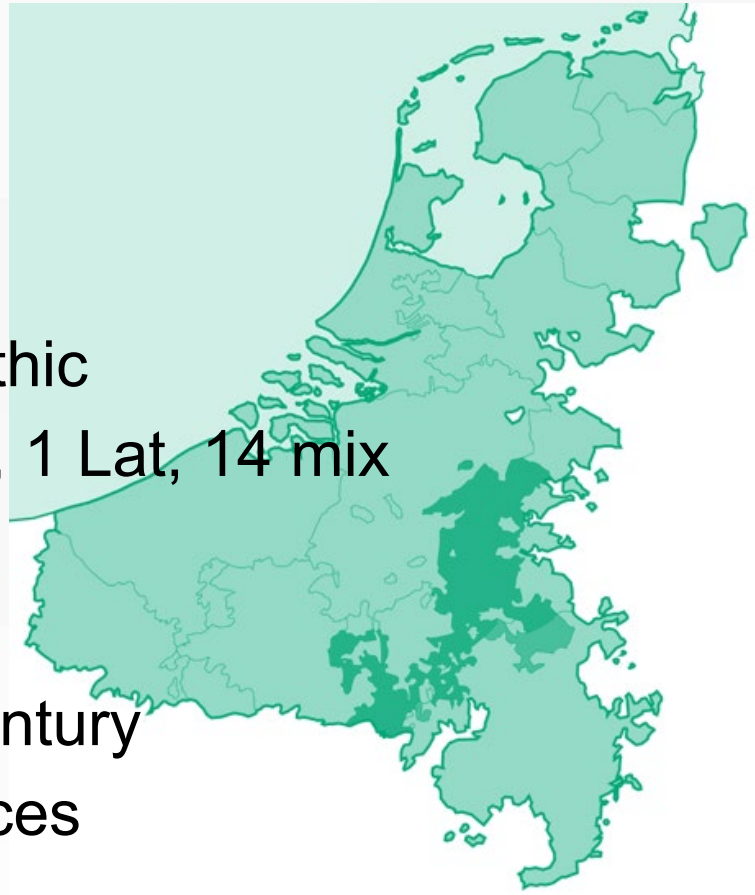
… Michel de Gruijter (Project Manager)

And help of many others!

Part 1. Segmentation of texts

Digitisation: Sources in numbers

Sources #	108
Font	88 roman, 20 gothic
Language	67 Dutch, 26 FR, 1 Lat, 14 mix
Pages#	75.000
Characters#	550 mln
Publication date	17th and 18th century
Region	Nearly all provinces



intro | LA: segmentation | Datafication | Automatic metadating | Conclusion

Teaching the computer to read and to segment texts

Transkribus

E Staten van StadtGroningen ende Ommelanden: DOEN TE WETEN. Alhoewel wy verhoopt hadden dat door het redres eñ correctie van de Lijſte van verpachtinge by ons in het laeſt verloopen Jaer 1622. gedaen ende alomme gepubliceert / eenighſins gheremedieert ende belet ſulden zijn geweeſt / de veelvuldige frauden ende dieverpen / Item liſtige ende ſchadelijcke practiquen / compoſitien eñ andere abuyſen / ſtreckende tot groten aff-breuck van de Generale Middelen / eñ merckelijcke præjuditie van alle goede getrouwe Ingeſetenen ende oprechte Liefhebbers van 't Vaderlant / dewelcke deur alſulcke publicque Dieven ende Pyanden van 't Gemeene beſte / niet alleene in haere neeringe becortet worden / maer oock de ſchade ende 't achterhepdt van 't Gemeene by den ſelven veroorſaeckt

Segmentation (the thing you are supposed to do first…)

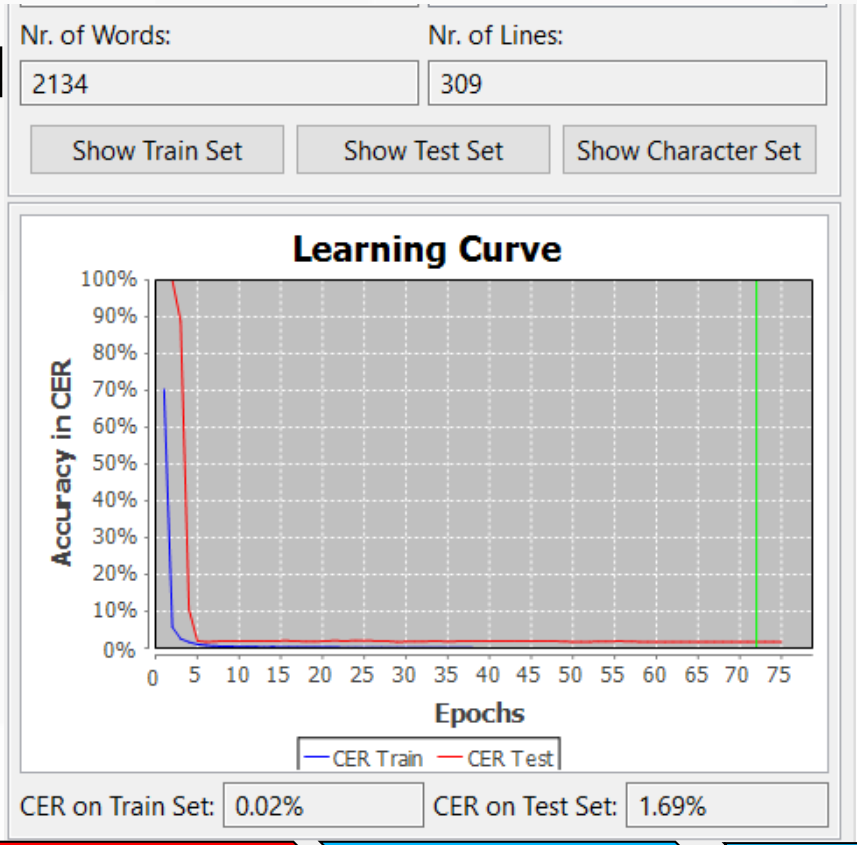
- Methods: Abbyy FineReader v.11 (✗/✓)
- P2PaLA (α-version) (✗/ ✓)
- NLE Document Understanding (AI-based) (✓)
- Rule-based approach (✓!)

Part 2. Datafication/ Text Recognition

intro | LA: segmentation | Datafication | Automatic metadating | Conclusion

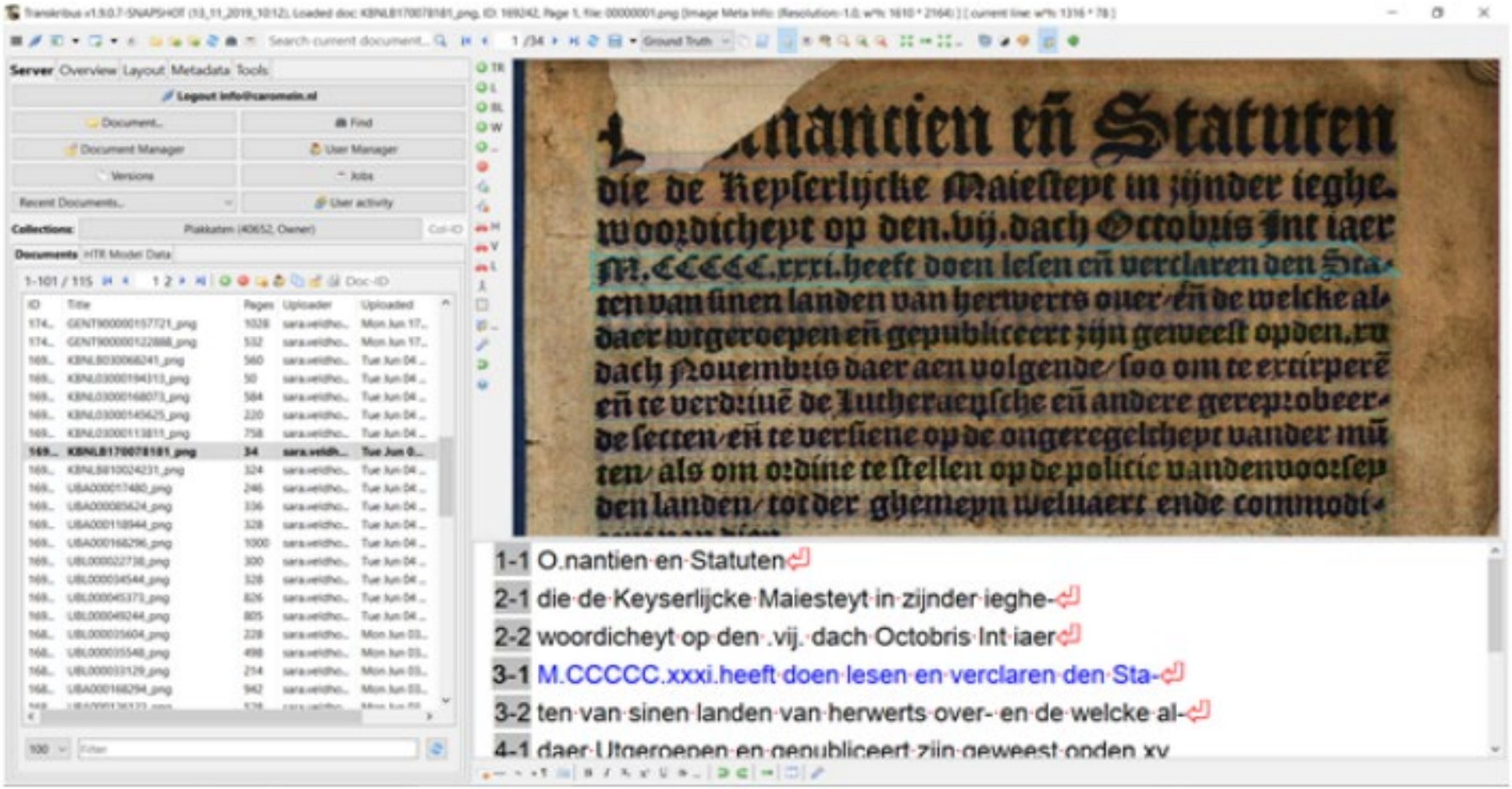
Improving the OCR-results throu


E Staten
van Stadt Gronin-
gen ende Ommelan-
den: DOEN TE
WETEN. Alhoewel
wy verhoopt hadden
dat door het redres eñ
correctie van de Lijste
van verpachtinge by
ons in het laest verlo-
pen Jaer 1622. gedaen
ende alomme republi-
ceert / eenighsins gheremedieert ende belet sulden zijn
geweest / de veelvuldige frauden ende dieverpen / Item
lickige ende schadelijcke practiquen / compositien eñ an-
dere abupsen / streckende tot groten aff-breuck van de
Generale Middelen / eñ merckelijcke præjuditie van alle
goede gerouwe Ingesetenen ende oprechte Liefhebbers
van 't Vaderlant / dewelcke deur alsulcke publicque Die-
ven ende Pyanden van 't Gemeene beste / niet alleene in
haere neeringe becortet worden / maer oock de schade
ende 't achterhepdt van 't Gemeene by den selven veroor-
saeckt



Transkribus… what's that?

https://www.youtube.com/watch?v=gS3GvvelN4o



intro | LA: segmentation | Datafication | Automatic metadating | Conclusion

Automatic Text Recognition

- Basically “Handwriting Text Recognition” but then applied to printed texts.
- It performed better than we had expected.

Model name [ID]	Training (CER)	Test (CER)	# Words (training)	# Lines (training)
Dutch_Gothic_Print [Model ID18944]	0.22%	1.71%	51143	7143
French_18thC_printed [Model ID19166]	0.33%	0.65%	38487	3883
Romantype_Dutch_Print [Model ID19423]	1.26%	1.17%	88105	13013

Bonus-output

- We created three HTR+-models within Transkribus that have been made public:
  - Dutch Gothic Print
  - French 18thC Print
  - Romantype Dutch Print

  (since October 2020, these are also available in PyLaia)

Part 3. Automatic metadating

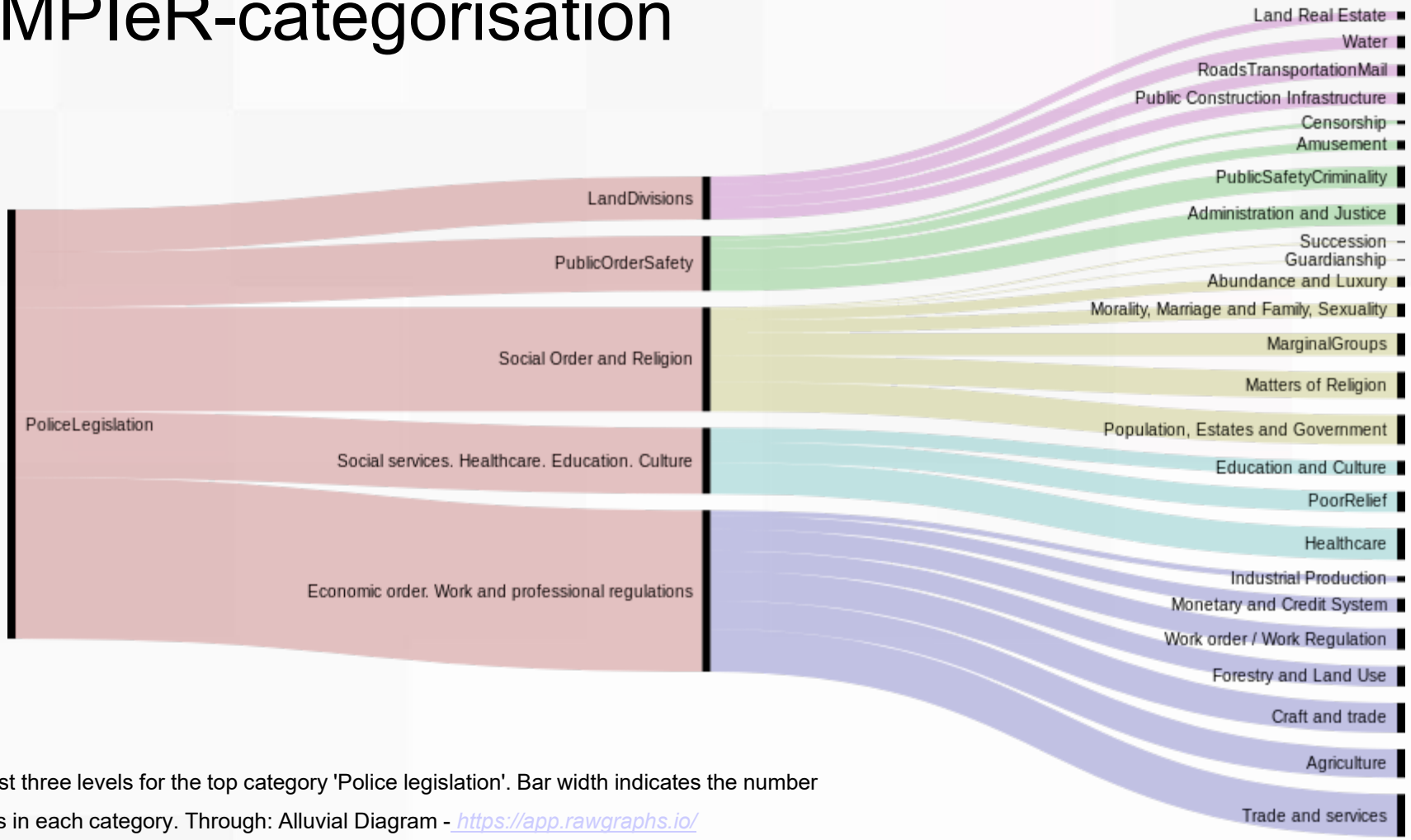
Method for automatic metadating

- Use of a controlled subject vocabulary
- This was developed within the projects ‘Repertorium der Policeyordnungen’ and ‘Gute Policey und Policeywissenschaft’.
  - Not perfect for all situations, but additions are possible

Method for automatic metadating

- 4 level deep hierarchical categorisation, but we needed to add a 5th level (police ordinance or not)
- So, at level 1: yes police ordinance or not
- Level 2: 5 options
- Level 3: 25 options
- Level 4: 163
- Level 5: 1584 options (currently)

MPIeR-categorisation



The first three levels for the top category 'Police legislation'. Bar width indicates the number of texts in each category. Through: Alluvial Diagram - https://app.rawgraphs.io/

SKOS: https://doi.org/10.5281/zenodo.3564586

Categorisation

Manually labelled ±3000 laws.

Used in the casestudy: 470.

Hierarchical categorisation – 4(+1) layers (MPIeR).

Converted to SKOS format (Short Knowledge Organisation System)

https://doi.org/10.5281/zenodo.3564586


annif

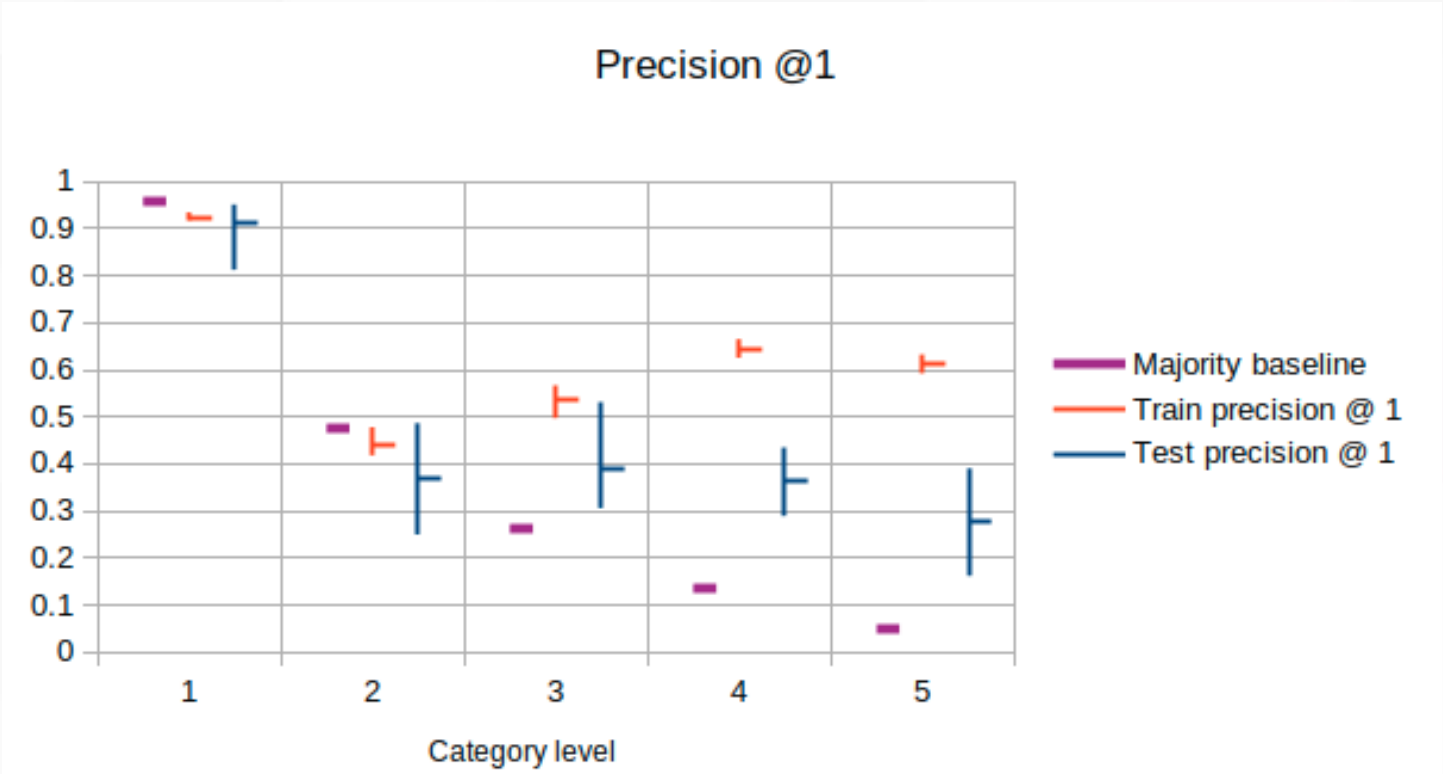
Training Annif

SET-UP

- Proof of concept: 470 labeled 'documents' (laws)
- <10 subjects each, 3.3 on average
- 69% and 28% of annotations at level 5 (deepest) and 4
- Training per category level
- 10-fold Cross Validation
- Backend: TF-IDF
- Limit = 4, threshold = 0.4

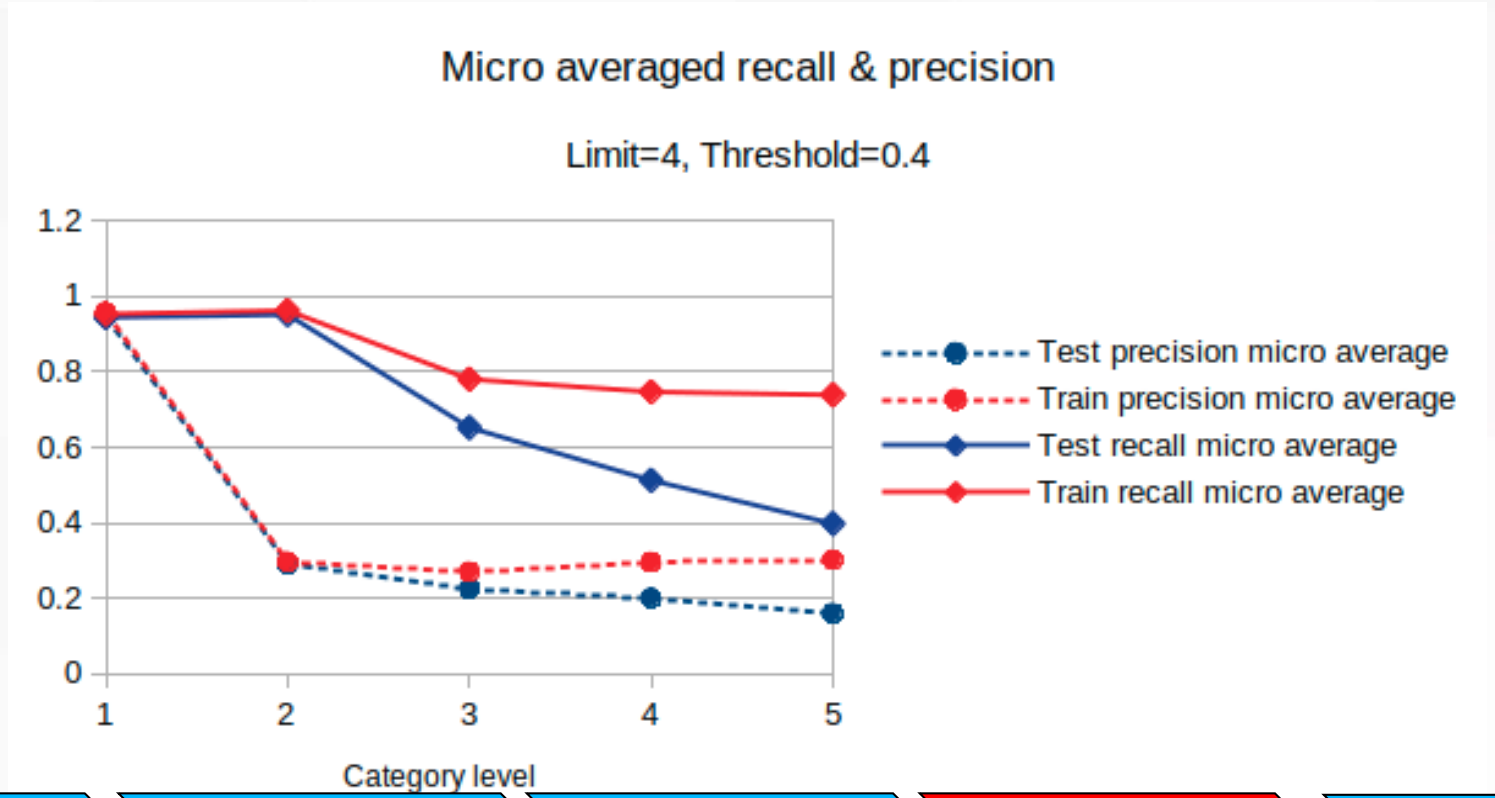
Training Annif

RESULTS

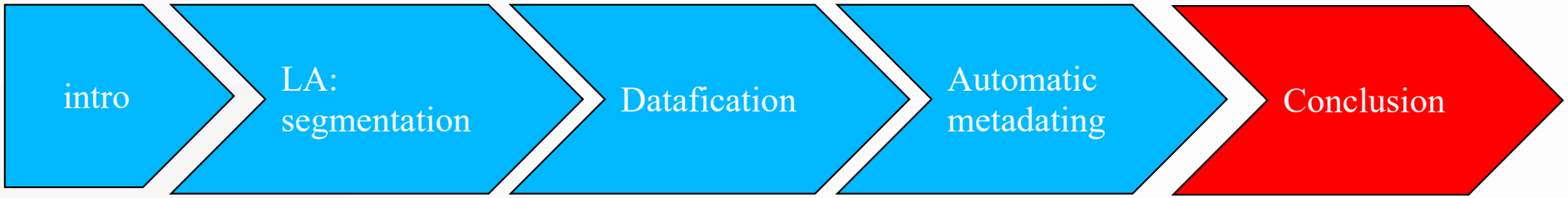


Training Annif

RESULTS



Conclusions



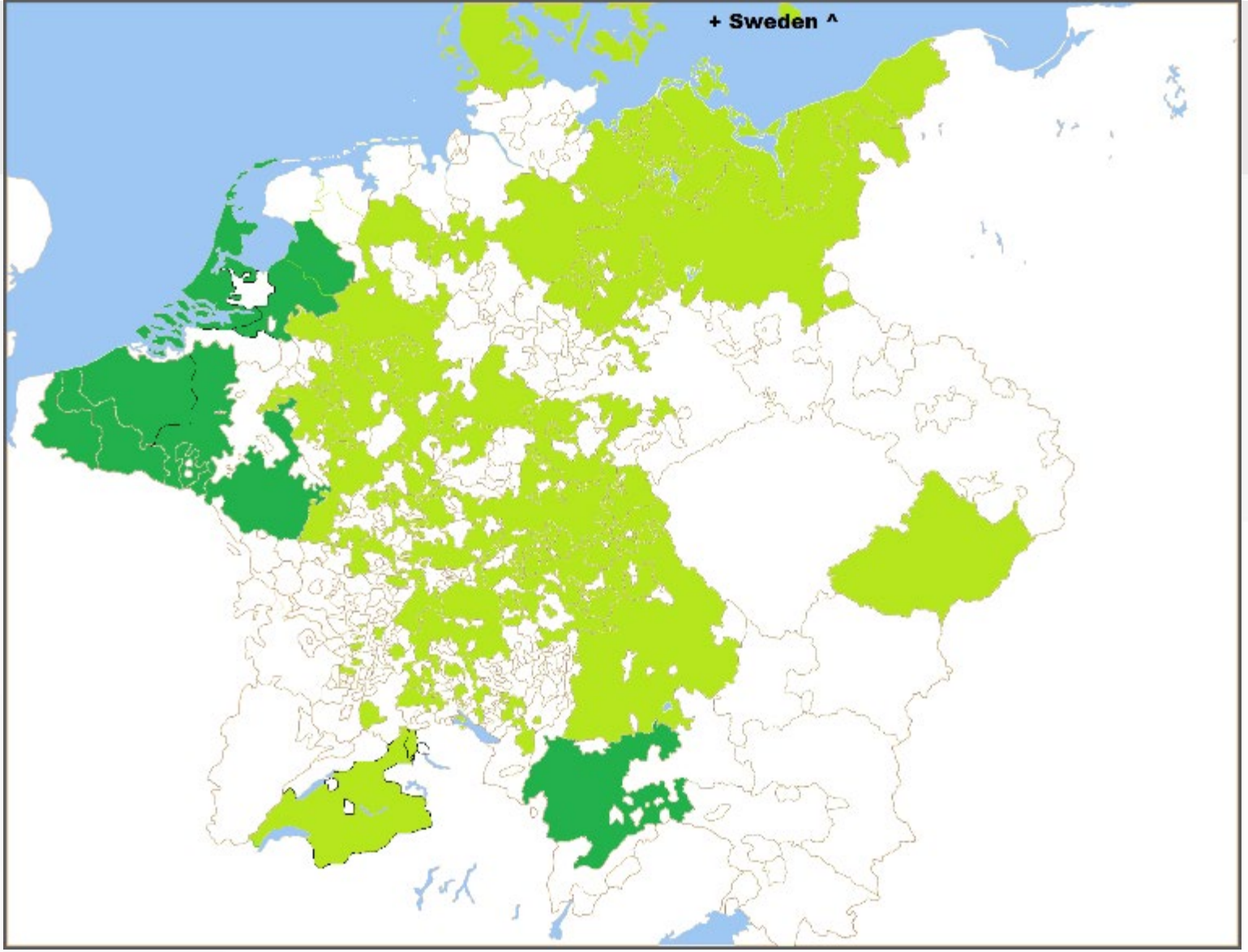
Conclusions

- Yes, the computer can automatically metadata the early modern legislation and it can be trained to become really good (but manual training will be necessary).
- No, we couldn’t automatically segment the texts just yet.
- No, we couldn’t prove the hypothesis to the fullest yet – application on all of the books was just too time consuming.

Training Annif

DISCUSSION

- Limited amount of data (470texts)
- Only 'dumb' TF-IDF – curious about other backends
- Impact of OCR quality?
- Stemmer: Snowball Dutch
  Impact of historical language & spelling variation?
- Making use of the hierarchy in categories



MPIeR *Repertorium and other initiatives (dark green), period 1500-1800*.

… future/ further research…

- In my current project I will try to test the use of Annif on German texts (Berne).
- I hope to apply for funding for future projects that will start connecting ‘the dots’ through Linked Data.
- If someone is interested in testing out Annif on early modern Swedish ordinances… let me know!

Further Reading

POINTERS

- DH Benelux journal publication: http://journal.dhbenelux.org/journal/issues/002/article-23-romein/article-23-romein.html
- Info on KB lab:

  Dataset: https://lab.kb.nl/dataset/entangled-histories-ordinances-low-countries

  Blogs: https://lab.kb.nl/about-us/blog/entangled-histories-bumps-road-and-bursts-success
- Other work with Annif at KB: https://zenodo.org/record/3899723#.X2B2NIbRZJd

FUTURE WORK

Further Reading

- Annemieke continues to work on early-modern legislation (https://en.huygens.knaw.nl/projecten/game-of-thrones/)

- At the KB, they're experimenting with Annif in a tool to aid the cataloguing department: 'Demosaurus'.
  See https://zenodo.org/record/3899723#.X2B6Q4bRZJc

Thank you for your attention.