Glyph Miner:
A System for Efficiently Extracting Glyphs from Early Prints in the Context of OCR

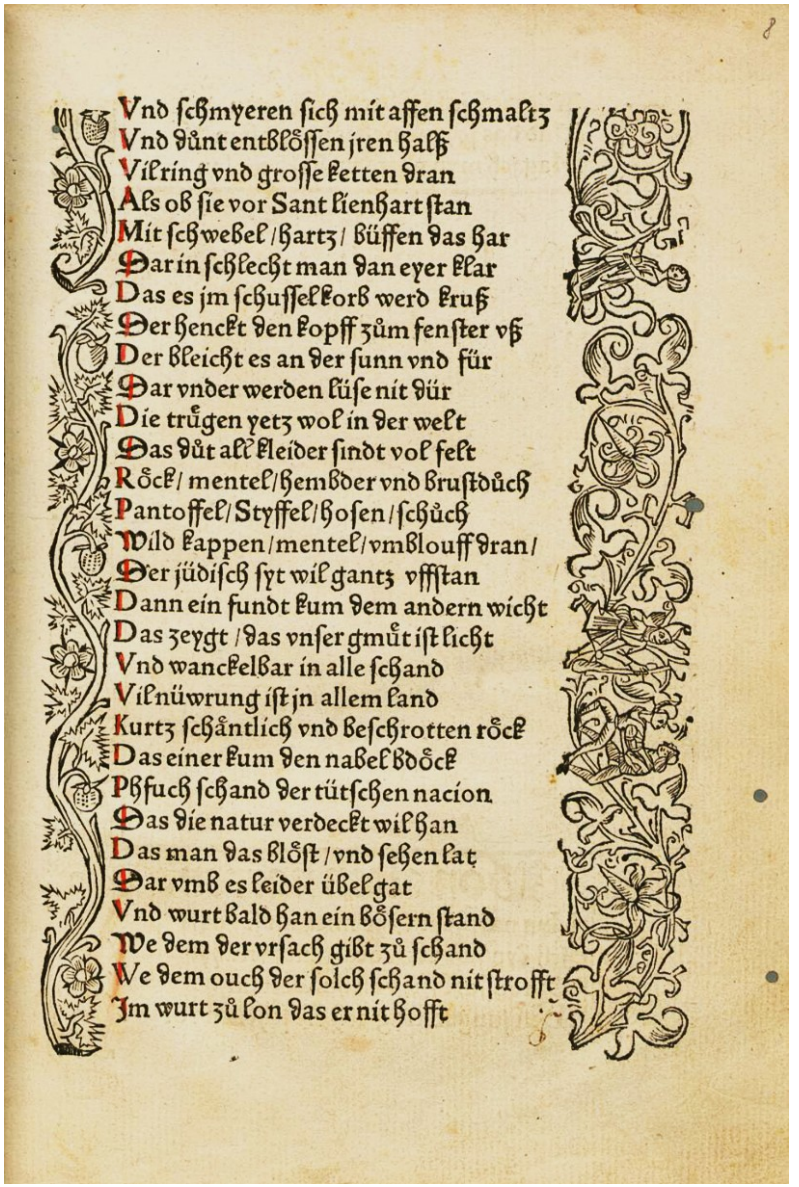
Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Benedikt Budig
Thomas C. van Dijk
Felix Kirchner

Incunables: What are they?

- Early prints (before 1500)
- Movable type
- Design resembles medieval handwritings

- Circa 30,000 editions

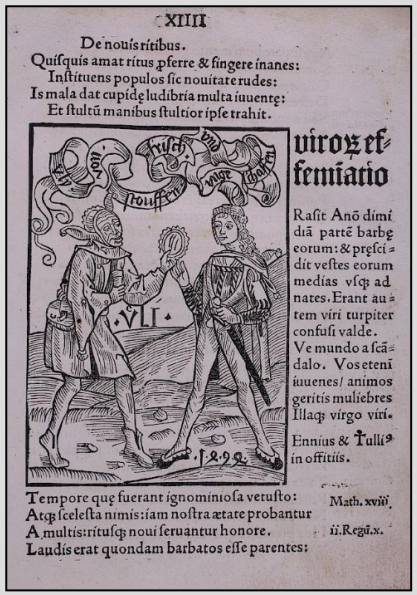


The *Ship of Fools*

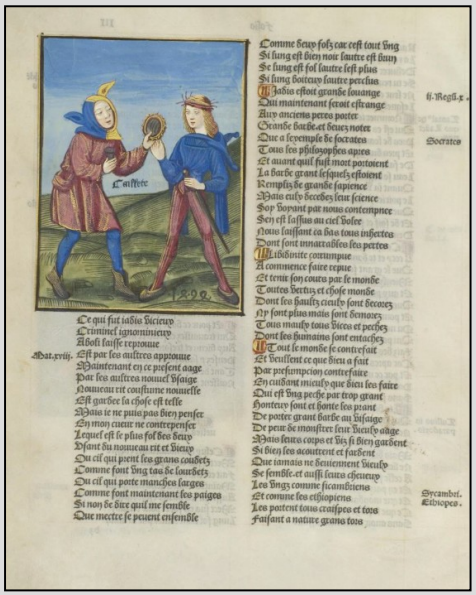
- Incunables of Sebastian Brant's *Narrenschiff*
- One of the most popular books in the early modern period

Basel 1494
German

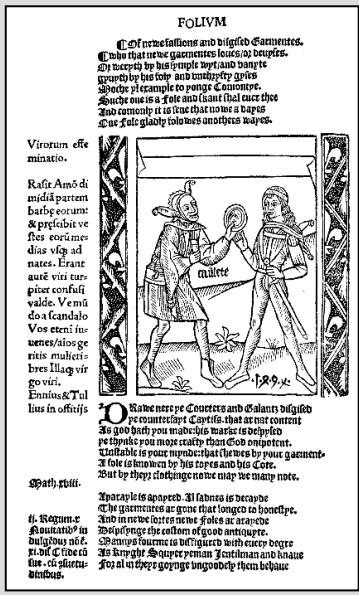
Nürnberg 1494
German



Basel 1498
Latin



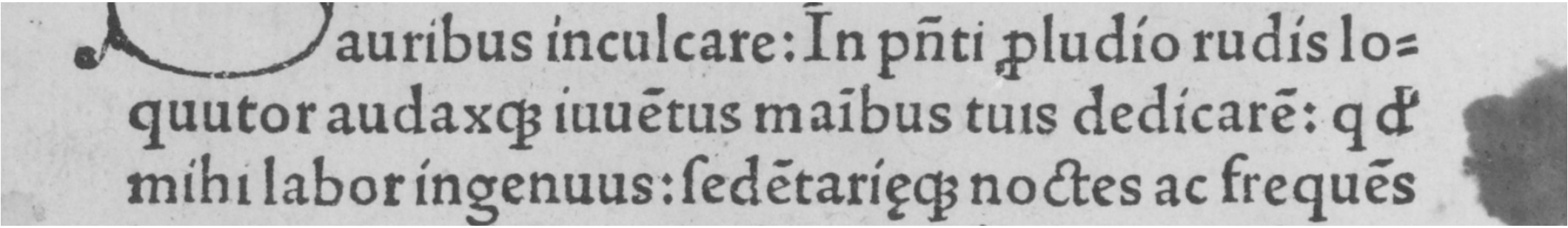
Paris 1497
French



London 1509
English

Challenges for OCR

- Many different glyphs: abbreviations, ligatures, diacritics
- Variance in printing, poor conservation state


auribus inculcare: In pñti ꝑludío rudís lo=
quutor audaxq; iuuẽtus maĩbus tuis dedícarẽ: q d'
mihi labor íngenuus: ſedẽtarięq; noctes ac frequẽs

- Off-the-shelf OCR software fails

  → Train general purpose OCR software (e.g. Tesseract or OCRopus)

requires **training data**

Introducing *Glyph Miner*

- Objective: obtain training data for OCR software
- Existing practice: e.g. Aletheia, Franken+, Gamera

- Mark example glyph on arbitrary page

  → get occurrences from all pages of the print
- Use occurrences to create realistic training images

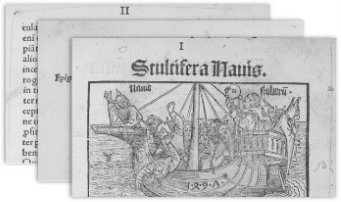
Das Narrenschiff

Collection

GW5042

Sebastian Brant — 1494

Select



Stultifera Navis

Collection

GW5061

Sebastian Brant — 1497

Select

Add new document

Title: Title

Subtitle: Subtitle

Author: Author

Year: Year

Signatur: Signature

Collectio:

Image (color): Select File

Image (b/w): Select File

Upload document

Add new collection

Title: Title

Subtitle: Subtitle

Author: Author

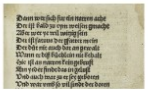
Year: Year

"Das Narrenschiff"

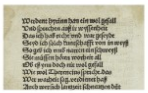
0001v
Size: 2015 x 3009
Select

0001r
Size: 2015 x 3010
Select

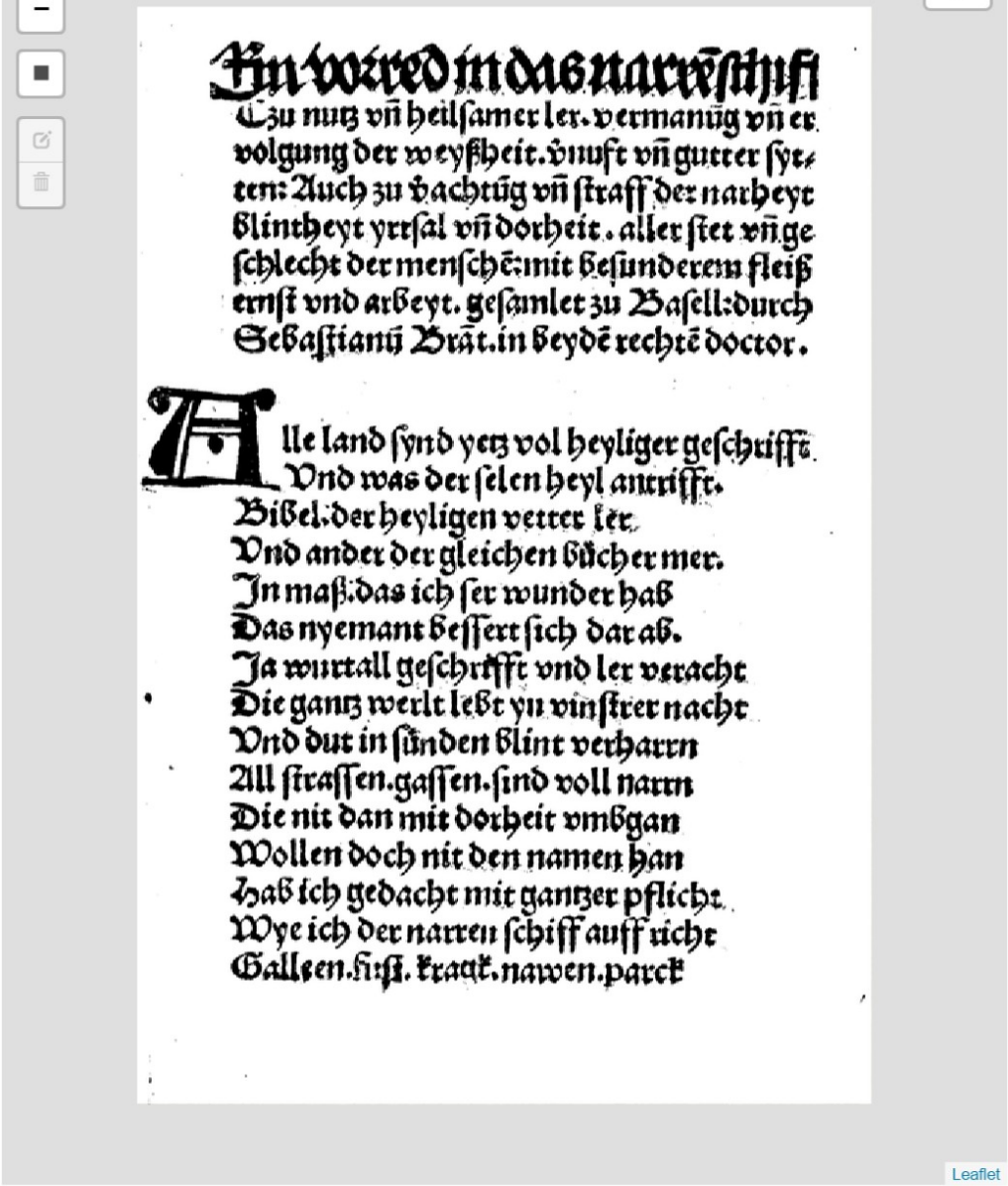
0002v
Size: 2015 x 3009
Select



0002r
Size: 2015 x 3010
Select



0003v
Size: 2015 x 3010
Select



Leaflet

New template | Existing templates

Top Left: x = 0, y = 0

Width = 0 px, Height = 0 px

Bottom Right: x = 0, y = 0

Character Unicode Character (

Pick new template | Search

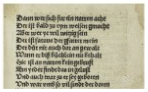
Detect Threshold

"Das Narrenschiff"

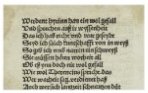
0001v
Size: 2015 x 3009
Select

0001r
Size: 2015 x 3010
Select

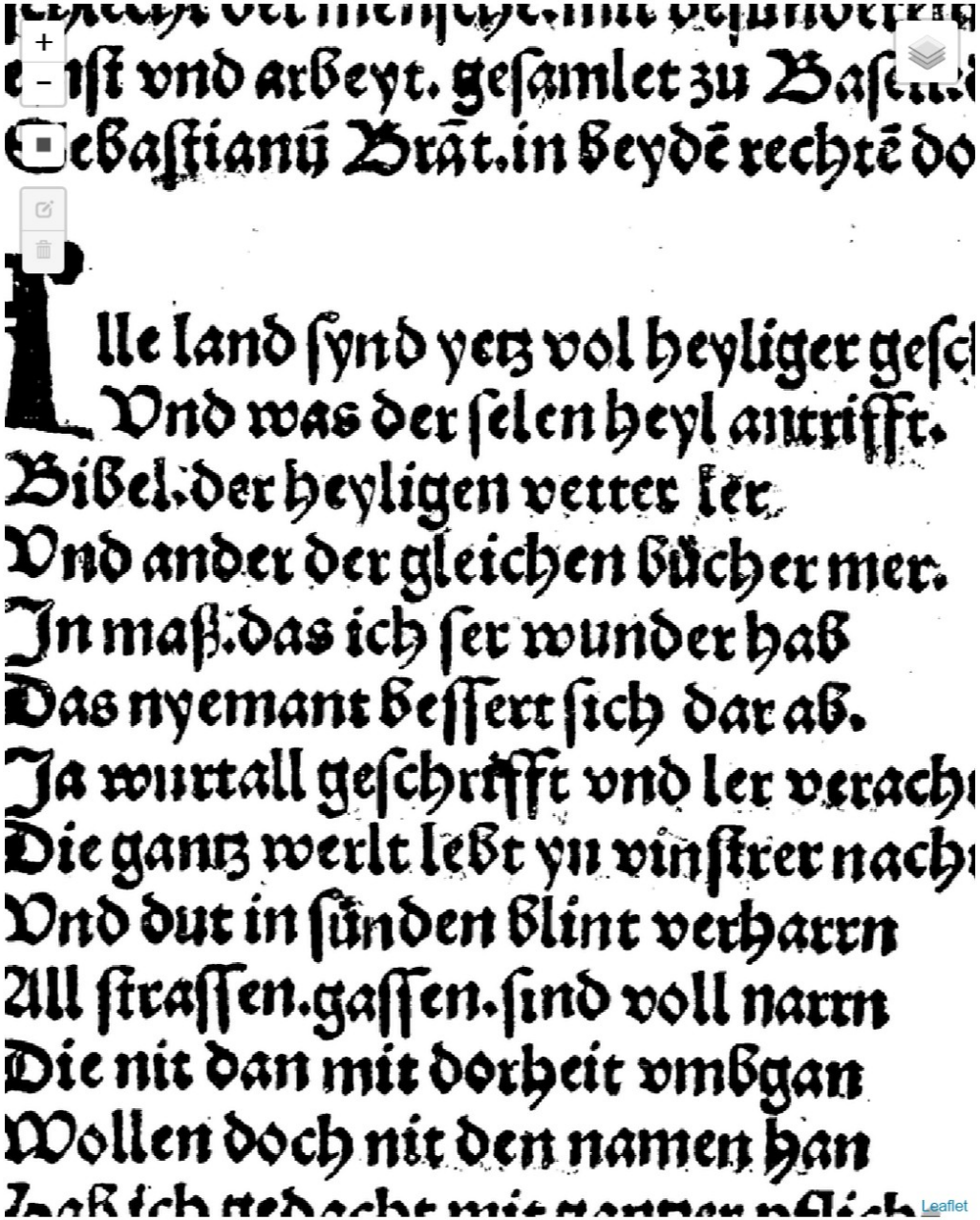
0002v
Size: 2015 x 3009
Select



0002r
Size: 2015 x 3010
Select



0003v
Size: 2015 x 3010
Select



New template | Existing templates

Top Left: x = 0, y = 0

Width = 0 px, Height = 0 px

Bottom Right: x = 0, y = 0

Character

Unicode Character (

Pick new template

Search

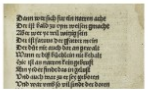
Detect Threshold

"Das Narrenschiff"

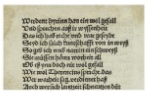
0001v
Size: 2015 x 3009
Select

0001r
Size: 2015 x 3010
Select

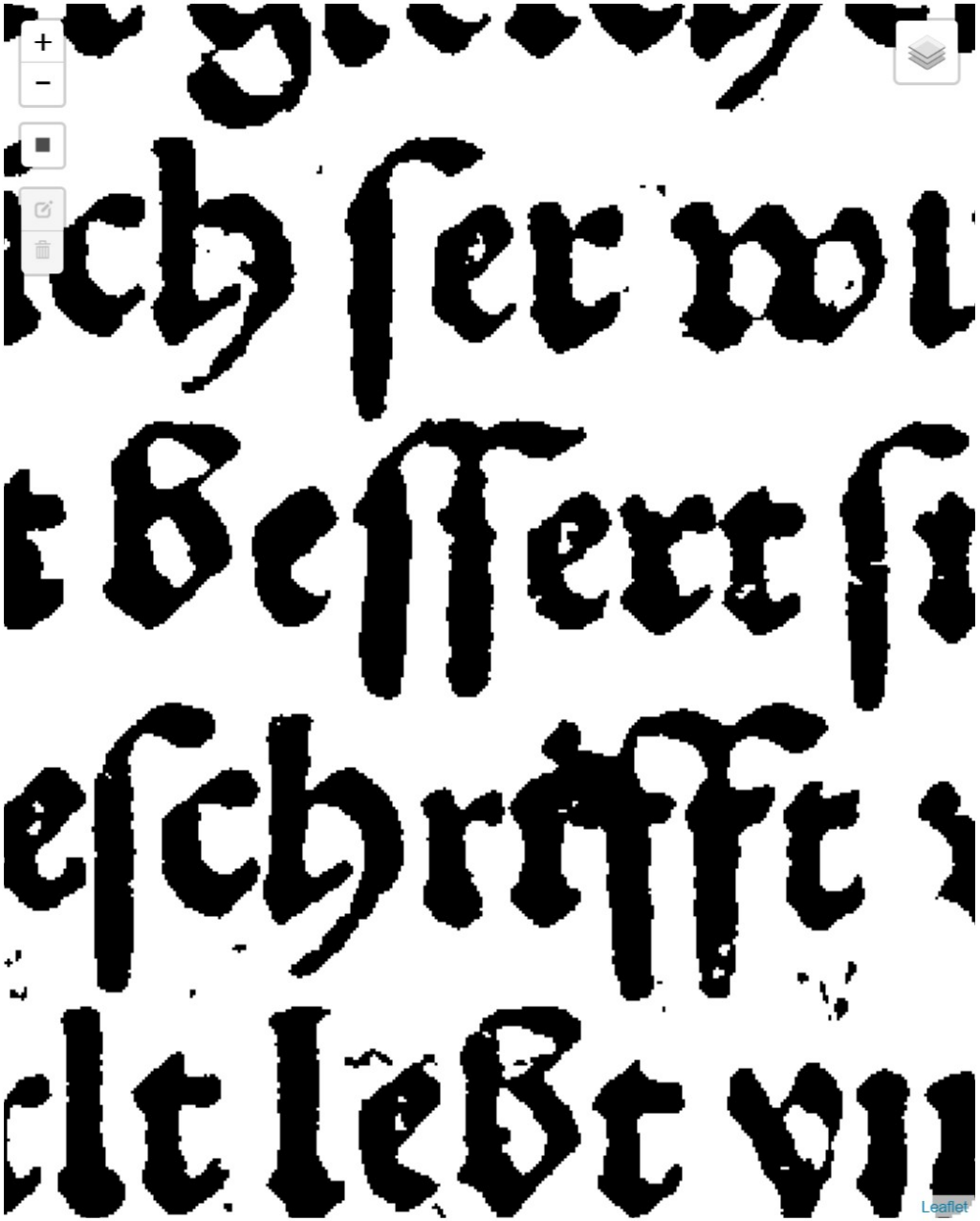
0002v
Size: 2015 x 3009
Select



0002r
Size: 2015 x 3010
Select



0003v
Size: 2015 x 3010
Select



New template | Existing templates

Top Left: x = 0, y = 0

Width = 0 px, Height = 0 px

Bottom Right: x = 0, y = 0

Character

Unicode Character (

Pick new template

Search

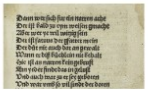
Detect Threshold

"Das Narrenschiff"

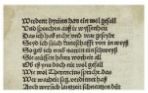
0001v
Size: 2015 x 3009
Select

0001r
Size: 2015 x 3010
Select

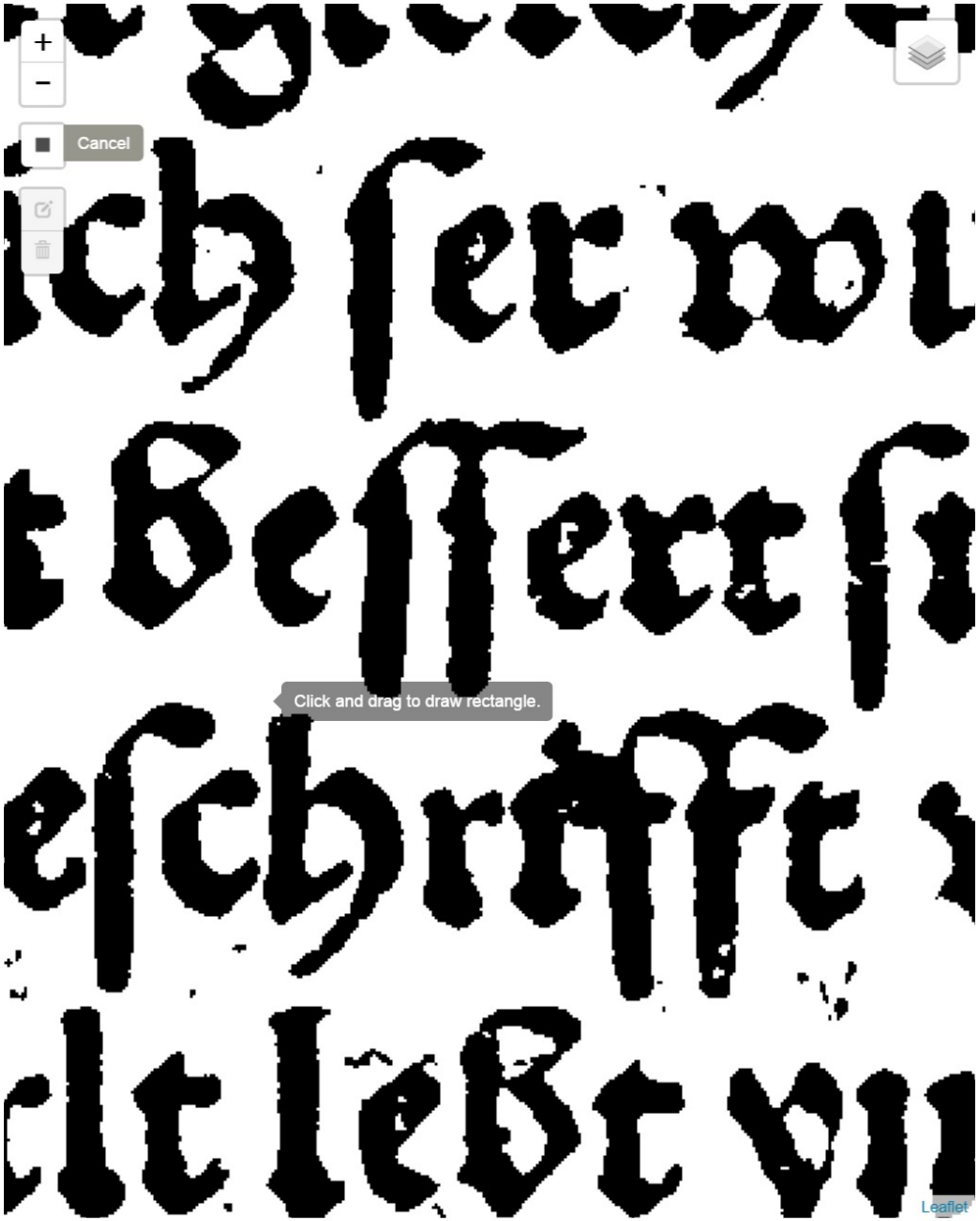
0002v
Size: 2015 x 3009
Select



0002r
Size: 2015 x 3010
Select



0003v
Size: 2015 x 3010
Select



New template | Existing templates

Top Left: x = 0, y = 0

Width = 0 px, Height = 0 px

Bottom Right: x = 0, y = 0

Character

Unicode Character (

Pick new template

Search

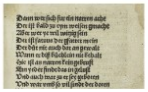
Detect Threshold

"Das Narrenschiff"

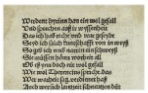
0001v
Size: 2015 x 3009
Select

0001r
Size: 2015 x 3010
Select

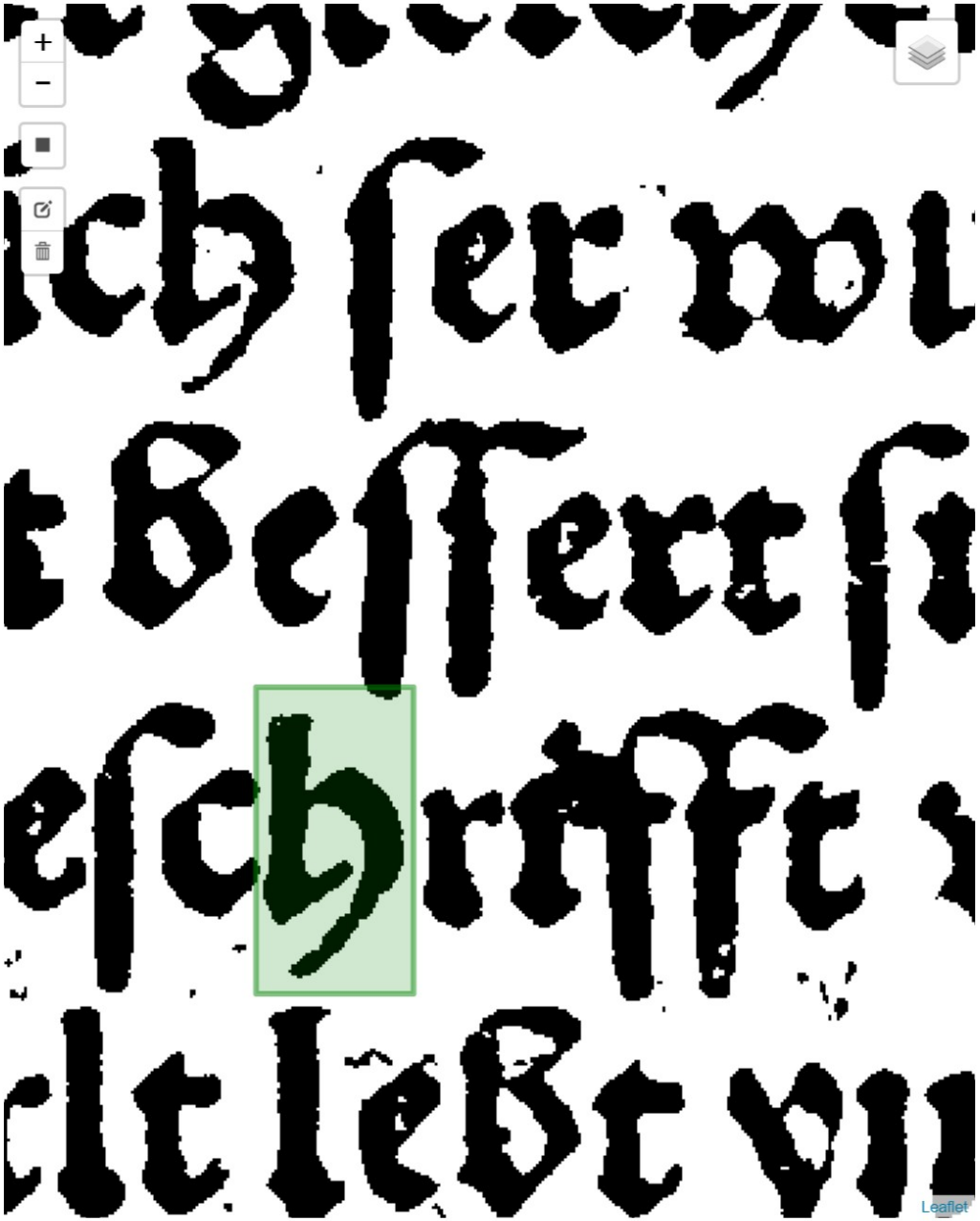
0002v
Size: 2015 x 3009
Select



0002r
Size: 2015 x 3010
Select



0003v
Size: 2015 x 3010
Select



New template | Existing templates

Top Left: x = 855, y = 1760

Width = 56 px, Height = 109 px

Bottom Right: x = 911, y = 1869

Character Unicode Character (

Pick new template | Search

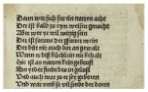
Detect Threshold

"Das Narrenschiff"

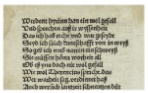
0001v
Size: 2015 x 3009
Select

0001r
Size: 2015 x 3010
Select

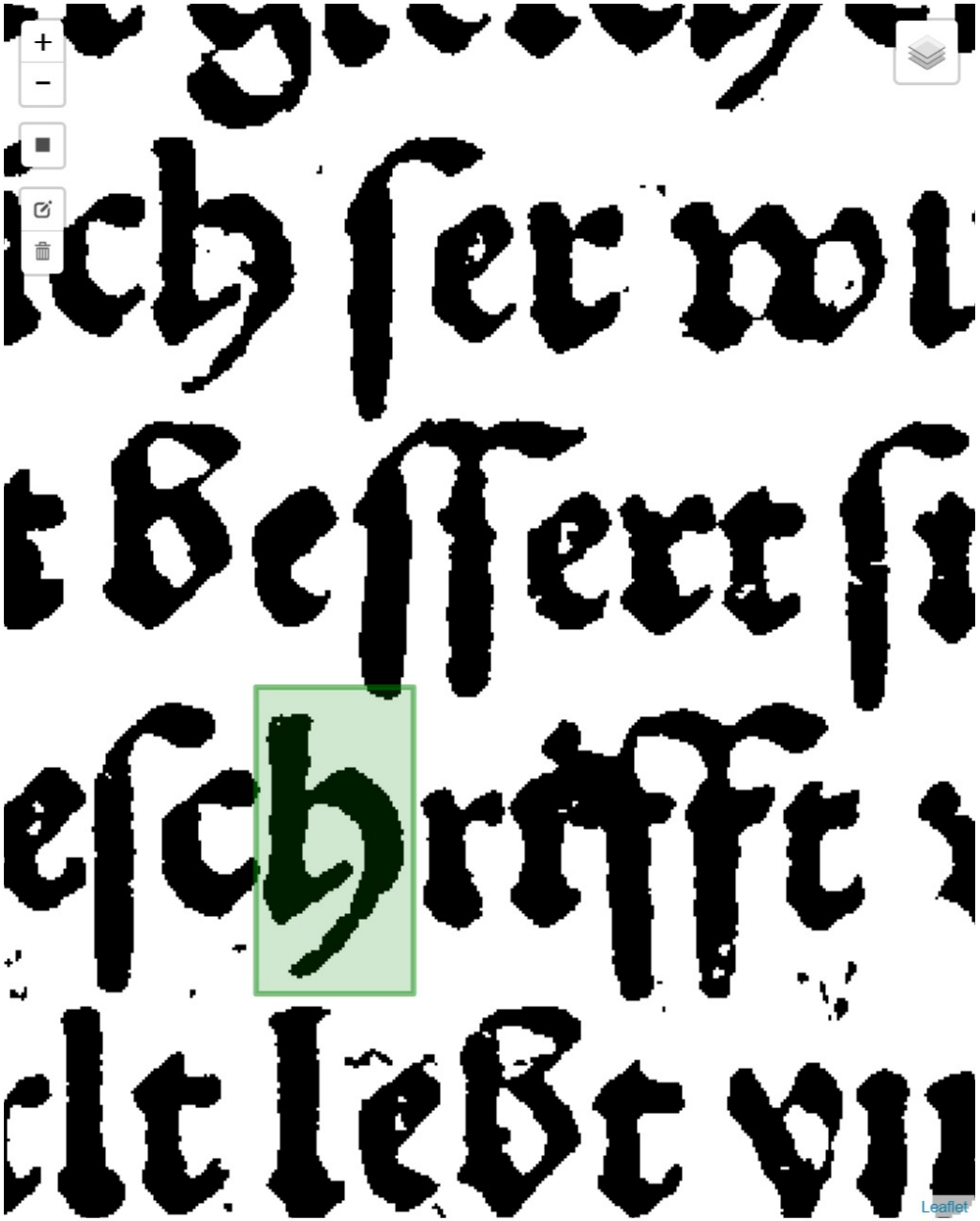
0002v
Size: 2015 x 3009
Select



0002r
Size: 2015 x 3010
Select



0003v
Size: 2015 x 3010
Select



New template | Existing templates

Top Left: x = 855, y = 1760

Width = 56 px, Height = 109 px

Bottom Right: x = 911, y = 1869

Character h

Pick new template | Search

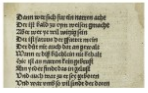
Detect Threshold

"Das Narrenschiff"

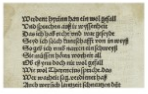
0001v
Size: 2015 x 3009
Select

0001r
Size: 2015 x 3010
Select

0002v
Size: 2015 x 3009
Select



0002r
Size: 2015 x 3010
Select



0003v
Size: 2015 x 3010
Select

New template | Existing templates

Top Left: x = 855, y = 1760

Width = 56 px, Height = 109 px

Bottom Right: x = 911, y = 1869

Character h

Pick new template | Search

Templ. 3 x

x:	855	y:	1760	rank:	0	score:	1		
x:	506	y:	2376	rank:	1	score:	0.950197		
x:	1114	y:	555	rank:	2	score:	0.948722		
x:	738	y:	245	rank:	3	score:	0.944954		
x:	1120	y:	1246	rank:	4	score:	0.942005		
x:	711	y:	2279	rank:	5	score:	0.941841		
x:	1179	y:	1144	rank:	6	score:	0.941514		
x:	1296	y:	1458	rank:	7	score:	0.93922		
x:	793	y:	2383	rank:	8	score:	0.937582		
x:	1152	y:	1661	rank:	9	score:	0.937418		

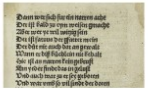
Detect Threshold (Template 3)

"Das Narrenschiff"

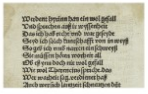
0001v
Size: 2015 x 3009
Select

0001r
Size: 2015 x 3010
Select

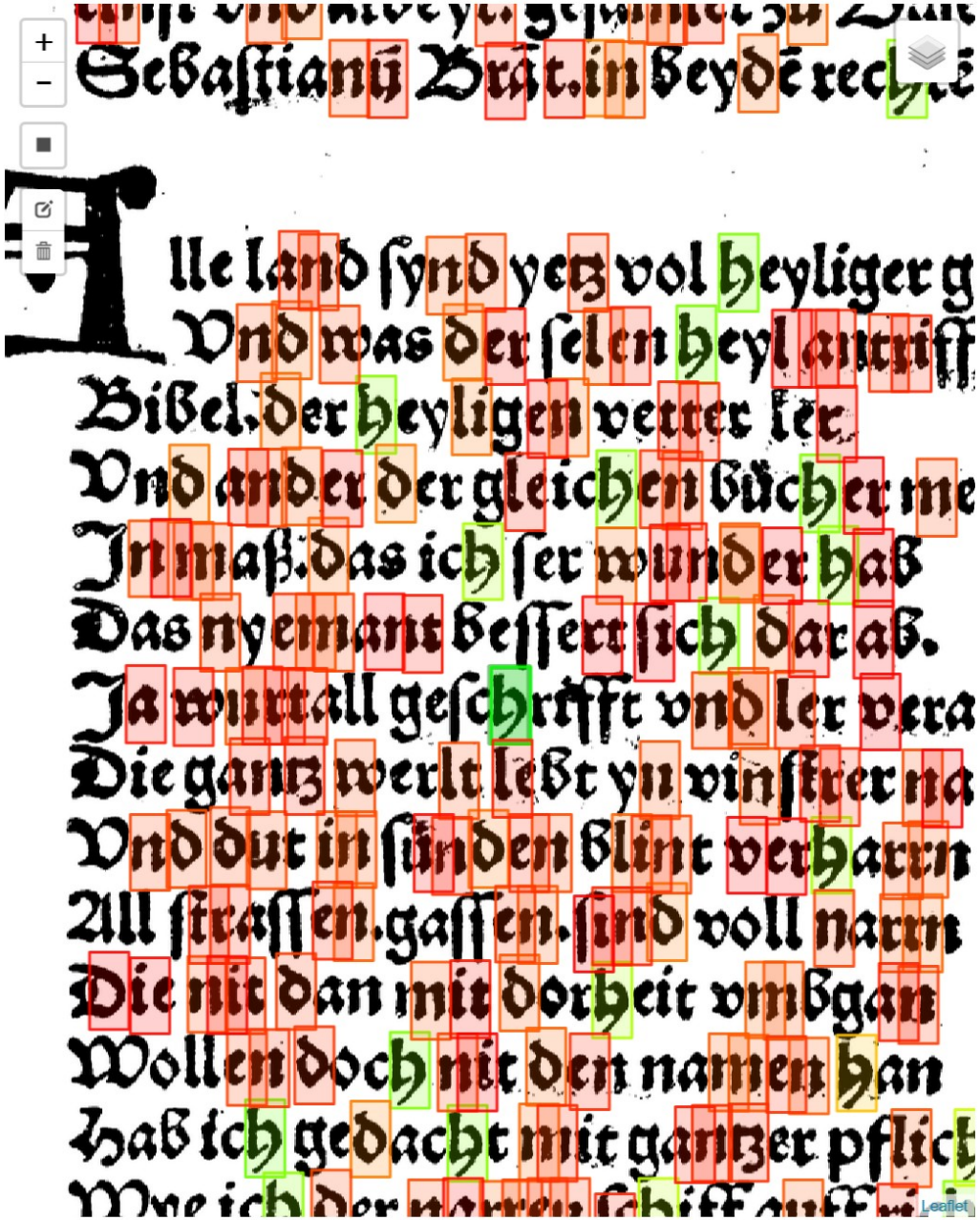
0002v
Size: 2015 x 3009
Select



0002r
Size: 2015 x 3010
Select



0003v
Size: 2015 x 3010
Select



New template | Existing templates

Top Left: x = 855, y = 1760

Width = 56 px, Height = 109 px

Bottom Right: x = 911, y = 1869

Character h

Pick new template | Search

Templ. 3 x

x:	855	y:	1760	rank:	0	score:	1
x:	506	y:	2376	rank:	1	score:	0.950197
x:	1114	y:	555	rank:	2	score:	0.948722
x:	738	y:	245	rank:	3	score:	0.944954
x:	1120	y:	1246	rank:	4	score:	0.942005
x:	711	y:	2279	rank:	5	score:	0.941841
x:	1179	y:	1144	rank:	6	score:	0.941514
x:	1296	y:	1458	rank:	7	score:	0.93922
x:	793	y:	2383	rank:	8	score:	0.937582
x:	1152	y:	1661	rank:	9	score:	0.937418

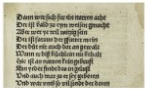
Detect Threshold (Template 3)

"Das Narrenschiff"

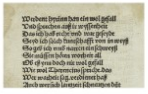
0001v
Size: 2015 x 3009
Select

0001r
Size: 2015 x 3010
Select

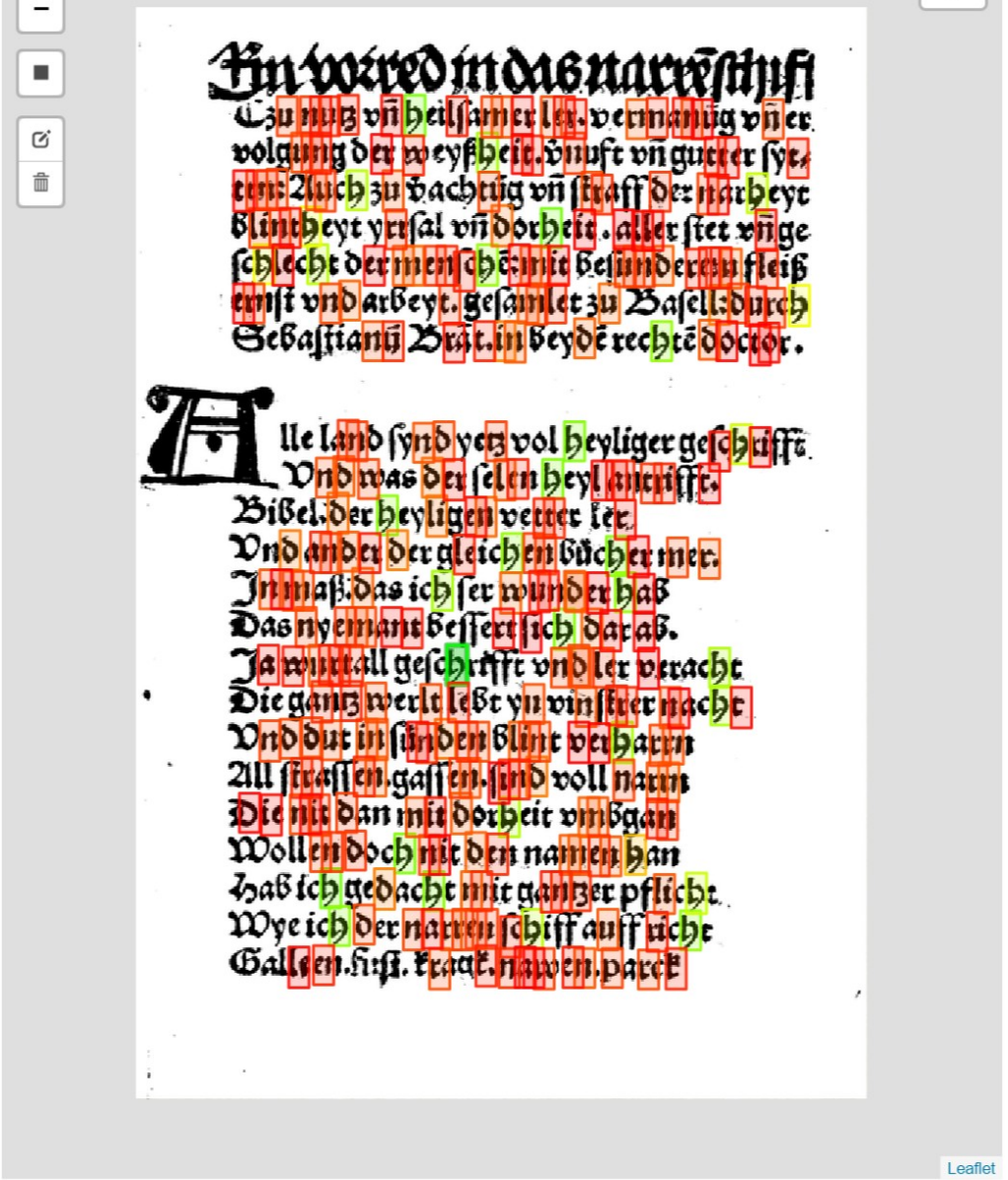
0002v
Size: 2015 x 3009
Select



0002r
Size: 2015 x 3010
Select



0003v
Size: 2015 x 3010
Select



Leaflet

New template | Existing templates

Top Left: x = 855, y = 1760

Width = 56 px, Height = 109 px

Bottom Right: x = 911, y = 1869

Character h

Pick new template | Search

Templ. 3 x

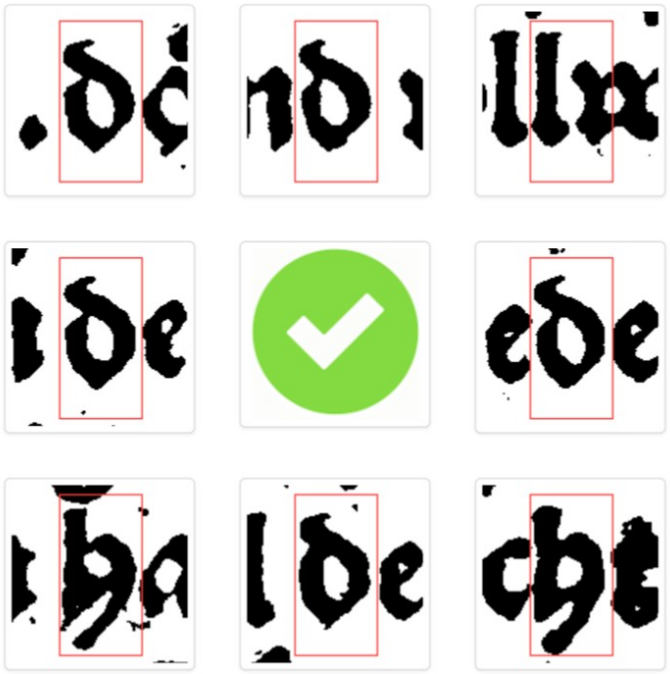
x:	855	y:	1760	rank:	0	score:	1
x:	506	y:	2376	rank:	1	score:	0.950197
x:	1114	y:	555	rank:	2	score:	0.948722
x:	738	y:	245	rank:	3	score:	0.944954
x:	1120	y:	1246	rank:	4	score:	0.942005
x:	711	y:	2279	rank:	5	score:	0.941841
x:	1179	y:	1144	rank:	6	score:	0.941514
x:	1296	y:	1458	rank:	7	score:	0.93922
x:	793	y:	2383	rank:	8	score:	0.937582
x:	1152	y:	1661	rank:	9	score:	0.937418

Detect Threshold (Template 3)

Searching for glyph: 'h'

Next

Searching for glyph: 'h'



Next

Searching for glyph: 'h'

Next

Searching for glyph: 'h'

Next

Training finished

We have now sampled enough data for this template.

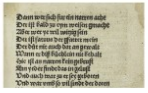
✔ Back to Document Viewer

"Das Narrenschiff"

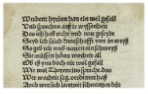
0001v
Size: 2015 x 3009
Select

0001r
Size: 2015 x 3010
Select

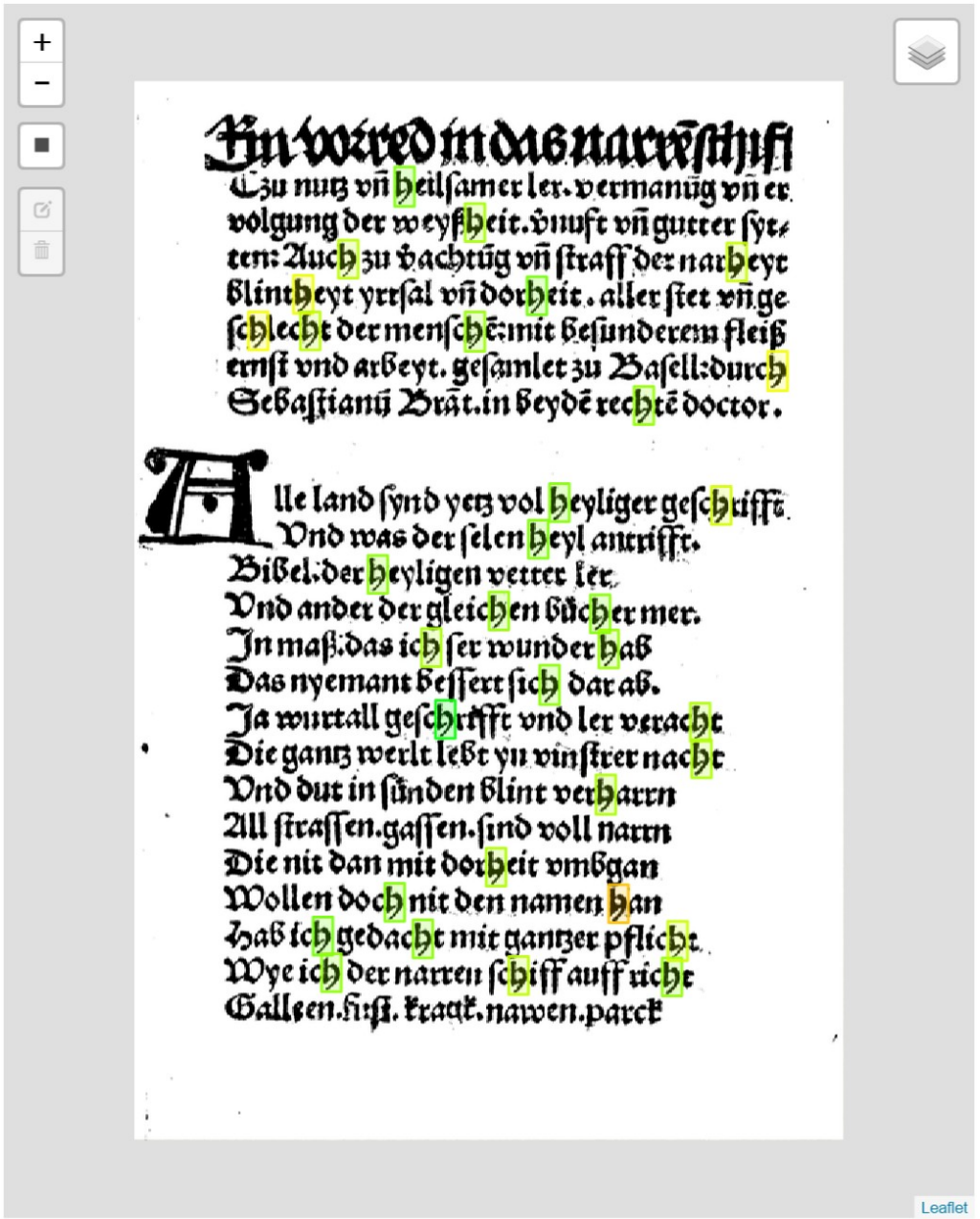
0002v
Size: 2015 x 3009
Select



0002r
Size: 2015 x 3010
Select



0003v
Size: 2015 x 3010
Select



Leaflet

New template | Existing templates

Top Left: x = 0, y = 0

Width = 0 px, Height = 0 px

Bottom Right: x = 0, y = 0

Character Unicode Character (

Pick new template | Search

Templ. 3 x

x:	855	y:	1760	rank:	0	score:	1
x:	506	y:	2376	rank:	1	score:	0.950197
x:	1114	y:	555	rank:	2	score:	0.948722
x:	738	y:	245	rank:	3	score:	0.944954
x:	1120	y:	1246	rank:	4	score:	0.942005
x:	711	y:	2279	rank:	5	score:	0.941841
x:	1179	y:	1144	rank:	6	score:	0.941514
x:	1296	y:	1458	rank:	7	score:	0.93922
x:	793	y:	2383	rank:	8	score:	0.937582
x:	1152	y:	1661	rank:	9	score:	0.937418

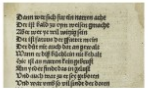
Detect Threshold (Template 3)

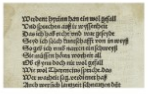
"Das Narrenschiff"

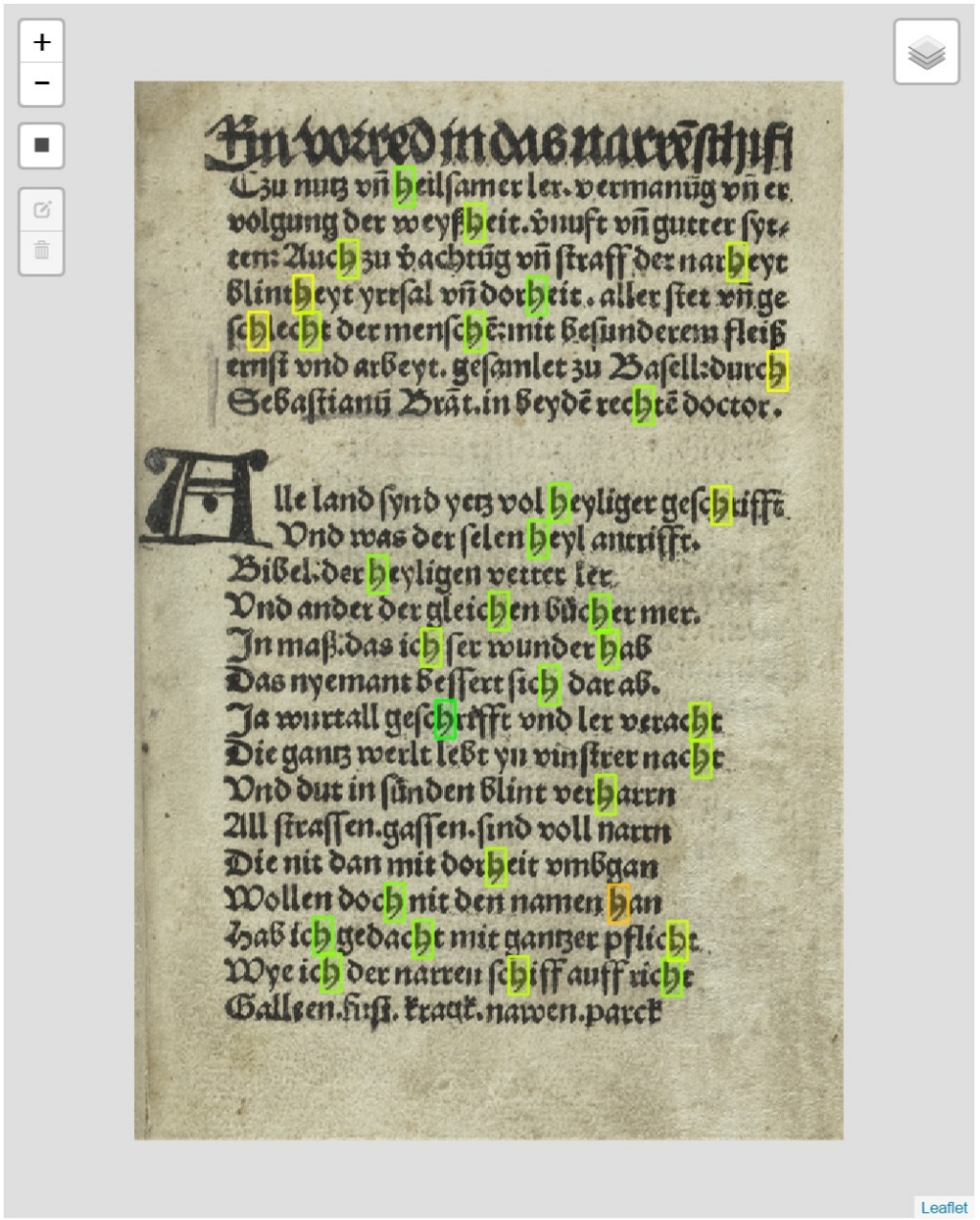

0001v
Size: 2015 x 3009
Select


0001r
Size: 2015 x 3010
Select


0002v
Size: 2015 x 3009
Select


0002r
Size: 2015 x 3010
Select


0003v
Size: 2015 x 3010
Select



New template | Existing templates

Top Left: x = 0, y = 0

Width = 0 px, Height = 0 px

Bottom Right: x = 0, y = 0

Character Unicode Character (

Pick new template | Search

Templ. 3 x

x:	855	y:	1760	rank:	0	score:	1		
x:	506	y:	2376	rank:	1	score:	0.950197		
x:	1114	y:	555	rank:	2	score:	0.948722		
x:	738	y:	245	rank:	3	score:	0.944954		
x:	1120	y:	1246	rank:	4	score:	0.942005		
x:	711	y:	2279	rank:	5	score:	0.941841		
x:	1179	y:	1144	rank:	6	score:	0.941514		
x:	1296	y:	1458	rank:	7	score:	0.93922		
x:	793	y:	2383	rank:	8	score:	0.937582		
x:	1152	y:	1661	rank:	9	score:	0.937418		

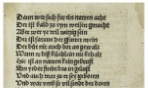
Detect Threshold (Template 3)

"Das Narrenschiff"

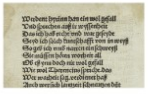
0001v
Size: 2015 x 3009
Select

0001r
Size: 2015 x 3010
Select

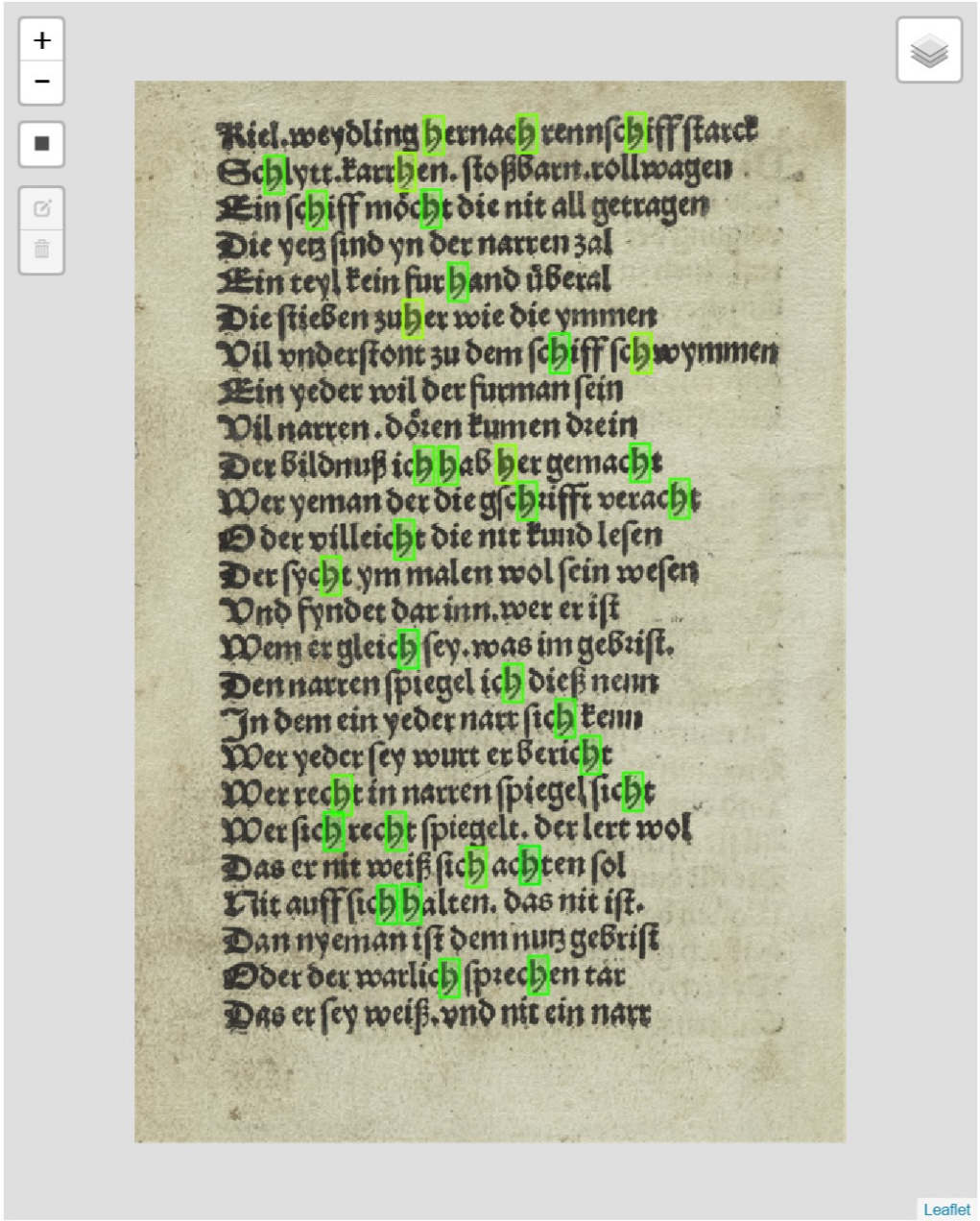
0002v
Size: 2015 x 3009
Select



0002r
Size: 2015 x 3010
Select



0003v
Size: 2015 x 3010
Select



Leaflet

New template | Existing templates

Top Left: x = 0, y = 0

Width = 0 px, Height = 0 px

Bottom Right: x = 0, y = 0

Character Unicode Character (

Pick new template | Search

Templ. 3 x

x:	1093	y:	2167	rank:	0	score:	0.944463
x:	689	y:	2278	rank:	1	score:	0.939056
x:	748	y:	1554	rank:	2	score:	0.938237
x:	534	y:	2072	rank:	3	score:	0.938073
x:	757	y:	2277	rank:	4	score:	0.935944
x:	864	y:	2486	rank:	5	score:	0.933814
x:	1176	y:	717	rank:	6	score:	0.93365
x:	1116	y:	2479	rank:	7	score:	0.933322
x:	1405	y:	1028	rank:	8	score:	0.932503
x:	1384	y:	1959	rank:	9	score:	0.93152

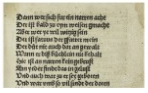
Detect Threshold (Template 3)

"Das Narrenschiff"

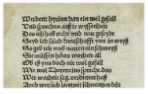
0001v
Size: 2015 x 3009
Select

0001r
Size: 2015 x 3010
Select

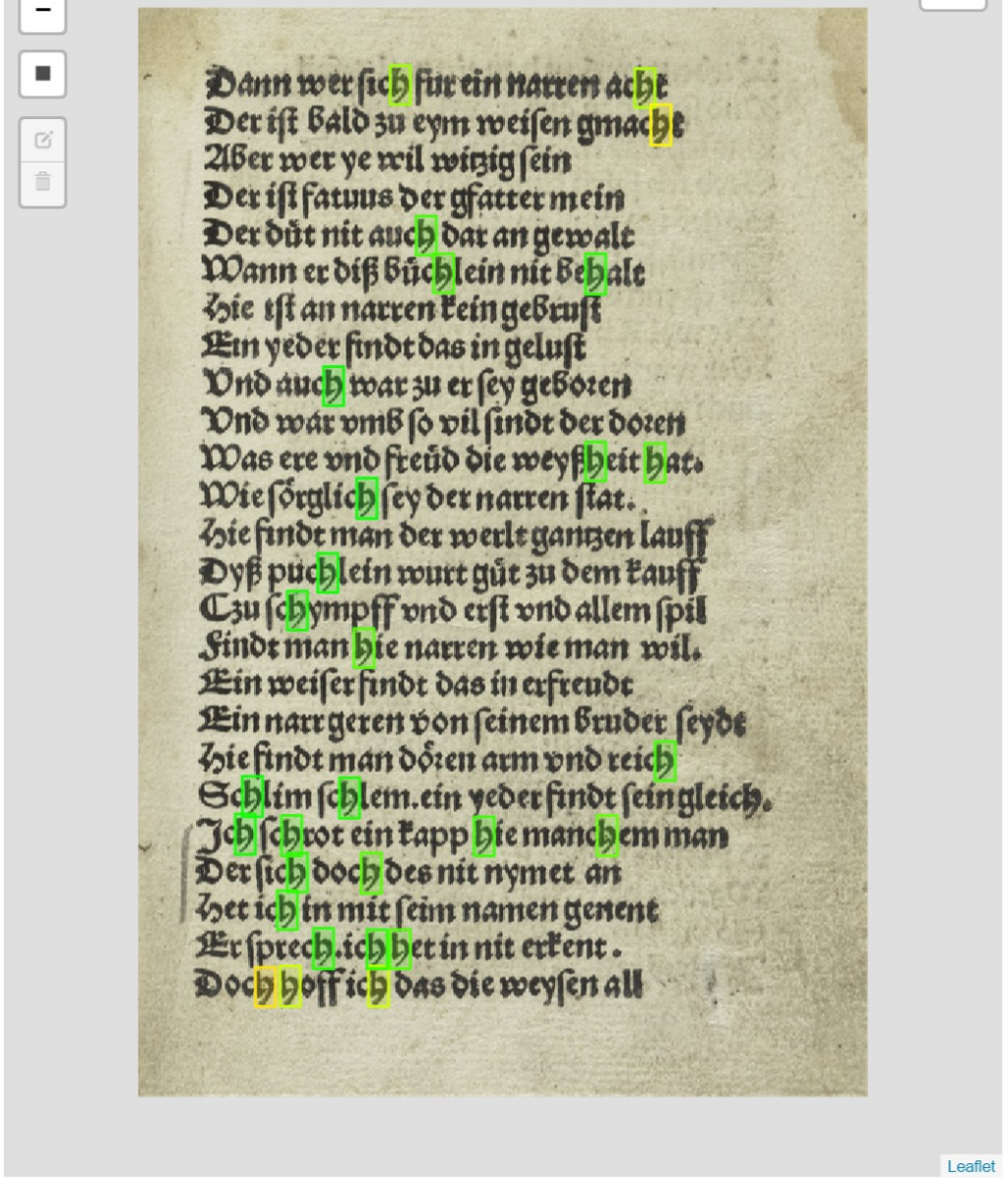
0002v
Size: 2015 x 3009
Select



0002r
Size: 2015 x 3010
Select



0003v
Size: 2015 x 3010
Select



Leaflet

New template | Existing templates

Top Left: x = 0, y = 0

Width = 0 px, Height = 0 px

Bottom Right: x = 0, y = 0

Character Unicode Character (

Pick new template | Search

Templ. 3 x

x: 604	y: 1301	rank: 0	score: 0.946265
x: 513	y: 990	rank: 1	score: 0.944954
x: 289	y: 2125	rank: 2	score: 0.94479
x: 496	y: 1508	rank: 3	score: 0.942661
x: 267	y: 2229	rank: 4	score: 0.941514
x: 412	y: 1611	rank: 5	score: 0.940203
x: 556	y: 2129	rank: 6	score: 0.93709
x: 484	y: 2545	rank: 7	score: 0.933978
x: 1237	y: 681	rank: 8	score: 0.93365
x: 397	y: 2233	rank: 9	score: 0.932995

Detect Threshold (Template 3)

Documents

Das Narrenschiff

Sebastian Brant – 1494

Collection

Page Creator

Select

Stultifera Navis

Sebastian Brant – 1497

Collection

Page Creator

Select

0001v

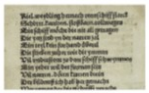
Size: 2015 x 3009

Select

0001r

Size: 2015 x 3010

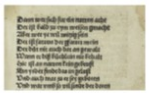
Select



0002v

Size: 2015 x 3009

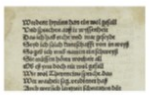
Select



0002r

Size: 2015 x 3010

Select



0003v

Size: 2015 x 3010

Select

R0001r

Size: 2166 x 3219

Select

R0001v

Size: 2166 x 3219

Select

Templates

Template 3 – "h"

Position: (855, 1760),

Size: 56 x 109

Das Narrenschiff

Select

Filter templates by character...

Glyphs

Template | User Positive | Predicted | Remaining | Group

Download

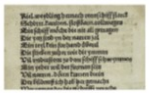
Documents

Das Narrenschiff
Sebastian Brant – 1494
Collection
Page Creator
Select

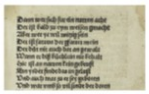
Stultifera Navis
Sebastian Brant – 1497
Collection
Page Creator
Select

0001v
Size: 2015 x 3009
Select

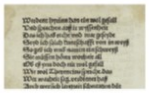
0001r
Size: 2015 x 3010
Select



0002v
Size: 2015 x 3009
Select



0002r
Size: 2015 x 3010
Select



0003v
Size: 2015 x 3010
Select

R0001r
Size: 2166 x 3219
Select

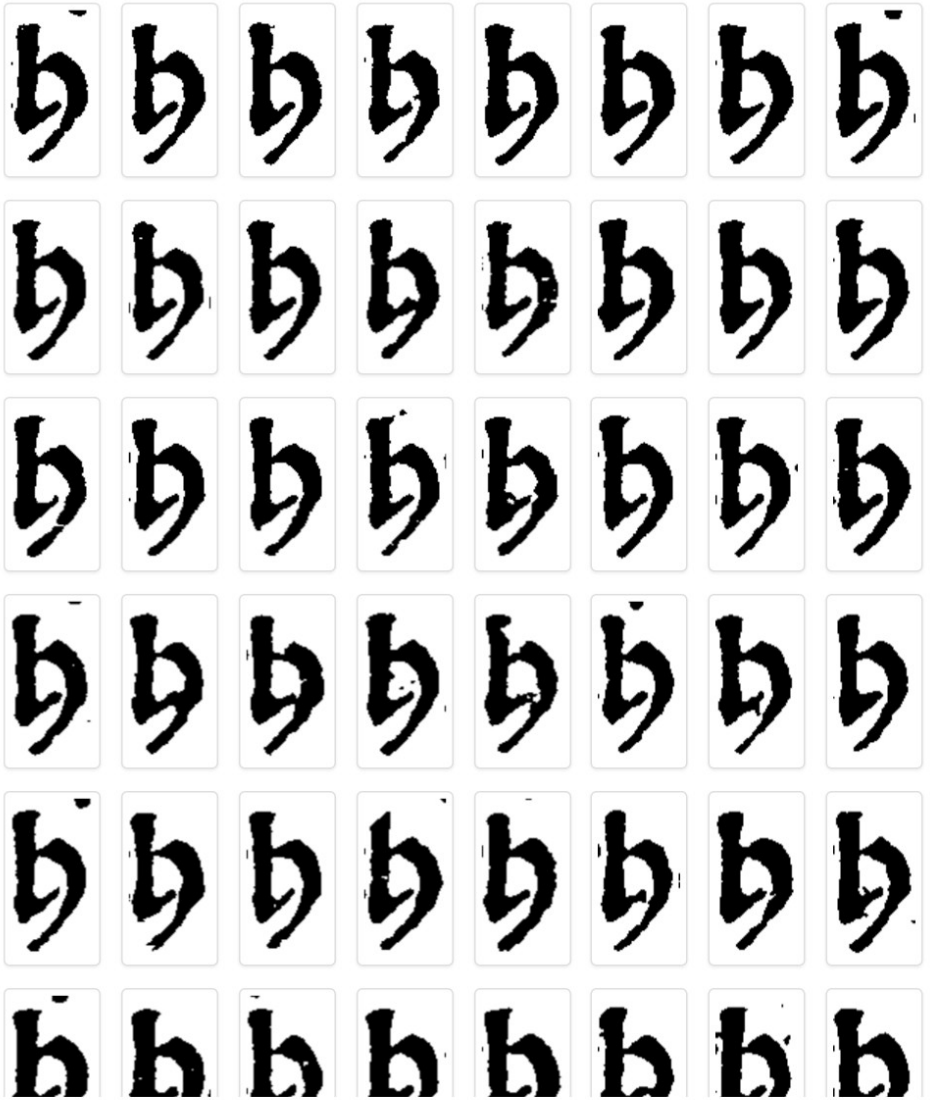
R0001v
Size: 2166 x 3219
Select

Templates

Template 3 – "h"
Position: (855, 1760),
Size: 56 x 109
Das Narrenschiff
Select

Filter templates by character...

Glyphs



Template | User Positive | Predicted | Remaining | Group

Download

Documents

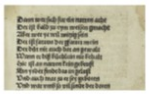
Das Narrenschiff
Sebastian Brant – 1494
Collection
Page Creator
Select

Stultifera Navis
Sebastian Brant – 1497
Collection
Page Creator
Select

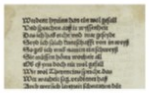
0001v
Size: 2015 x 3009
Select

0001r
Size: 2015 x 3010
Select

0002v
Size: 2015 x 3009
Select



0002r
Size: 2015 x 3010
Select



0003v
Size: 2015 x 3010
Select

R0001r
Size: 2166 x 3219
Select

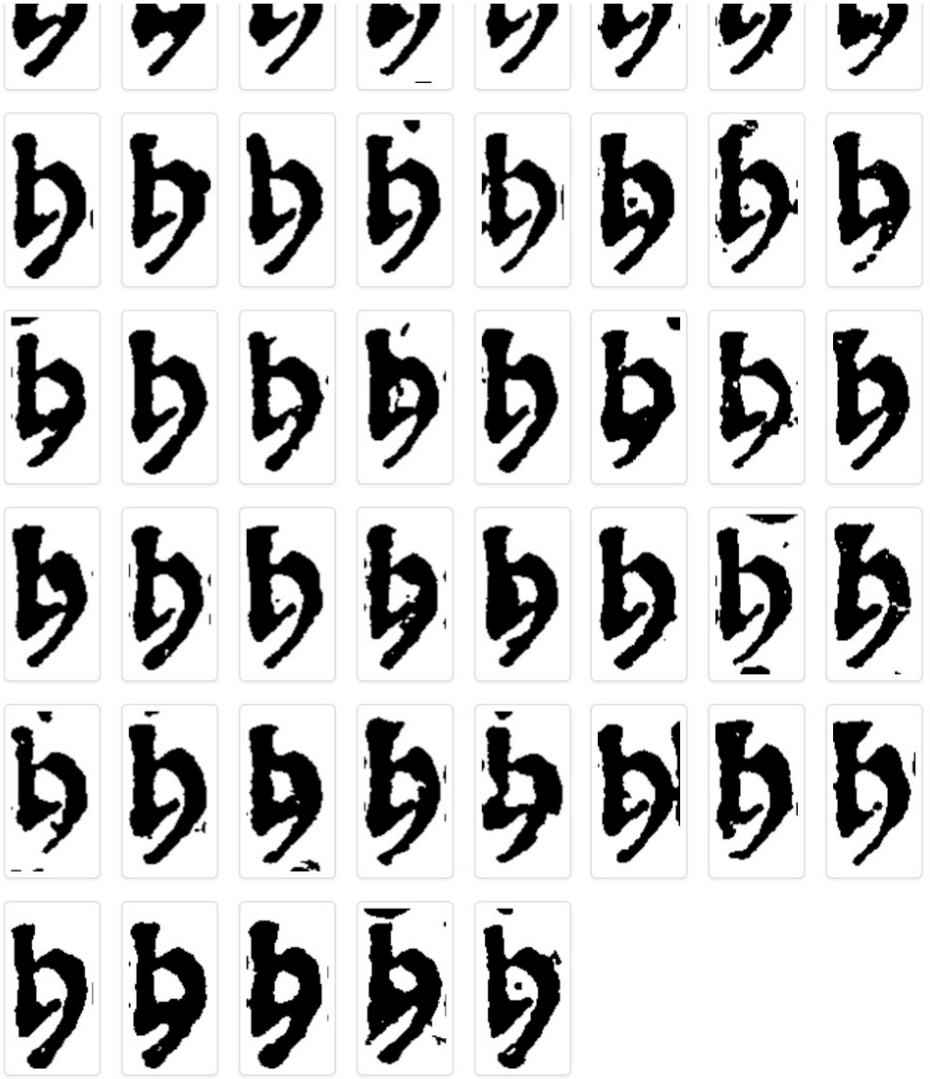
R0001v
Size: 2166 x 3219
Select

Templates

Template 3 – "h"
Position: (855, 1760),
Size: 56 x 109
Das Narrenschiff
Select

Filter templates by character...

Glyphs



Template | User Positive | Predicted | Remaining | Group

Download

Glyph Miner: Two Ingredients

Ingredient 1: Template Matching

- Find approximate repeat-occurrences of an example image
- Here: black-and-white, only translation

Ingredient 2: Active Learning

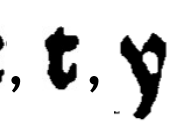
- Distinguish matches that are semantically correct from the rest
- Efficient user interaction

Preliminary Case Study

- On 20 out of 320 pages from Latin *Narrenschiff*
- Experiment 1: precision/recall/F1 score > 94%
- Experiment 2: Glyph Miner vs. manually collecting glyphs by OCR engineer (for 45 minutes each)

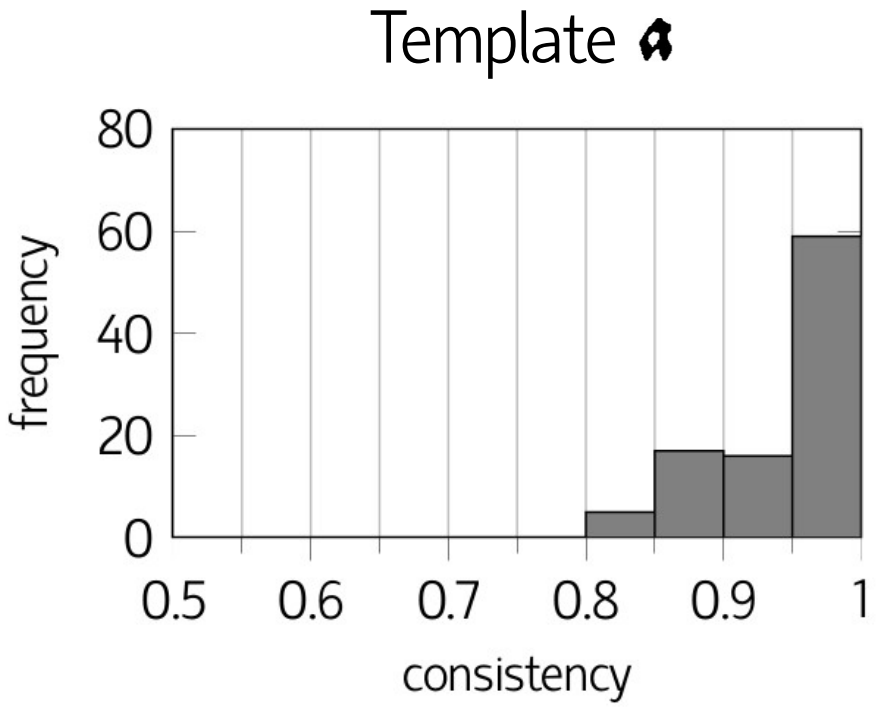
	Aletheia	**Glyph Miner**
detected occurrences	1,251	17,426
different glyphs	65	26
occurences per glyph (median)	7	498
glyphs with support > 10	25	26

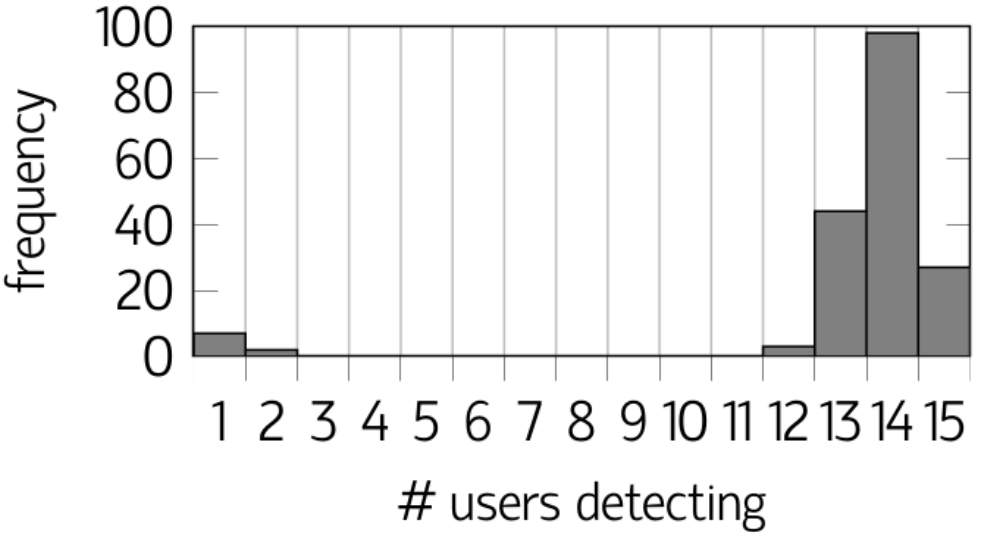
User Study at *philtag 2016*

- Participants from Digital Humanities, OCR and linguistics
- Hands-on session, five pages from German *Narrenschiff*
- 30 minutes to process 

User Study: Outcomes

- 59 templates, 5000 labels,
  17 questionnaires
- User evaluation:
  - enjoyable to work with
  - would use it in daily work
- Reliability:
  - label consistency is high
  - classifier consistency is high

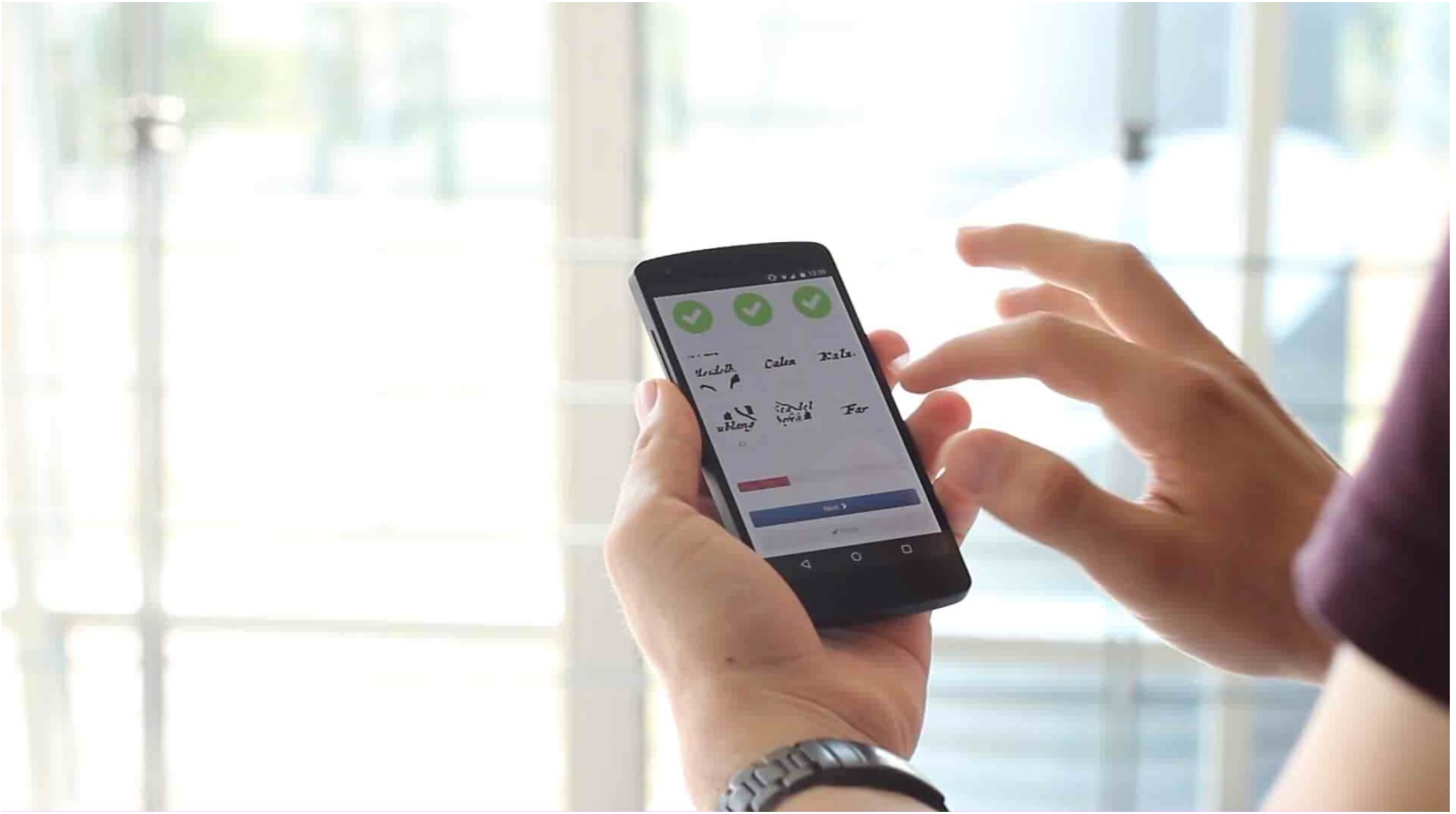




Conclusion & Future Work

- Incunables are interesting, but hard to OCR
- Human effort is necessary → smart interactions!
- Need for good training data

- Crowdsourcing?



Conclusion & Future Work

- Incunables are interesting, but hard for OCR
- Human effort is necessary → smart interactions!
- Need for good training data

- Crowdsourcing! But how exactly?
- In-depth evaluation with Tesseract and OCRopus

Das Narrenschiff

Template 8 – "c"
Position: (498, 2074),
Size: 37 x 111
Select

Template 9 – "d"
Position: (633, 1134),
Size: 48 x 120
Select

Template 4 – "e"
Position: (878, 1966),
Size: 38 x 112
Select

Template 3 – "h"
Position: (855, 1760),
Size: 56 x 109
Select

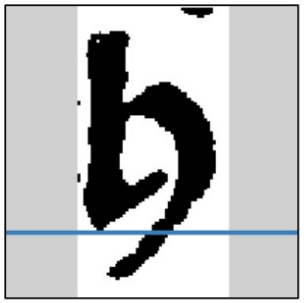
Template 5 – "l"
Position: (985, 1250),
Size: 34 x 93
Select

Template 6 – "o"
Position: (1086, 1145),
Size: 50 x 106
Select

Template 11 – "r"

Filter templates by character...

Typesetting



Baseline 84 px

Left crop 0 px

Right crop 0 px

Save changes

Letter Spacing px

Horizontal Margin px

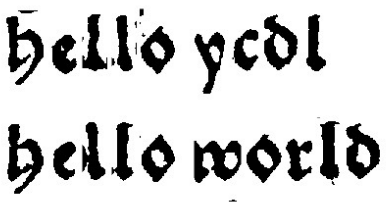
Word Spacing px

Vertical Margin px

Baseline Skip px

hello ycdl
hello world

Preview



Leaflet

Conclusion & Future Work

- Incunables are interesting, but hard for OCR
- Human effort is necessary → smart interactions!
- Need for good training data

- Crowdsourcing! But how exactly?
- In-depth evaluation with Tesseract and OCRopus

- Check our demo video on YouTube: