



Glyph Miner:

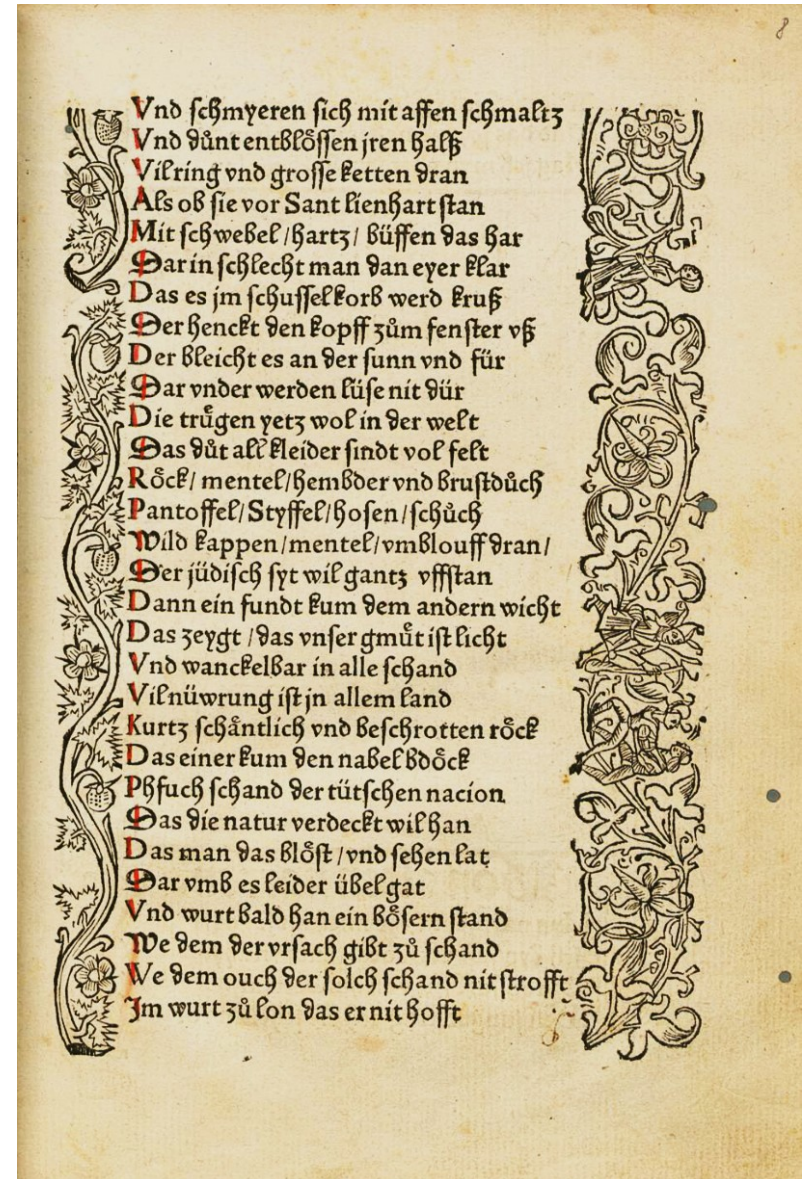
A System for Efficiently Extracting Glyphs from Early Prints in the Context of OCR



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



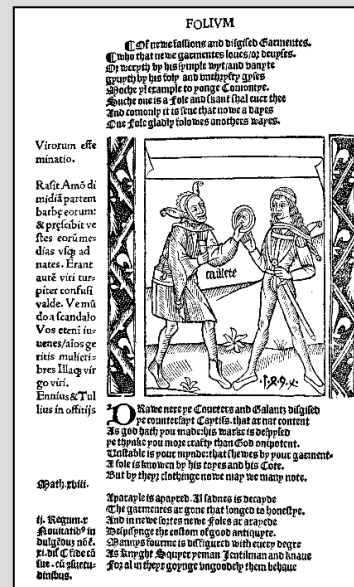
Nürnberg 1494
German



Basel 1498
Latin



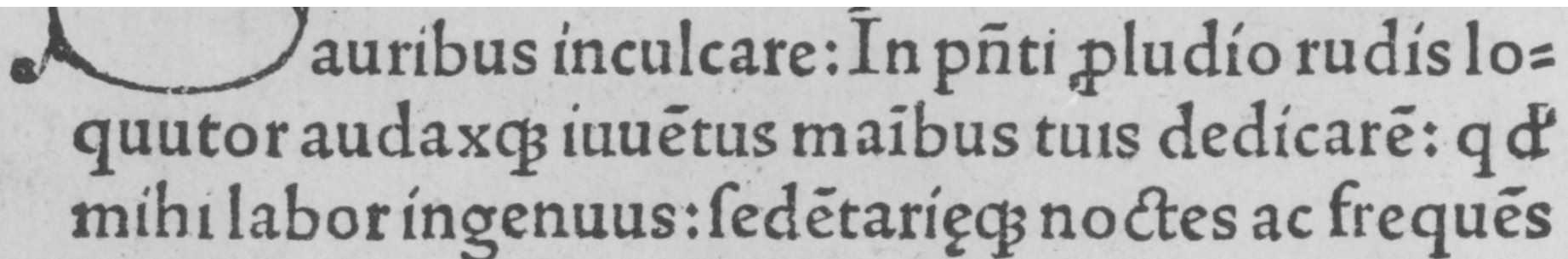
Paris 1497
French



London 1509
English

Challenges for OCR

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



auribus inculcare: In pñti pludío rudís lo=
quutor audaxq iuuētus maībus tuis dedícarē: q d
mīhi labor íngenuus: sedētaríeq noctes ac frequēs

- 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+, Camera
- 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

Subtitle

Author

Year

Signature

Collection

Image (color)

Image (b/w)

Add new collection

Title

Subtitle

Author

Year

Wo edert beyden das ein wol gefall
 Das gedenck auff die verfluchte
 Das ich auß mich noch man gefall
 Heyn ich hab den hantichafft von in weyß
 Das ich ich noch ausen in die weyß
 Die edelheit beyden weyß die
 Das ich beyden die wol gefall
 Wie man beyden die wol gefall
 Wie man beyden die wol gefall
 Auch beyden die wol gefall

Select

Leaflet

"Das Narrenschiff"



0001v

Size: 2015 x 3009

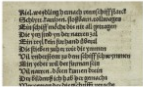
Select



0001r

Size: 2015 x 3010

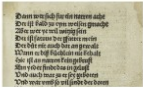
Select



0002v

Size: 2015 x 3009

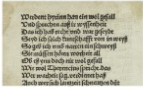
Select



0002r

Size: 2015 x 3010

Select



0003v

Size: 2015 x 3010

Select

...et vnd arbeit. gesamlet zu Basel
Gebastianū Brät.in beyde rechte do

Alle land synd yetz vol heyliger gesch
Vnd was der selen heyl antrifft.
Bibel:der heyligen vetter ler
Vnd ander der gleichen bücher mer.
In maß:das ich ser wunder haß
Das nyemant bessert sich dar ab.
Ja wurtall geschrifft vnd ler verach
Die ganz werlt lebt yn vinstreer nach
Vnd dut in sünden blint verharrn
All strassen.gassen.sind voll narren
Die nit dan mit dorheit ombgan
Wollen doch nit den namen han
Zu ich gedacht mit narren nlich

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

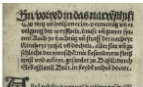
"Das Narrenschiff"



0001v

Size: 2015 x 3009

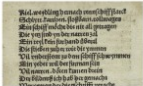
Select



0001r

Size: 2015 x 3010

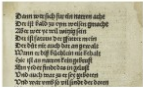
Select



0002v

Size: 2015 x 3009

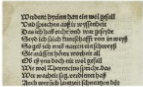
Select



0002r

Size: 2015 x 3010

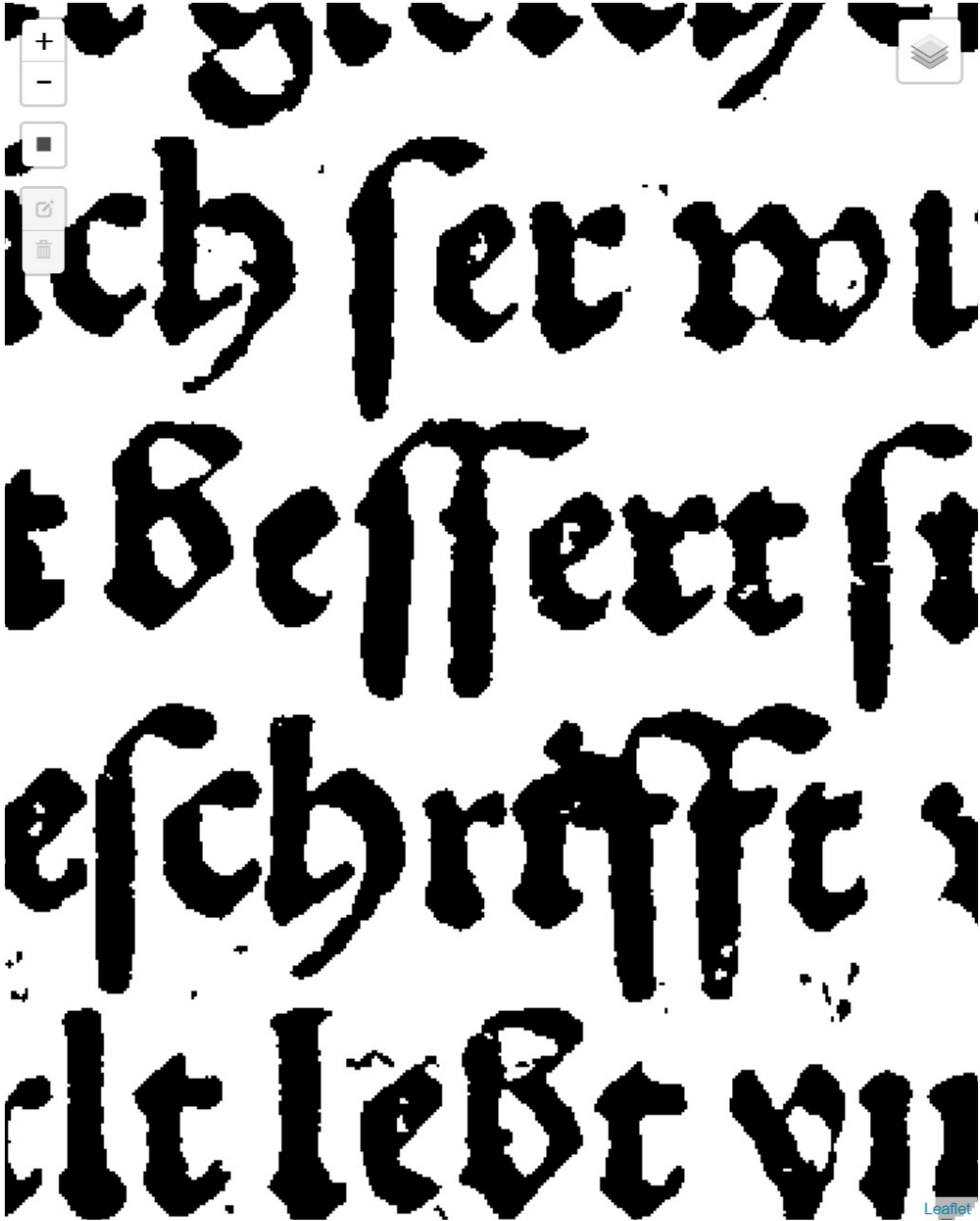
Select



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

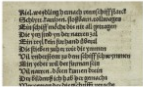
Select



0001r

Size: 2015 x 3010

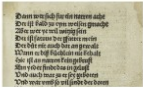
Select



0002v

Size: 2015 x 3009

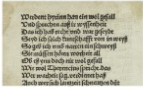
Select



0002r

Size: 2015 x 3010

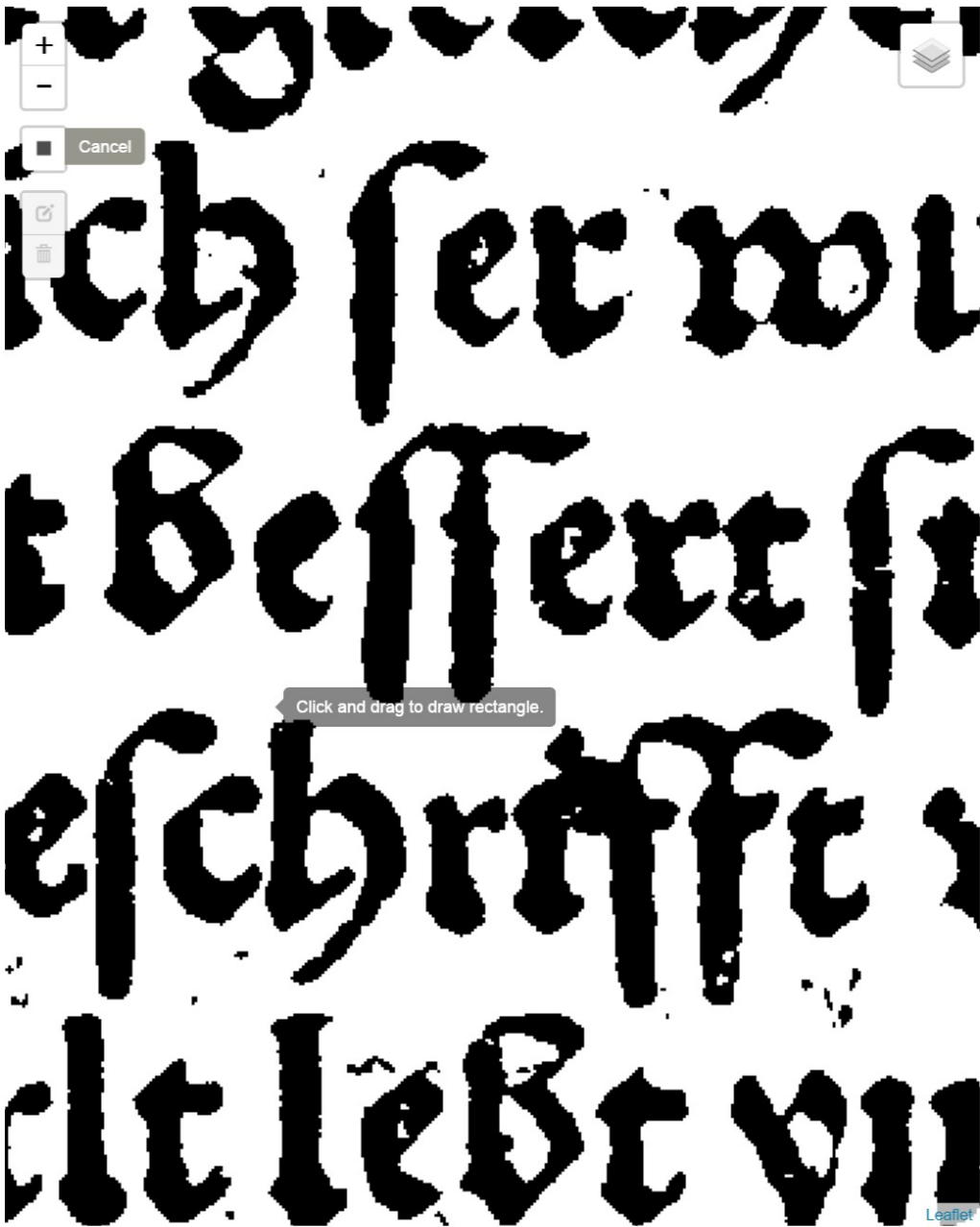
Select



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"



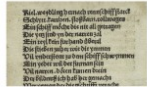
0001v
Size: 2015 x 3009

Select



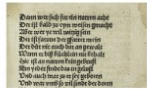
0001r
Size: 2015 x 3010

Select



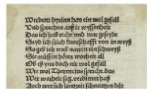
0002v
Size: 2015 x 3009

Select



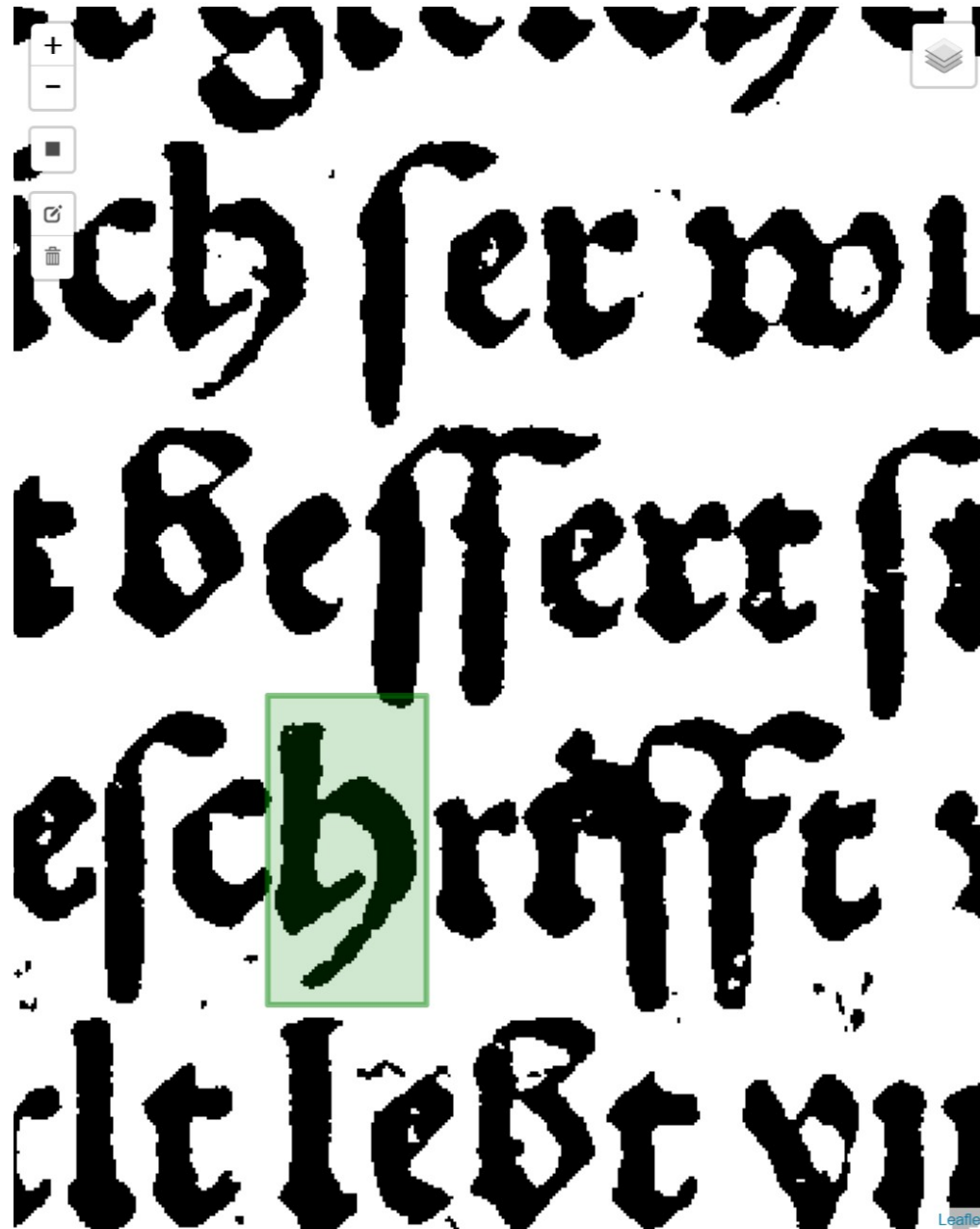
0002r
Size: 2015 x 3010

Select



0003v
Size: 2015 x 3010

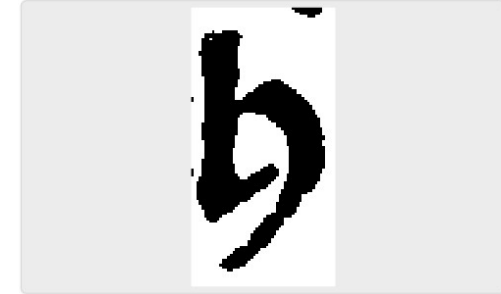
Select



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"



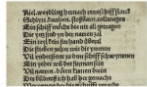
0001v
Size: 2015 x 3009

Select



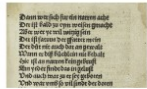
0001r
Size: 2015 x 3010

Select



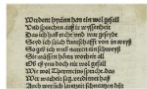
0002v
Size: 2015 x 3009

Select



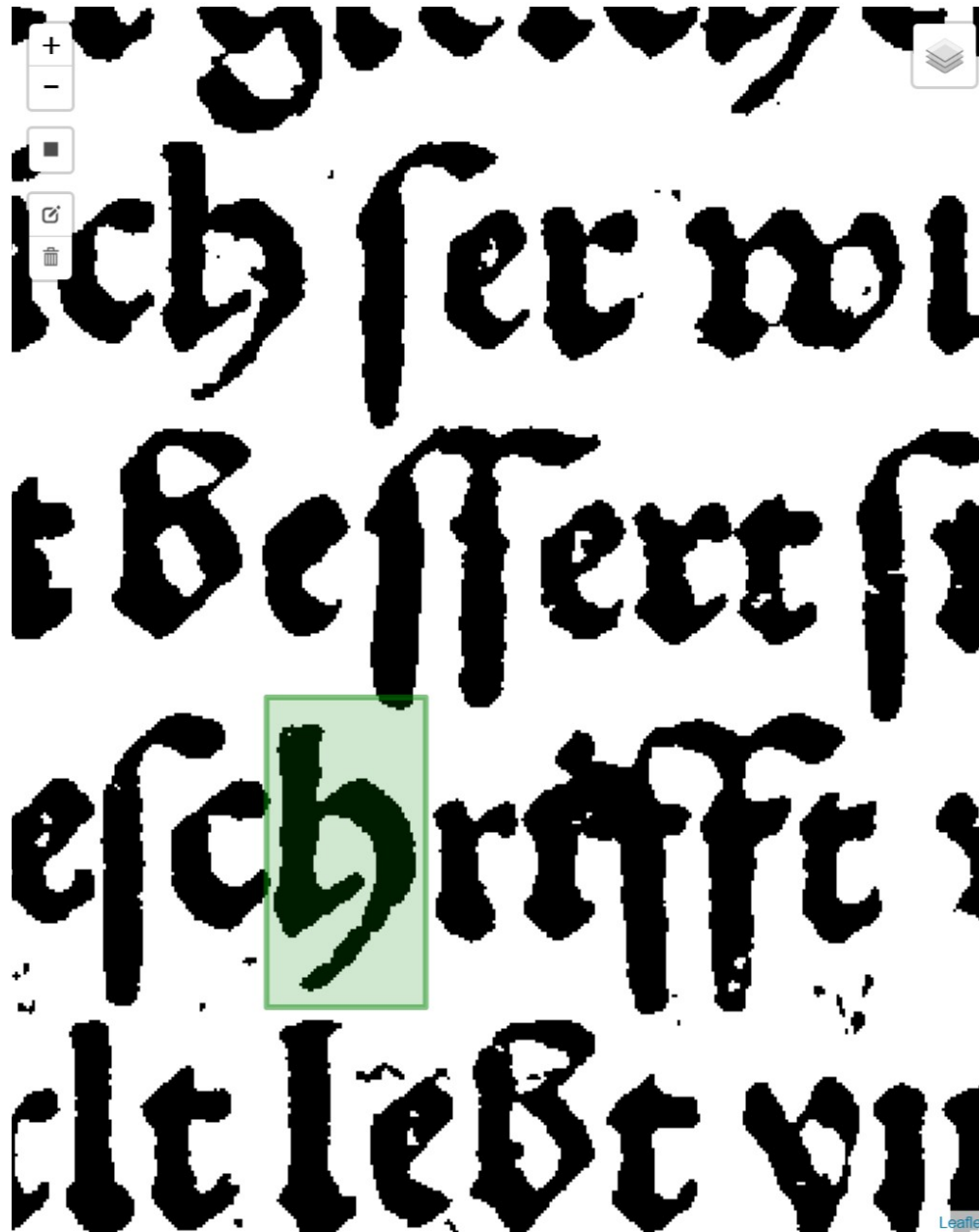
0002r
Size: 2015 x 3010

Select



0003v
Size: 2015 x 3010

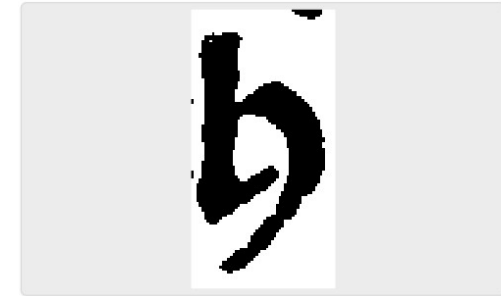
Select



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

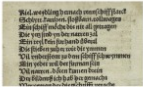
Select



0001r

Size: 2015 x 3010

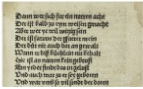
Select



0002v

Size: 2015 x 3009

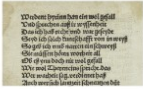
Select



0002r

Size: 2015 x 3010

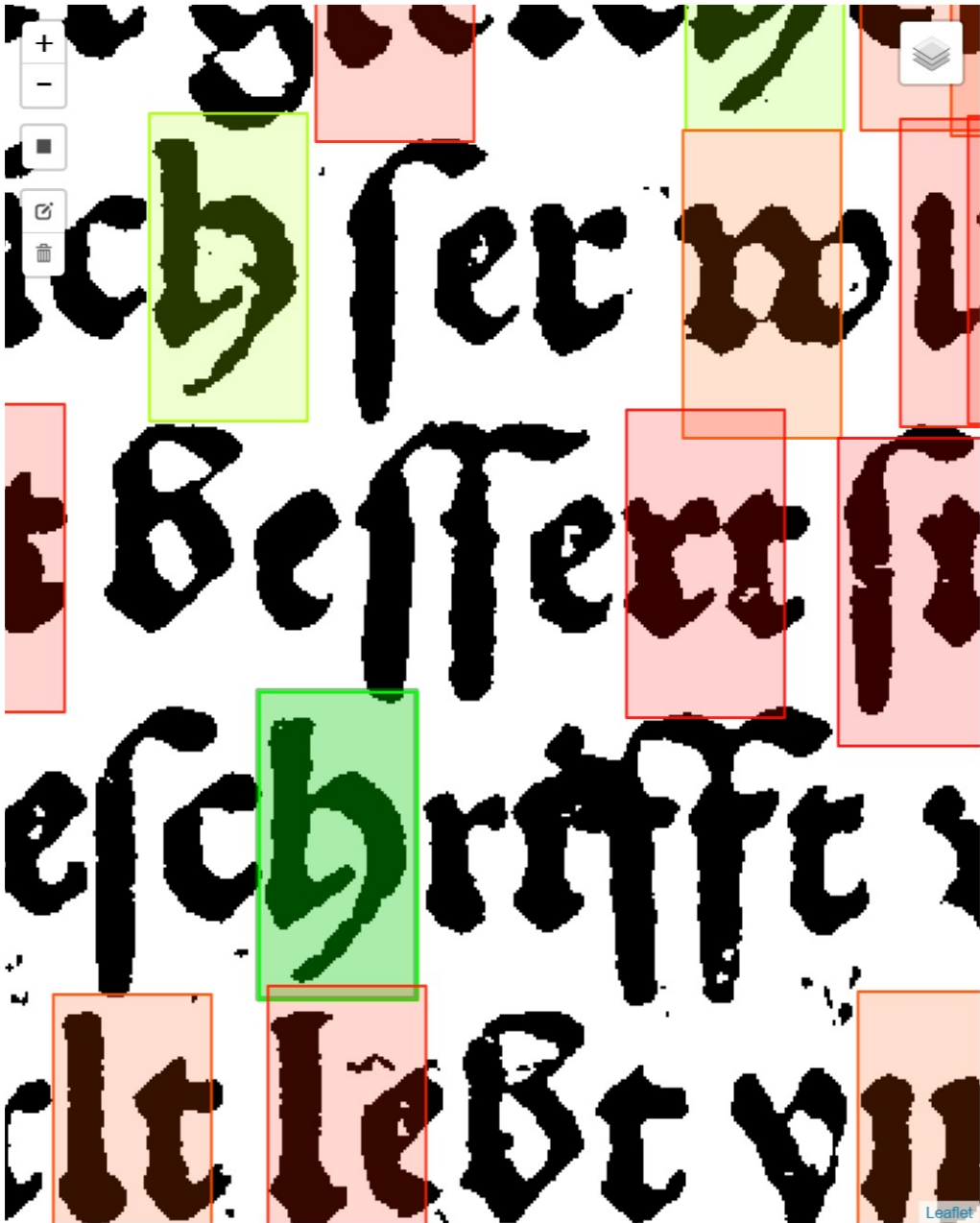
Select



0003v

Size: 2015 x 3010

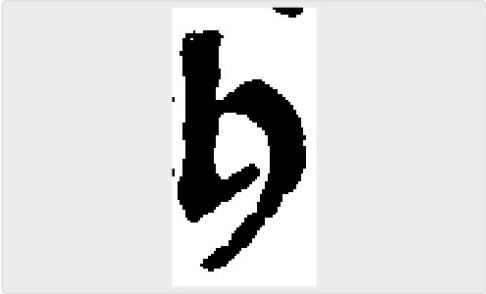
Select



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

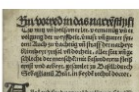
x: 855	y: 1760	rank: 0	score: 1
x: 506	y: 2376	rank: 1	score: 0.950197
x: 1114	y: 555	rank: 2	score: 0.948722
x: 738	y: 245	rank: 3	score: 0.944954
x: 1120	y: 1246	rank: 4	score: 0.942005
x: 711	y: 2279	rank: 5	score: 0.941841
x: 1179	y: 1144	rank: 6	score: 0.941514
x: 1296	y: 1458	rank: 7	score: 0.93922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

Detect Threshold (Template 3)

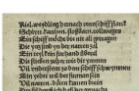
"Das Narrenschiff"



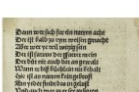
0001v
Size: 2015 x 3009
Select



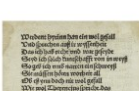
0001r
Size: 2015 x 3010
Select



0002v
Size: 2015 x 3009
Select



0002r
Size: 2015 x 3010
Select



0003v
Size: 2015 x 3010
Select

+

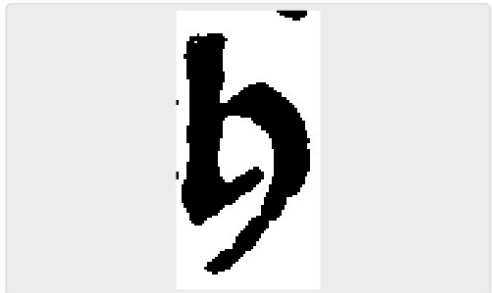
-

Sebastianij Brät.in beyde rech

Ille land synd yetz vol heylicher g
Vnd was der selen heyl antriff
Bibel,der heyiligen vetter ler
Vnd ander der gleichen bücher me
Inmaß, das ich ser wunder hab
Das nyemant bessert sich dar ab.
Ja wortall geschrfft vnd ler vera
Die ganz werlt lebt yn vinstre na
Vnd dut in sünden blint verharrn
Zill strassen, gassen, sind voll narn
Die nit dan mit dorheit ombgan
Wollen doch nit den namen han
Hab ich gedacht mit ganzer pflicht
Mne ich der narren schiff auff

New templateExisting templates

Top Left: x = 855, y = 1760



Width = 56 px, Height = 109 pxBottom 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)

Wo rident lymnen dat die mol gefall
 Dat foudens an is wiffend
 Die ich loof ende mit den ghefte
 Die ich feld hangheffen van in weff
 Die ich mit minen in fchueren
 Die an fien jolue wende al
 Die of you doch nu wel gefall
 Wie mol foudens an is wiffend
 Wie wende die erden mit
 Die wech in den fteuen foudens die

Select



Alle land synd yetz vol heyliger geschrifte.
Vnd was der selten heyl antwiffe.
Bibel. der heyligen vetter ler.
Vnd ander der gleichen bacher mer.
In mass. das ich ser wunder hab.
Das nyemant bessert sich darab.
Ja vnt all geschrifte vnd ler veracht.
Die ganz werlt lebt yn vnsuer nacht.
Vnd daz in sinden blint verharren.
All strassen. gassen. sind voll narren.
Die nit dan mit dorheit vmbgan.
Wollen doch nit den namen han.
Hab ich gedacht mit ganzer pflicht.
Wye ich der narren schiff auff richte.
Gallen. huff. kraut. nawen. parel.

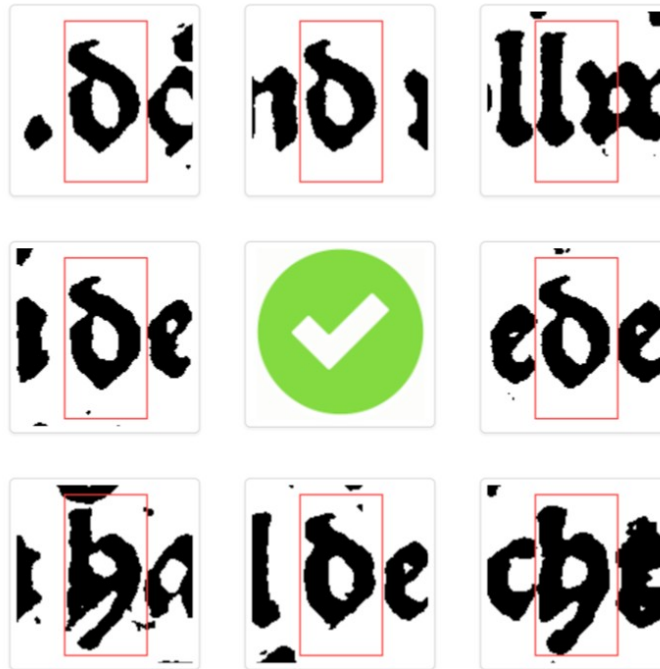
Detect Threshold (Template 3)

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](#)

[illegible]

Select

Leaflet

[illegible]

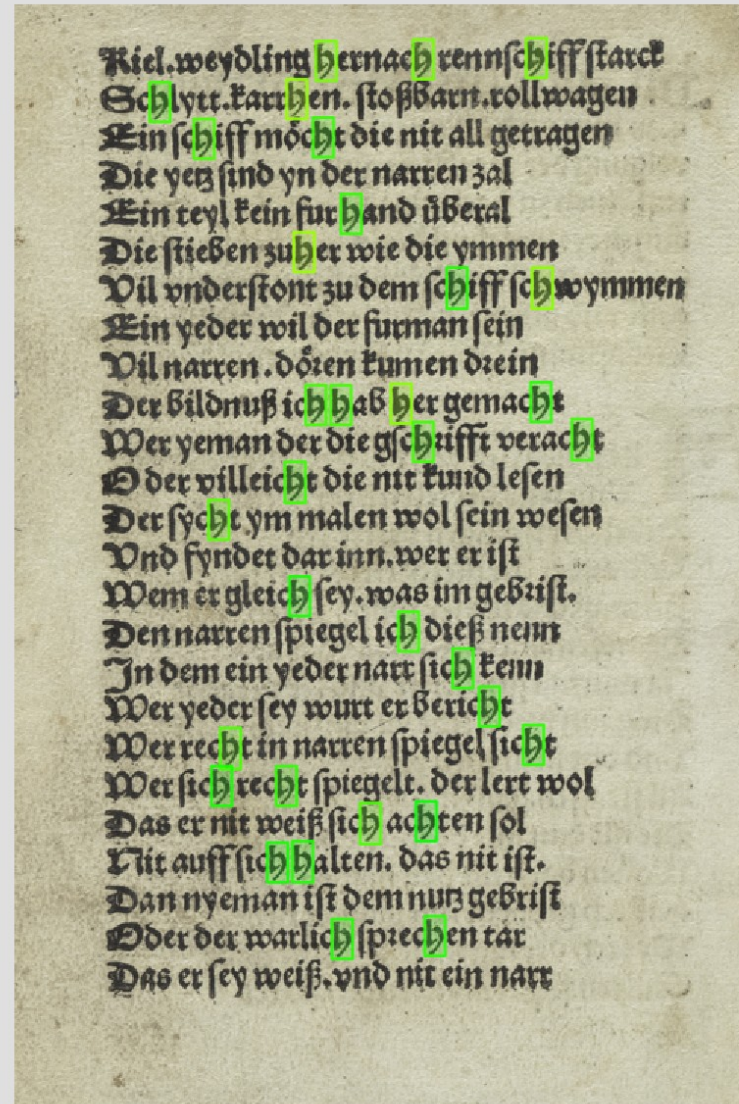
Select



Detect Threshold (Template 3)

Wadent dyne den ein wol gefall
 Das fuchen auff in verftend
 Das ich hab end mit mir gefinde
 Heyd ich fuch hangen auff in der weyß
 Das ich mit mir ausen in fchneyß
 Die auff den helen wend als
 Als of you den ein wol gefall
 Wie wol dreyen den die den
 Wie wadent die in der weyß
 Auch wadent die in der weyß

Select



Detect Threshold (Template 3)

[illegible]

Select



Documents



Das Narrenschiff
Sebastian Brant – 1494
[Collection](#)
[Page Creator](#)

Select



Stultifera Navis
Sebastian Brant – 1497
[Collection](#)
[Page Creator](#)

Select



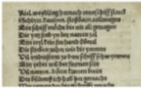
0001v
Size: 2015 x 3009

Select



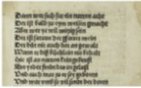
0001r
Size: 2015 x 3010

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

Glyphs

Filter templates by character...

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



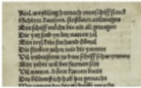
0001v
Size: 2015 x 3009

Select



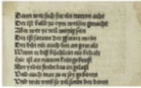
0001r
Size: 2015 x 3010

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

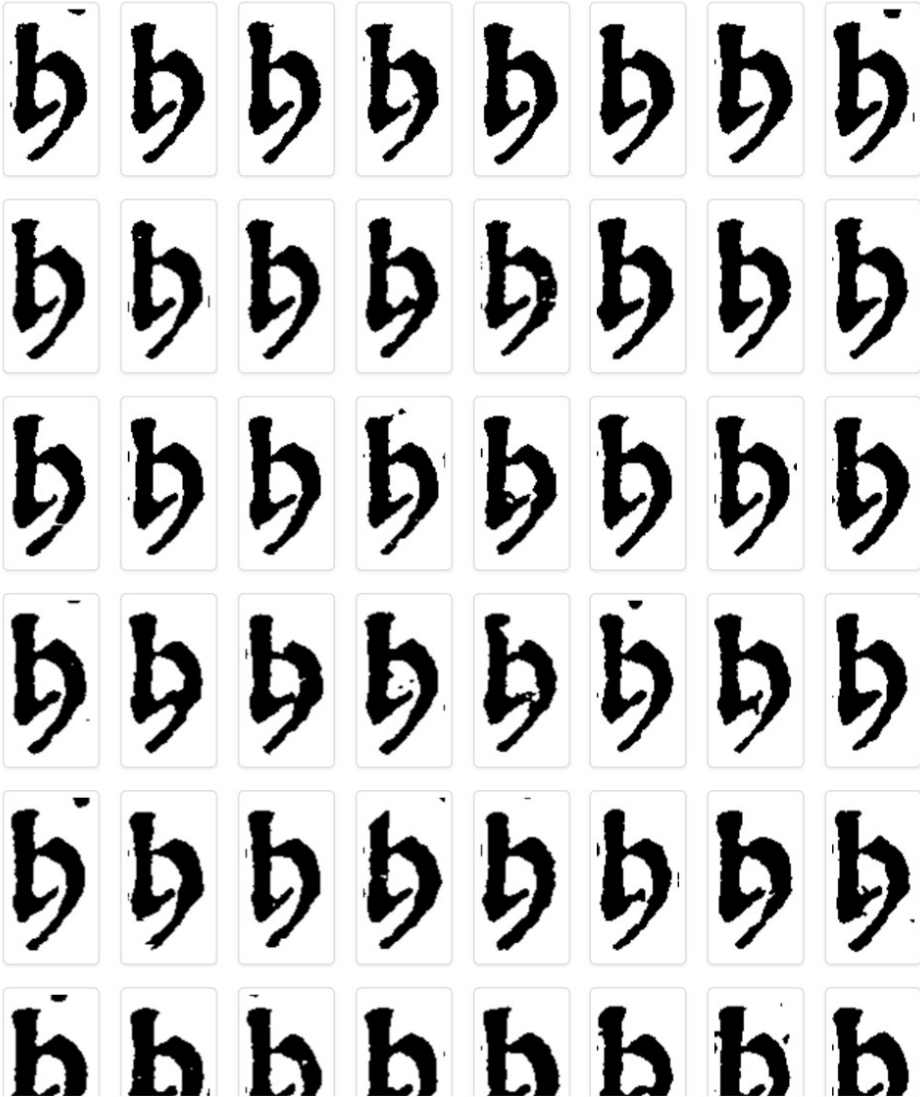
Templates



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

Select

Glyphs



Filter templates by character...

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



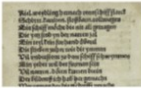
0001v
Size: 2015 x 3009

Select



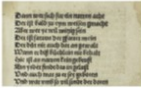
0001r
Size: 2015 x 3010

Select



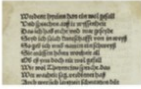
0002v
Size: 2015 x 3009

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

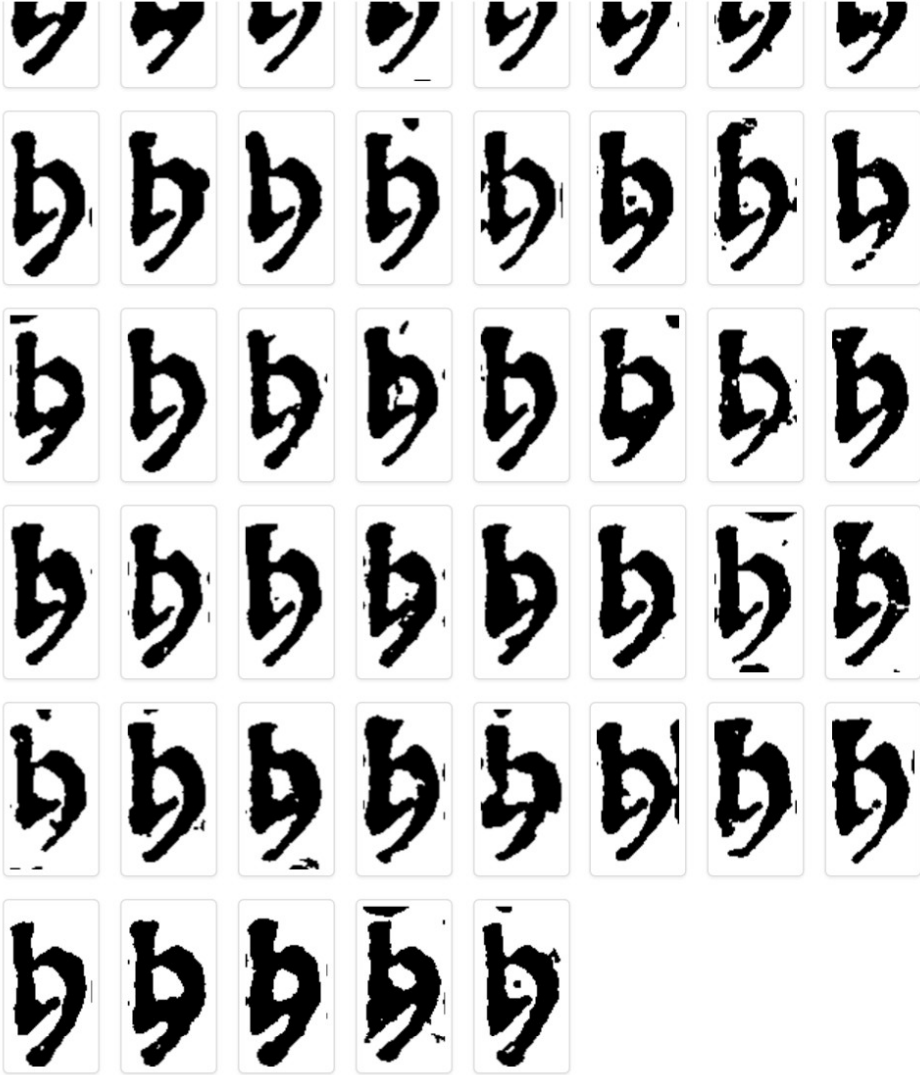


Template 3 – "h"
Position: (855, 1760),
Size: 56 x 109
Das Narrenschiff

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

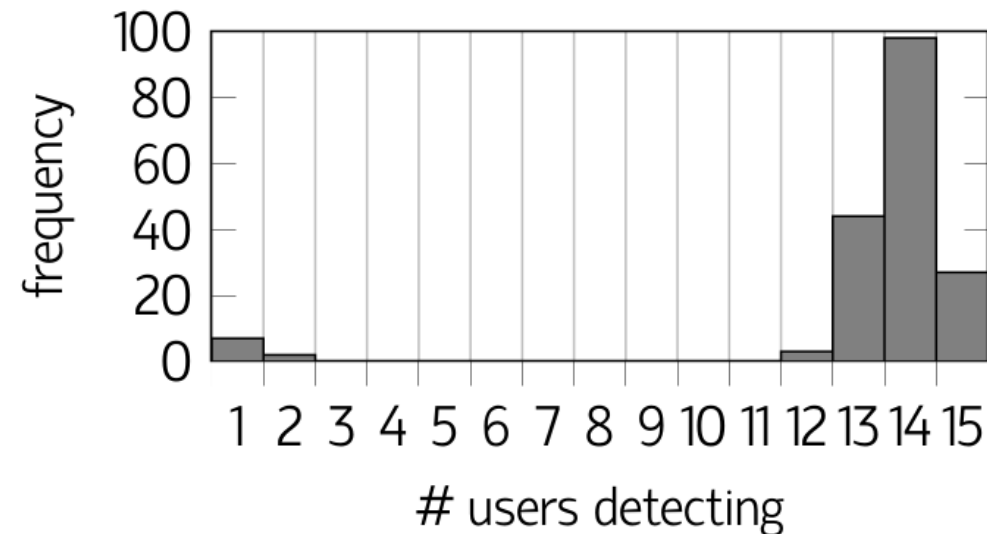
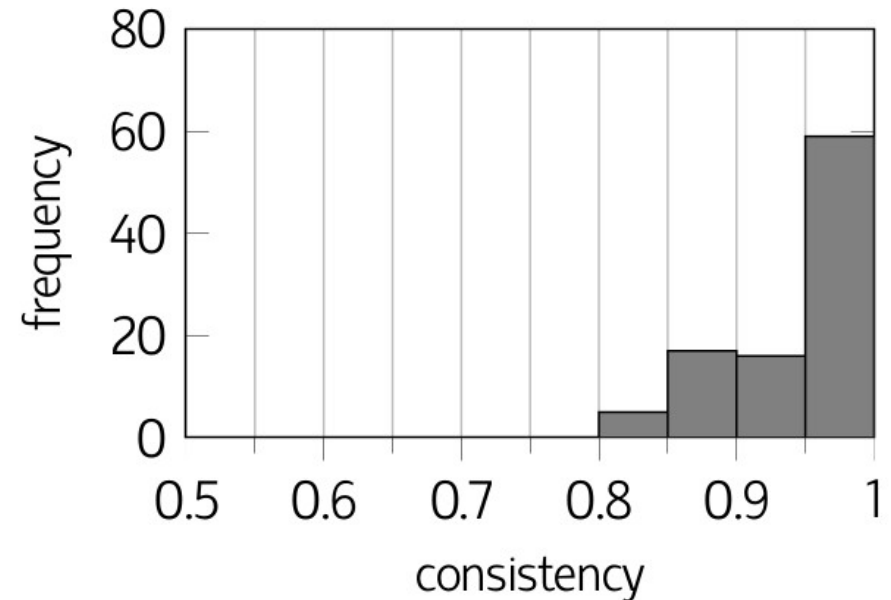
- Participants from Digital Humanities, OCR and linguistics
- Hands-on session, five pages from German *Narrenschiff*
- 30 minutes to process **a, d, e, t, y**



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

Template 



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

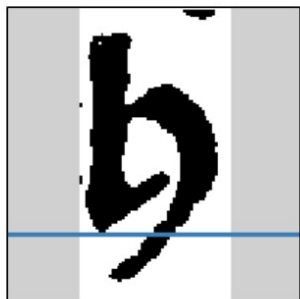
Select

Template 11 – "r"

Select

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 vcdl
hello world

Preview



hello vcdl
hello world

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:

