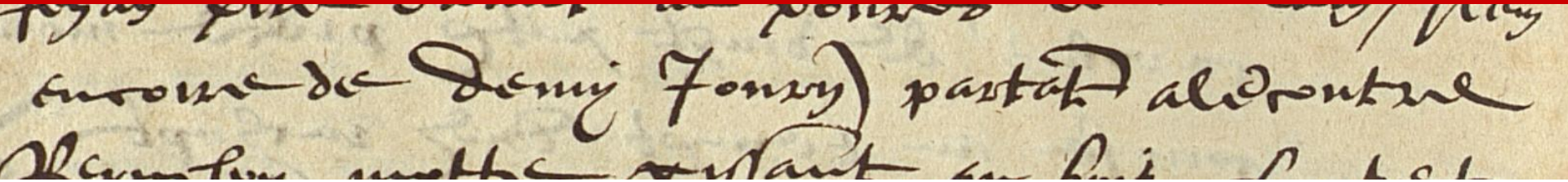
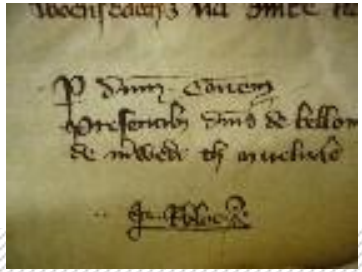
university of groningen

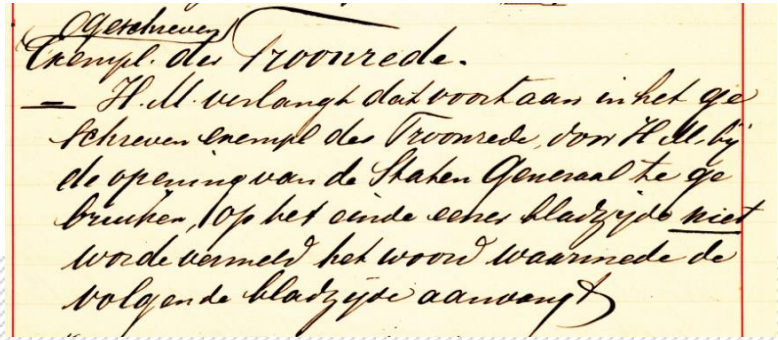
1614

The birth of a massive search engine for historical and multi-cultural handwritten collections

prof. dr. Lambert Schomaker





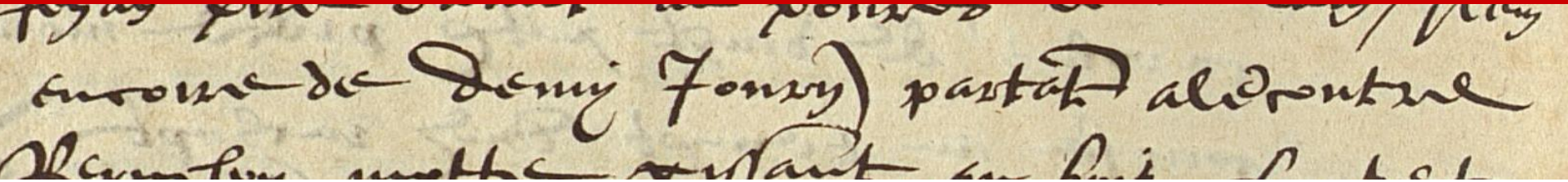


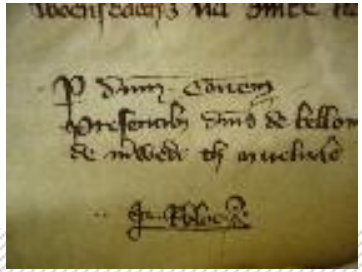
university of groningen

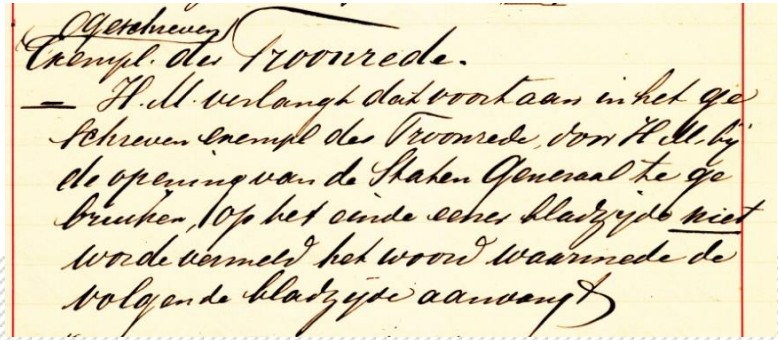
1614

Continuous machine learning in Big Data

prof. dr. Lambert Schomaker





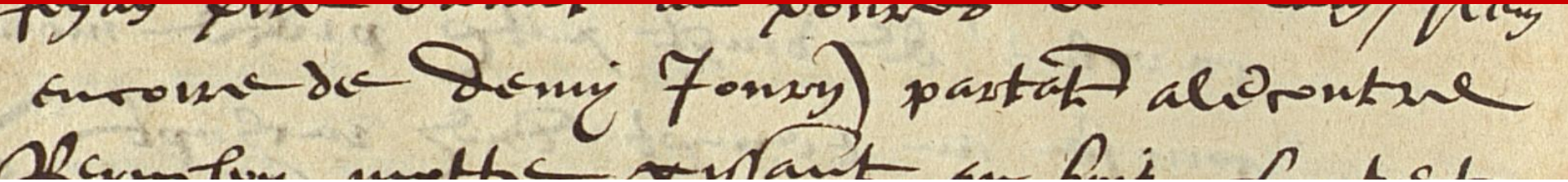


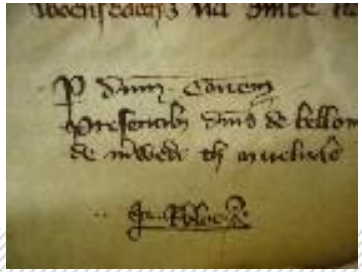
university of groningen

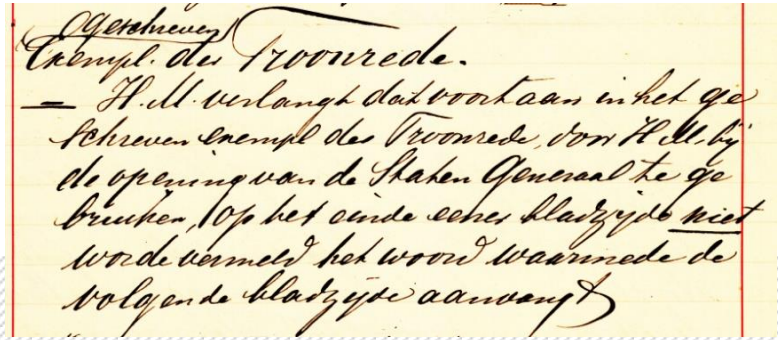
1614

The Monk System

prof. dr. Lambert Schomaker

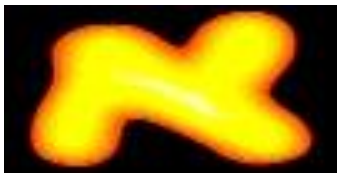






Monk e-Science web service addressing these **questions**:

- What? Word retrieval by 24/7 machine learning
  - Ambition: A European Google for handwriting
  - Server storage 1PB, filled 200 TB, with 37 collections
  - Qumran, Medieval, 17th century, 20th century, etc.
  - 400 M word images, 500 k labeled words, 30 k classes
- When? Medieval manuscript dating
  - On a server (MPS), uploading of charters from 1300-1550
- Who? Writer identification
  - On Monk server over internet
  - Using GIWIS Windows tool



Monk

Amsterdam-

Groningen

Strafgefangenis

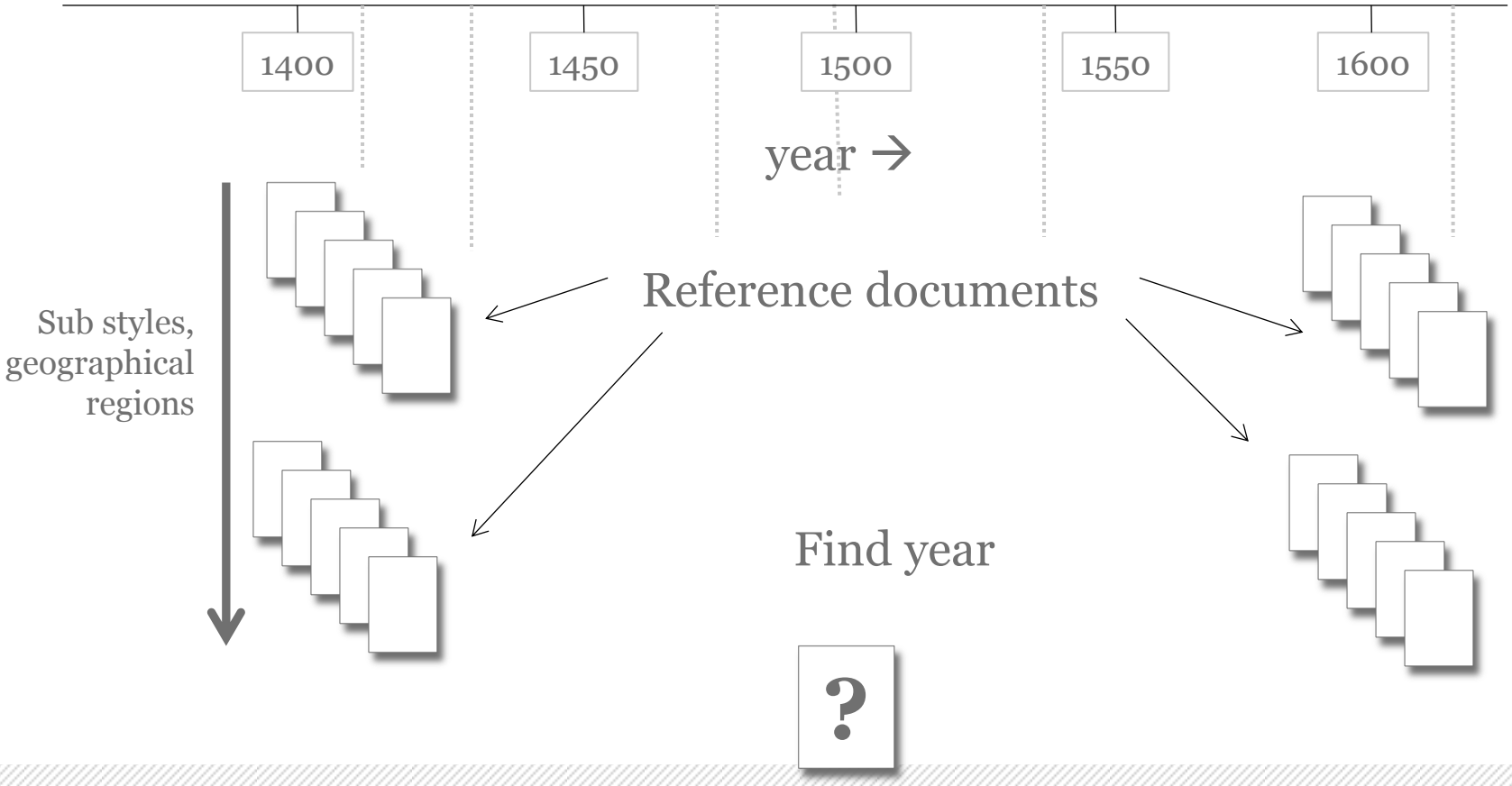
Colors

Coop Furniture

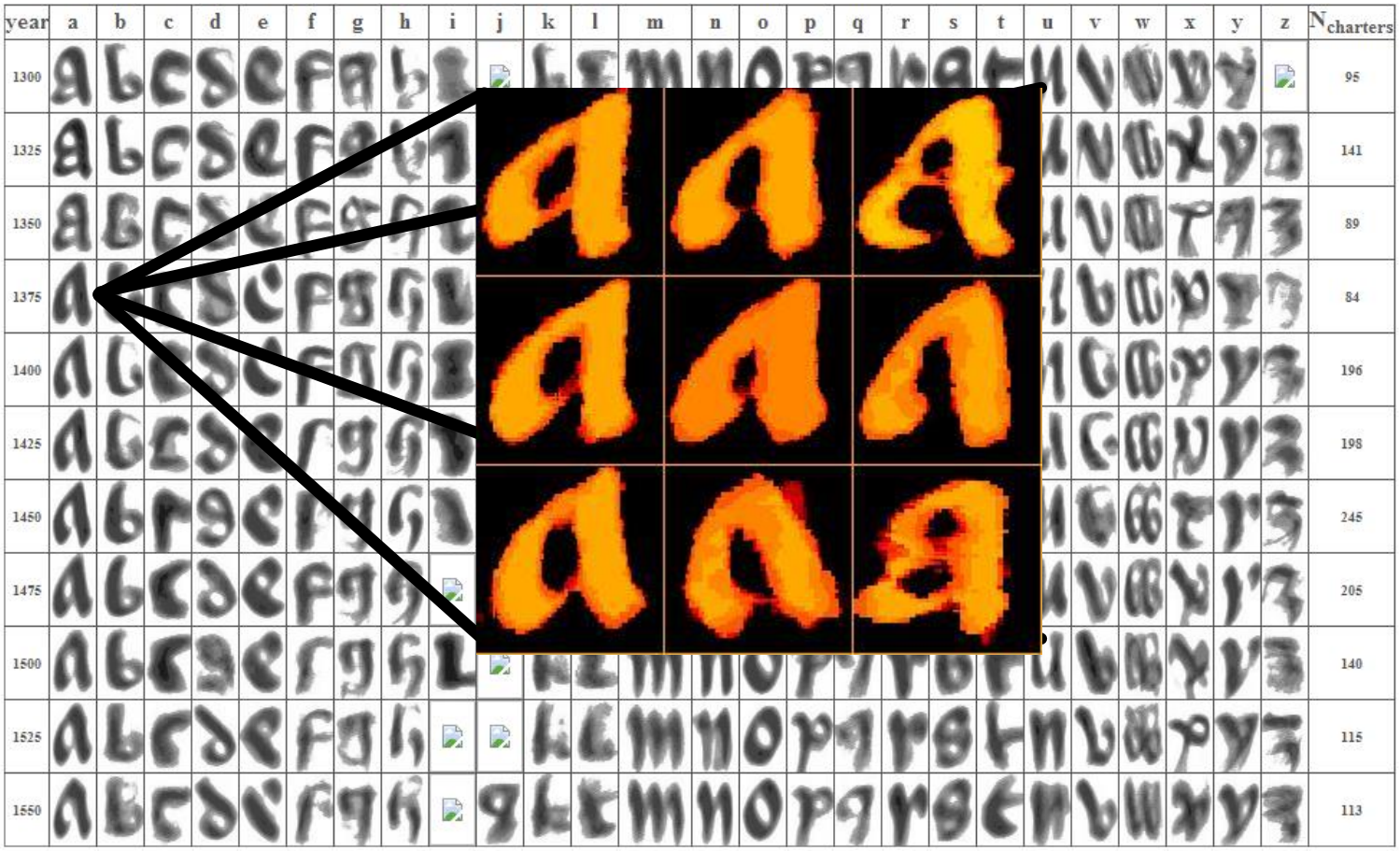
What?

When?

Medieval Paleographical Scale (MPS)
with prof. Jan Burgers (paleographer, Huygens Inst.)



Medieval Paleographic Scale (under dev.)



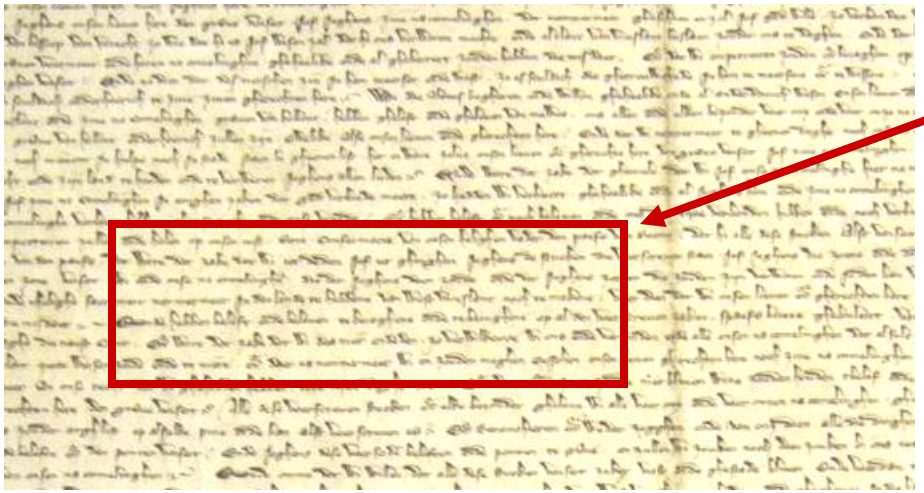
Writer identification (1:N) and verification (1:1)

1. Interactive, fully manual

2. Automatic,

3. OCR based

ROI based



Who?

Region Of Interest

Researchers

- Marius Bulacu, Axel Brink, Katrin Franke
- Ralph Niels, Louis Vuurpijl, Jean-Paul van Oosten, Sheng He, Jan Burgers, Petros Samara, Olarik Surinta
- "The Nijmegen Handwriting Group 1984-1993"
- Netherlands Forensic Institute: Ton Broeders, Wil Fagel, Elisa van den Heuvel
- *Isabelle Guyon, Rejean Plamondon*
- **Users (humanities researchers)**:
  - Jinna Smit, Mark Aussems, Masahiro Niitsuma, Mladen Popovic, Daniel Stoekl, Jetze Touber, Grace Fong, Elaine Treharne, …

Overview

- Digitisation of handwritten manuscripts
- Image processing challenges
- Text retrieval
  - Monk system: continuous learning
- Difficult material & the role of Knowledge
- Conclusions

Digitisation

- Int. Conf. on Document Analysis & Recognition 2009, Barcelona
  - Representative of British Library: we have computed that we would need three petabytes (three thousand disks of a terabyte, i.e., 3 PB = 3 million gigabytes)
  - … therefore we are obliged to use 'lossy' compression techniques for storing images …
- ? ?
- We are using **10 PB** in Groningen just for some astronomers, manuscript archives and biomedical researchers !

Digitisation

- Current disk capacities allow for a full exploitation of al digital renderings of an object:
  - Gray scale
  - Color
  - Multi-spectral
  - 3D scanner (for Swedish haellristningar, Roman and Greek stone inscriptions, identification of the ‘cisel’)
- 1 PB for Monk (‘in the noise’, for astronomers)

Digitisation

- Digitisation prevents wear and tear of the books
- … keep the books and documents in storage
- Quality considerations
  - Color mode: RGB/gray/black-and-white
  - Dots per inch? Bits per pixel?
  - Mechanical safety of scanning system
- Cost? (scanning, maintenance in digital life of object)
- Speed?
- Usage: **How to make the digital access effective?**

Case: Optical Character Recognition

- "The process of segmenting a text image into individual c h a r a c t e r images and classifying each as being a letter in an alphabet"
- Impressive results on contemporary printed text in machine fonts: with some linguistic postprocessing results are close to 100%

- Solved?

OCR ?

- ‘OCR’ on historic documents does not work well
- On *handwritten* manuscripts it doesn’t work, at all!
- Problems:
  - image quality
  - unknown character shapes
  - unknown statistical language models
- However: pattern recognition and machine learning make enormous progress these days!
- Which methods? How to apply them?

Current technology: neat text/known language

- Why is 'OCR', i.e., letter by letter transcription on handwriting so difficult?
- Machine print: per character, per font, 8000 training examples are needed, typically
- E. Barney-Smith: 200k instances of printed **c** vs **e**
- Adress reading: reduced lexicon, zip codes etc., **help**
- In linguistic modeling: 20th century newspaper corpora do very little for 15th century acts
- Literary text, acts and charters each need their own knowledge models in order for OCR to work

Handwriting recognition: eat this!

- Many languages, scripts
- Over historical periods
- Contractions of letters
- 'Suggested' sloppy letter shapes
- Individual writer styles
- Image problems

➔ Sliding window for character search usually problematic:

exit 'OCR'

Monk - Design considerations

1. Don't promise perfection
2. Don't promise 'transcription'
3. Don't promise exhaustive coverage (as in databases)
4. Make use of human trainers, volunteers

› Word spotting:

- "a Google for handwritten documents"

Monk - Design considerations

- Word spotting:
  - "a Google for handwritten documents"
- The word is a reliable chunk of information with many shape features: redundancy

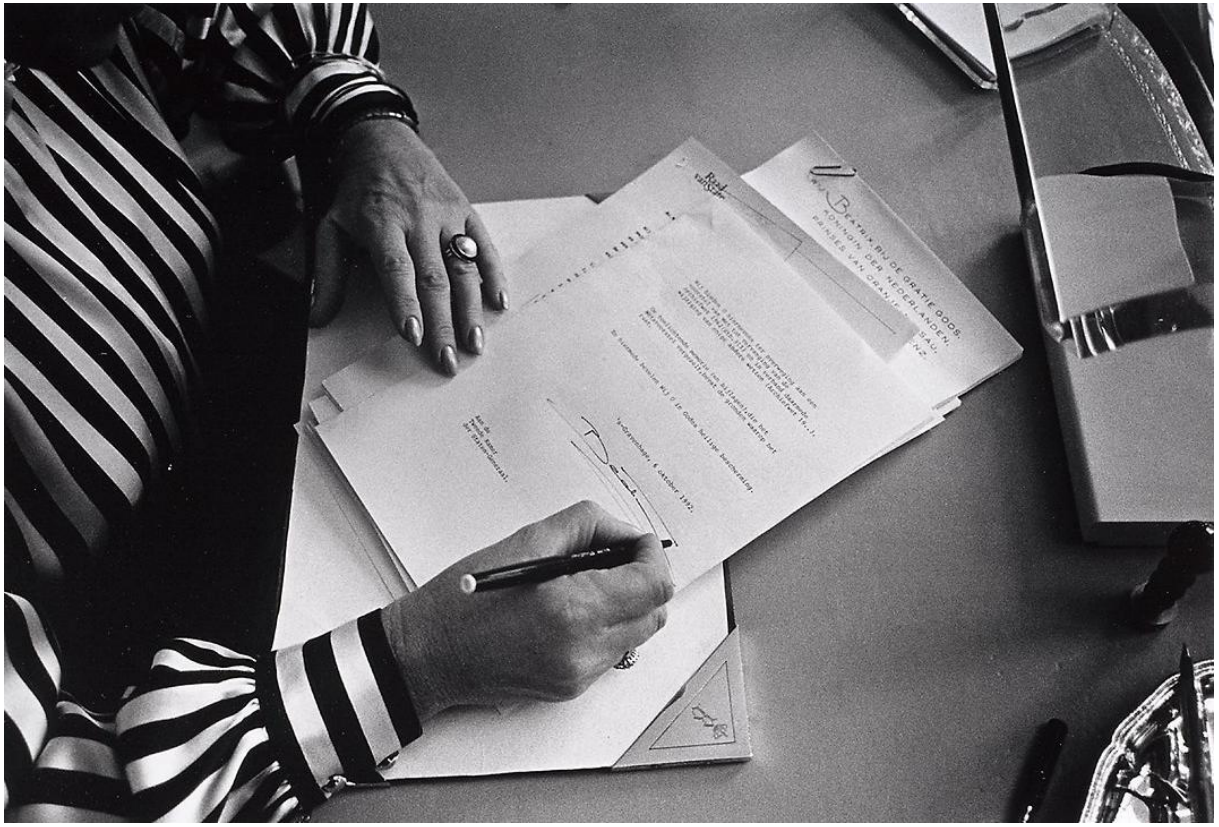
minimum

- Big Data: With sufficient data, there is always a reasonable response on a query

Monk's world model:

- Institutes
  - Collections
    - Books (i.e., documents)
      - Pages
        - Paragraphs
          - Lines
            - Word zones and characters
              - Pixels

Example collection: Cabinet of the Queen (KdK)



› the Queen is the head of state

› the head of state is a member of the government and signs all laws and many government decisions

KdK Archival system

› all laws, decrees and correspondence are kept:

- in a chronological order
- and then in a numerical order

the Queen's Cabinet

- KdK Archive
  with a total extent of

  (era 1798-1988):

  - 3,250 linear meter of shelves

- consisting of:

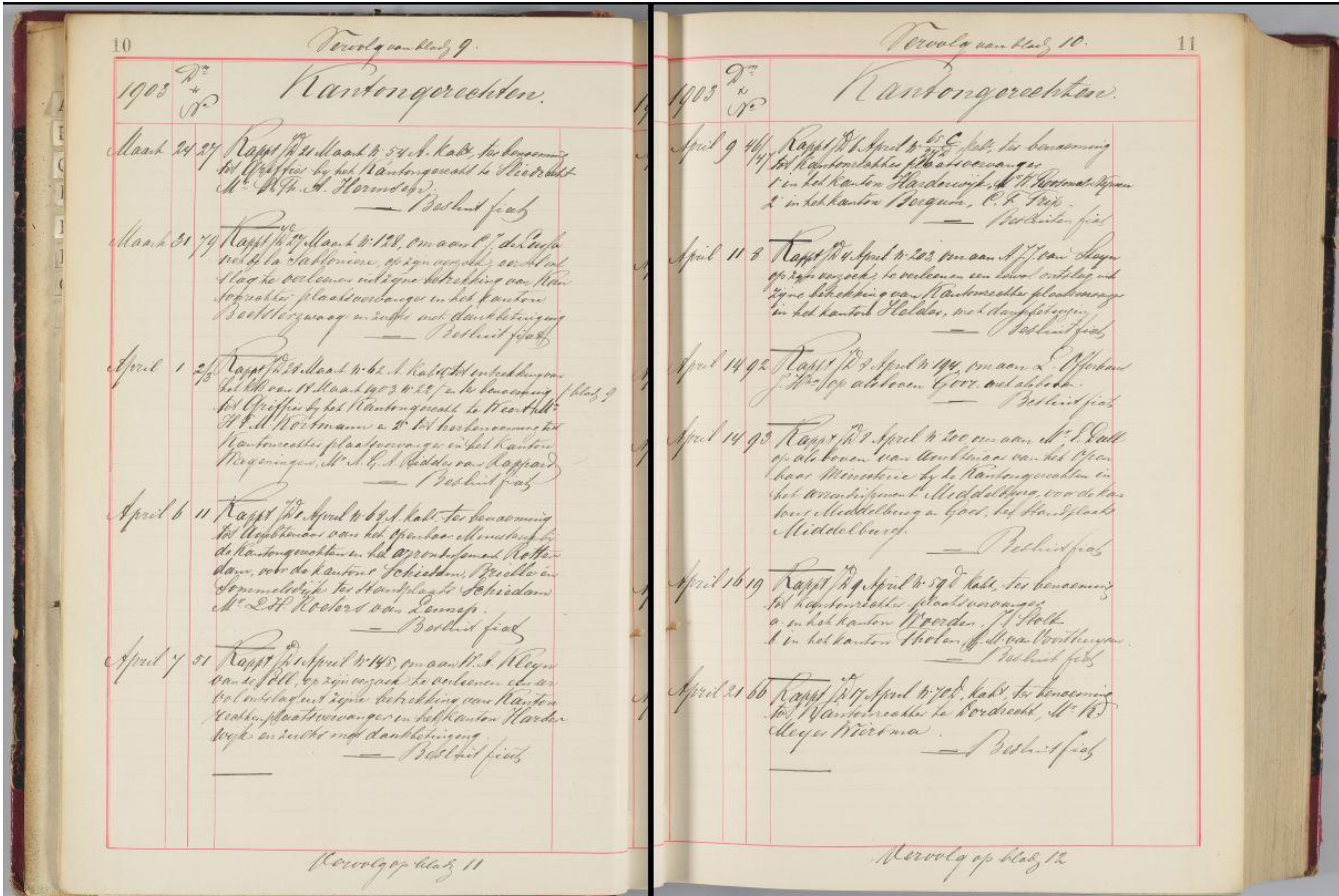
  - 28,000 boxes
  - average 1,000 pages per box
  - → 28,000,000 pages
  - Of which Monk analyses the handwritten *Indices*, > 60k pages

Functions of Monk from user perspective

- Pages (read, annotate)
- Lines (read, annotate, search)
- Words (cut ROI, labeling, inspect hit lists)

Human	Machine
Cut out region of interest with the mouse	Segmentation of image into objects
Label a ROI with a text label: `region`	Classify image object

- General search functions, 'google' over several collections



Vervolg van blad 9.

1903	Dn en No		Kantongerechten.
Maart	24	27	Rappt Jd 21 Maart No 54 A. Kabt, ter benoeming tot Griffier by het Kantongerecht te Sliedrecht Mr W. Fr. A. Hoorndert — Besluit fiat
Maart	31	79	Rappt Jd 27 Maart No 128, om aan C. J. de Luyfa ver of la Sablonière, op zijn verzoek eervol ontslag te verleenen uit zijne betrekking van Kantonrechter plaatsvervanger in het Kanton Beetsterzwaag en zulks met dankbetuiging — Besluit fiat
April	1	2/3	Rappt Jd 28 Maart No 62 A. Kabt tot intrekking van het Kb van 18 Maart 1903 No 22 / en te benoeming tot Griffier by het Kantongerecht te Weert Mr H. J. M. Wortmann en 2e tot herbenoeming tot Kantonrechter plaatsvervanger in het Kanton Wageningen, Mr A. G. A. Ridder van Rappard — Besluit fiat
April	6	11	Rappt Jd 2 April No 64 A. Kabt. ter benoeming tot Ambtenaar van het Openbaar Ministerie bij de Kantongerechten in het arrondissement Rotterdam, voor de kantons Schiedam, Brielle en Sommelsdijk ter standplaats Schiedam Mr L. H. Kooters van Lennep. — Besluit fiat
April	7	51	Rappt Jd 2 April No 145, om aan W. A. Kleyn van de Poll, op zijn verzoek te verleenen eervol ontslag uit zijne betrekking van Kantonrechter plaatsvervanger in het Kanton Harderwijk en zulks met dankbetuiging — Besluit fiat

blad 9

Vervolg op blad 11

Vervolg van blad 10.

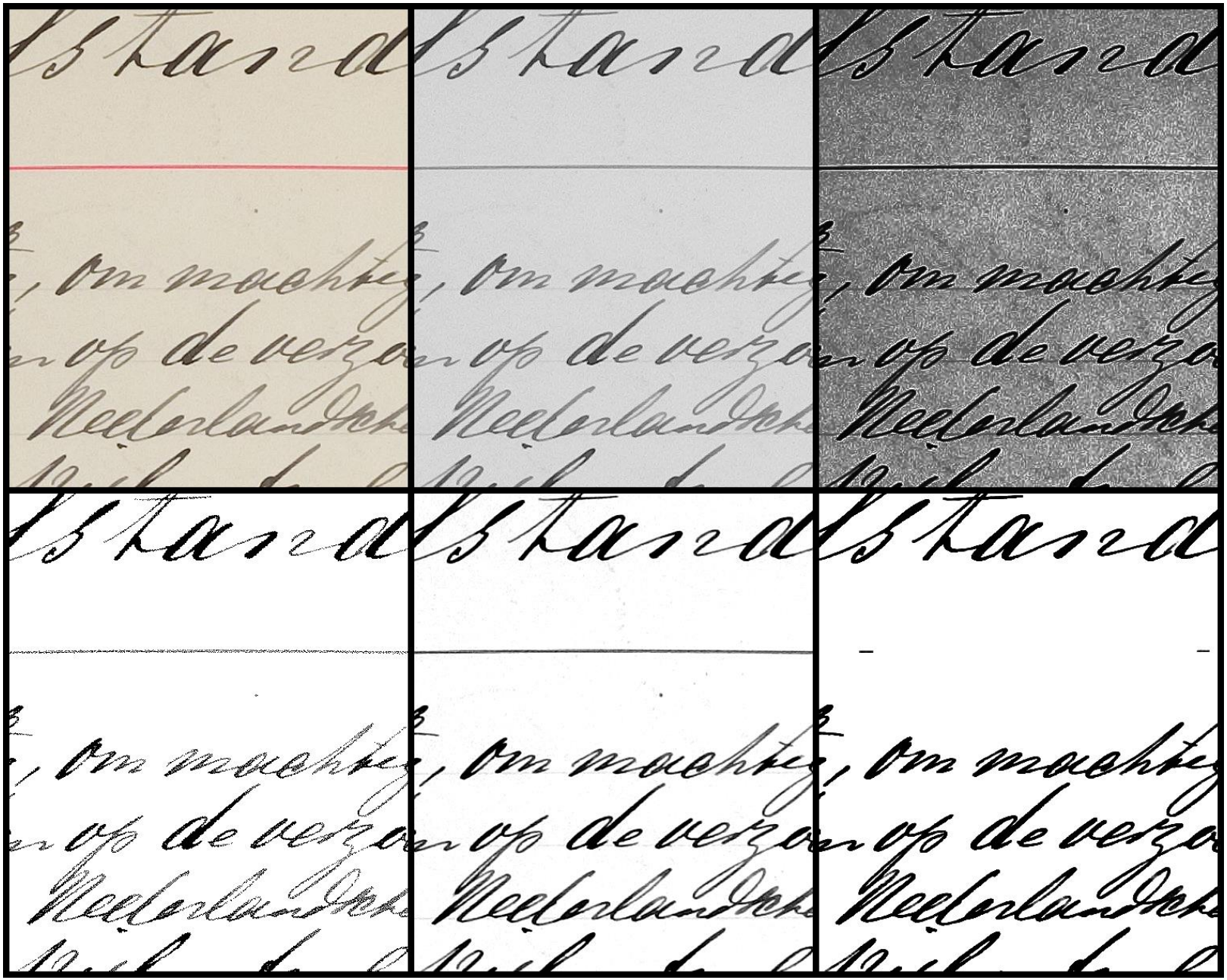
1903	Dn en No		Kantongerechten.
April	9	46/147	Rappt Jd 6 April No 65 C / 272 ... ter benoeming tot Kantonrechter plaatsvervanger 1 in het kanton Harderwijk, Mr W. Roosmale Nepveu 2 in het kanton Bergum, C. F. Trip — Besluiten fiat
April	11	8	Rappt Jd 4 April No 203 om aan A. J. J. van Slooten op zijn verzoek, te verleenen een eervol ontslag uit zyne betrekking van Kantonrechter plaatsvervanger in het kanton Helder, met dankbetuiging — Besluit fiat
April	14	92	Rappt Jd 8 April No 194, om aan L. Offerhaus J. Wzn op als boven Corr. met als boven. — Besluit fiat
April	14	93	Rappt Jd 8 April No 200 om aan Mr E. Dall op als boven van Ambtenaar van het Openbaar Ministerie by de Kantongerechten in het arrondissement Middelburg, voor de kantons Middelburg en Goes, ter Standplaats Middelburg. — Besluit fiat
April	16	19	Rappt Jd 9 April No 59 d Kabt, ter benoeming tot Kantonrechter plaatsvervanger a. in het kanton Woerden, J. J. Stolk b in het kanton Tholen, J. M. van Voorthuysen — Besluit fiat
April	21	66	Rappt Jd 17 April No 70 d. Kabt, ter benoeming tot Kantonrechter te Dordrecht, Mr R. Meyer Wiertsma. — Besluit fiat

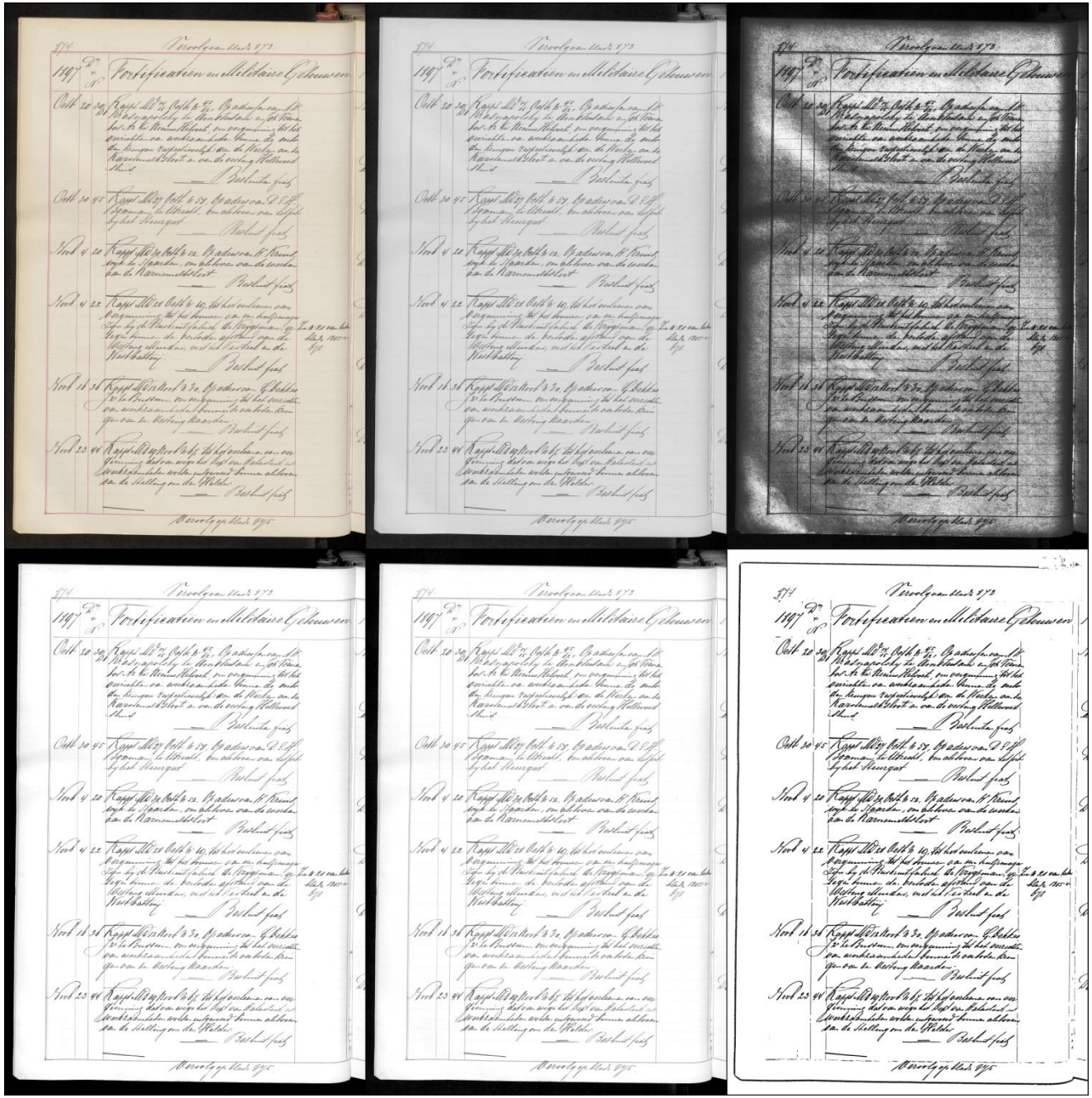
Vervolg op blad 12

KdK 1903

The Monk system / Lambert Schomaker

The Monk system / Lambert Schomaker





KdK 1897 Scan Company B

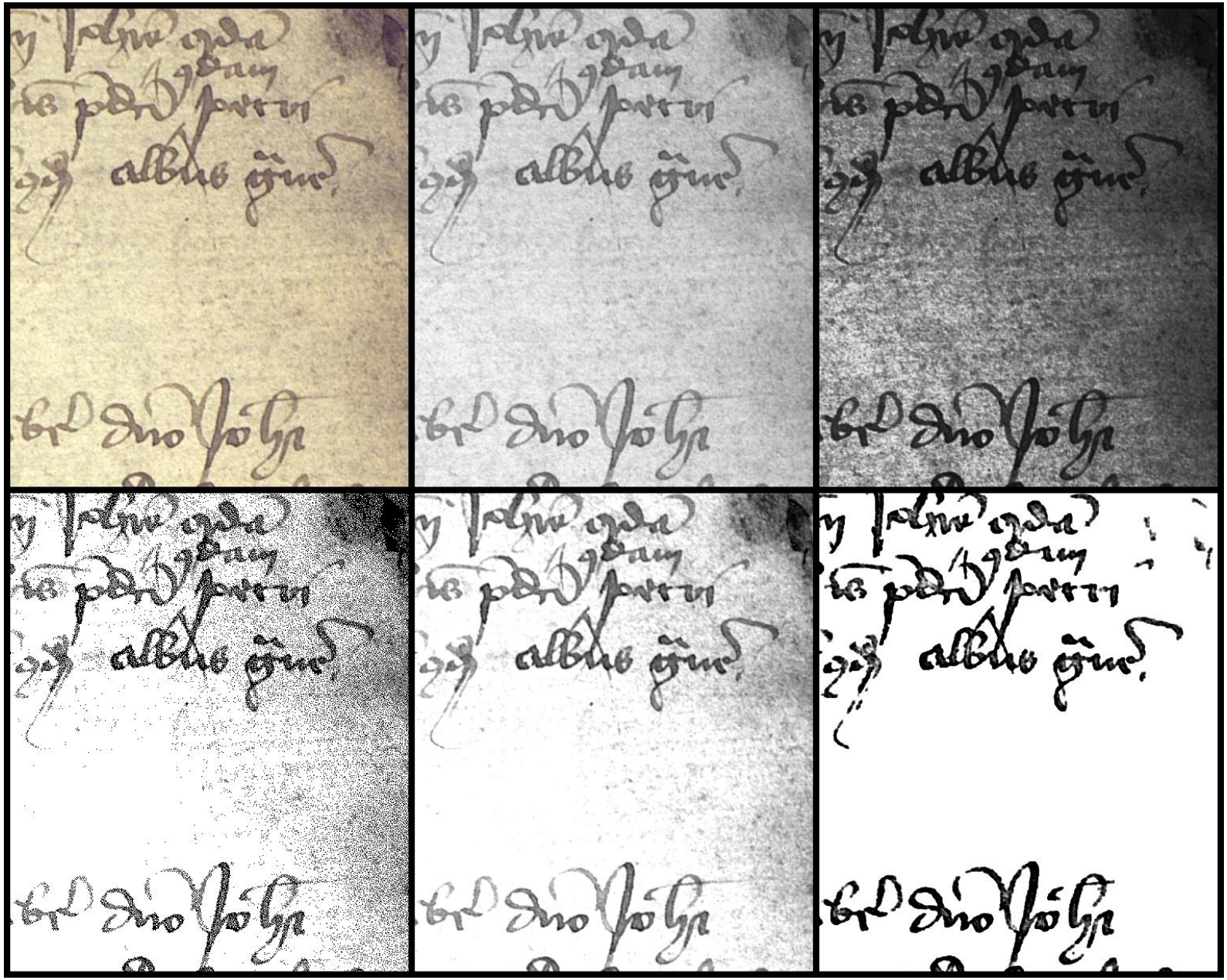
The Monk system / Lambert Schomaker

KdK 1897 Scan Company B

The Monk system / Lambert Schomaker

SAL 1421 Company C

The Monk system / Lambert Schomaker



The Monk system / Lambert Schomaker

1421

1

1421

1457

1559

on eadmodnesse se wuniende ge hy wan noldon þæt
hi þonne geswedan his mihte swa micele swyðlicor
swa micclum swa hi nu heora swuran to his gewylde
nellað gebigan. Þas word synd gecwedene be þam
wiðercorenum ac her filiað þa word þe þa gecore
nan gesweswiað. Se hælend cwæð. þonne þas
wundra onginnað ahebbað þonne eowre heafda
⁊ behealdaþ for þan ðe eower alysednes genealæcð.
Swylce he sweotelice his gecorenan manode. þonne mid-
dan eardes wita gelomlæcað þonne oga þæs mycclan
domes bið æteowed. ahebbað þonne eowre heafda. þ is
gladiað on eowrum mode for þi þonne se middaneard bið
geendod þe ge nu lufedon þonn bið gehende seo alysednes
þe ge sohton. On halgū gewritū bið gelome heafod geset
for þæs mannes mode. for ðan þe þ heafod gewissað
þam oðrū limum. swa swa þ mod gediht þa geþohtas. we a
hebbað ure heafda þonne we ure mod arærað to gefean
þær heofenlican æðeles. þa ðe god lufiað. hi synd gema-

Scanning and image preprocessing

- Quality of scanning companies varies considerably even if 'strict' standards (NL: Metamorfoze) are applied
- OCR has **very different demands** from web access and human reading requirements
- 'Simple operations' such as ink/paper separation are not at all solved fundamentally in image-processing science due to a pervasive 'chicken and egg' problem:

"In order *to know the text,*
a system needs a good segmentation, but

a perfect segmentation into lines and words is only possible knowing the text "

Coarse overview current contents Monk, 71 documents, 53000 page scans

- KdK Dutch administrative 1893-1906 20 books
- Dutch Admirality 1760-1823 5 books
- Printed, Elzevirium, 1616
- Qumran scrolls (2463 scans)
- Middelduits (example document)
- Accounts, 1425 Gelria
- Schepenbank Louvain, 1421-1559, 3 books
- Colonial diary (1932)
- Municipal year report 1855
- 20k illuminated initials
- Russian handwritten newspaper 1672
- Scholarly correspondence 1674-1682
- Chronicon Boemorum 1201
- Homiliarum Opatovicense 1150
- Resoluties Staten Generaal 1627
- Medieval charters 1300-1550 in 25yr periods
- Witch trial 1605
- 'Beowulf' and related, 4 books
- Ming Qing poetry , 7 books
- Arabic document, 291 scans
- Charlotte Perkins-Stetson diaries 1883- (682p.)
- Wittenbergsches (fraktur machine print)

Cooperations with:

- *Huygens institute, Sorbonne,*
- *Harvard,*
- *Stanford,*
- *Czech National Library,*
- *Dutch National Library, McGill Univ.,*
- *Utrecht Univ.,*
- *Univ. Uppsala*
- *City archive Louvain,*

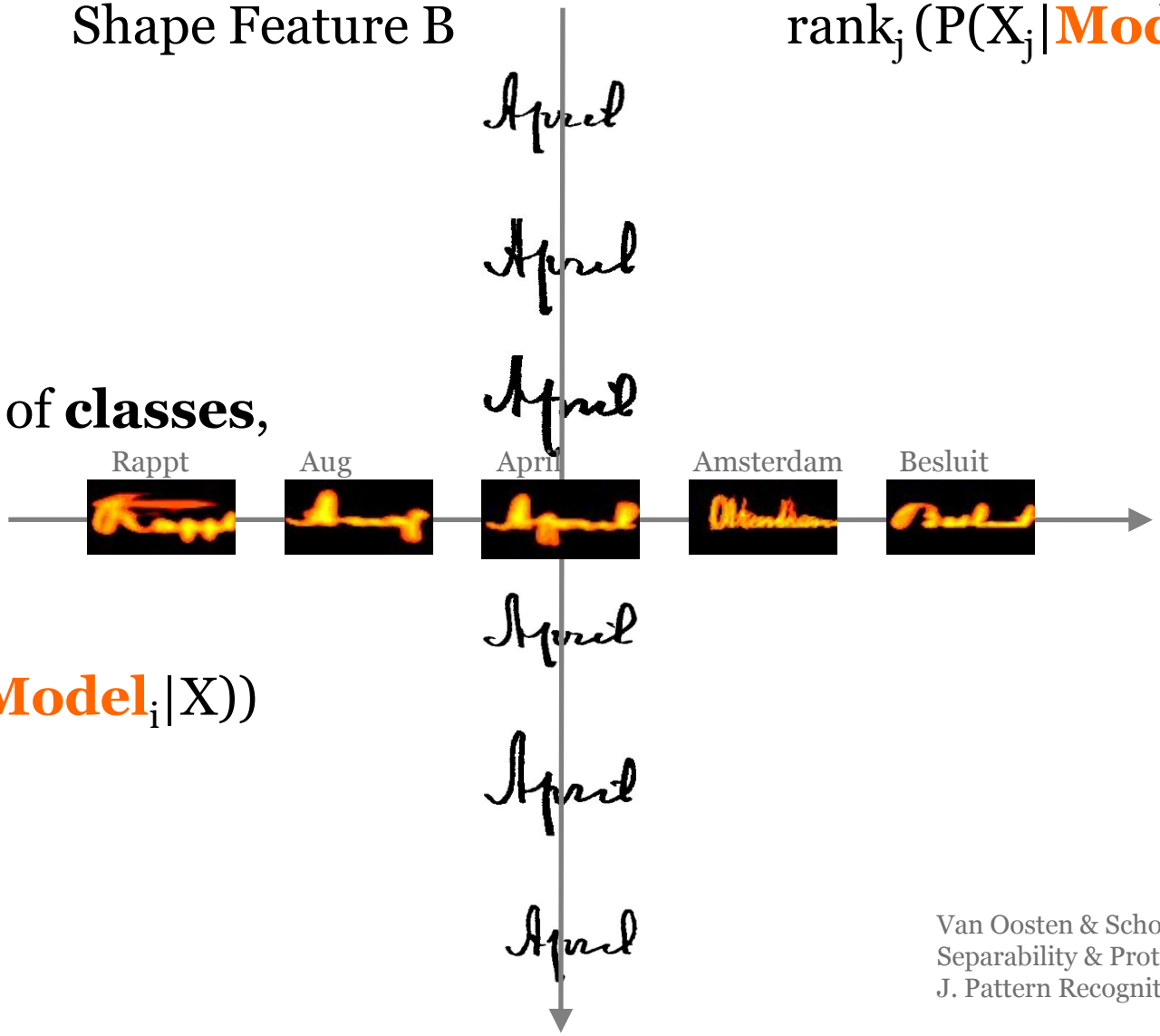
Pattern recognition and machine learning

- Several years of experimentation
  - started in 2005
  - Monk was switched on, to autonomous mode, in 2009
- For developing and optimizing two functions:
- **Retrieval**: return images for a given keyword
- **Recognition**: return the most likely word given an image

Retrieval of **instances**,

Shape Feature B

$rank_j\,(P(X_j|\textbf{Model})$



Recognition of **classes**,

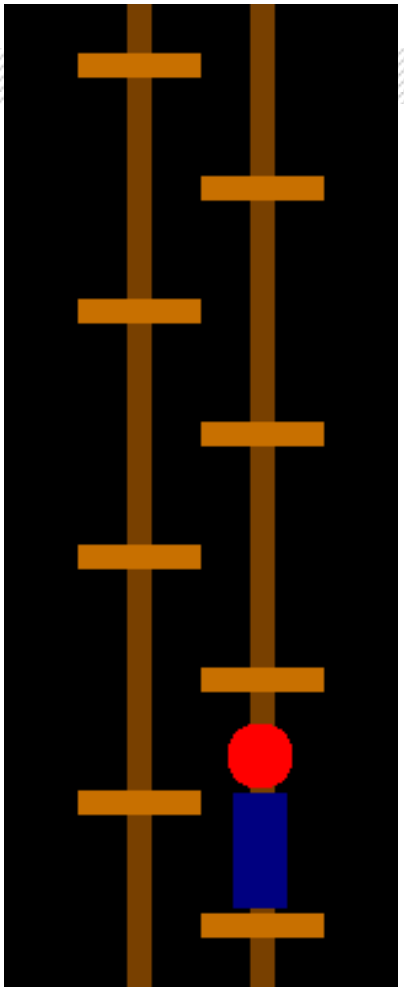
Feature A

$argmax_i\,(P(\textbf{Model}_i|X))$

Van Oosten & Schomaker (2013)
Separability & Prototypicality ...
J. Pattern Recognition

Boosting performance

- Old mine-shaft elevator principle: *Fahrkunst*
- After having trained with method A to its 'max'
- An orthogonal method B can reach a higher performance
- Then method A again, etc.
- Until the real asymptote is reached



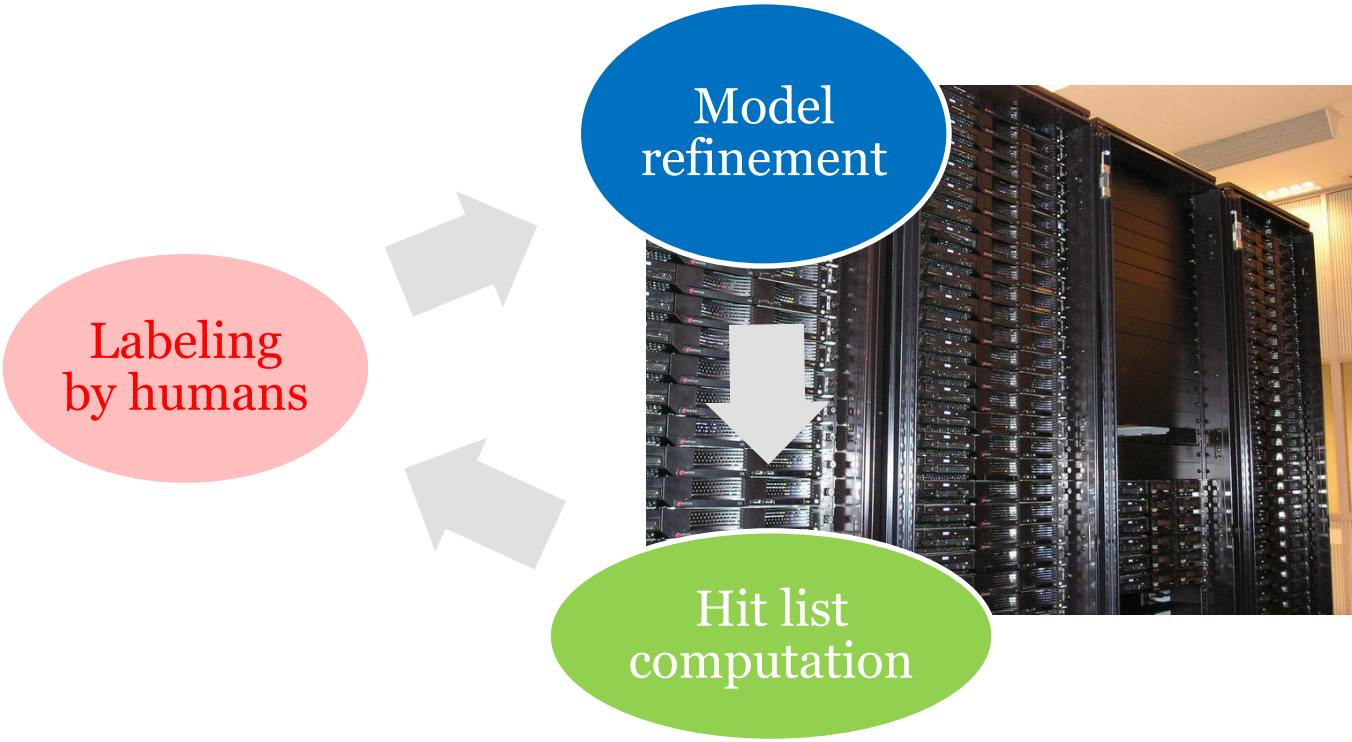
Method A – Method B

THE SOLUTION:

24/7 learning over internet

RuG HPC cluster

“Interactive Supercomputing”

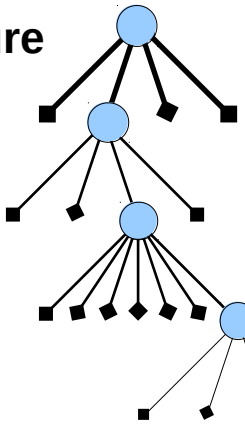


Monk architecture

Data types:

- Image files
- Text (labels, tags) files
- Key-Value index files
- Matrices (lists of labeled vectors)

Most of the files are Wordzone images



Data

Collections *x20*

Books *x20*

Pages *x1000*

Lines *x40*

Annotation, Labeling (.txt)

Wordzones *x100*

Vectorial matching

Word Models

Human

Human

Human

Human Admin

Abbot

Queue:
Sor dex "Amst er dam",
Recog "book- 20",
I ndex "book- 12",
etc,

Novice-001 Node-001

Novice-002 Node-001

Novice-003 Node-005

Novice-011 Node-022

Use

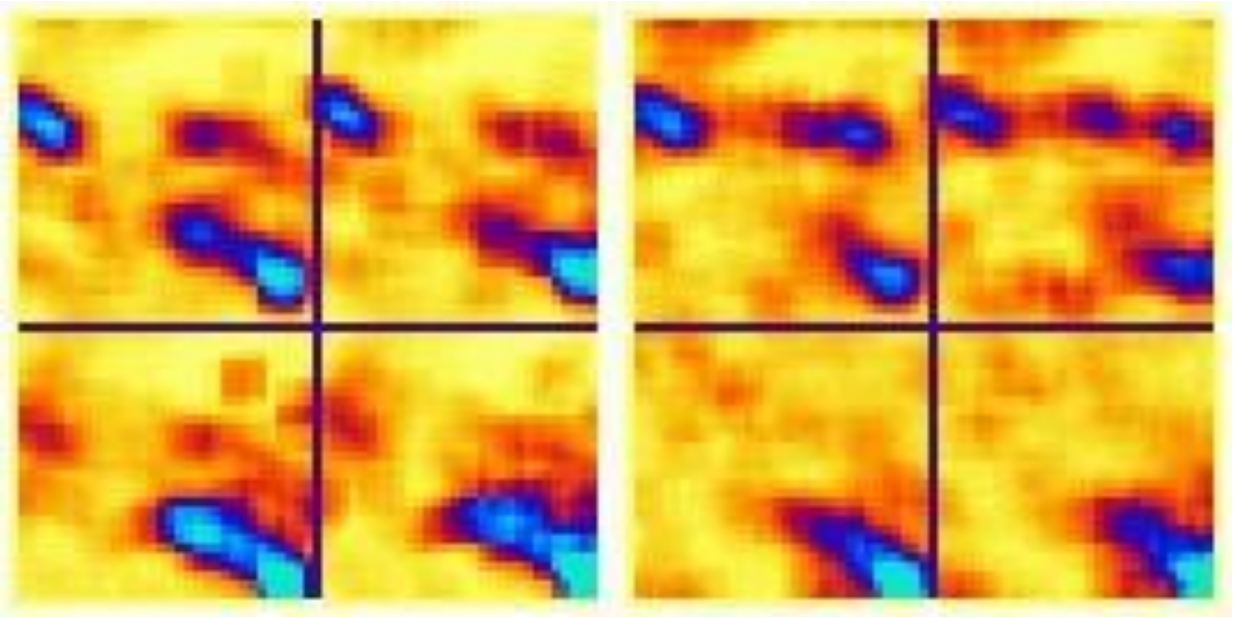
Labeling
Browsing
Searching
(Control)

Computing

Lessons learned during Monk development

- A shape feature which is powerful for Retrieval may not be strong in Recognition!
- Feature B: hit list should provide nice, intuitive ranking in a satisfying 'hit list'
- Feature A: target word class should survive competition with the other word classes (emerging needle from the heterogeneous hay stack)

Examples of feature patterns for two words



Amsterdam Groningen

van der Zant. Schomaker, Haak (2008). IEEE PAMI

Distance Measures for nearest-neighbour matching

- Hamming:

$$d = \sum_i |p_i - q_i|$$

- Euclid

$$d = \sqrt{\sum_i (p_i - q_i)^2}$$

- Minkowski

$$d = \left(\sum_i |p_i - q_i|^n \right)^{\frac{1}{n}}$$

- Hausdorff

$$d = \max_i \left(|p_i - q_i| \right)$$

- Chi Squared (χ^2)

$$d = \sum_i \frac{(p_i - q_i)^2}{p_i + q_i}$$

- Bhattacharyya

$$d = \sqrt{1 - \sum_i p_i \cdot q_i}$$

Schomaker, Bulacu & Franke (2004) 9th IWFHR

Monk search - keywo...

http://aps.ai.rug.nl/cgi-bin/wrdlist?db=*&cmd=scroogle&keyword=Tilburg

Google | CNN.com - Breaki... | Overview of LS's t... | nu.nl | Het laatste... | Other Bookmarks

monk: No dBASE db=*

Monk - steekwoord zoeken

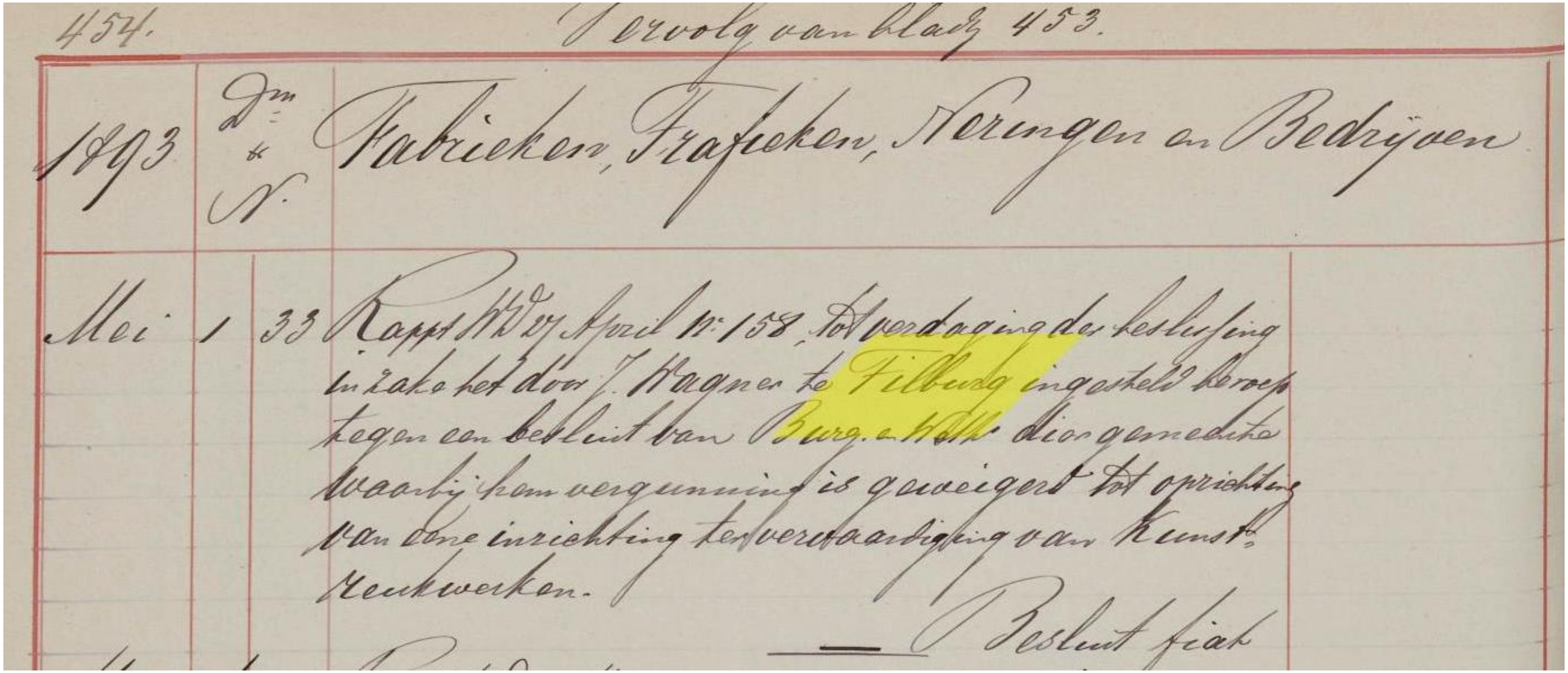
Tilburg

Search

- KdK 1903 [Tilburg] bladz. 668 regel 8 0668 [L] Lager Onderwijs - 1903/Novb/28 #30
- KdK 1893 [Tilburg] bladz. 148 regel 25 0148 [L] Spoorwegen - 1893/
- KdK 1893 [Tilburg] bladz. 148 regel 25 0148 [L] Spoorwegen - 1893/
- KdK 1893 [Tilburg] bladz. 454 regel 4 0454 [L] Fabrieken, Trafieken, Neringen en Bedrijven - 1893/
- KdK 1893 [Tilburg] bladz. 716 regel 24 0716 [L] Procureurs - 1893/

Aantal [*Tilburg*] gevonden: 5

Tilburg:



454. Vervolg van blad 453.

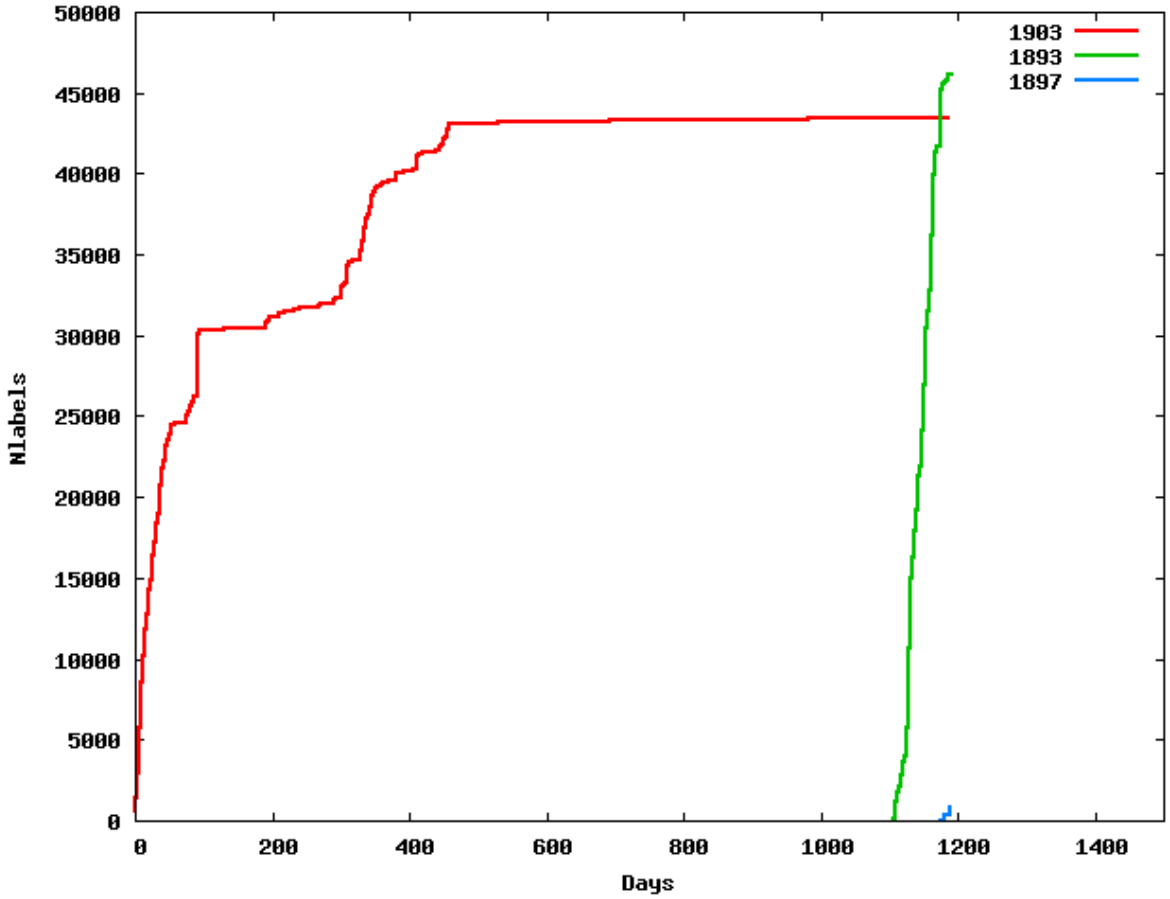
1893	D^m & N.	Fabrieken, Trafieken, Neringen en Bedrijven
Mei	1 33	Rapp. W.D. 27 April No. 158, tot verdaging der beslissing in zake het door J. Wagner te Tilburg ingesteld beroep tegen een besluit van Burg. en Weth. dier gemeente waarbij hem vergunning is geweigerd tot oprichting van eene inrichting ter vervaardiging van Kunst- [illegible]. — Besluit fiat

Does it work?

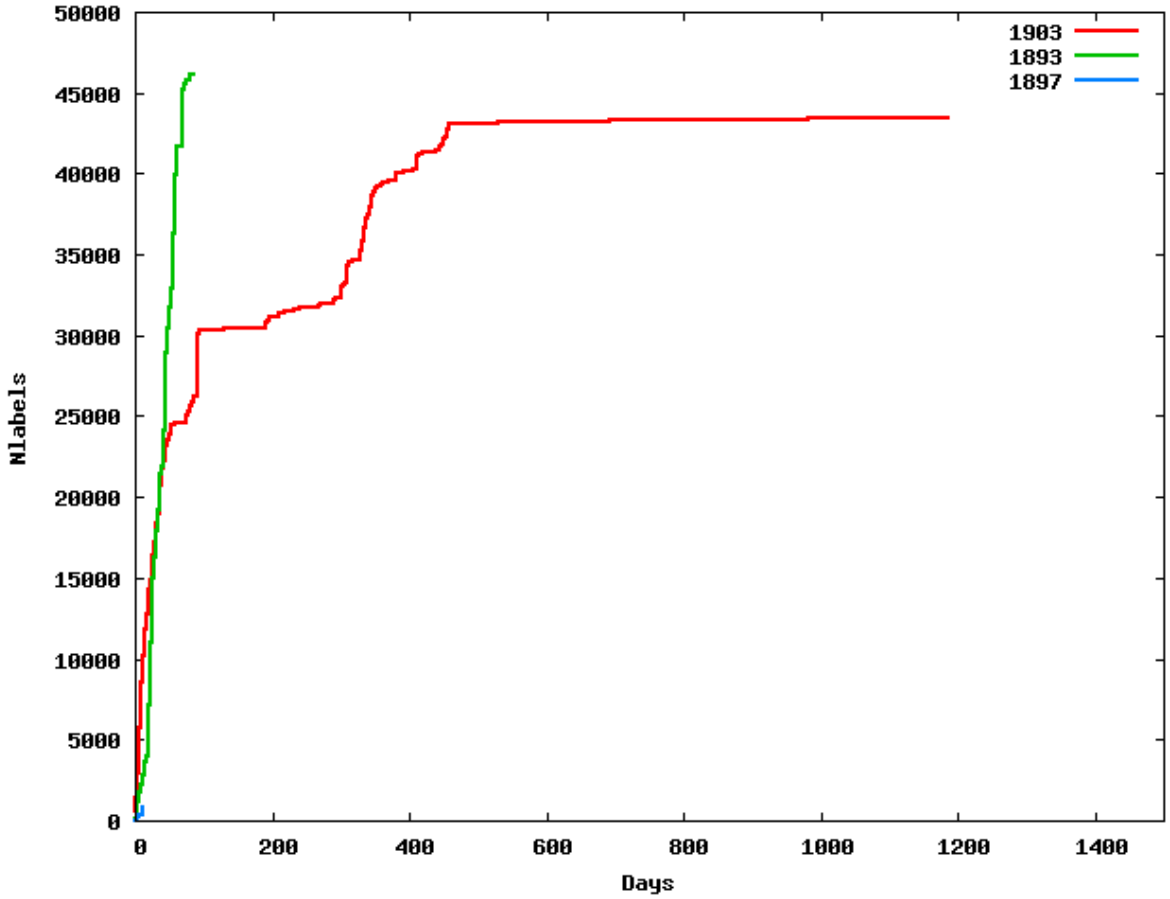
- Do you need to label continuously?
  - Then transcription would have been better/easier
- How good is the pattern recognition?

- Do you need a supercomputer?
- How many people are working 'back office'?
- Can I send my collection to Monk?
- What is the benefit, what is the output of the process?

#Labels, total as a function of 'days on-line', KdK 1903,1893,1897



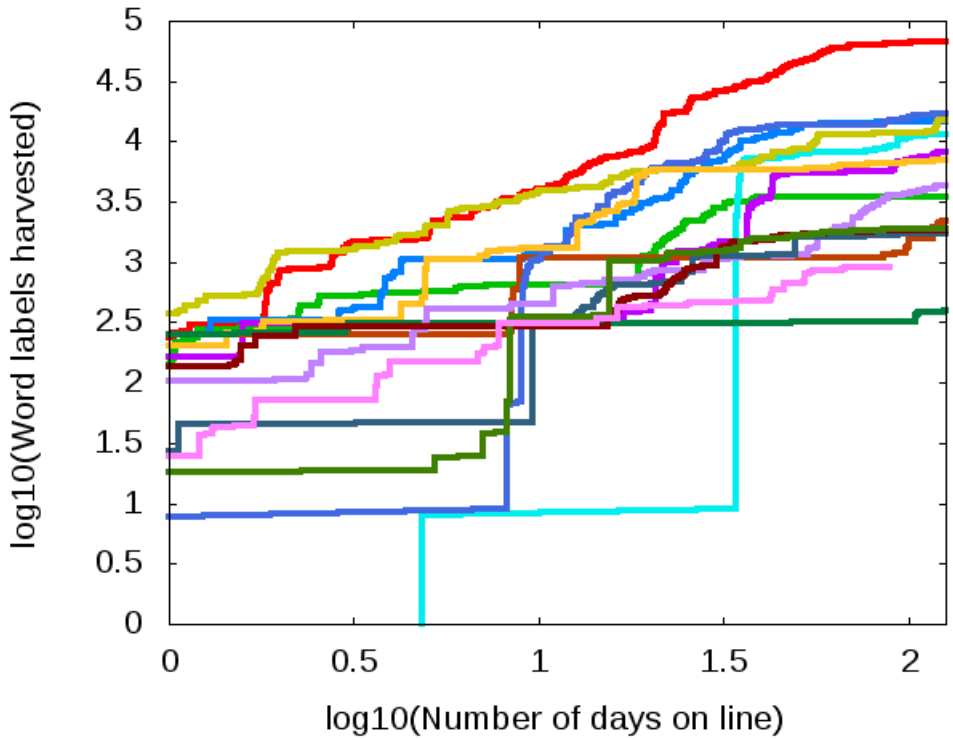
#Labels, total as a function of 'days on-line', KdK 1903,1893,1897



start t=0
synchronized

Monk

Jumps in the number of harvested word labels (y-axis) coincide with reaching a critical training dataset size (time axis)



Note the log() scale: these are '**snowball effects**'

Performance

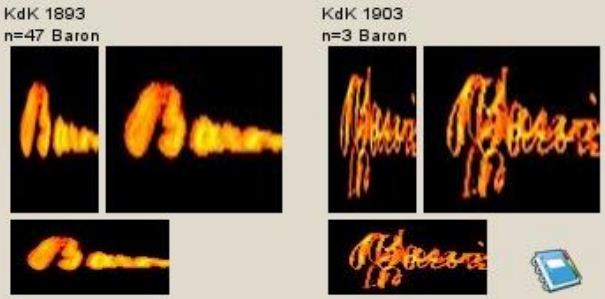
- **Retrieval precision** increases monotonously with number of word examples

#training examples	What to expect in Top 50
1	Correct instance is good luck
5	A few correct hits
20	Many useful hits

- Lexicon of 11k words: > 81% **recognition rate**, KdK
- Lexwords with 50+ examples: > 95% correct
- Idem, Cliwoc captain's logs: 93%

Good results: *bramzeijls* (top gallant's) hit list after training on just 2 examples

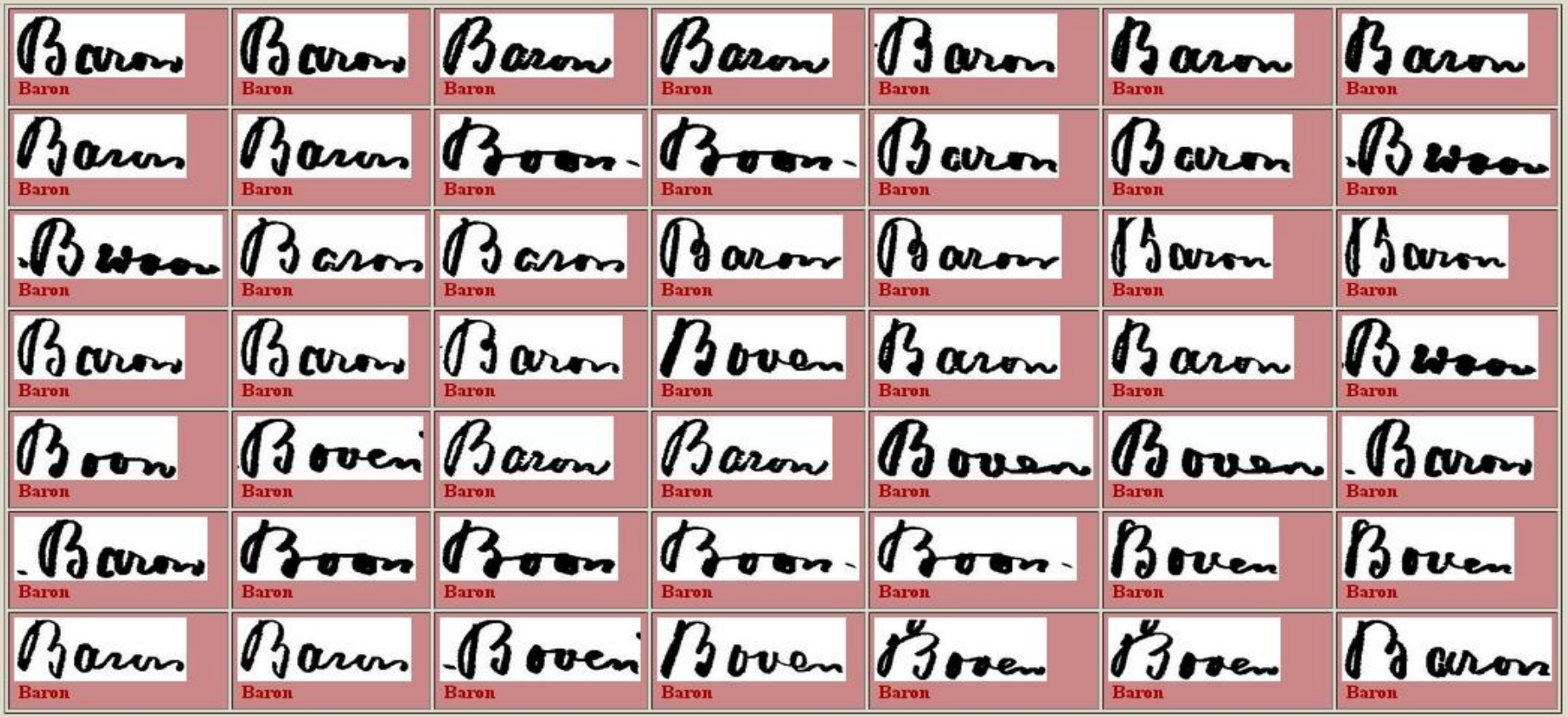
	Boek: KdK 1893		
Hoofdstuktitel:			
Woordherkenner:	ToDo		
Woord:	Baron	Selectie uit hit list:	Niet (HUMAN)
Aantal regels in hit list:	10	Eerste itemnr. in hit list:	0
Weergave:	Postzegels	Vernieuw:	Onderstaand overzicht



ToDo / Baron / (no comment)

Items in hit list of Baron numbered: 0-2999. (now showing: 0-48)

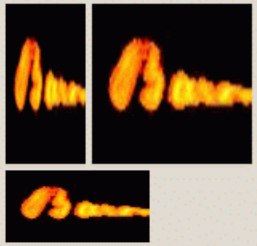
Begin Prev Reload Next End



Boek:	KdK 1893		
Hoofdstuktitel:			
Woordherkenner:	ToDo		
Woord:	Baron	**Selectie uit hit list:**	Alles
Aantal regels in hit list:	49	**Eerste itemnr. in hit list:**	0
Weergave:	Postzegels	**Vernieuw:**	Onderstaand overzicht

-# RECOMPUTE

KdK 1893
n=47 Baron

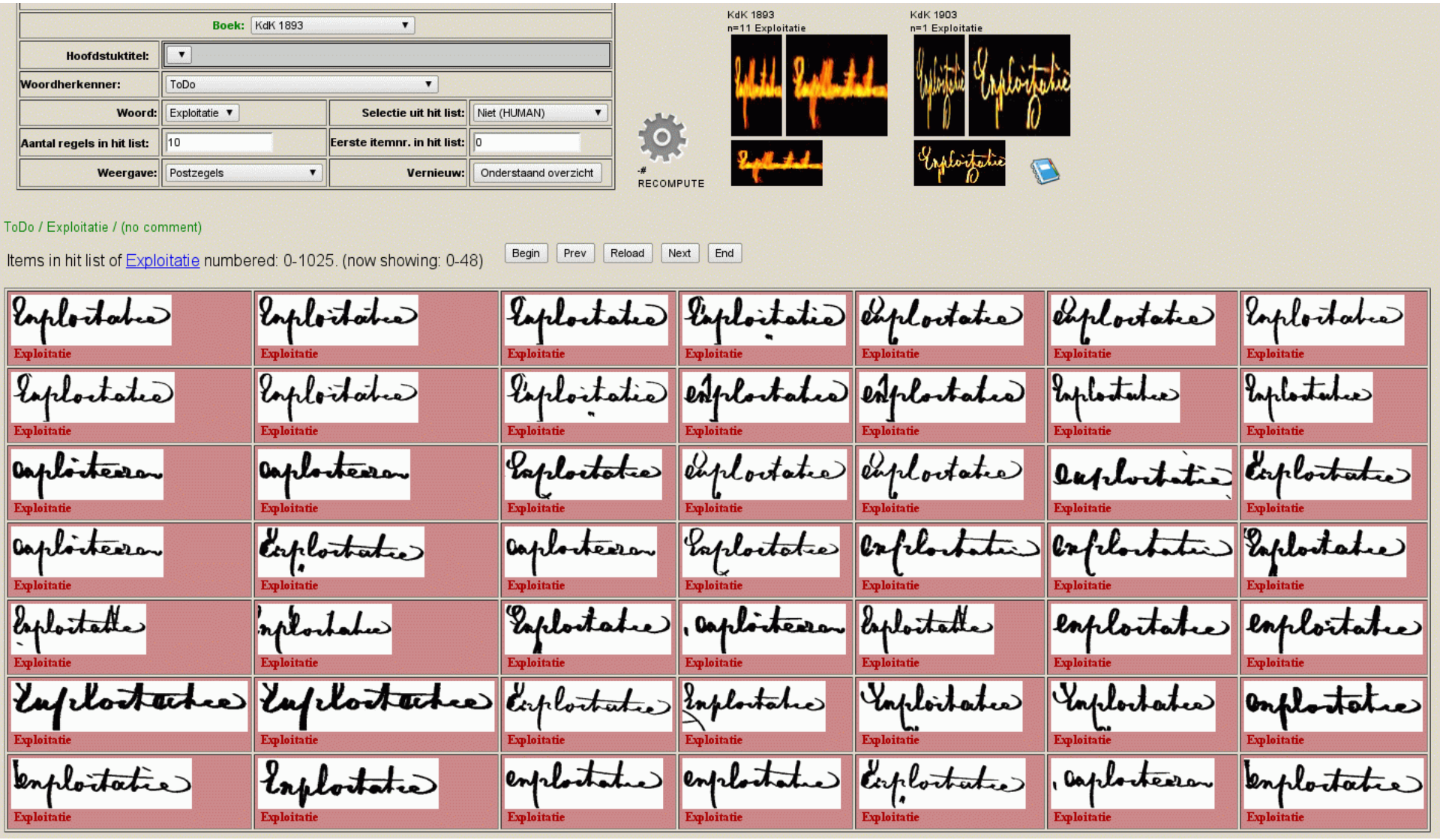


KdK 1903
n=3 Baron

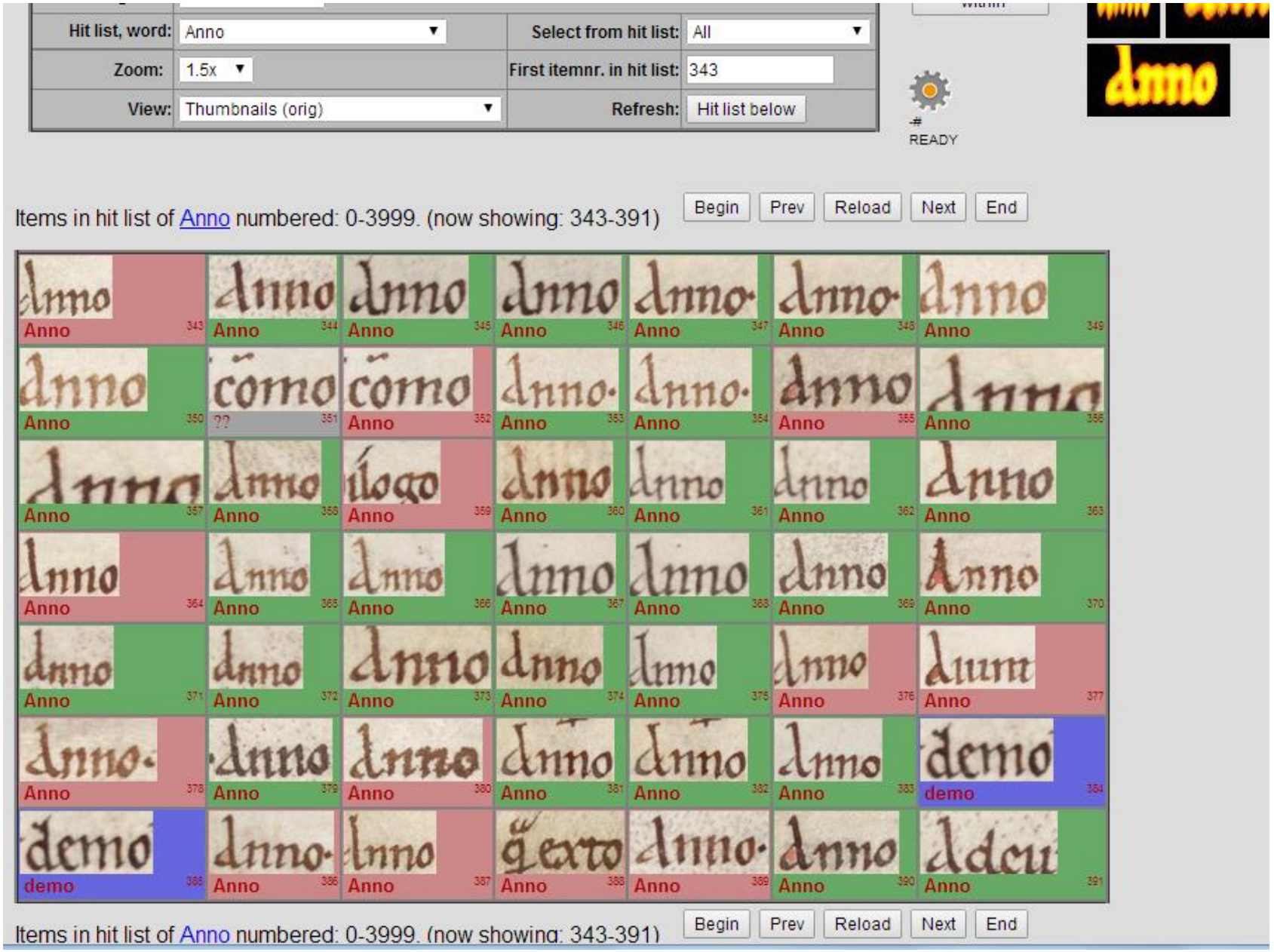
ToDo / * / (no comment)

Items in hit list of Baron numbered: 0-2999. (now showing: 0-48)

Begin Prev Reload Next End


Boek: KdK 1893
Hoofdstuktitel:
Woordherkenner: ToDo
Woord: Exploitatie
Selectie uit hit list: Niet (HUMAN)
Aantal regels in hit list: 10
Eerste itemnr. in hit list: 0
Weergave: Postzegels
Vernieuw: Onderstaand overzicht
RECOMPUTE
KdK 1893
n=11 Exploitatie
KdK 1903
n=1 Exploitatie
ToDo / Exploitatie / (no comment)
Items in hit list of Exploitatie numbered: 0-1025. (now showing: 0-48)
Begin
Prev
Reload
Next
End
Exploitatie

At rank 343 in the list: pink newlings



Monk - * (Overvi | Gnuplot | Monk - * (Overvi | Monk - Octb (Hi | Bulk-labeling-se | And the world's | Merecedes E250

application02.target.rug.nl/cgi-bin/monkweb?cmd=similarwords&db=1002&wrdid=Octb.linlist&key=Octb&annot=all&trainedwordmethod=Alldex&dispmode

tems in hit list of Octb numbered: 0-3999. (now showing: 0-99)

Begin Prev Reload Next End

. . .

tems in hit list of Octb numbered: 0-3999. (now showing: 0-99)

Begin Prev Reload Next End

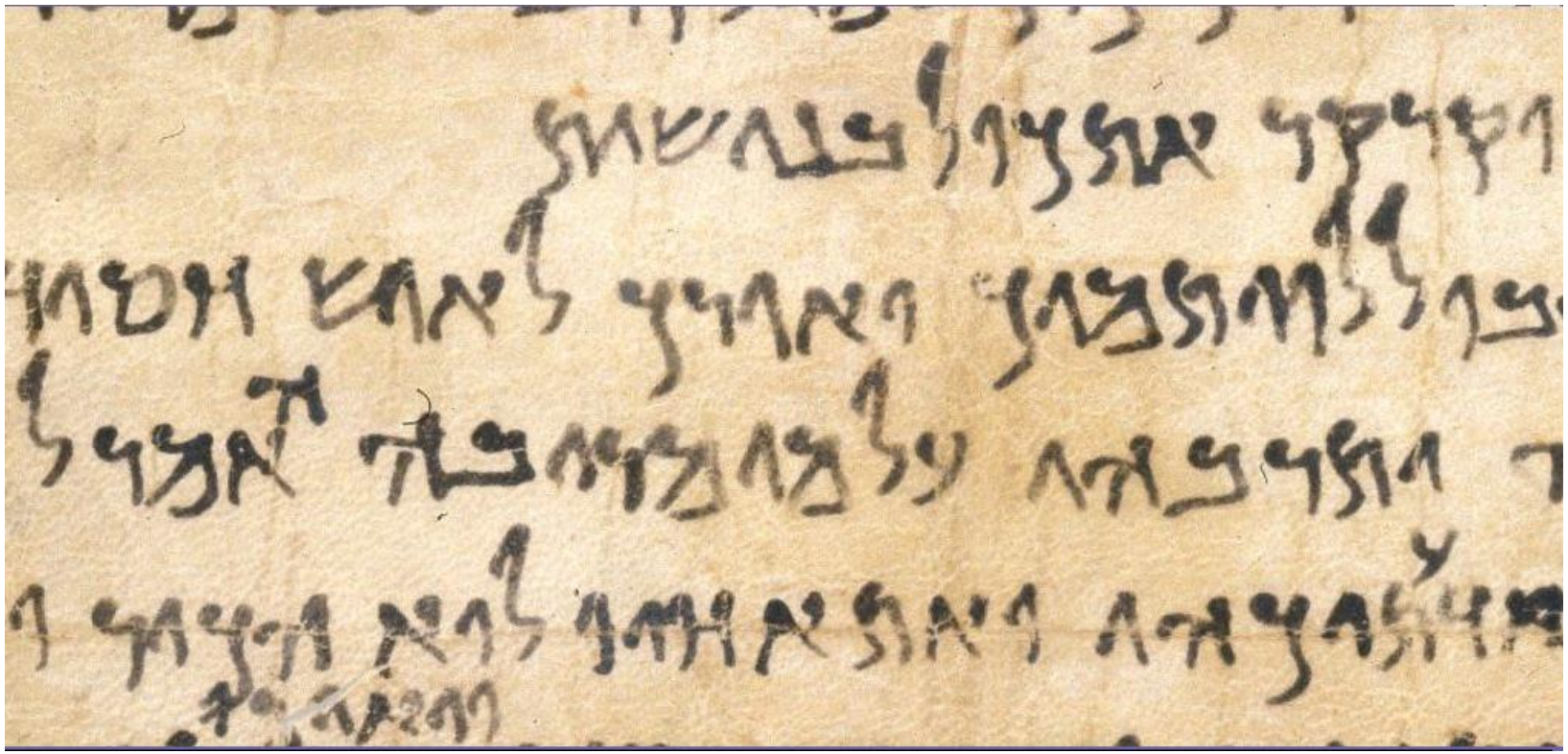
ok

Bulk labeling of all instances of [Octb] above

Items in hit list of ende numbered: 0-3999. (now showing: 0-48) Begin Prev Reload Next End

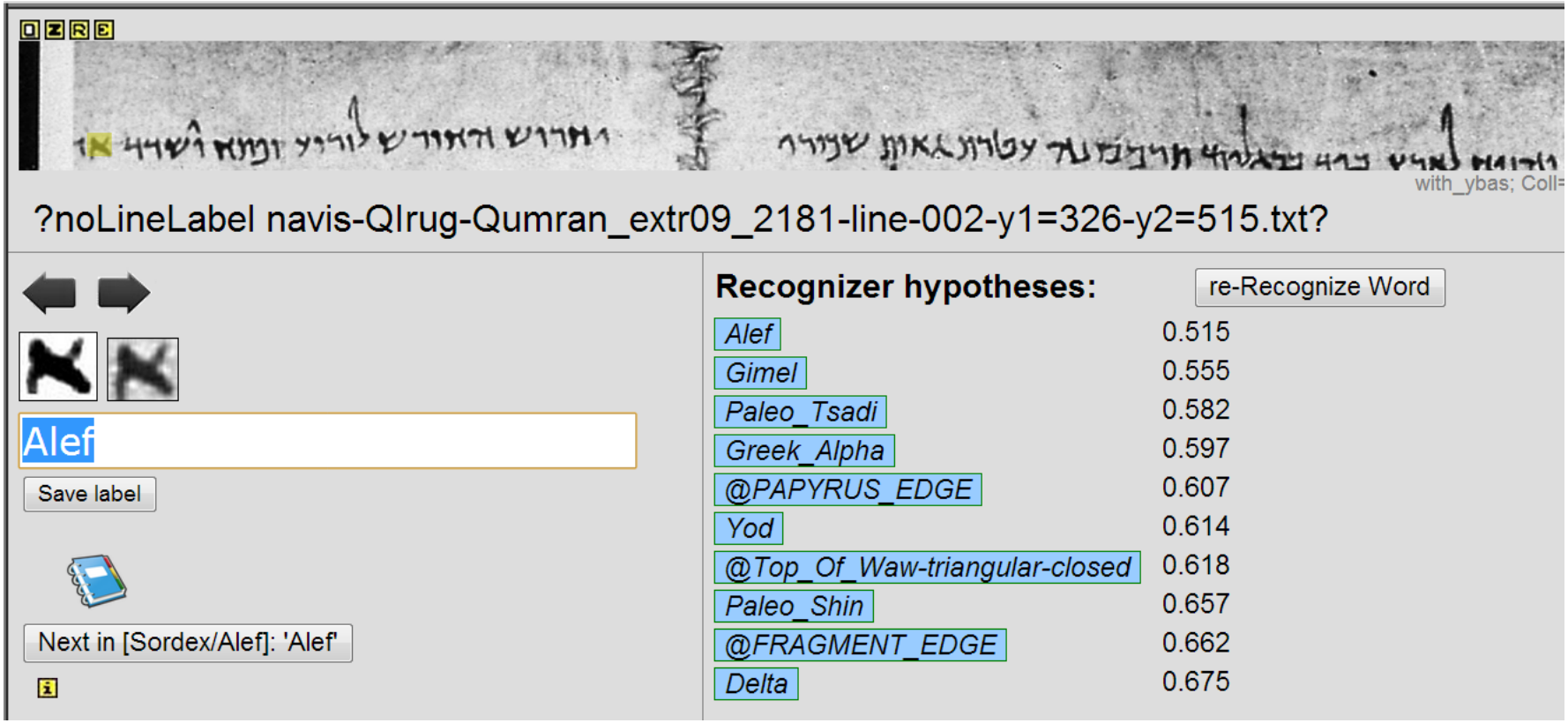
Items in hit list of ende numbered: 0-3999. (now showing: 0-48) Begin Prev Reload Next End

Monk usage example: May 2013 Qumran scrolls: Daniel Stoekl (Sorbonne), Mladen Popovics (Groningen)



Qumran scrolls: 2400 photographs

- Using **Monk** for character labeling
- With Daniel Stoekl and Mladen Popovics
- Using its 24/7 machine-learning cycle:
  - Label ➔ Train
    ➔ Label some More ➔ Train
      ➔ Easily label Many
- **Thousands** of characters 'mined' out of the Qumran collection of photographs in just **two weeks**, with very **little effort** in human labeling


OZRE
with_ybas; Coll=
?noLineLabel navis-QIrug-Qumran_extr09_2181-line-002-y1=326-y2=515.txt?
Alef
Save label
Next in [Sordex/Alef]: 'Alef'
Recognizer hypotheses:
re-Recognize Word
Alef 0.515
Gimel 0.555
Paleo_Tsadi 0.582
Greek_Alpha 0.597
@PAPYRUS_EDGE 0.607
Yod 0.614
@Top_Of_Waw-triangular-closed 0.618
Paleo_Shin 0.657
@FRAGMENT_EDGE 0.662
Delta 0.675

Back to overview of trained words

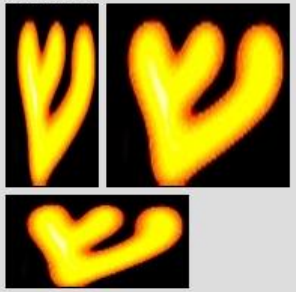
Book: Qumran: Qumran scrolls

Word recognizer:	Sordex		
Hit list, word:	Shin	Select from hit list:	All
Zoom:	1.0x	First itemnr. in hit list:	0
View:	Thumbnails (orig)	Refresh:	Hit list below

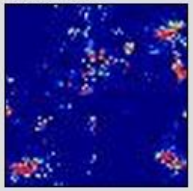
set recompute 'within'

-# READY

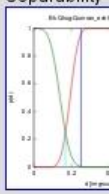
Qumran scrolls n=2311 Shin



Qumran scrolls n=2311 Shin



Separability



Items in hit list of Shin numbered: 0-3999. (now showing: 0-99)

Begin Prev Reload Next End

. . .


Shin
Alef
He
Mem-medial
Bet
Taw
Ayin
Tsadi-medial
Nun-medial
Mem
Samekh
Gimel
Tet
Het
Lamed
Qof


Paleo_Alef
Paleo_Waw_var01
Paleo_Yod
Paleo_Dalet
Paleo_He-closed
Paleo_Waw
Paleo_He
Paleo_Yhw
Paleo_yhwh
Paleo_Resh
Paleo_Zayin
Paleo_Mem
Paleo_yh
Paleo_Bet
Paleo_Var004
Paleo_Shin

university of groningen

1614

/ artificial intelligence / HWR group

Allograph harvest in Qumran collection

university of groningen

1614

/ artificial intelligence / HWR group

Allograph harvest 33k labels in NWO/MPS project with Huygens institute

Back to overview of trained words

? Book: 1300refset: MPS 1300 [*]

Word recognizer: YearLetters

Hit list, word: @a_charter_1300 | Select from hit list: All, hide the label

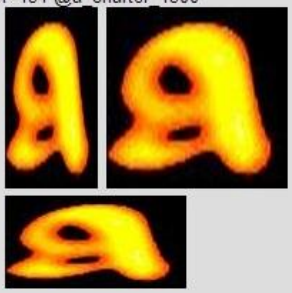
Zoom: 1.0x | First itemnr. in hit list: 0

View: Thumbnails (orig) | Refresh: Hit list below

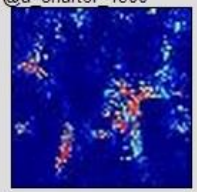
set recompute 'within'

#
READY

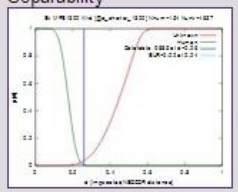
MPS 1300 [*]
n=151 @a_charter_1300



MPS 1300 [*]
n=151
@a_charter_1300



Separability

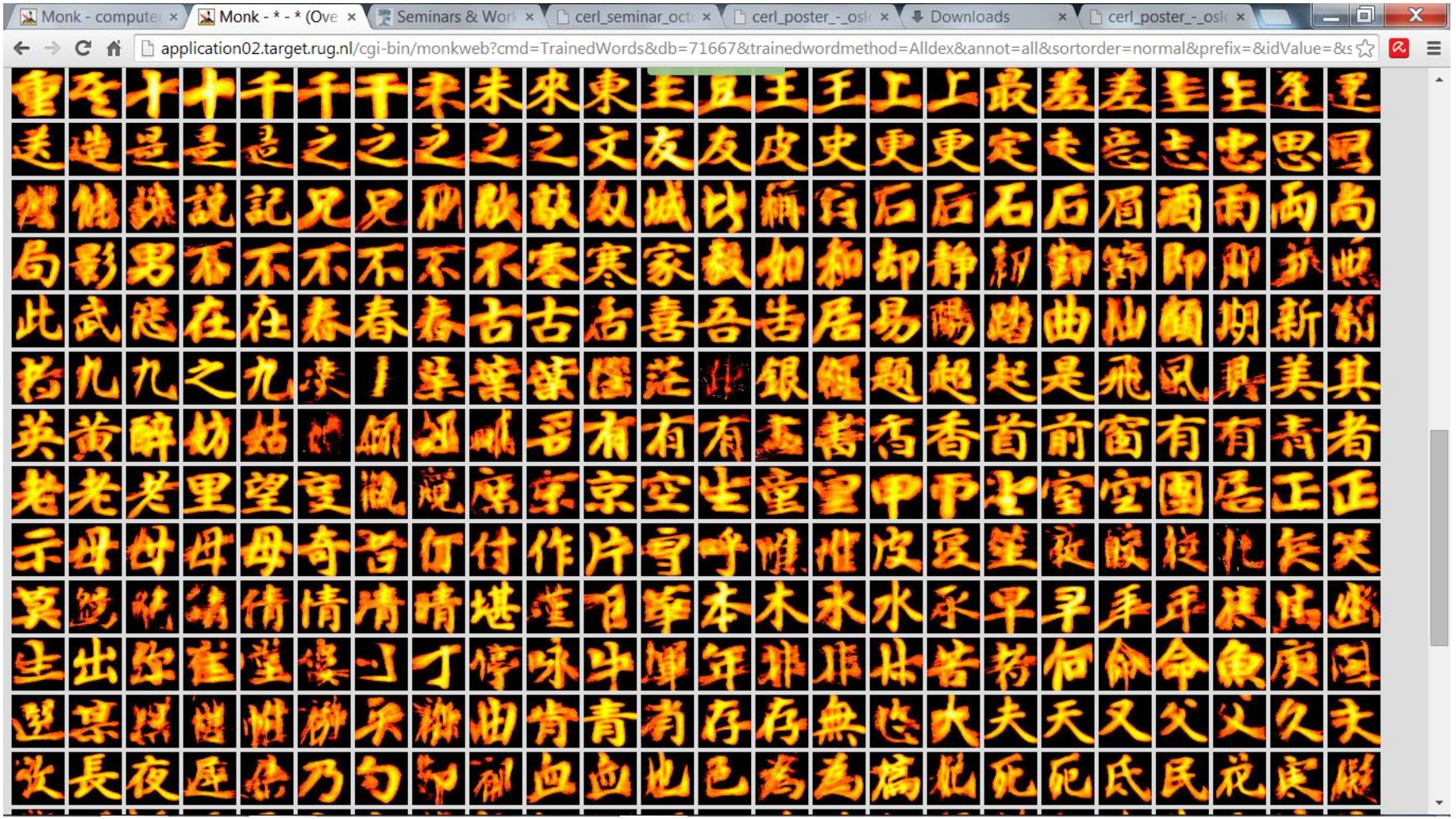


Items in hit list of @a_charter_1300 numbered: 0-3999. (now showing: 0-99)

Begin | Prev | Reload | Next | End

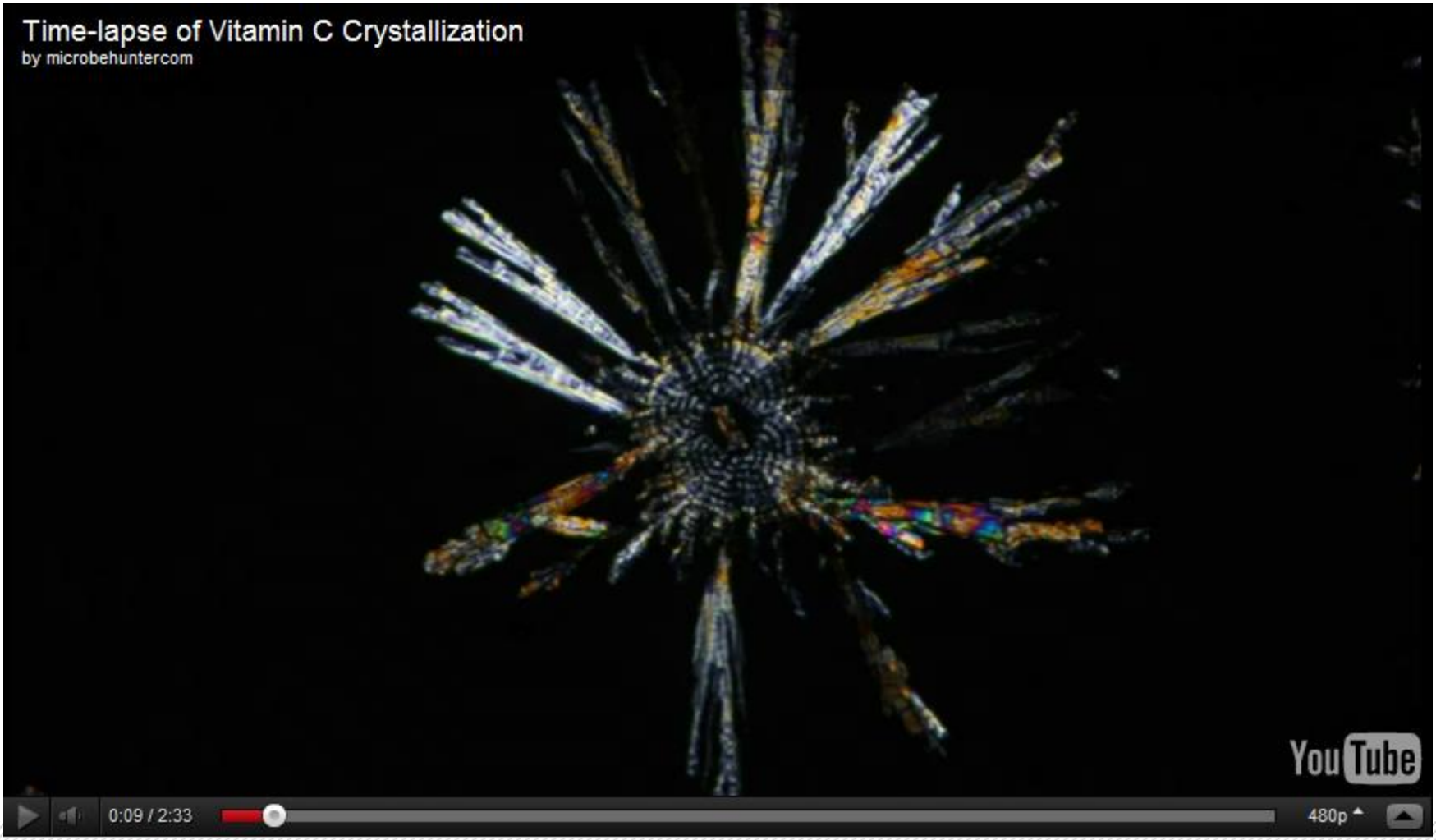
... harvested during March 2014

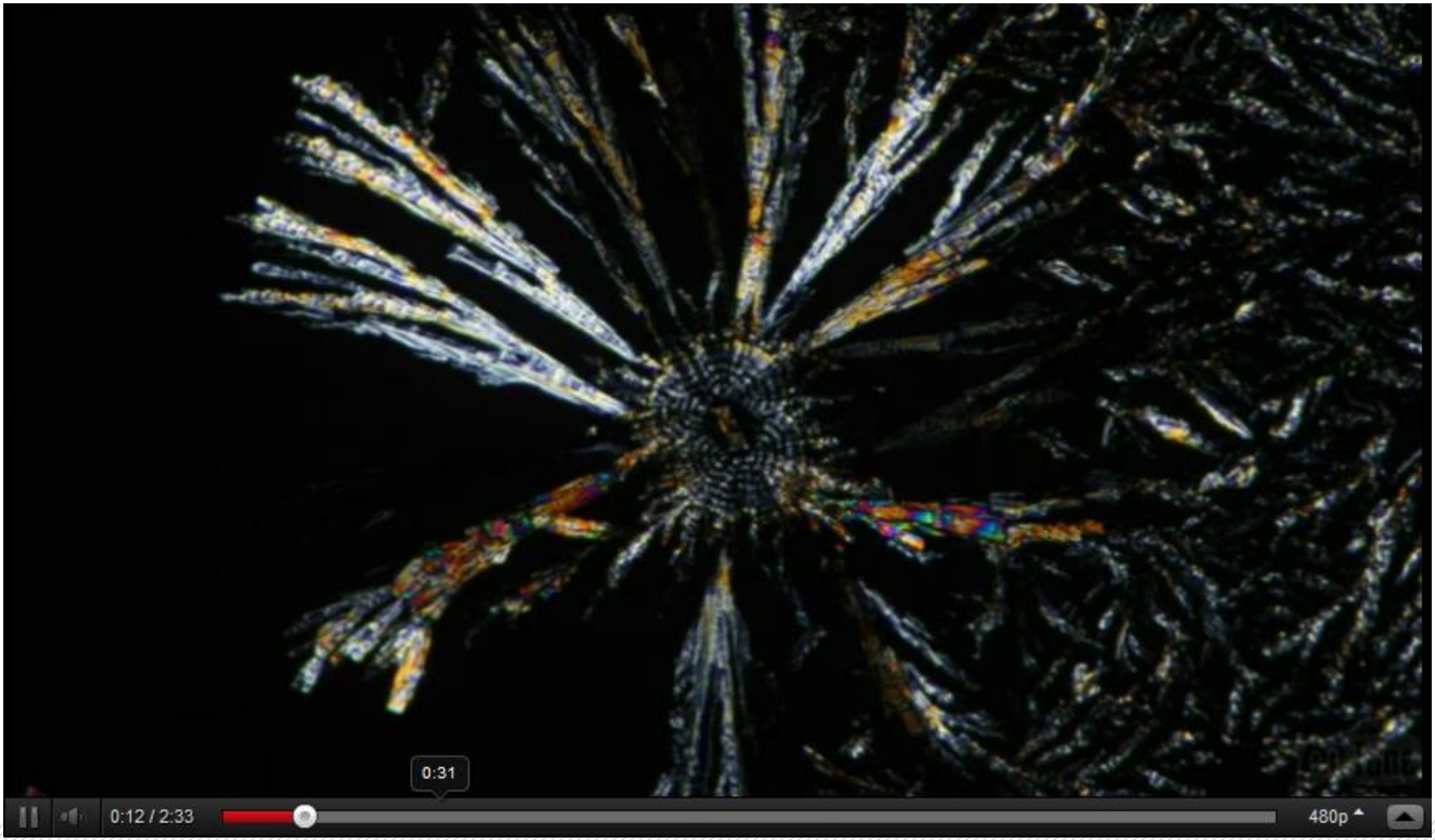
Yes, also Chinese characters (handwritten & woodblock,
Grace Fong & Harvard Yenchin collection)



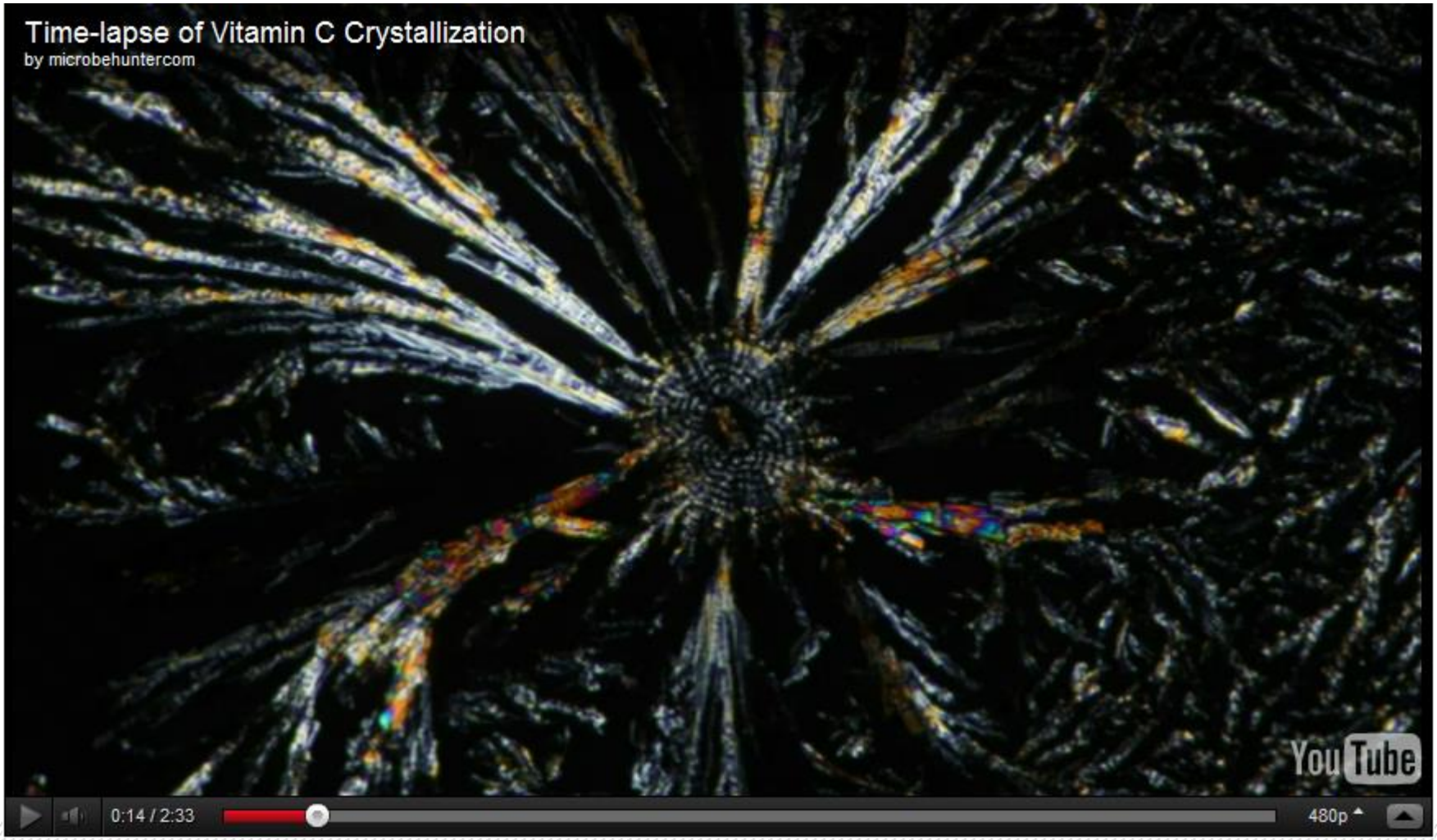
Necessary metaphor for the Monk process

- Not: a stream-lined mechanical factory
- Yes: crystal-like growth, arborization,
- Randomly over a large surface
- Energy comes from:
  - human labels
  - machine re-compute actions


Time-lapse of Vitamin C Crystallization
by microbehuntercom
0:09 / 2:33
480p
YouTube



The Monk system / Lambert Schomaker


Time-lapse of Vitamin C Crystallization
by microbehuntercom
YouTube
0:14 / 2:33
480p

dBid	Year	Short name	Pages	Scans	BookID (url base)	.reindex	.files	.IDs+txt	/Index	/Sordex	Words	Days
1	1893	*KoK 1893 [*] blz. 1-1316*	1-1384	1384	navis-NL-HaNA_2.02.04_3960/	1 days	1 months	1 months	1 days	1 days	12984 (=)	205.3 (+1.0)
2	1897	*KoK 1897 blz. 781-1628*	781-1632	852	navis-NL-HaNA_2.02.04_3965/	5 days	1 months	1 months	1 weeks	1 months	10514 (=)	264.4 (+1.0)
3	1898	*KoK 1898 blz. 1-842*	1-842	842	navis-NL-HaNA_2.02.14_7813/	5 days	1 months	1 months	1 weeks	1 weeks	8939 (=)	225.8 (+1.0)
4	1899	*KoK 1899 blz. 1-922*	1-922	922	navis-NL-HaNA_2.02.14_7815/	1 days	1 months	1 months	1 months	1 months	9210 (=)	205.0 (+1.0)
5	1899	*KoK 1899 blz. 862-1782*	862-1794	933	navis-NL-HaNA_2.02.14_7816/	5 days	1 months	1 months	1 weeks	1 weeks	9136 (=)	183.7 (+1.0)
6	1901	*KoK 1901 blz. 1-1040*	1-1099	1099	navis-NL-HaNA_2.02.14_7819/	5 days	1 months	1 months	1 weeks	2 days	9999 (=)	195.2 (+1.0)
7	1901	*KoK 1901 blz. 1041-2024*	1041-2038	998	navis-NL-HaNA_2.02.14_7820/	5 days	2 weeks	2 weeks	1 weeks	2 weeks	9286 (=)	181.0 (+1.0)
8	1902	*KoK 1902 blz. 974-2065*	982-2073	1092	navis-NL-HaNA_2.02.14_7822/	5 days	1 months	1 months	2 months	1 weeks	6485 (=)	197.3 (+1.0)
9	1903	*KoK 1903 blz. 1-1094*	1-1094	1094	navis-H2_7823_0001-1094/	4 days	1 months	1 months	4 days	4 days	9013 (=)	504.3 (+0.9)
10	1903	*KoK 1903 blz. 1299-1346*	1299-1346	48	navis/	1 years	9 months	9 months	1 weeks	1 weeks	1411 (=)	116.3 (+1.0)
11	1903	*KoK 1903 blz. 1348-1401*	1348-1401	54	navis2/	1 years	9 months	9 months	3 months	3 weeks	1861 (=)	116.3 (+1.0)
12	1904	*KoK 1904 blz. 1-998*	1-1064	1064	navis-NL-HaNA_2.02.14_7825/	4 days	1 months	1 months	4 days	4 days	7624 (=)	171.5 (+1.0)
13	1904	*KoK 1904 blz. 999-2008*	999-2110	1112	navis-NL-HaNA_2.02.14_7826/	5 days	1 months	1 months	1 weeks	4 days	8662 (=)	168.7 (+1.0)
14	1905	*KoK 1905 blz. 1-1120*	1-1190	1190	navis-NL-HaNA_2.02.14_7827/	5 days	1 months	1 months	1 months	1 weeks	7808 (=)	174.0 (+1.0)
15	1905	*KoK 1905 blz. 1121-2247*	1121-2257	1137	navis-NL-HaNA_2.02.14_7828/	9 hours	1 months	1 months	1 weeks	1 weeks	5895 (=)	145.7 (+1.0)
16	1906	*KoK 1906 blz. 1161-2330*	1161-2334	1174	navis-NL-HaNA_2.02.14_7830/	1 days	1 months	1 months	2 weeks	4 days	14891 (=)	260.2 (+0.8)
17	1809	*KoK 1809, scannr. 130-739*	525-912	388	navis-nl-hana_h26506_0556_0001-0739/	1 years	2 weeks	2 weeks	1 years	1 years	-	-
18	1810	*KoK 1810, scannr. 100-822*	1-624	624	navis-nl-hana_h26506_0557_0001-0822/	1 years	2 weeks	2 weeks	2 months	1 months	687 (=)	77.1 (+0.9)
19	1903	*KoK 1903, reg.+blz. 1-1094 XML-layout*	1-1094	1094	navis-H2_7823_0001-1094b/	NonEx	NonEx	NonEx	NonEx	NonEx		
20	1779	*Adm 1779 - 177 1177 [*]*	1-538	538	cliwoc-Adm_177_1177/	9 hours	2 weeks	2 weeks	2 weeks	1 months	1192 (=)	179.6 (+1.0)
21	1779	*Adm 1779 - 177 1189*	1-205	205	cliwoc-Adm_177_1189/	1 years	9 months	9 months	1 weeks	1 weeks	295 (=)	97.2 (+1.0)
22	1783	*Adm/Fam 1136_38 jaar 1783*	1-90	90	cliwoc-Fam_1136_38/	8 months	8 months	8 months	2 months	5 days	1325 (=)	85.4 (+1.0)
23	1760	*Adm_177_1157 jaar 1760*	1-558	558	cliwoc-Adm_177_1157/	9 months	8 months	8 months	3 months	1 days	1055 (=)	101.7 (+1.0)
24	1823	*Mar_D23905_4063 jaar 1823*	1-416	416	cliwoc-Mar_D23905_4063/	1 months	4 months	4 months	1 months	5 days	670 (=)	88.1 (+1.0)
25	1616	*Ubbo Emmius RFH*	1-70	70	ubrug-Ubbo_Emmius_RFH_1616_Omslag_p72/	2 years	2 weeks	2 weeks	NonEx	NonEx		
26	1947	*Qumran scrolls*	1-2463	2463	Qirug-Qumran_extr09/	5 days	2 months	2 months	5 days	5 days	423 (=)	72.9 (+1.0)
27	1400	*Middeldults*	1-4	4	navis-medieval-text-Leuven/	7 months	4 months	4 months	1 months	3 weeks	811 (=)	133.0 (+0.9)
28	1425	*Gelderse rekening 1425*	1-20	20	GeldArch-rekeningen-1425/	4 months	4 months	4 months	2 weeks	1 weeks	1645 (=)	122.4 (-3.2)
29	1421	*Schepenbank Leuven 1421 [*]*	1-784	784	SAL7316/	1 days	1 months	1 months	1 months	3 weeks	4044 (=)	227.6 (+1.0)
30	1457	*Schepenbank Leuven 1457*	1-860	860	SAL7751/	7 months	6 months	6 months	2 months	2 weeks	1433 (=)	107.5 (+1.0)
31	1559	*Schepenbank Leuven 1559*	1-1373	1373	SAL7453/	7 months	6 months	6 months	3 months	3 days	1460 (=)	80.5 (+1.0)
32	1779	*Adm 1779 - (test)*	100-272	173	cliwoc-Adm_177_1177b/	1 months	1 months	1 months	2 months	11 months	274 (=)	727.7 (+1.0)
33	1932	*P.V.Astro - 1932 [*]*	1-3	3	PV-Astro-1932/	1 days	1 months	1 months	1 weeks	4 months	275 (=)	129.5 (+1.0)
34	1855	*GV.Appingedam - 1855 [*]*	1-239	239	GV-Appingedam_8149/	6 days	1 months	1 months	2 months	2 weeks	1807 (=)	295.2 (+1.0)
35	1625	*Sleninitialen*	1-20627	20627	Sleninitialen/	1 years	2 weeks	2 weeks	NonEx	NonEx		
36	1672	*F155 [*]*	1-262	262	MoscowA-F155-1N7-1/	4 months	4 months	4 months	2 months	1 weeks	319 (=)	67.4 (+1.0)
37	1674	*Cuper_Braun*	1-91	91	Cuper_Braun/	3 weeks	2 months	2 months	2 weeks	5 days	1405 (=)	101.3 (+1.0)
38	1682	*Cuper_Altri*	1-922	922	Cuper_Altri/	1 months	1 months	1 months	1 months	1 weeks	1985 (=)	62.2 (+1.0)
39	1201	*Chronicon Boemorum 1-151*	1-151	151	KNMP-VIII_F_69_____2C2O/	1 days	3 hours	3 hours	1 days	1 days	1179 (=)	26.1 (-5.0)
40	1627	*RSG 1627 [*] scan 516 f.f.*	1-96	96	RSG_1627/	3 weeks	4 weeks	4 weeks	4 weeks	1 days	1754 (=)	10.7 (+1.0)
41	1300	*MPS 1300 [*]*	1-95	95	MPS1300/	13 hours	1 days	1 days	9 hours	10 minutes	807 (+3)	9.3 (+0.7)
42	1325	*MPS 1325*	1-141	141	MPS1325/	19 hours	1 days	1 days	9 hours	36 minutes	958 (=)	11.4 (+1.0)
43	1350	*MPS 1350*	1-89	89	MPS1350/	25 seconds	13 hours	13 hours	9 hours	1 days	1091 (+18)	4.7 (+0.9)
44	1375	*MPS 1375*	1-84	84	MPS1375/	13 hours	1 days	1 days	9 hours	4 minutes	710 (+1)	9.9 (+1.0)
45	1400	*MPS 1400*	1-196	196	MPS1400/	7 hours	8 hours	8 hours	8 hours	2 hours	990 (+16)	9.3 (+0.9)
46	1425	*MPS 1425*	1-198	198	MPS1425/	24 seconds	18 hours	18 hours	9 hours	10 hours	991 (+10)	8.3 (+0.9)
47	1450	*MPS 1450*	1-245	245	MPS1450/	4 hours	1 days	1 days	9 hours	2 seconds	973 (+2)	9.9 (+0.9)
48	1475	*MPS 1475*	1-205	205	MPS1475/	1 days	15 hours	15 hours	9 hours	1 days	1015 (+1)	9.8 (+0.9)
49	1500	*MPS 1500*	1-141	141	MPS1500/	0 seconds	5 hours	5 hours	5 hours	16 hours	996 (=)	9.9 (+1.0)
50	1525	*MPS 1525*	1-115	115	MPS1525/	0 seconds	2 hours	2 hours	2 hours	3 minutes	985 (+13)	10.6 (+0.9)
51	1550	*MPS 1550*	1-113	113	MPS1550/	24 seconds	1 days	1 days	9 hours	17 minutes	987 (+23)	9.8 (+0.7)
52	1605	*HulsBergh_0214_7268*	1-76	76	HulsBergh_0214_7268	1 weeks	1 weeks	1 weeks	1 weeks	1 weeks	558 (=)	4.8 (+1.0)

Monk collection on 02-04-14

52 'books'

and growing fast

On 28-10-14 72 'books'

Static index example: Chronicon Boemorum (1201 AD, Cosmas)

Monk search engine for handwritten words - KNMP collection -

KNMP/CB - Chronicon Boemorum, Cosmas, p. 1-151
boek KNMP-VIII_F_69_____2C2O(1201).

L. Schomaker, ALICE, Rijksuniversiteit Groningen/ Tomas Klimek, National Library of the Czech Republic, Monasterium (Last update of these static tables: Thu Jan 30 17:41:00 CET 2014)

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z @ a b c d e f g h i j k l m n o p q r s t u v w x y z

@

@ampersand

A

Adhec
Adhuc
Adque
Anno
Annoncanta
Aut

B

Babenberk
Bawarios
Benedict
Boemie
Boemii
Boemo

Piura
Placet
Plura
Podium
Polonias
Polonie
Poloniensi
Polonios
Pomora
Post
Postera
Praga
Pridie

Q

Qua
Quare
Que
Qui
Quid
Quinto

aspicias
aspicio
assistebant
auctoritas
auctoritate
auderem
audientes
aura
aure
auro
auru
ausis
aut
auta

b

barba
bawaria
bello
bene
benediccione

dedisse
dedita
defensio
defraudare
delphi
denario
denegare
depilet
deposuit
descendere
deunde
diaboli
diadema
dicens
dicere
dictare
dicto
dicunt
didit
die
dignitas
dignitate

familia
famula
fatis
fecerit
fecunt
felicit
felix
femina
ferata
ferre
ferro
feruus
festina
festis
fiat
fide
fidele
fideliores
fidelitate
fides
filia
filio

huc
hui
huic
huma
humani
humanitatis
hunc
hystoriaca

i

iacet
ibi
ille
illi
illicitis
illos
illuc
illustri
impetrare
impetu
infacto

memorat
memorati
memoria
memorie
menia
meo
minor
mittere
modis
molestia
monens
monete
mons
mora
mordax
mortalis
morte
mortis
mortuos
mortuus
mox
moxq

patro
patroni
patronis
patu
pauca
paucis
pauco
pax
pectore
pectu
pecunia
pede
pena
penas
peticione
pia
pla
placeat
placent
placita
placuit
plaustris

romano
rota
rubentes

s

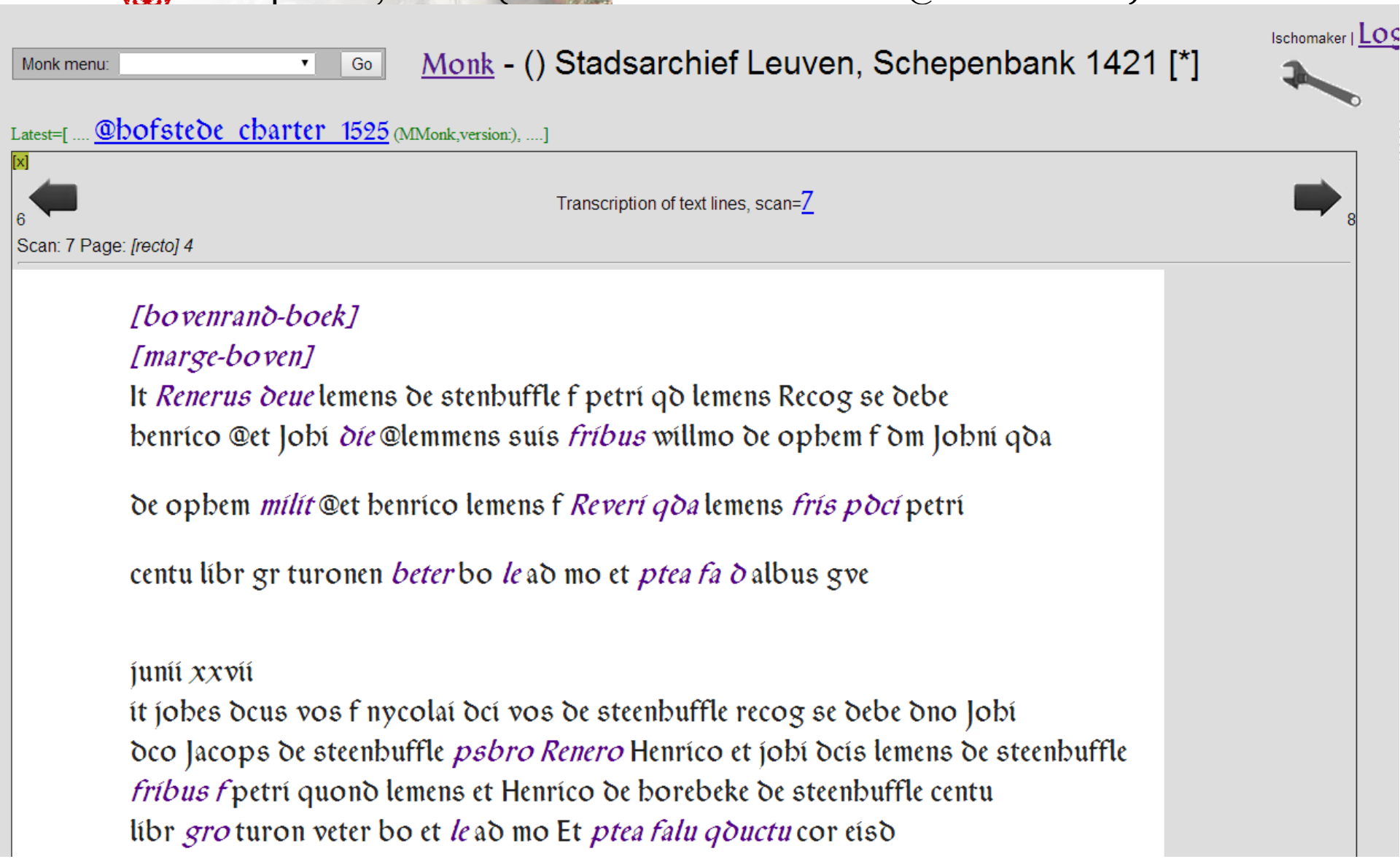
sacra
sacris
salutat
salute
sanati
sapiens
sciencia
scient
scientes
scriptura
scripture
seccessit
secunda
sed
sede
sedis

uolo
urbe
urbes
urbio
urbis
usus
utero
utro

v

vocabit
vundat

Line transcriptions , actual (literal and text-rendered @Monk-codes)



Note: some @Monk-codes for word shapes have not been codified here, the user can click on them to do that

Line transcriptions , decontracted, literal and text-rendered @Monk-codes

[bovenrand-boek]
[marge-boven]
•Item *Renerus deveri* Lemmens de Stenhuffle filius Petri quondam Lemmens recognovit se debere
Henrico et Johanni *deveri* Lemmens suis *fratribus* Willelmo de Ophem filius dominus Johanni quondam

de Ophem *militer* et Henrico Lemmens filius *Reveri quandam* Lemmens *fratris predicti?* Petri

centum libras grossorum Turonensium *veterum* bonas *legales* ad monicionem et *preterea famulum conductum* Albus Graven

junii xxvii
•Item Johannes dictus Vos filius Nycolai dicti Vos de Steenhuffle recognovit se debere domino Johanni
dicto Jacops de Steenhuffle *presbytero Renero* Henrico et Johanni dicitis Lemmens de Steenhuffle
fratribus filiis Petri quondam Lemmens et Henrico de Horenbeke de Steenhuffle centum
libras *grossorum* Turonensium veterum bonas et *legales* ad monicionem et *preterea famulum conductum* coram eisdem
•Item Johannes Vos pistor in presentia promisit Quintinum de Valkenborch pistorem
ac domum et curtem cum suis pertinentibus eiusdem Quintini site in Dorpstrata inter bona Willelmi Nagel
et bona Johannis de Troembeke *exeuntes* cum una parva curte in Bakeleynstrata inter

Line transcriptions , word-by-word translation from Latin to Dutch

[bovenrand-boek]
[marge-boven]
•Item *Renerus deveri* Lemmens de|van Stenhuffel zoon van Peter voorheen Lemmens heeft erkend zich,dat hij,dat zij moeten
Hendrik en (aan)Johannes *deveri* Lemmens zijn(bez.) *fratribus* Wim de|van Ophem zoon heer van Johannes voorheen

de|van Ophem *militer* en Hendrik Lemmens zoon *Reveri quandam* Lemmens *fratris predicti?* van Peter

100 ponden grote van Tours *veterum* goede|goederen *legales* bij|tot|naar|op waarschuwing en *preterea famulum conductum* Albus Graven

juni 27
•Item Johannes genoemd Vos zoon Nycolai genoemd Vos de|van Steenhuffle heeft erkend zich,dat hij,dat zij moeten aan de heer (aan)Johannes
genoemd Jacops de|van Steenhuffle *presbytero Renero* Hendrik en Johannes genoemd(pl.) Lemmens de|van Steenhuffle
fratribus filiis van Peter voorheen Lemmens en Hendrik de|van Horenbeke de|van Steenhuffle 100
ponden *grossorum* van Tours oude goede|goederen en *legales* bij|tot|naar|op waarschuwing en *preterea famulum conductum* hoofdelijk aan dezelfden
•Item Johannes Vos bakker in aanwezigheid heeft beloofd Quintinus de|van Valkenborch bakker
ac huis en hof|tuin met zijn(bez.) aanhorigheden van dezelfde Quintinus gelegen in Dorpstraat tussen goed|goederen Willem Nagel
en goed|goederen van Johannes de|van Troembeke *exeuntes* met een klein|kleine hof|tuin in Bakeleynstraat tussen
goed|goederen van dezelfde Willem Nage en goed|goederen aan meester Hubertus voorheen de|van *aquis presbyteri indempnes penitus*
bewaren en *relevare* de|van *illo modio siliginis mensuris* Leuvense *annue pensionis quam* heer
Johannes Sampain *presbyter quo* bij|tot|naar|op leven *suam* heeft bij|tot|naar|op goed|goederen voornoemde en bij|tot|naar|op *maiorem*
securitatem genoemd Quintinus de|van *premissa facundia prefatus* Johannes Vos 3 gouden

Word-shape codification: attributes of a literal or @Monk-code

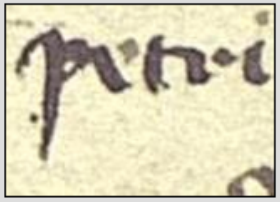
Monk menu: ▾ Go

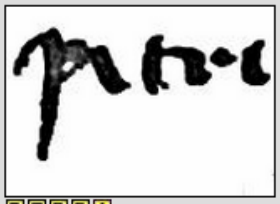
Monk - Word-shape codification () *

Latest=[.... hebben (MMonk,version:), ...]

Labels for words, characters, syllables or special symbols need to be described in detail. You entered this page because you have typed an hitherto unknown word label or because you wanted to modify the current definition.

Example [navis-SAL7316_0019-line-0 zone-HUMAN-x=1784-y=0057-w=0146 nink=2452-segm=COCOS6cocos]





User	Ischomaker
Unique Symbolic Monk Code (label)	petri
Taboo this label?	☐
Take this ZonID as example?	☐
ASCII characters, intended spelling	Petri
Characters, variant spelling*	Petri
LaTeX encoding	Petri
Html diacritics code	Petri
Translation	van Peter
Visual type	Handwriting ▾
Category	Proper (given) name (e.g., John) ▾
Comment	Guys, the second stroke of the e and the dot on the letter i may have been removed by the despeckler but the word is still quite unique so I don't expect many problems.
Web site (URL) with info	

Test this URL

Back Save

Rules

1. Please do not try to be hypercorrect by entering diacritics in the main label: it is 7-bit USASCII only.
2. Ligatures and special pattern codes start with '@'
3. Do not enter ASCII special characters, write them out in @CODE notation:
   '/' should be @SLASH

provisional transcriptions. The text shown comes either from human line transcription or (human) confirmed recognized words (see ...)

Cuper Braun 0014

6 [R] IOANNI BRAVNIO (continued) ut patet ex Liv. ...

... et ... 3 ...
tintinnabula illa summi sacerdotis ex ore fuisse, quae tü fuerunt ex auro

... B ... quibus ... u ...
vocant et quia oris in saevis Bracchii, in quibus et nutrices @GR_ chalko-
... quia ... qui ... a ...
@GR_ drutai sacro sine dubio noie vocantur; plurimus usus, strepiturque
... g ... ipse_doces ... p ... fuisse ...
undem Deü coli= et ipse doces p 96. Solenne Ethicus fuisse Bacchi
sacra celebrare, oneis crepidis et cymbalis ejusmode calciamenta
habuisse Empedoclem, teste Tertulliano c.q. de pallio; i Andque probas
... @Epsilon ... s ...
auctoritate Plutarchi; apud quem tü l. 6.Sym. quë laudas, nihil tale
auctoritata Plutarchi, apud quem tü l.6.Sym, que laudas, nihil tale
repperi. Sane tintinnabulorü in vestibus, ut apud Judaeos ...., usum
apud veteres fuisse nullibi reperio. Lego quidem apud M. Paulum
venetum l.2.c.23 cursores Tartarü Regium habere, cinetoria
... nihil ... ad ...
sonoris tintinnabulis plena gestantes, sed hoc nihil ad rem
... n ... a ...
unde o abs re est, quod putem comparatione H m
... @o_- ... a ... quod ... ut ... u ... m ...
strepituum, quos .... tintinnabulis, Bacchantes oreis cymbalis edunt
... quo ... s ... n ...
... b ...
instituere, nec illü dixisse Sacerdotë Bacchi et tintinnabula habuisse,
illa autem @GR_ ho deiknumenos en tois enantiois tou meteooron thursos
... a ... bene ... n ...

Conclusions

- Great progress is made in handwritten text retrieval in recent years
- The concept of iterative refinement by human labeling & computing really works
  ➔ interactive machine learning over Internet
- Image preprocession and layout analysis are important, still crude and not fundamentally solved
- Image-quality standards (NL norm: *Metamorfoze*) do not guarantee high-performance handwritten text recognition
- Approach works for several historical periods (1200, 1400, 1600, 1800): but each needed some minor tweaking of the image preprocessing
- Think big!

How to improve on human and machine transcriptions?

The legal transactions in the acts of the Leuven Alderman's rolls, 1421: Towards a formal semantic model of acts (Marianne Ritsema van Eck, august 2011). Presented at Digital Humanities 2012

➔ experts fight over the transcription
➔ students need to know more than paleography
➔ the machine needs a model for top-down reasoning

Semantic modeling and layout of acts: example: 'IOU'

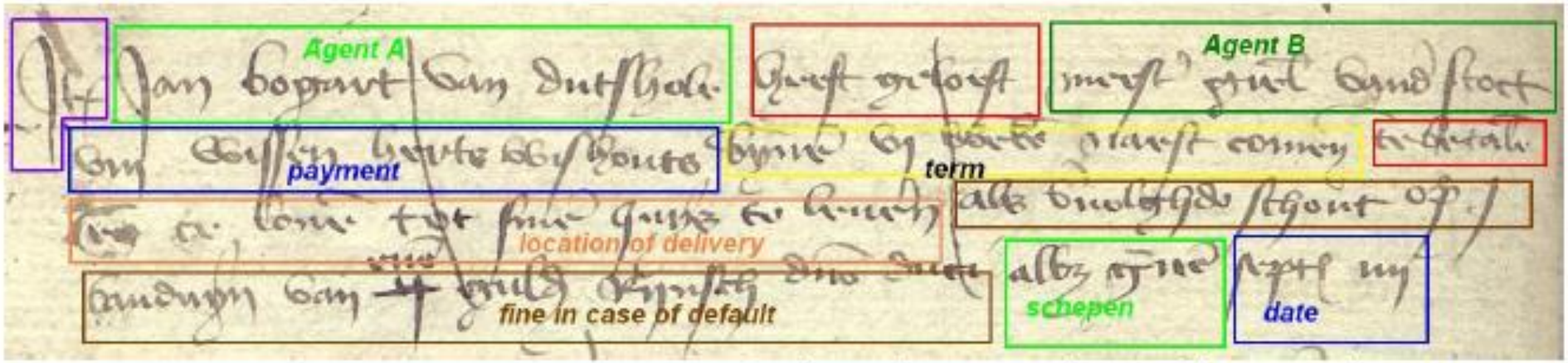
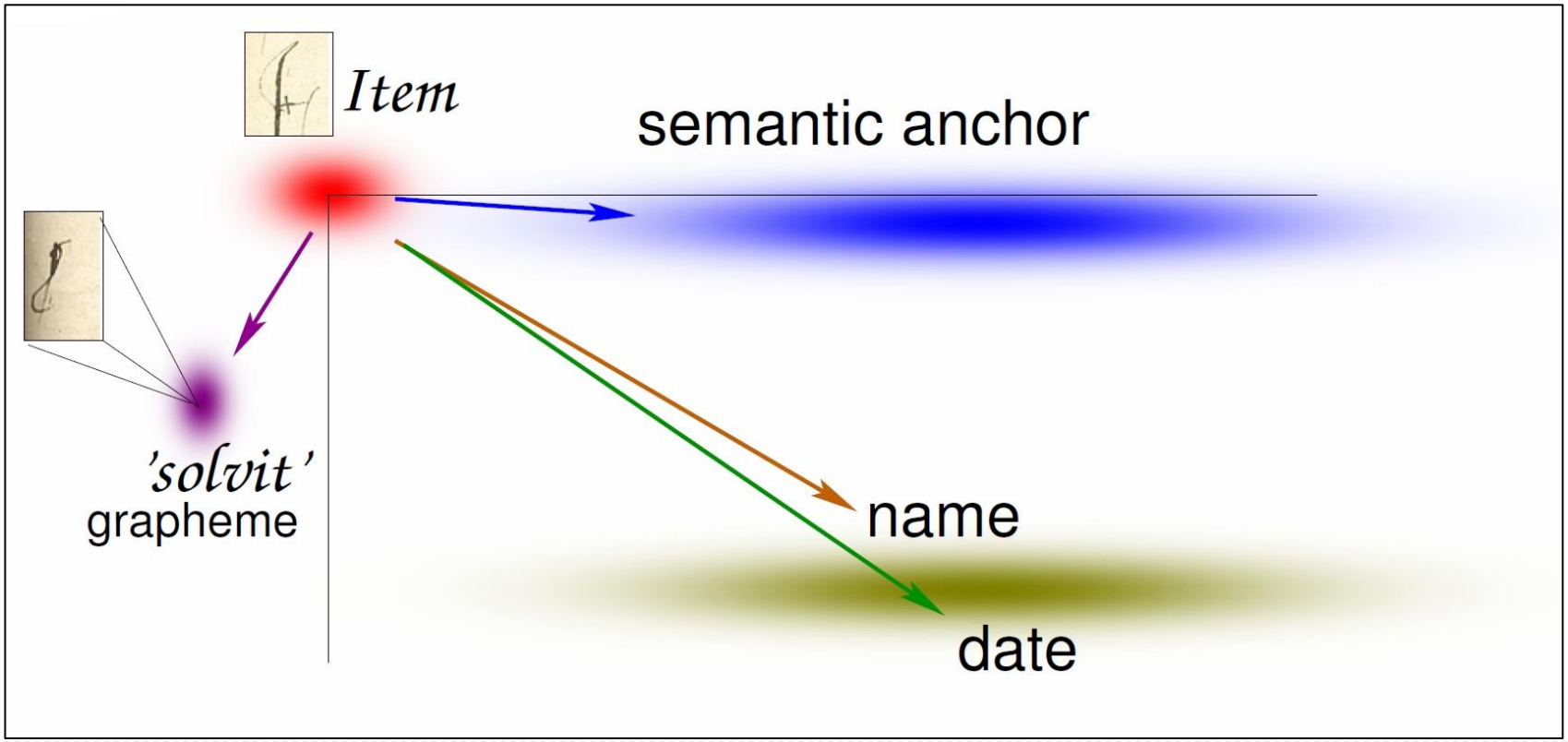
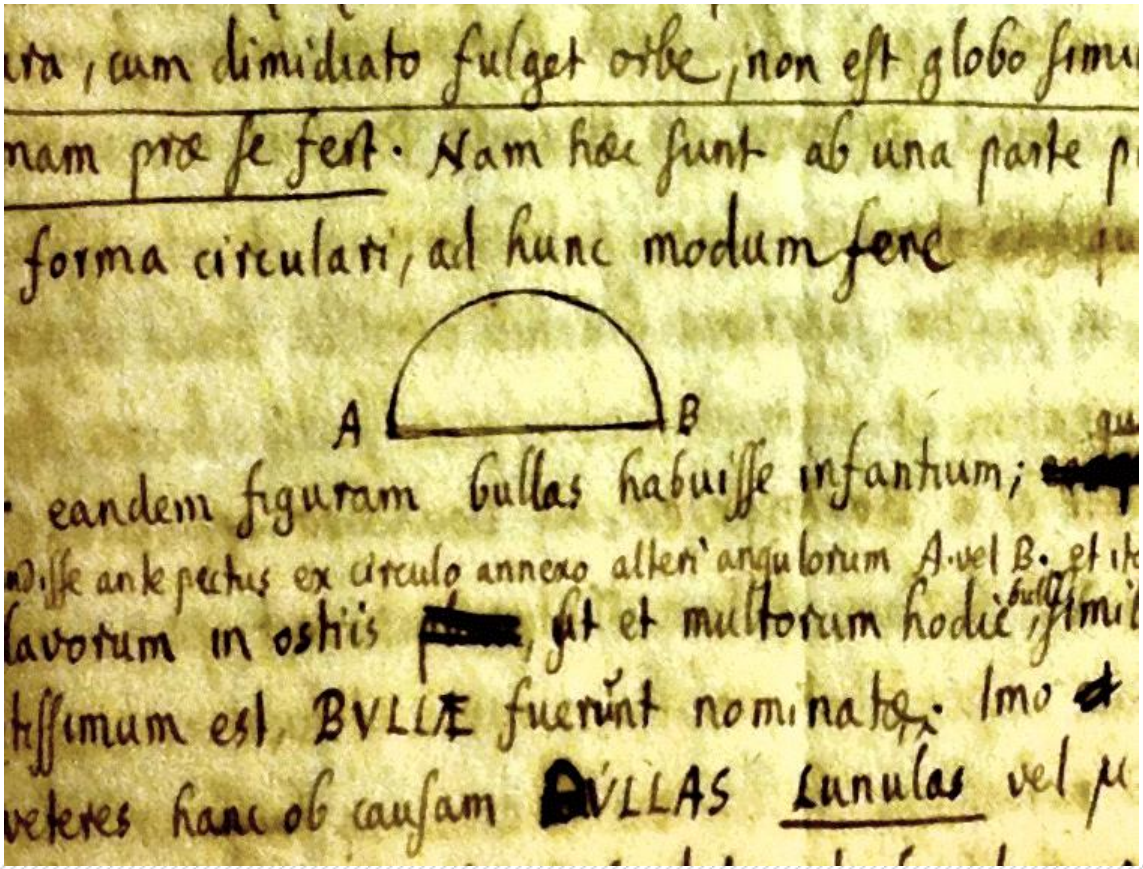


Fig. 4.9 (MS page V64.1 / scan 128) The relevant regions of interest, that correspond to the structural elements the act is composed of, are indicated on the MS page with bright colours.

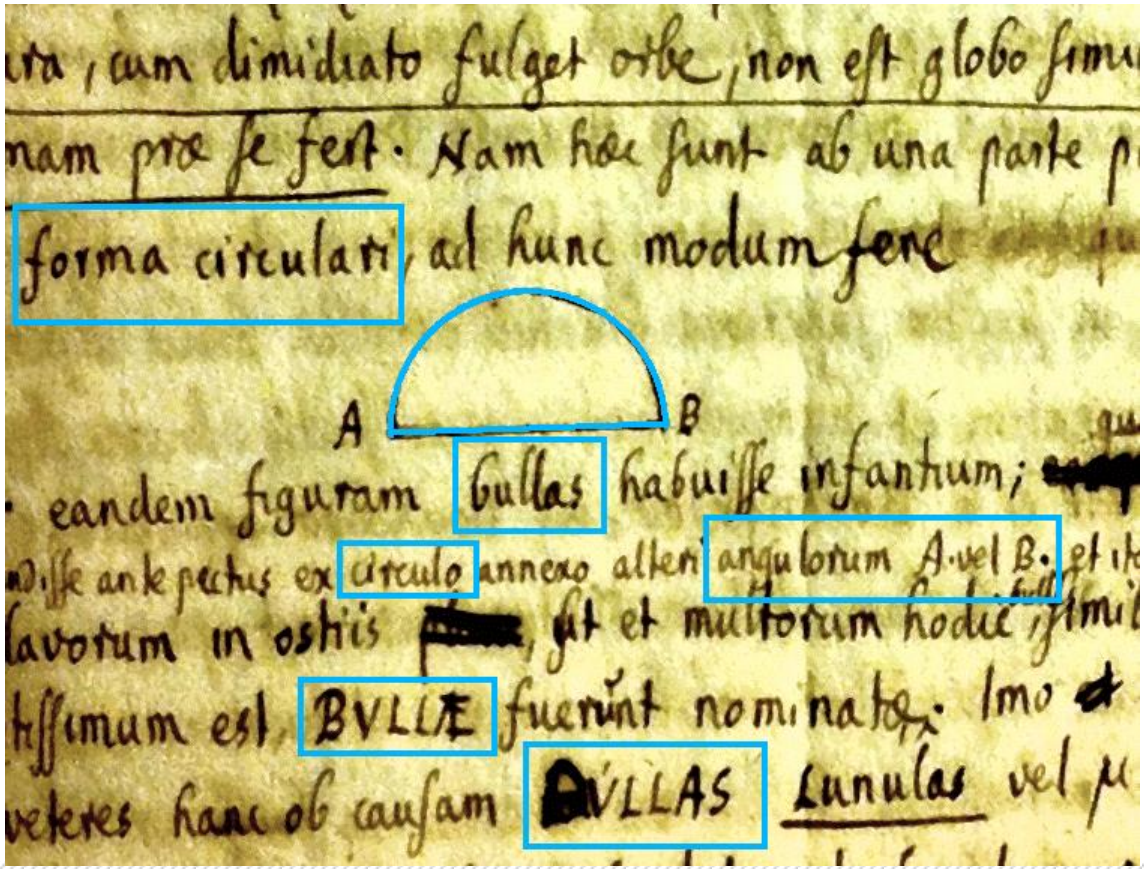
Spatial modeling: Layout vs Semantics in the acts of the Schepenbank



Digging into Data / Global currents -- US/Canada/NL project with
Mohamed Cheriet of Ecole de technologie supérieure(Montréal)
Andrew Piper (McGill University) (principal investigator)
Elaine Treharne (Stanford University)



Digging into Data / Global currents -- US/Canada/NL project with
Mohamed Cheriet of Ecole de technologie supérieure(Montréal)
Andrew Piper (McGill University) (principal investigator)
Elaine Treharne (Stanford University)



Conclusions

- Great progress is made in handwritten text retrieval in recent years
- The concept of iterative refinement by human labeling & computing really works
  ➔ interactive machine learning over Internet
- Image preprocession and layout analysis are important, still crude and not fundamentally solved
- Image-quality standards (NL norm: *Metamorfoze*) do not guarantee high-performance handwritten text recognition
- Approach works for four historical periods (1200, 1400, 1600, 1800): but each needed some minor tweaking of the image preprocessing
- Think big!