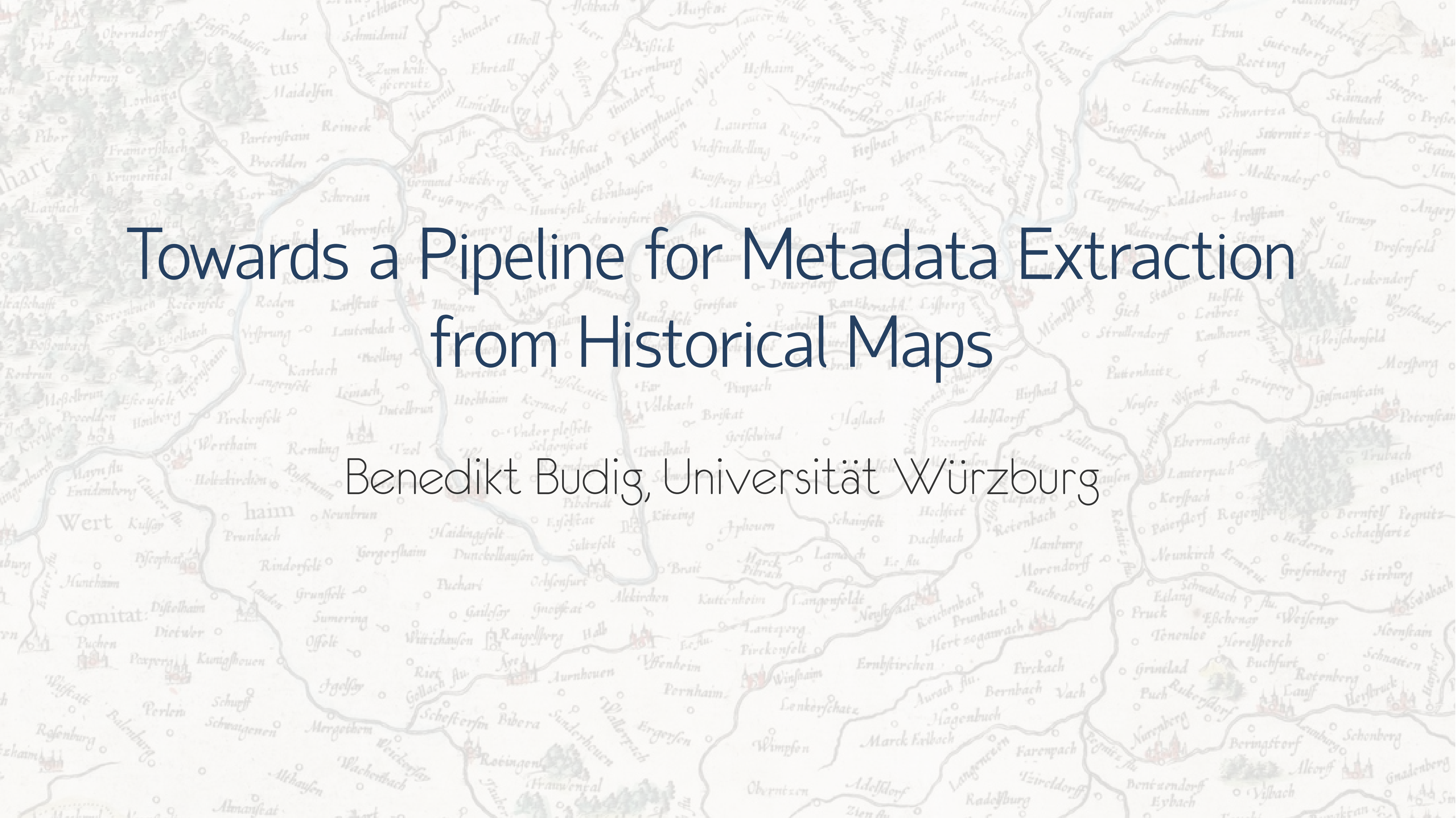


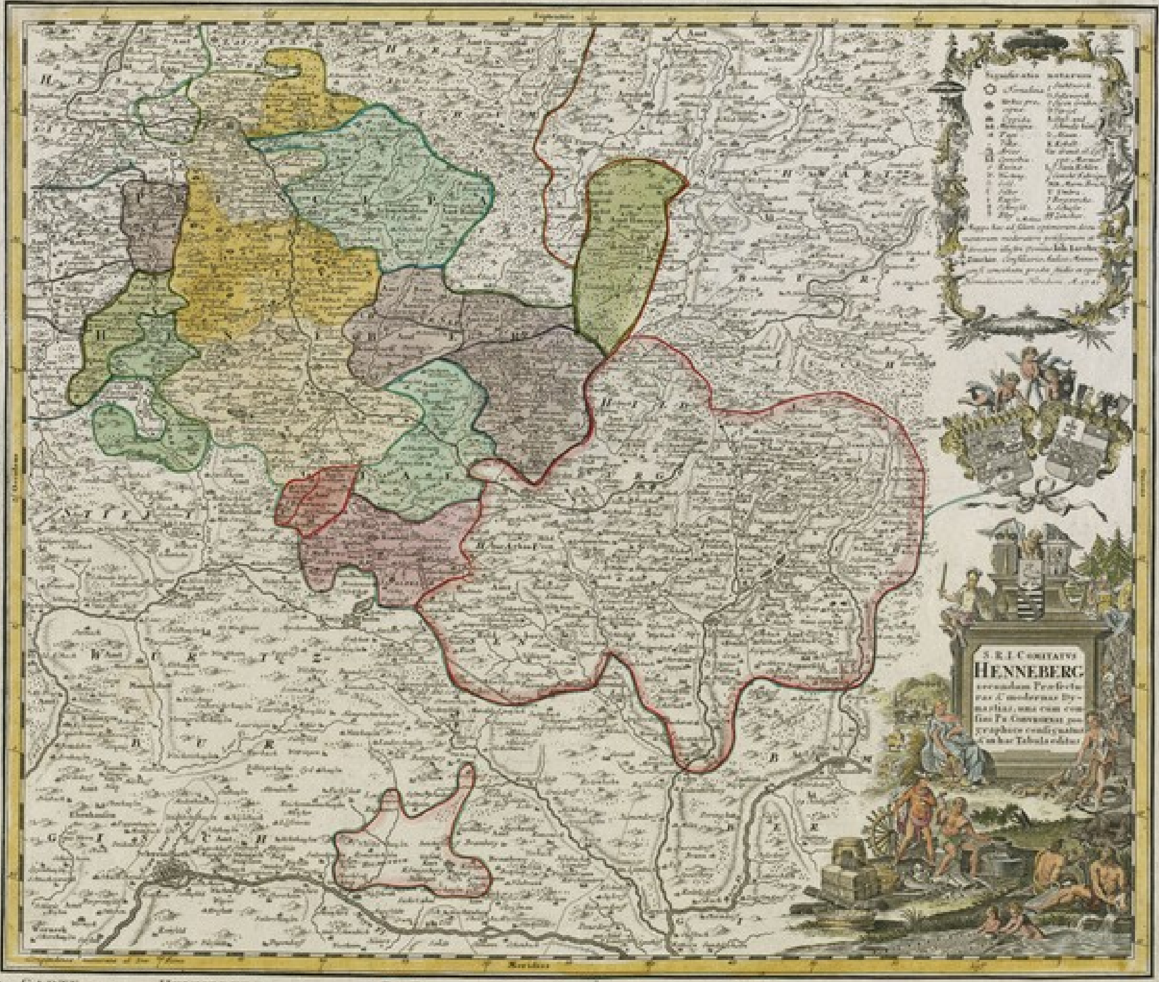
Towards a Pipeline for Metadata Extraction from Historical Maps

Benedikt Budig, Universität Würzburg

Overview

- Historical Maps: what and why?
- Sketch of a Pipeline
  - from bitmap image to georeferenced metadata
- Open Questions & Future Work

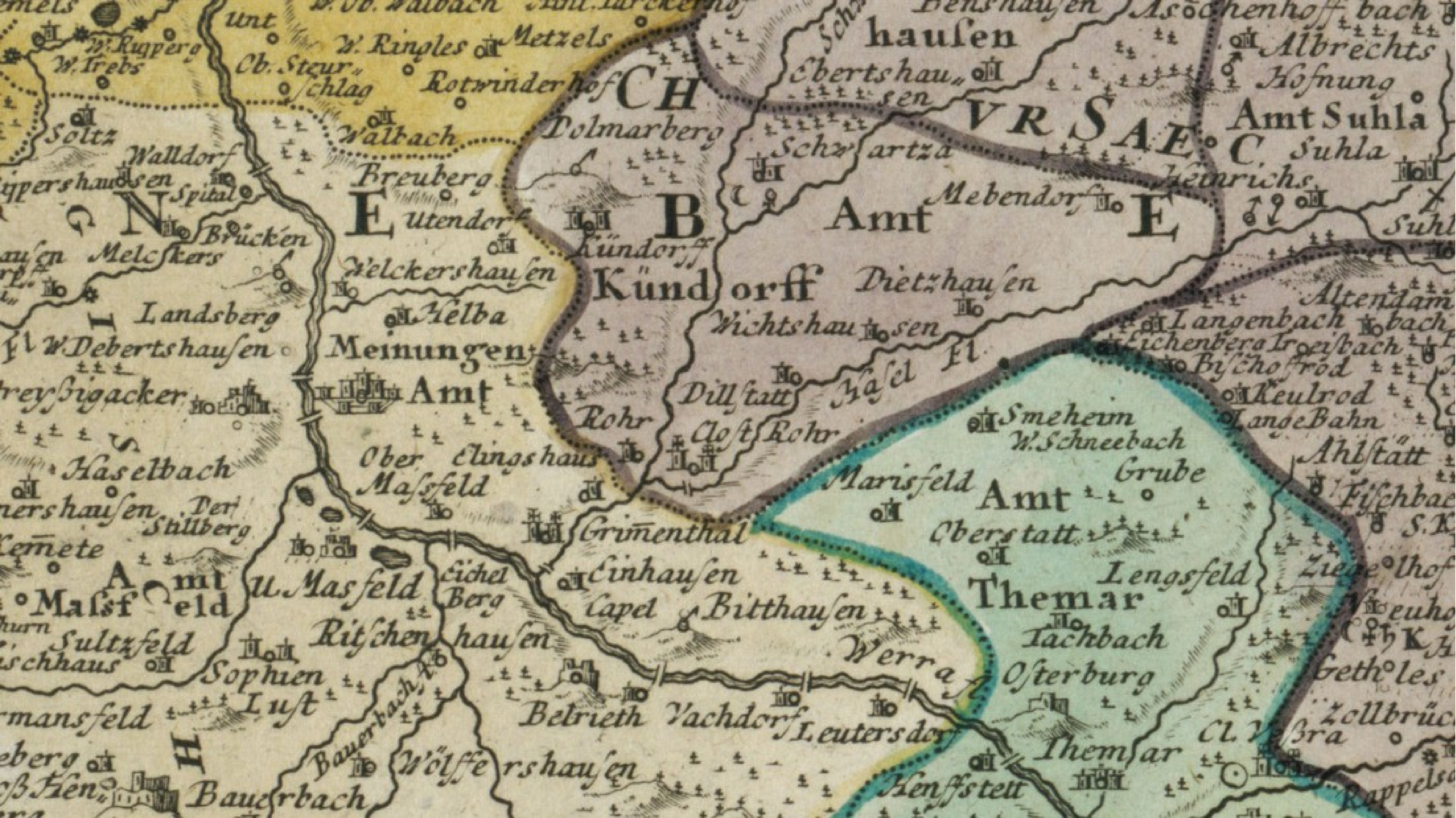
Karte der Gefürsteten GRAFSCHAFT HENNEBERG, mit dem angränzenden FIRSTENTH. COBURG u. andern GRAENZLAENDERN, nach authentischen Documenten
und Nachrichten verfertiget, und in ihre Aemter eingetheilet. Herausgegeben von Homännischen Erben in Nürnberg A. 1743

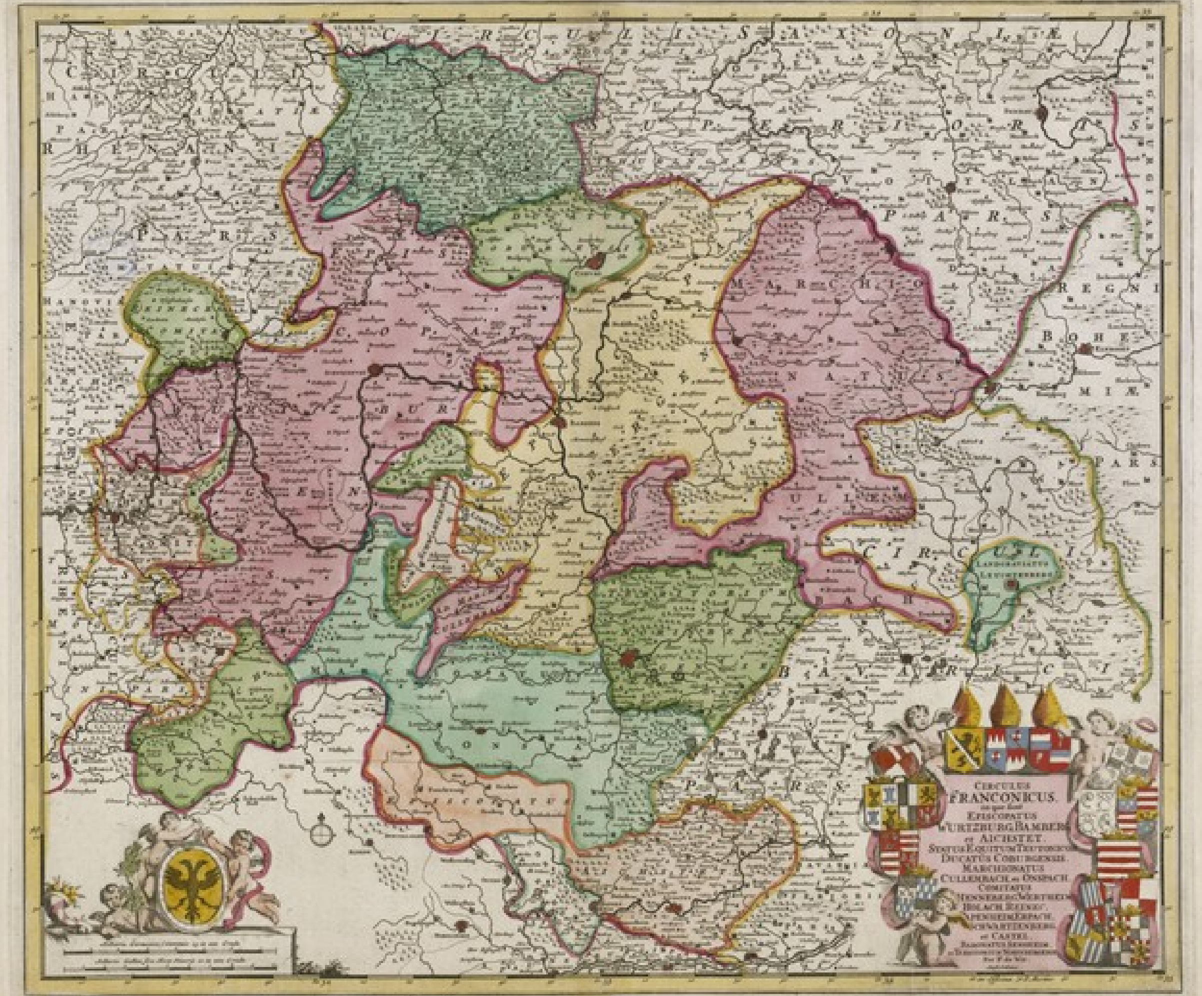

Septentrio
Meridies
HESSEN
FULDA
HENNEBERG
STIFFT
WURTZBURG
BAMBERG
S. R. I. COMITATUS
HENNEBERG
secundum Praefecturas & modernas Dynastias, una cum confinis Pr. COBURGENSIS geographice consignatus Cum hac Tabula editus
Significatio notarum

La CARTE du Comté de HENNEBERG avec les pais voisins du Principauté de COBURG le tout subdivisé en ses Baillages et dressé selon les memoires les plus authentiques
Tres Milliaria Germanica quorum 15 in Gradu
A Nuremberg chez les Heritiers de Homann A. 1743.
Duae Milliaria Gall. 20 in Gradu

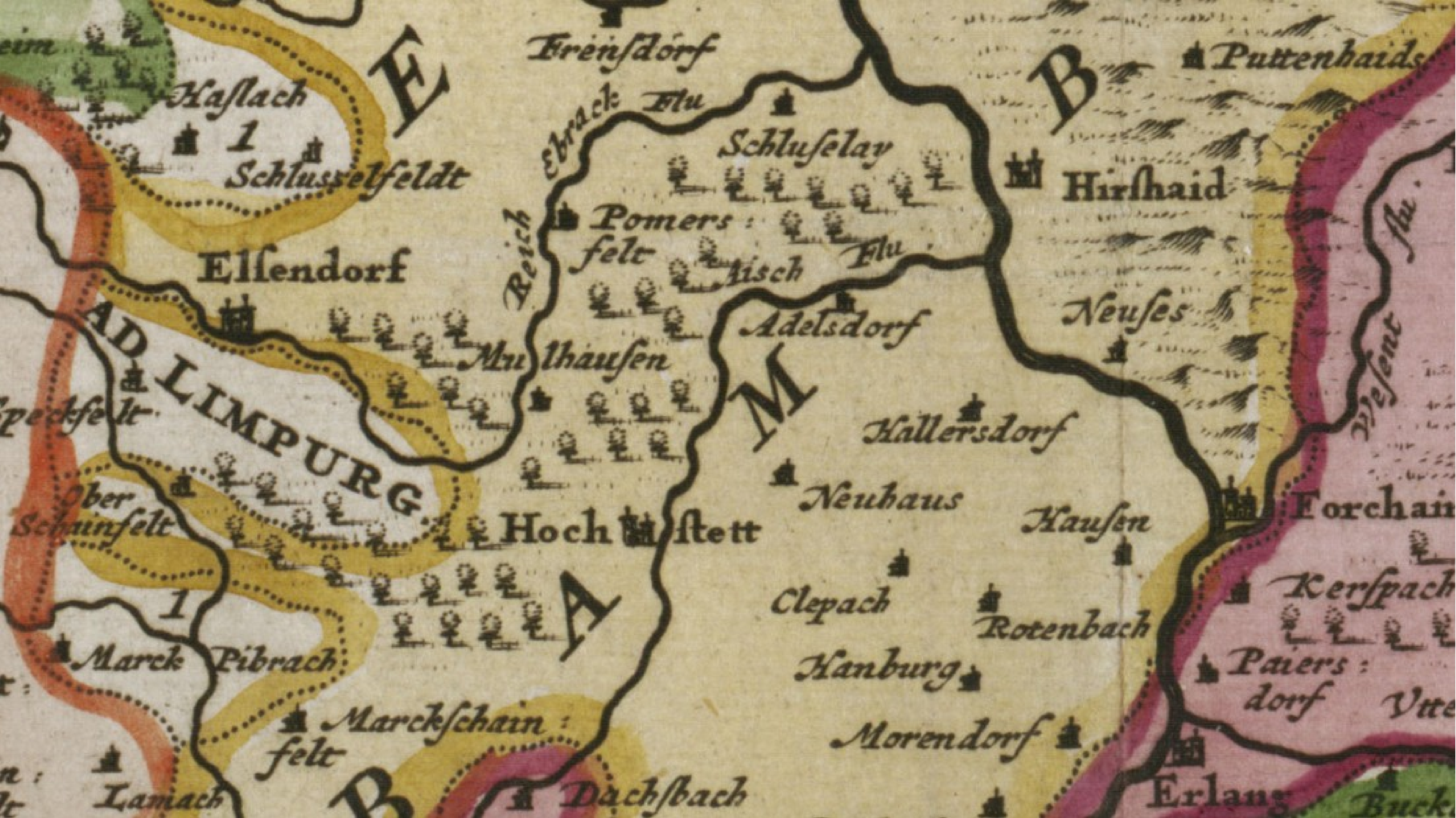
5

ut ſubdiviſé en ſes B
Homan̄. A. 1743.


W. Ob. Walbach
Benshauſen
Aſchenhoff
bach
unt
W. Rugberg
W. Kingles
Metzels
hauſen
Albrechts
W. Irebs
Ob. Steur ſchlag
Rotwinderhof
CH
Ebertshauſen
Hofnung
Schw
Soltz
Walbach
Dolmarberg
VR
S
AE
Amt Suhla
Walldorf
Schwartza
C.
Suhla
ppershauſen
Spital
Breuberg
Heinrichs
N
G
E
Utendorf
B
Amt
Mebendorf
E
Suhl
Brücken
Melckers
Kündorff
Welckershauſen
Kündorff
Dietzhauſen
Landsberg
Helba
Wichtshauſen
Altendambach
Langenbach
W. Debertshauſen
Meinungen
Haſel Fl.
Eichenberg
Ireiſbach
Biſchofrod
reyſsigacker
Amt
Dillſtatt
Keulrod
Lange Bahn
Rohr
Cloſt. Rohr
Smeheim
W. Schneebach
Ahlſtatt
Ober Maſsfeld
Elingshauſ
Haſelbach
Marisfeld
Amt
Grube
Fiſchbach
nershauſen
Der Stillberg
Grimenthal
Oberſtatt
S. K
emete
Einhauſen
Lengsfeld
Amt Maſsfeld
U. Masfeld
Eichel Berg
Capel
Bitthauſen
Themar
Sultzfeld
Ritſchenhauſen
Werra
Tachbach
K
ſchhaus
Sophien Luſt
Oſterburg
Bauerbach
rmansfeld
Belrieth
Vachdorf
Leutersdorf
Cl. Veſsra
Zollbrück
eberg
H
Wölfershauſen
Themar
Henſstett
Rappels
Bauerbach

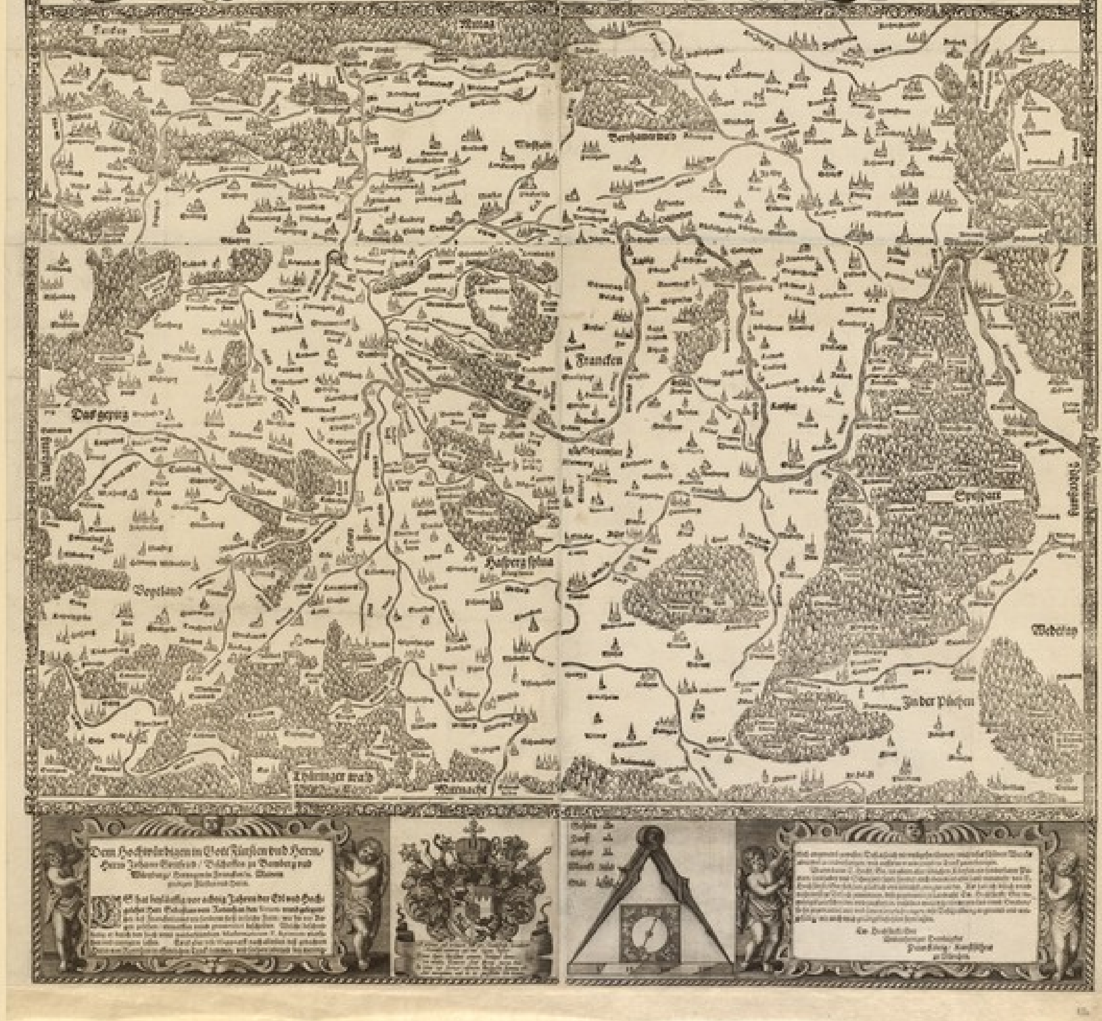

CIRCULUS FRANCONICUS, in quo sunt EPISCOPATUS WURTZBURG BAMBERG et AICHSTET. STATUS EQUITUM TEUTONICUS DUCATUS COBURGENSIS MARCHIONATUS CULLENBACH et ONSPACH COMITATUS HENNEBERG WERTHEIM HOLACH REINEC PENNEBERG ERBACH SCHWARTZENBERG et CASTEL BARONATUS SEINSHEIM
CIRCULI SAXONIÆ SUPERIORIS PARS
CIRCULI RHENANI PARS
EPISCOPATUS
MARCHIONATUS CULLENBACH
REGNI BOHEMIÆ PARS
CIRCULI BAVARICI PARS
LANDGRAVIATUS LEUCHTENBERG
ONSPACH
EPISCOPATUS

36 A 1 17



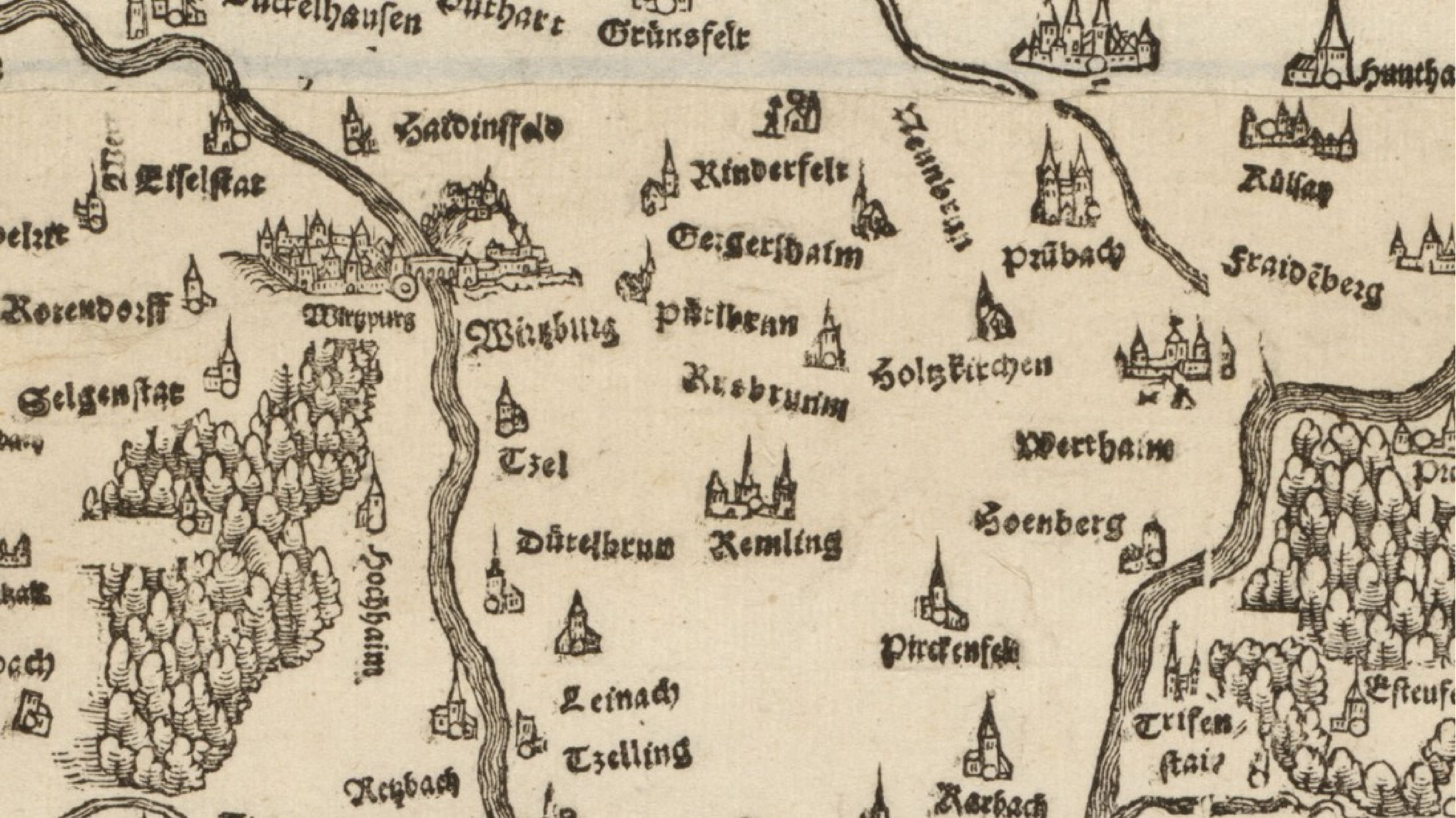
Frenſdorf
E
Haſlach
1
Schlusselfeldt
Ebrack Flu
B
Puttenhaids
Schluſelay
Hirſhaid
Reich
Pomers felt
Elſendorf
Aisch Flu
Adelsdorf
Neuſes
Wiſent flu
AD LIMPURG
Mulhauſen
Speckfelt
M
Hallersdorf
Ober Schwinfelt
Neuhaus
Hauſen
Hochſtett
Forchain
Kerſpach
Clepach
A
Rotenbach
1
Marck
Pibrach
Hanburg
Paiers dorf
Vtte
Marckſchain felt
Morendorf
Lamach
B
Dachſbach
Erlang
Buck

Das Francken Landt Chorographi Franciæ Orie.



Dem Hochwürdigen in Gott Fürſten vnd Herrn / Herrn Johann Gottfried / Biſchoffen zu Bamberg vnd Würtzburg / Hertzogen in Francken [illegible]

Es hat beyläuffig vor achtzig Jahren der Edel vnd Hochgelehrt Herr [illegible]


Buckelhauſen
Puchart
Grünsfelt
Hunthaſ
Haidinſfeld
Eiſſelſtat
Rinderfelt
Zemnbern
Rülſay
elzet
Geigerſhaim
Prübach
Fraideberg
Rorendorff
Wirtzpurg
Wirtzburg
Pütlbrun
Holtzkirchen
Roſbrunn
Selgenſtat
Werthaim
Tzel
Pro
Hoenberg
Hochhaim
Dürelbrun
Remling
kach
Pirckenfelt
bach
Leinach
Eſteuf
Triſen-
ſtain
Tzelling
Reybach
Karbach

Study historical maps: why?

- Many libraries have large collections of historical maps
- Relevant for the (digital) humanities
  - History of cartography
  - General history
  - Specific example: onomastics

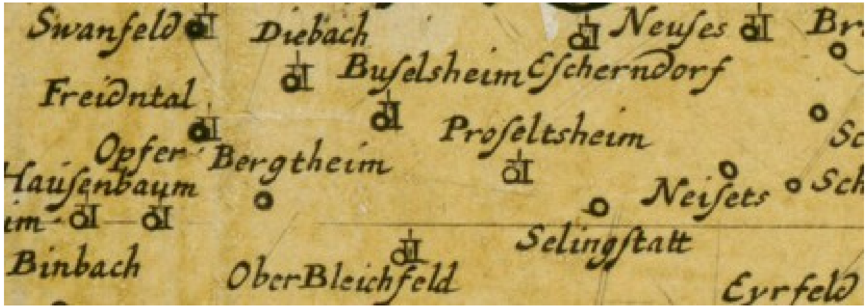
What happens with historical maps?

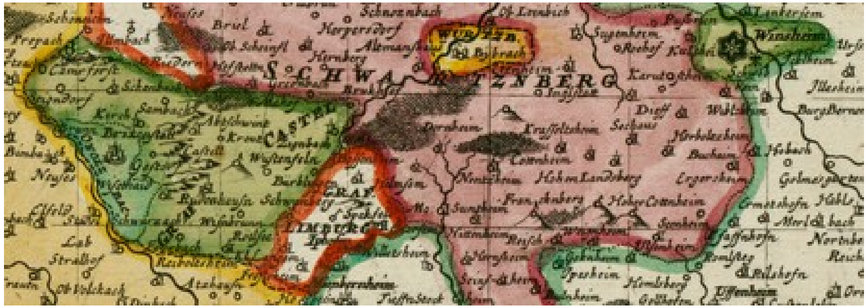
- Stored in a library basement
  - Retrievable by bibliographic information
- High-quality bitmap scans, online catalogue
  - Browsable by bibliographic information
- Useful queries?
  - In actual research practice
  - By interested laypeople

→ not bibliographic information, but metadata on actual contents

Metadata: what?

- Contained settlements
- Landscape topography
- Geopolitical features
- ...



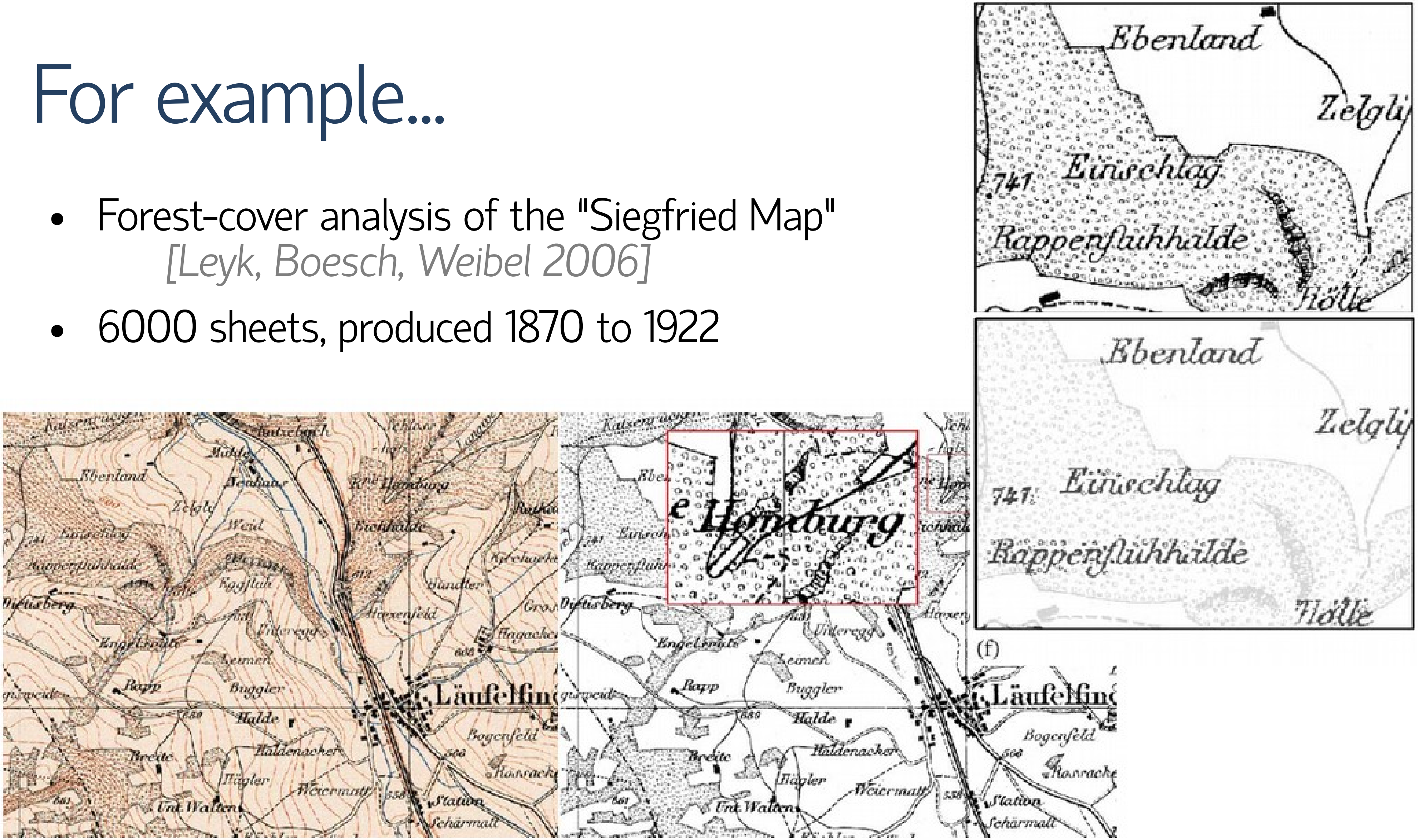


Metadata: how?

- Do it by hand
- Software: usability improvements *e.g. [Simon et al. 2011, 2015]*
  - Gains in efficiency are limited
- Software: computer vision *[Chiang 2014]*
  - No panacea, but can work well for restricted corpora
  - Significant custom R&D effort every time

For example...

- Forest-cover analysis of the "Siegfried Map"
  [Leyk, Boesch, Weibel 2006]
- 6000 sheets, produced 1870 to 1922



(f)

Our scope

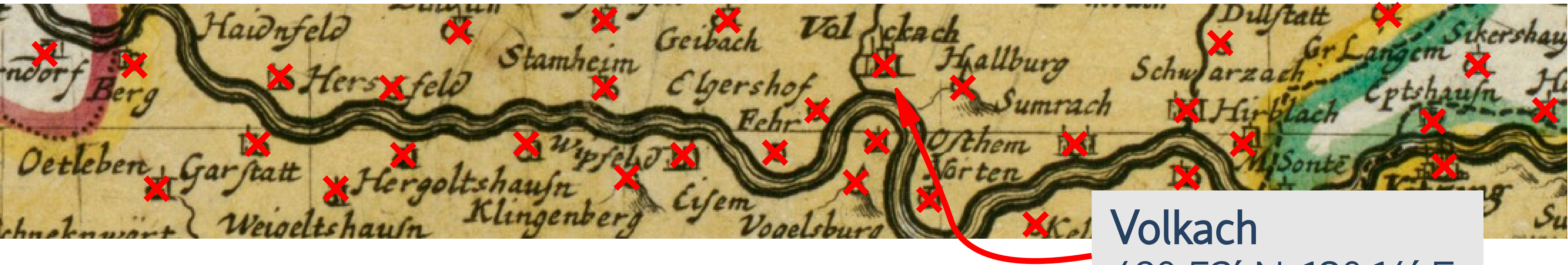
- We consider maps from early modern period forward
- Unique graphical styles, different fonts, handwriting
- Different cartographic conventions, heavy distortions

Goal: extract and georeference metadata

Note: georeference *metadata*,
not just map sheets

Deep Georeferencing

- Georeference individual elements contained in a map

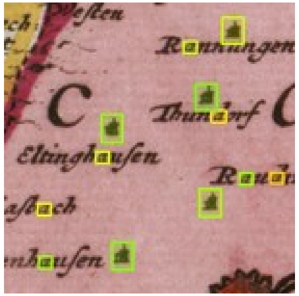


Volkach
49º 52′ N, 10º 14′ E

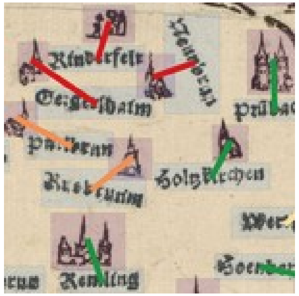
- Extraction strategy:
  - Locate map element and its corresponding label
  - Read label to identify and georeference element

So what now?

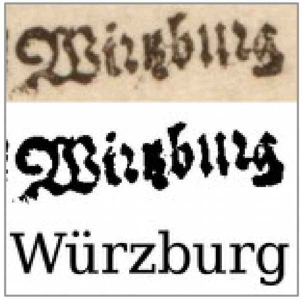
- Split problem into smaller goals
- Design a modular pipeline



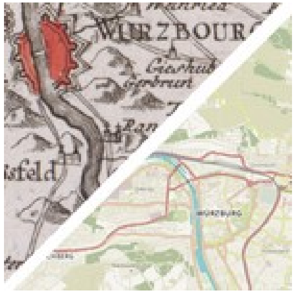
Segmentation



Clustering and Matching



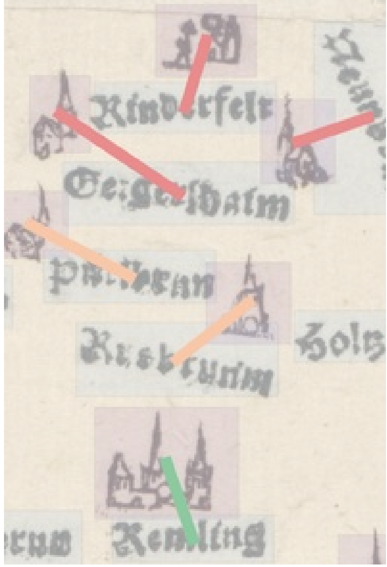
Understanding Text



Georeferencing

Weſten
Rannungen
C
Thundorf
Eltinghauſen
aſbach
enhauſen

Segmentation

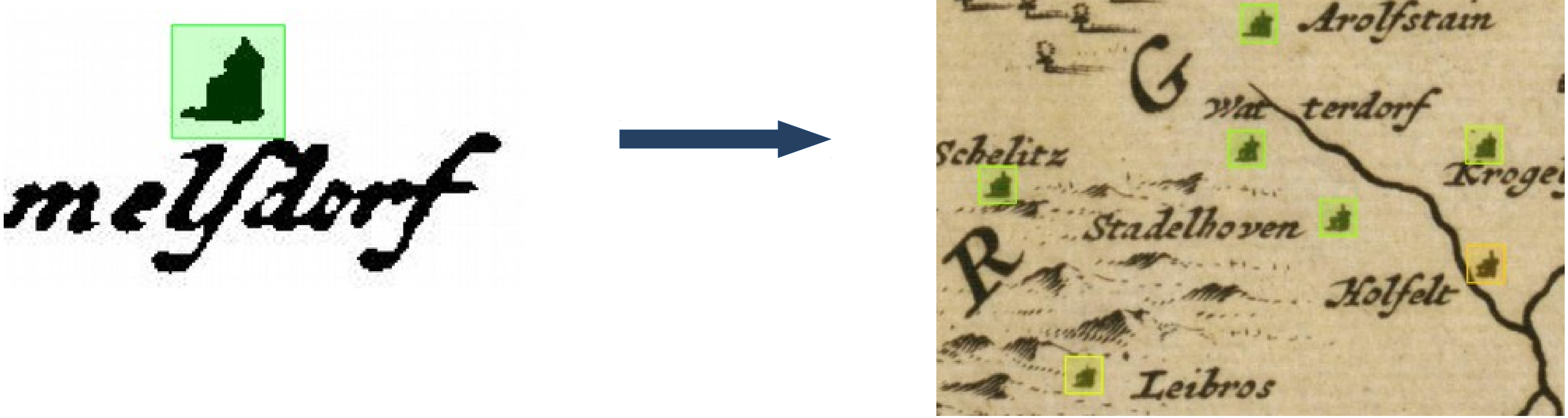


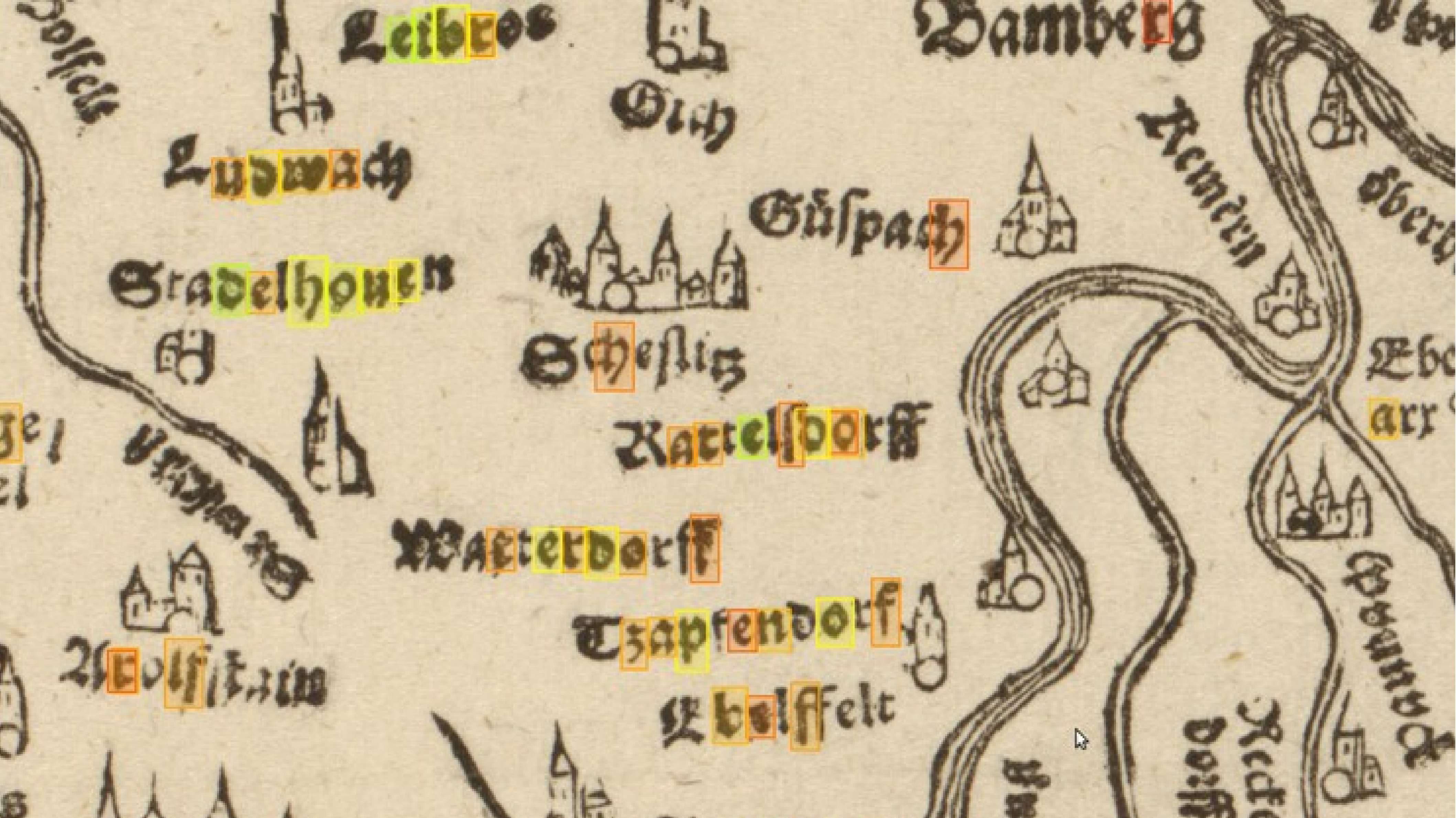
Clustering
Matchi

Segmentation

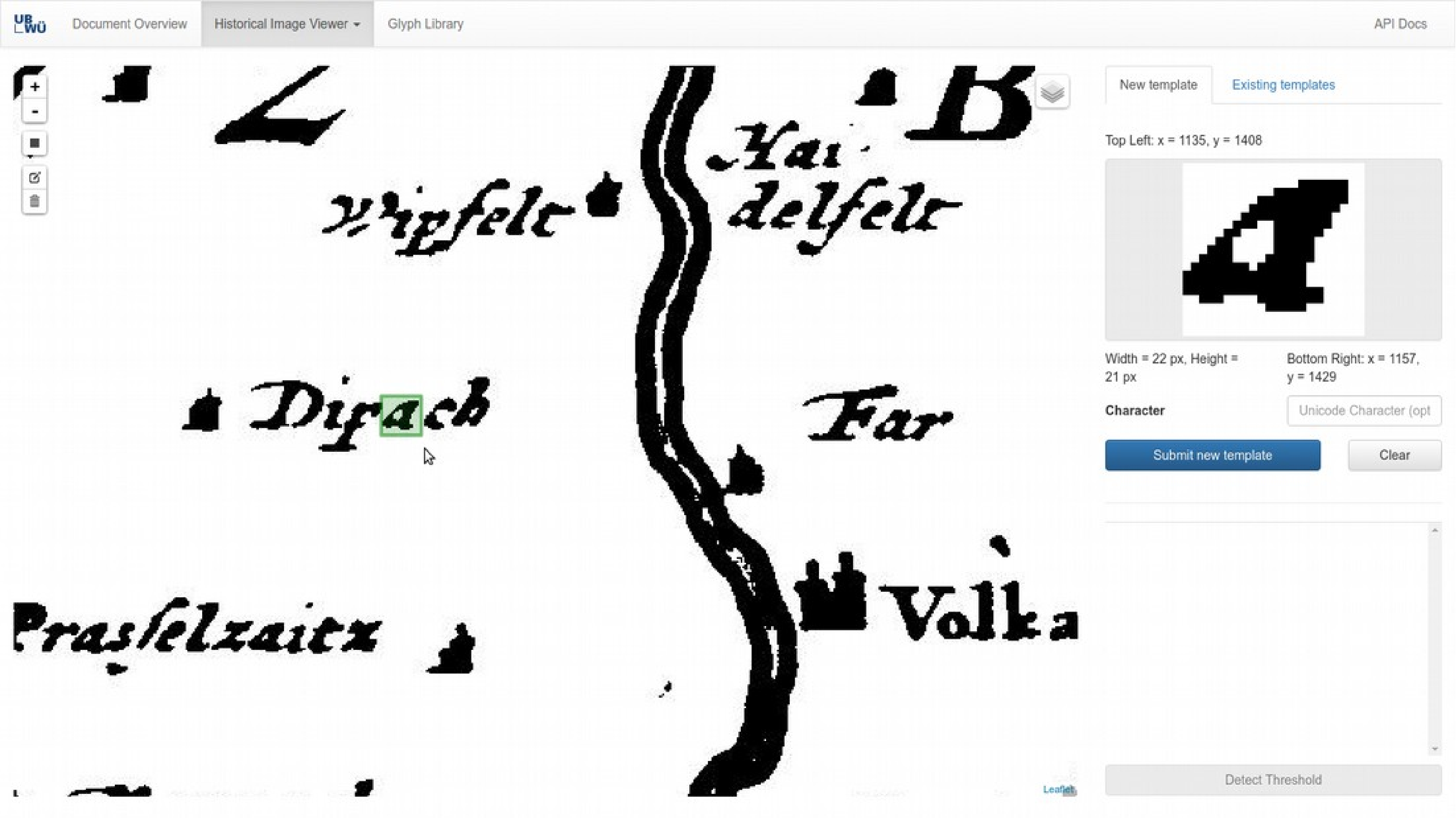
- Smaller goals
- Look for one particular element on one map

[Budig and Van Dijk 2015]





Leibros
Bamberg
Gich
Ludwach
Gradelhouen
Gůſpach
Scheßlitz
Remern
Ratteldorff
Watterdorff
Tzapfendorf
Ebelſfelt
Wolfſtein
Eb
arr

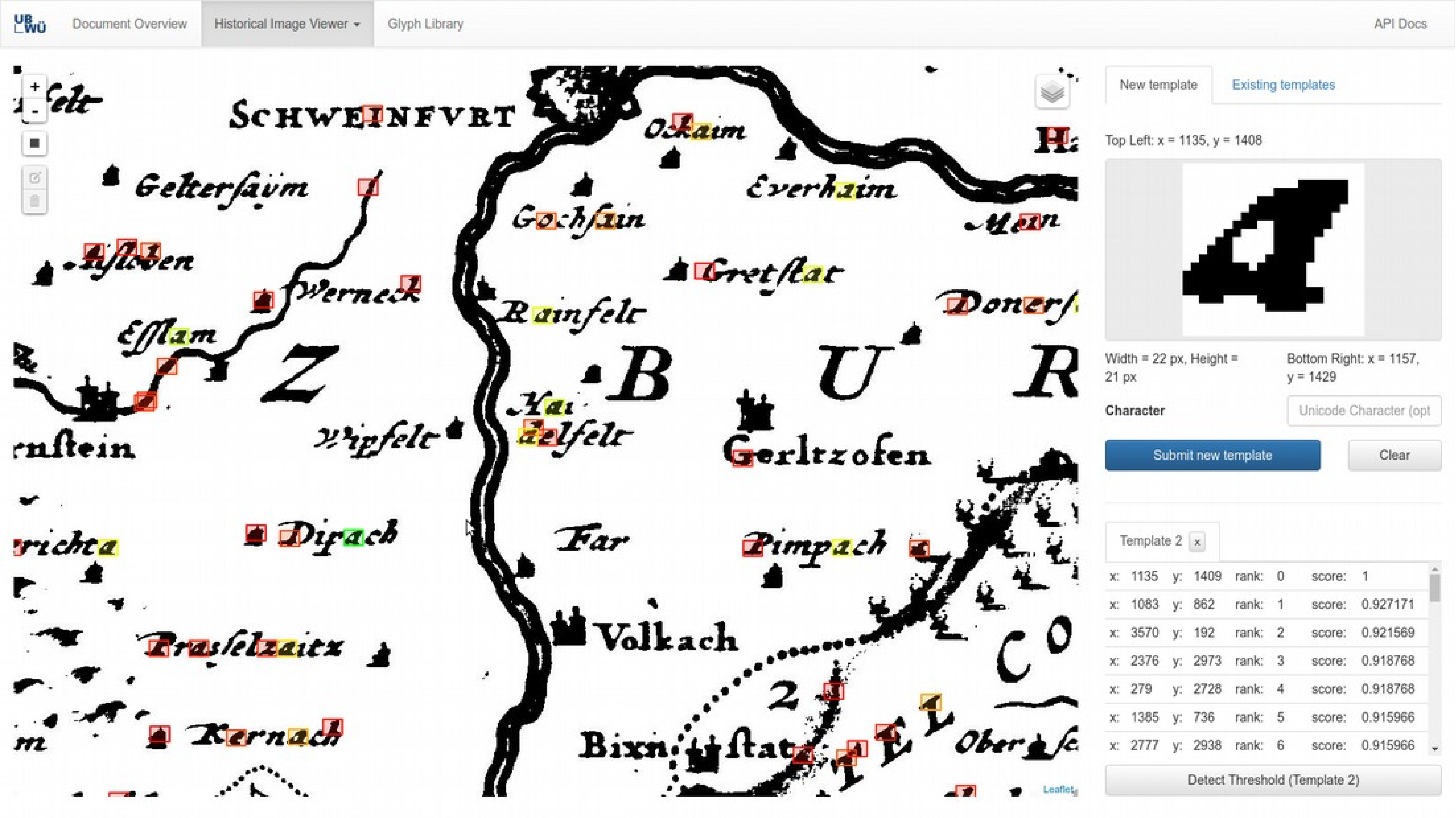


Top Left: x = 1135, y = 1408

Width = 22 px, Height = 21 px

Bottom Right: x = 1157, y = 1429

Character: Unicode Character (opt



New template | Existing templates

Top Left: x = 1135, y = 1408

Width = 22 px, Height = 21 px

Bottom Right: x = 1157, y = 1429

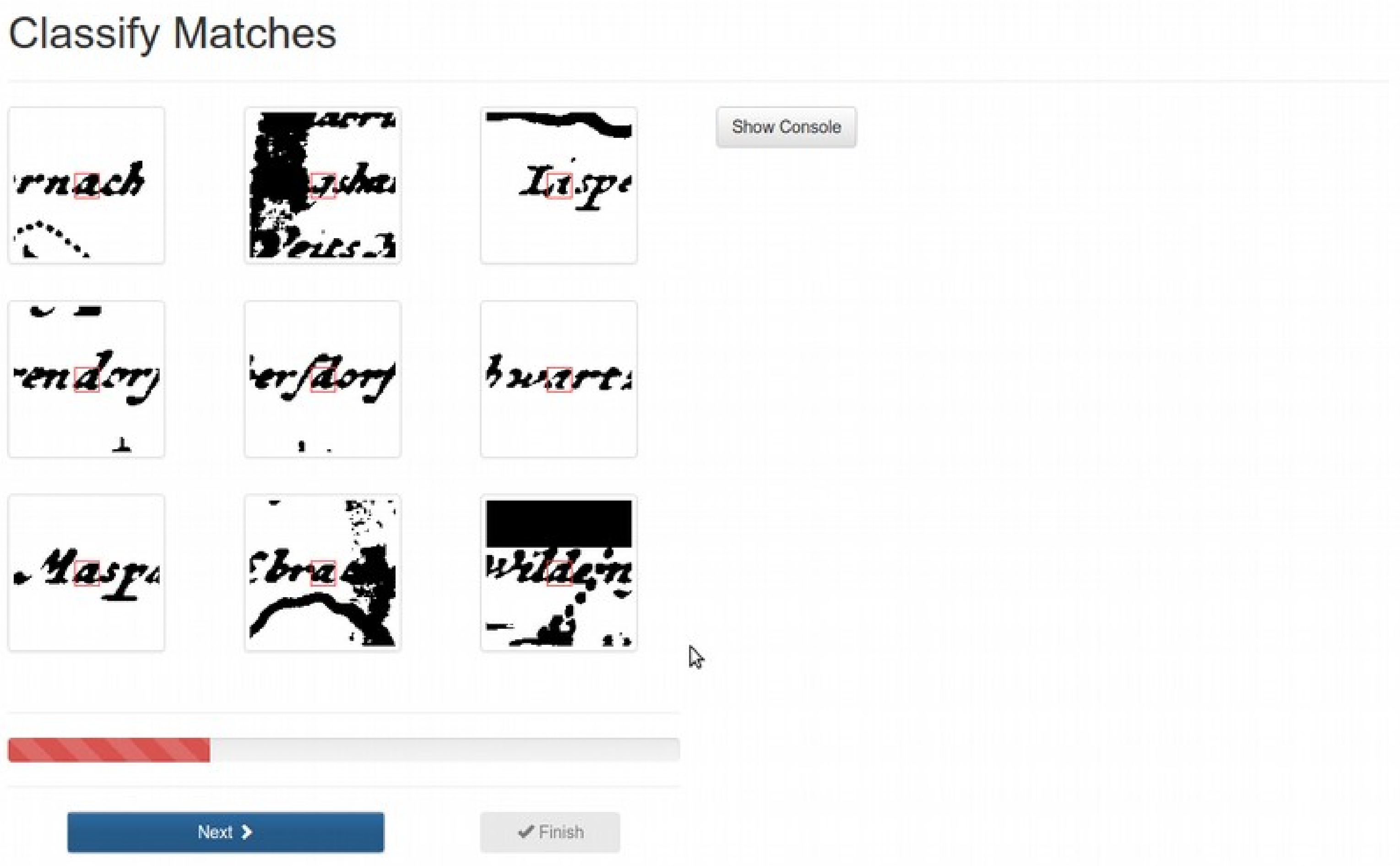
Character: Unicode Character (opt

Submit new template | Clear

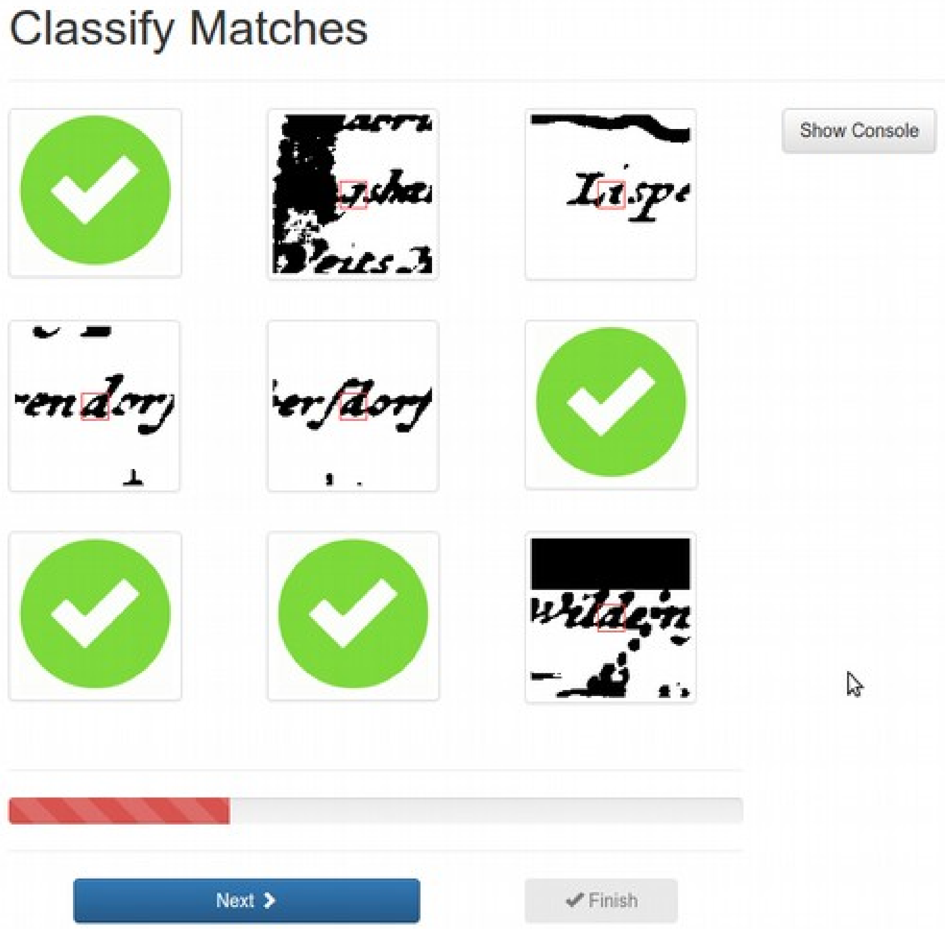
Template 2 x

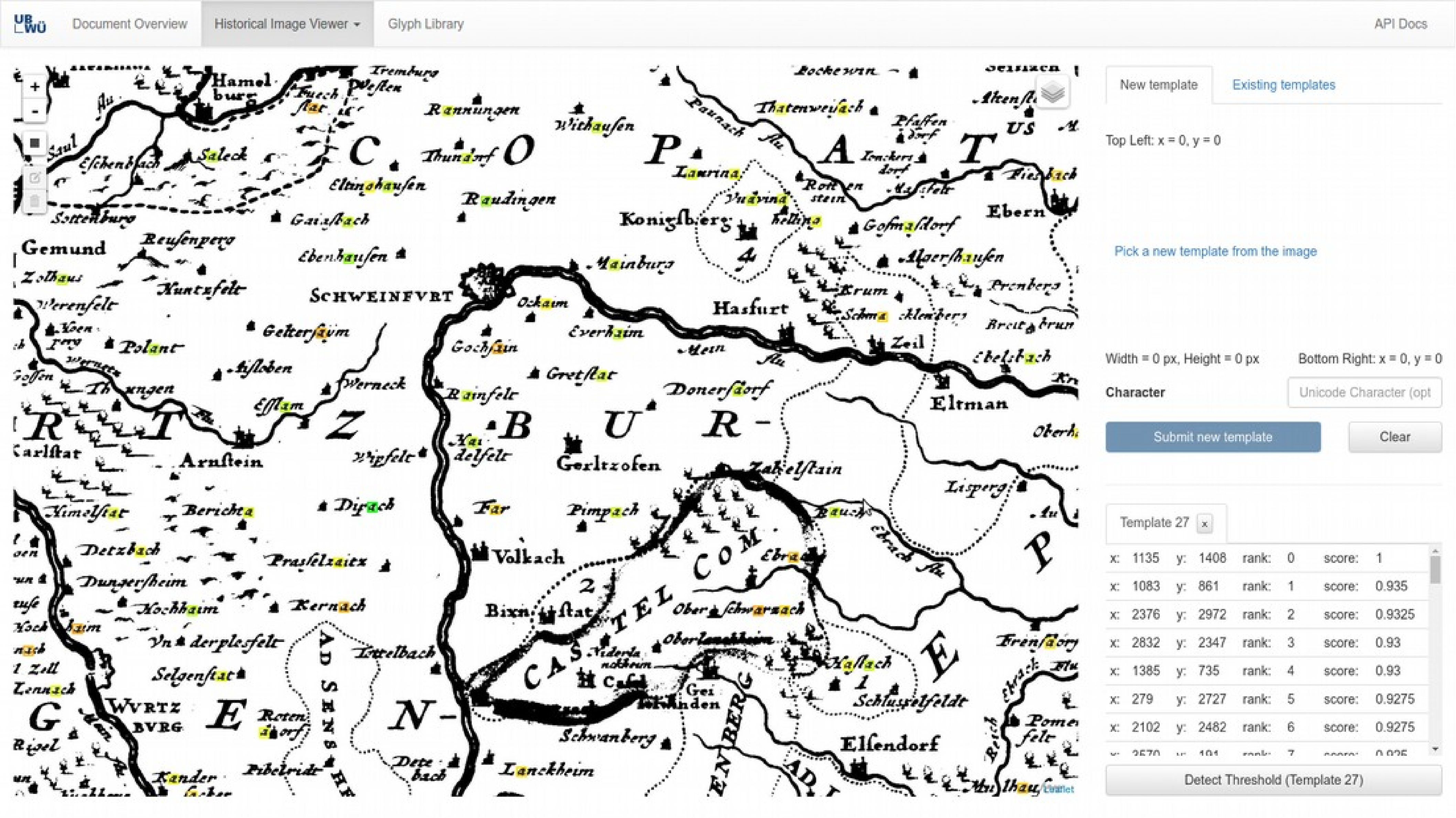
x:	1135	y:	1409	rank:	0	score:	1
x:	1083	y:	862	rank:	1	score:	0.927171
x:	3570	y:	192	rank:	2	score:	0.921569
x:	2376	y:	2973	rank:	3	score:	0.918768
x:	279	y:	2728	rank:	4	score:	0.918768
x:	1385	y:	736	rank:	5	score:	0.915966
x:	2777	y:	2938	rank:	6	score:	0.915966

Detect Threshold (Template 2)



Classify Matches


Classify Matches
Show Console
Next ›
✓ Finish



New template

Existing templates

Top Left: x = 0, y = 0

Pick a new template from the image

Width = 0 px, Height = 0 px

Bottom Right: x = 0, y = 0

Character: Unicode Character (opt

Submit new template

Clear

Template 27 x

x:	1135	y:	1408	rank:	0	score:	1
x:	1083	y:	861	rank:	1	score:	0.935
x:	2376	y:	2972	rank:	2	score:	0.9325
x:	2832	y:	2347	rank:	3	score:	0.93
x:	1385	y:	735	rank:	4	score:	0.93
x:	279	y:	2727	rank:	5	score:	0.9275
x:	2102	y:	2482	rank:	6	score:	0.9275
x:	2570	y:	101	rank:	7	score:	0.925

Detect Threshold (Template 27)

Segmentation: two ingredients

Ingredient 1: Template Matching

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

Segmentation: two ingredients

Ingredient 1: Template Matching

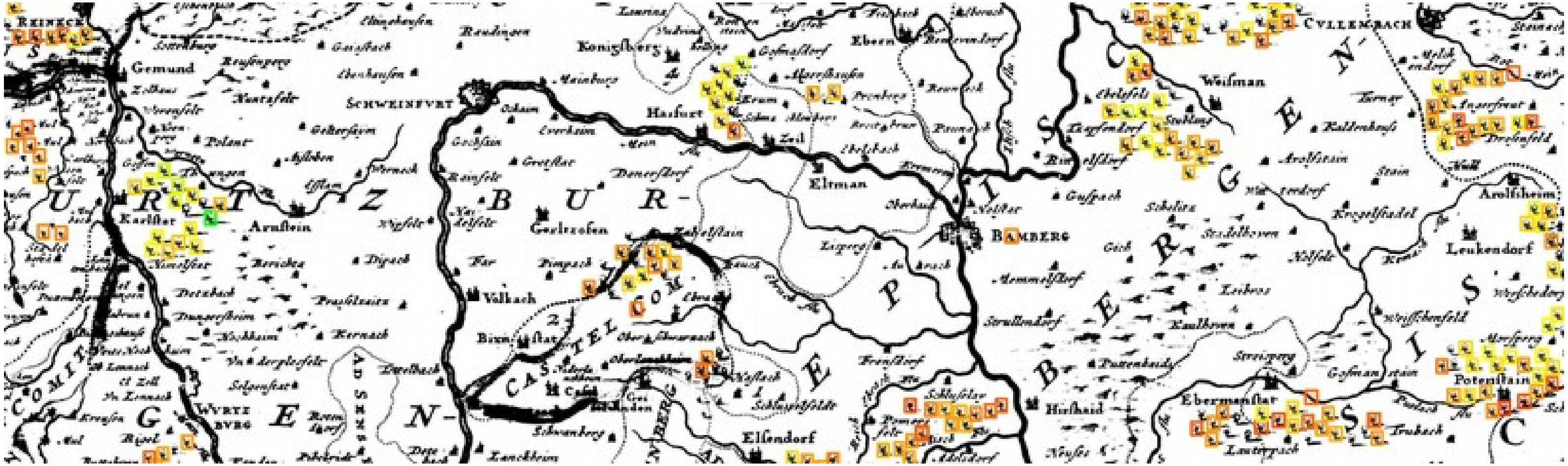
- Find approximate repeat-occurrences of an example image
- Here: black-and-white, only translation

Ingredient 2: Active Learning

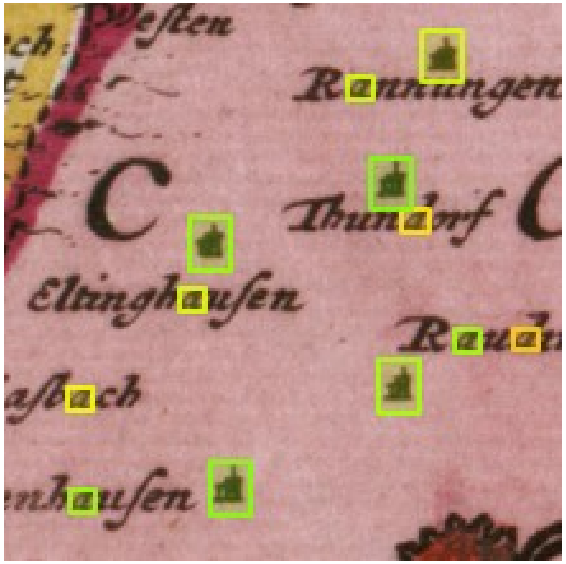
- Distinguish matches that are semantically correct from the rest
- Efficient user interaction

Segmentation: open questions

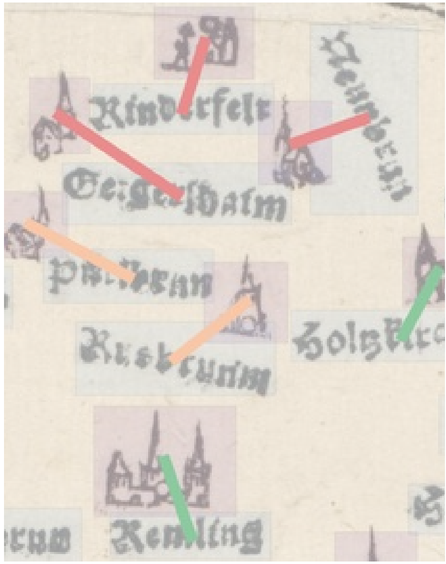
- How to locate landscape topography?
  - Template matching works for some features (on some maps)



- How to locate geopolitical features?



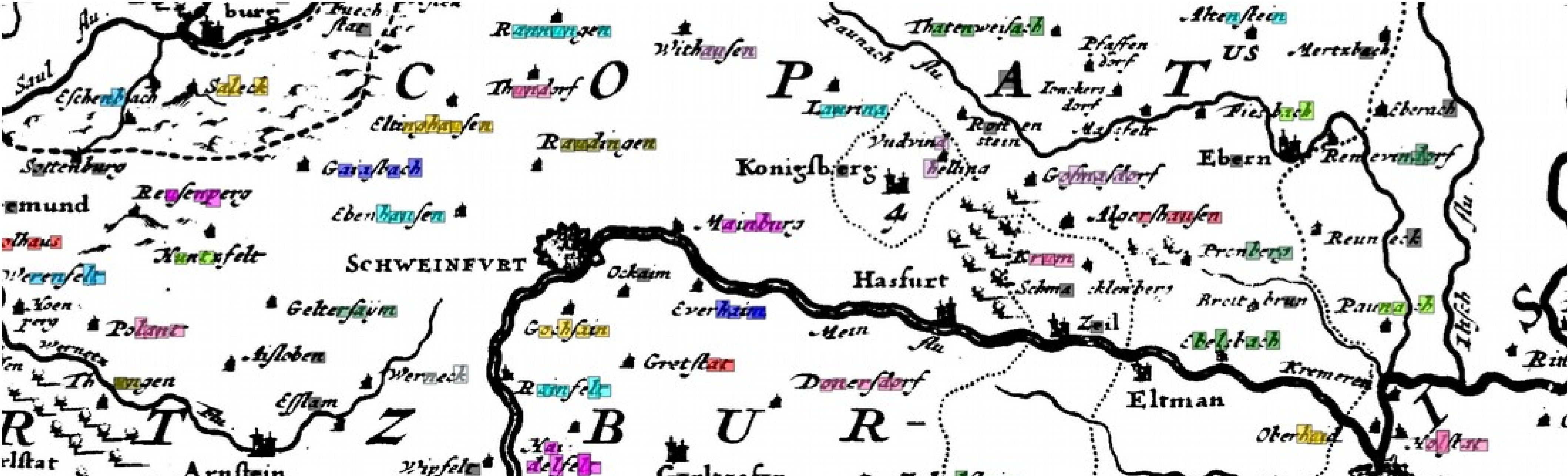
Segmentation



Clustering a
Matching

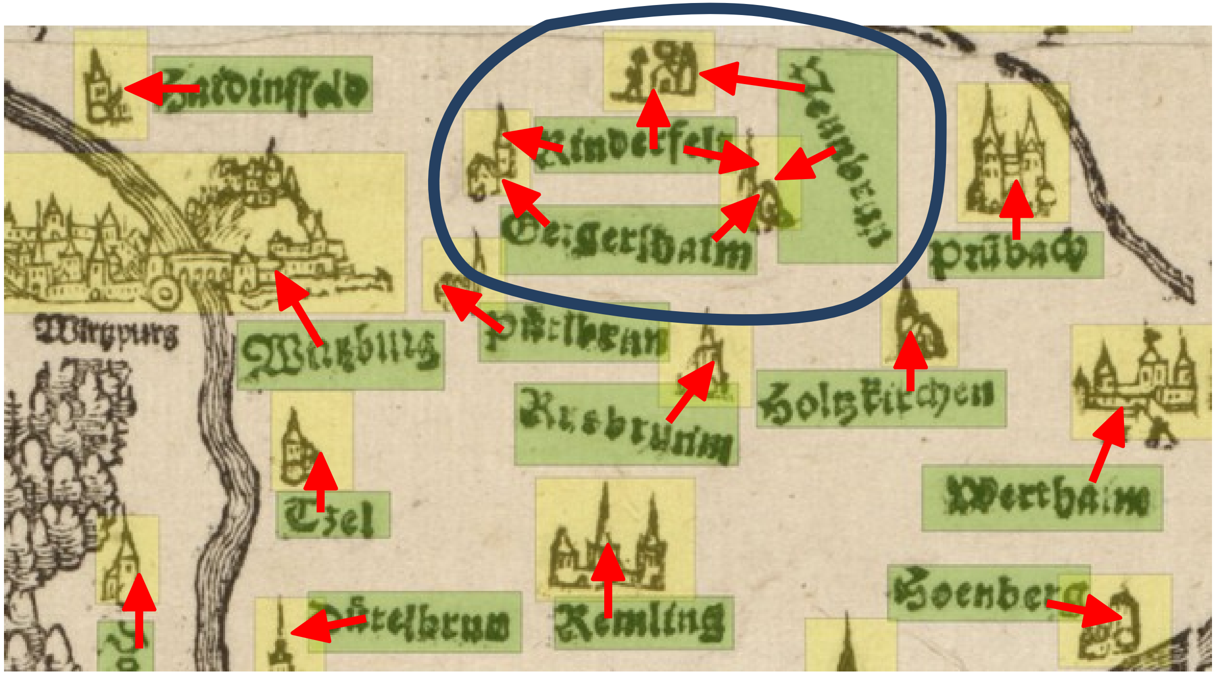
Clustering and Matching: open question

- Given matches of characters, how can we get labels?
  - Use clustering algorithms like DBSCAN?
  - Take the image into account (using approaches from computer vision)?



Matching Labels and Place Markers

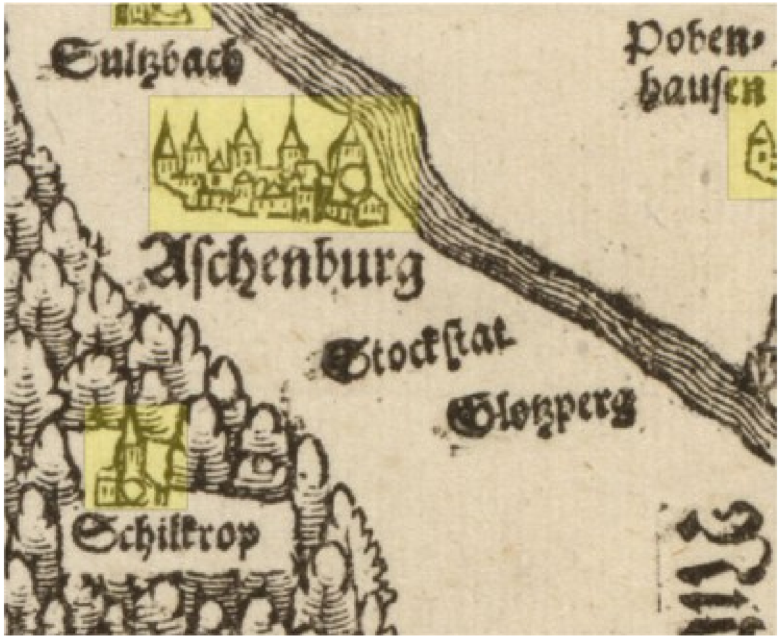
- Assumption: labels and markers already detected
- Match the corresponding ones [Budig, Van Dijk, Wolff, 2014]



Wanted: a Matching

- Find a *matching* of labels and place markers
- No 1-to-1 assignment possible
- Basic assumption: labels are *near* their corresponding markers
- Greedy strategy?

  → does not work well!
- Model as optimization problem



Experimental Results

- Franckenlandt (1533)
  - 539 markers, 524 labels
  - our algorithm: error rate 3.5%
  - greedy algorithm: error rate 17.8%

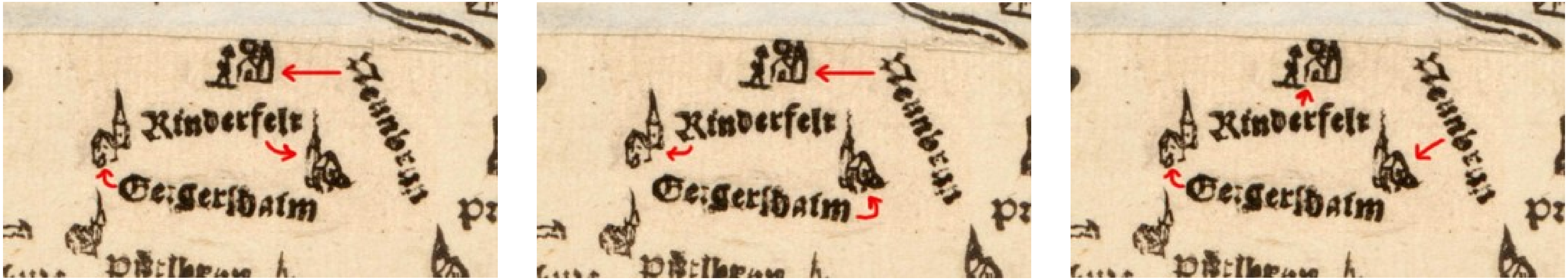
Rinderfelt
Remling

- Circulus Franconicus (1706)
  - 1663 markers, 1669 labels
  - our algorithm: error rate 1.3%
  - greedy algorithm: error rate 5.9%

Romhilt
Königſhove
Alſtat
G.Eibſtat
S
C
O
Sulzdorf

What now?

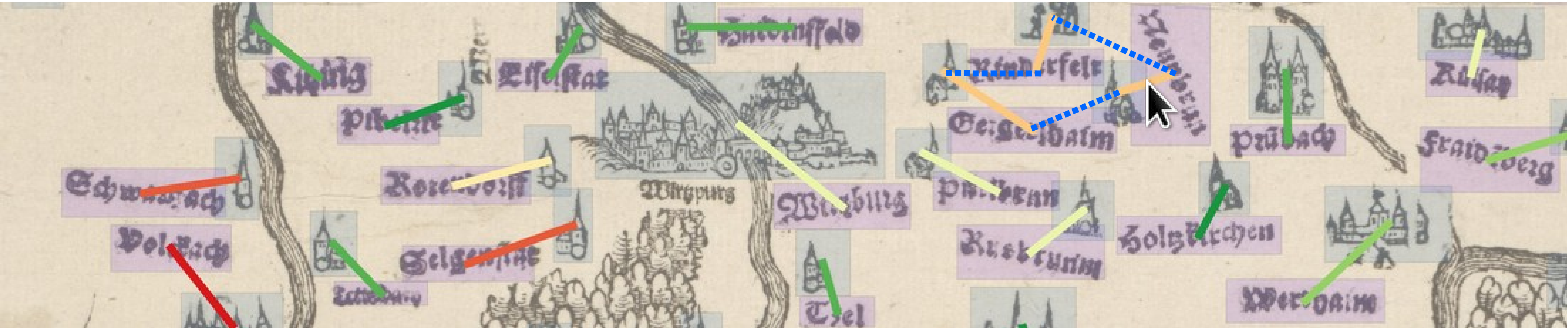
- Error rates in experiments: 1.3% and 3.5%
- Unclear situations:

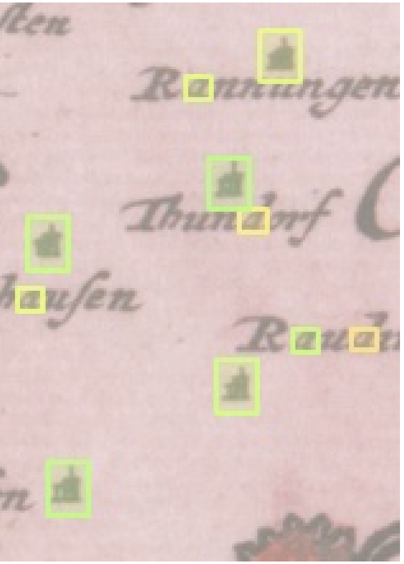


- Manual verification or correction necessary

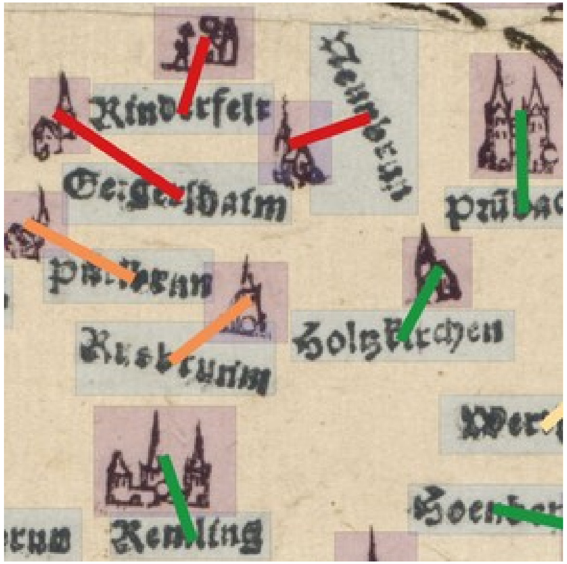
Sensitivity

- Calculate sensitivity analysis for the matching
- Only show assignments our algorithm is *uncertain* about





gmentation



Clustering and Matching

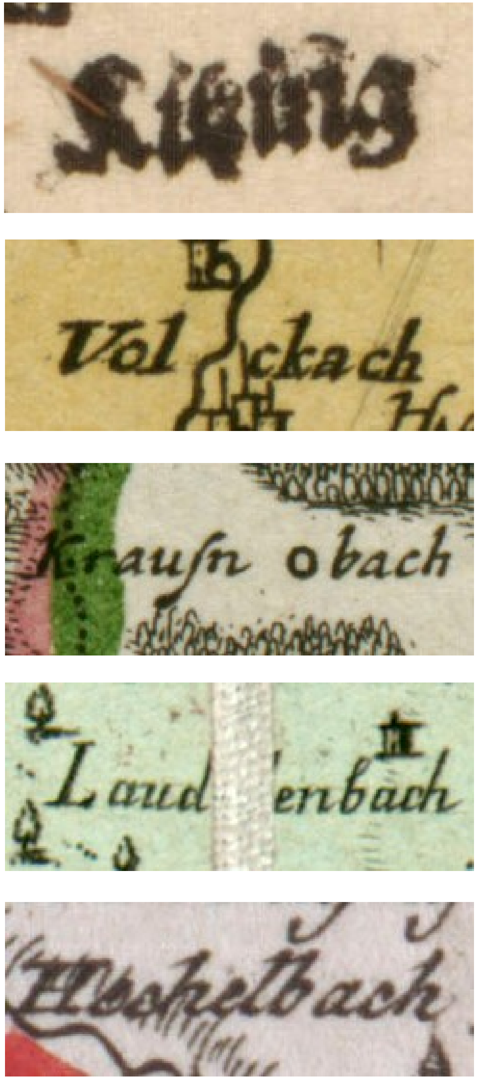
Understand Text

Understanding Text

Challenges:

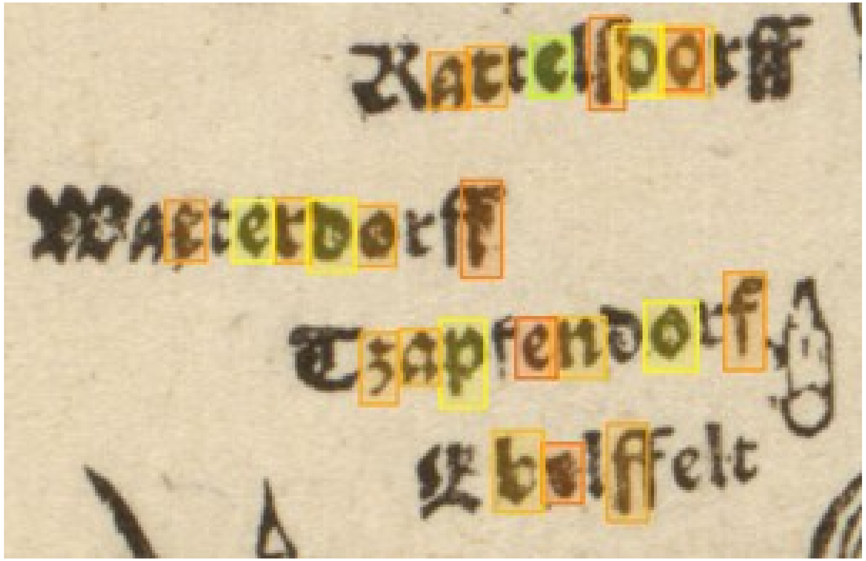
- Handwritten
- Poor conservation state
- Difficult layout, background noise

→ Off-the-shelf OCR software not suitable

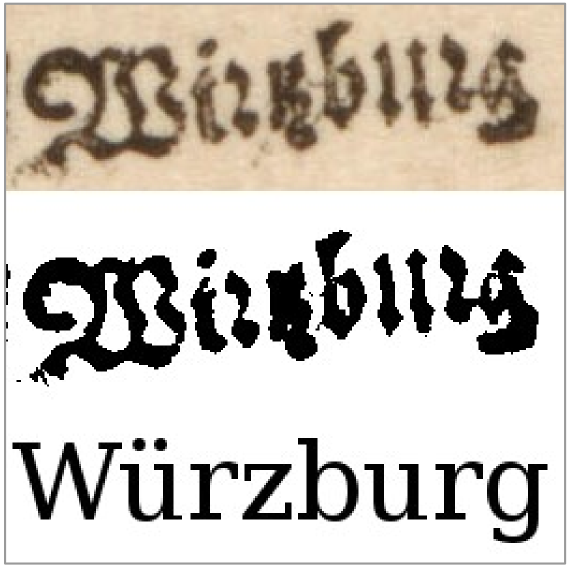


Understanding Text: open questions

- Train OCR engine, e.g. Tesseract or OCRopus?
  - But limited training data, unless generated synthetically
- Derive text directly from template matches?
  [Caluori and Simon 2013]
- Use gazetteers (with historic spellings)?



stering and
Matching



Understanding
Text

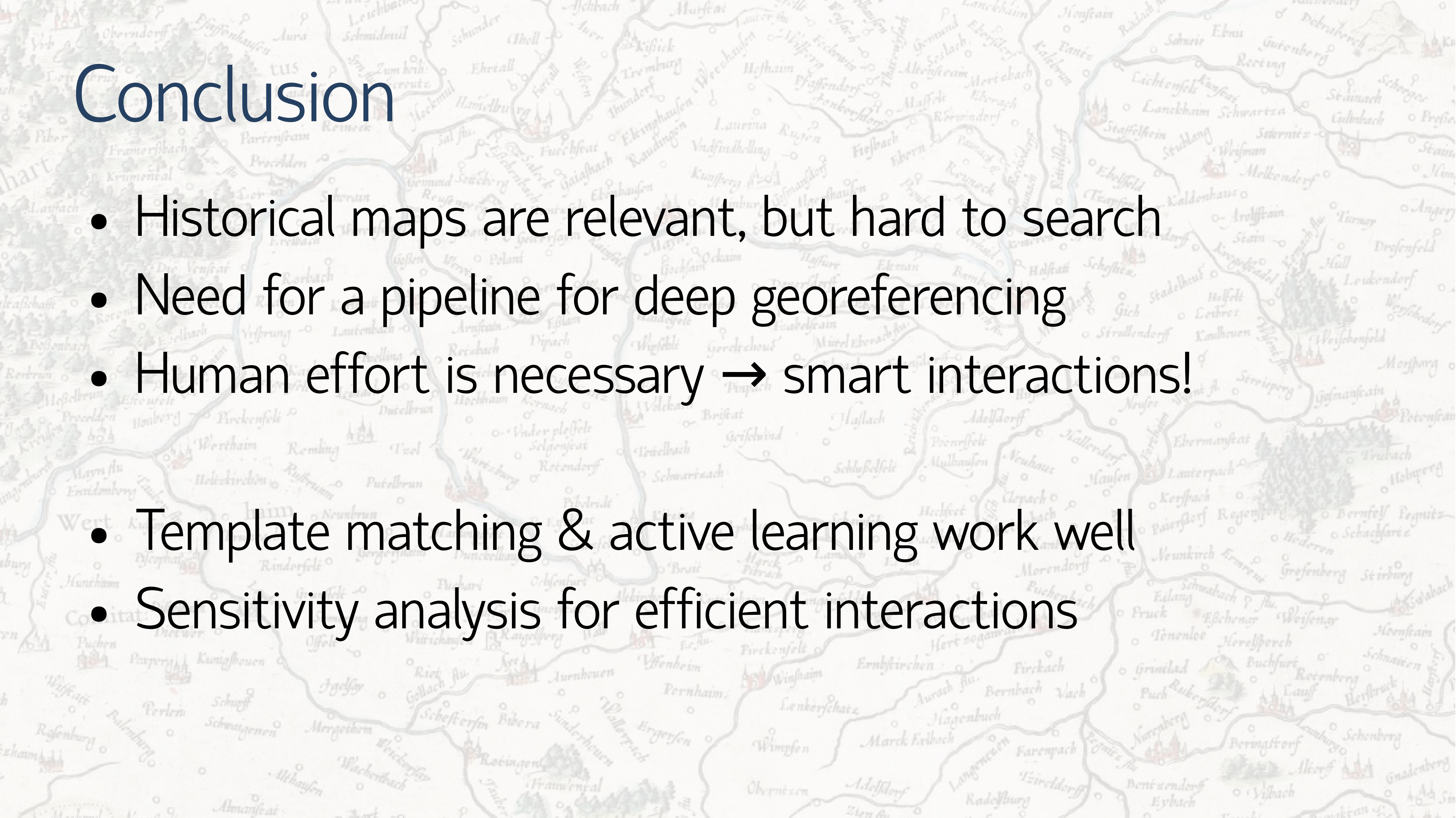
Georeferen

Georeferencing: open questions

Challenges:

- Spelling variations
- Potential errors in the previous steps

- Use gazetteers? Phonetic algorithms? *[Höhn et al. 2013]*
- Use modern maps?
- Geometric reasoning?



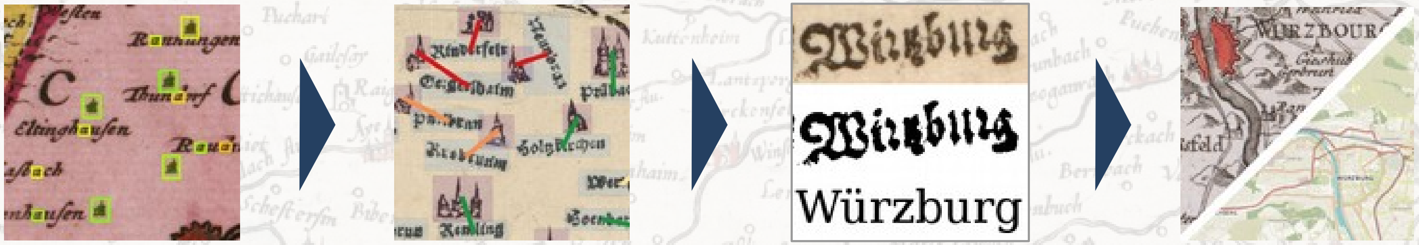
Conclusion

- Historical maps are relevant, but hard to search
- Need for a pipeline for deep georeferencing
- Human effort is necessary → smart interactions!

- Template matching & active learning work well
- Sensitivity analysis for efficient interactions

Open Questions & Future Work

- Solve more small goals from the pipeline, then integrate
  - Cluster template matches (e.g. into labels)
  - Use already collected information for OCR
  - Georeferencing, ...
- Should the pipeline really be sequential?
- Crowdsourcing?



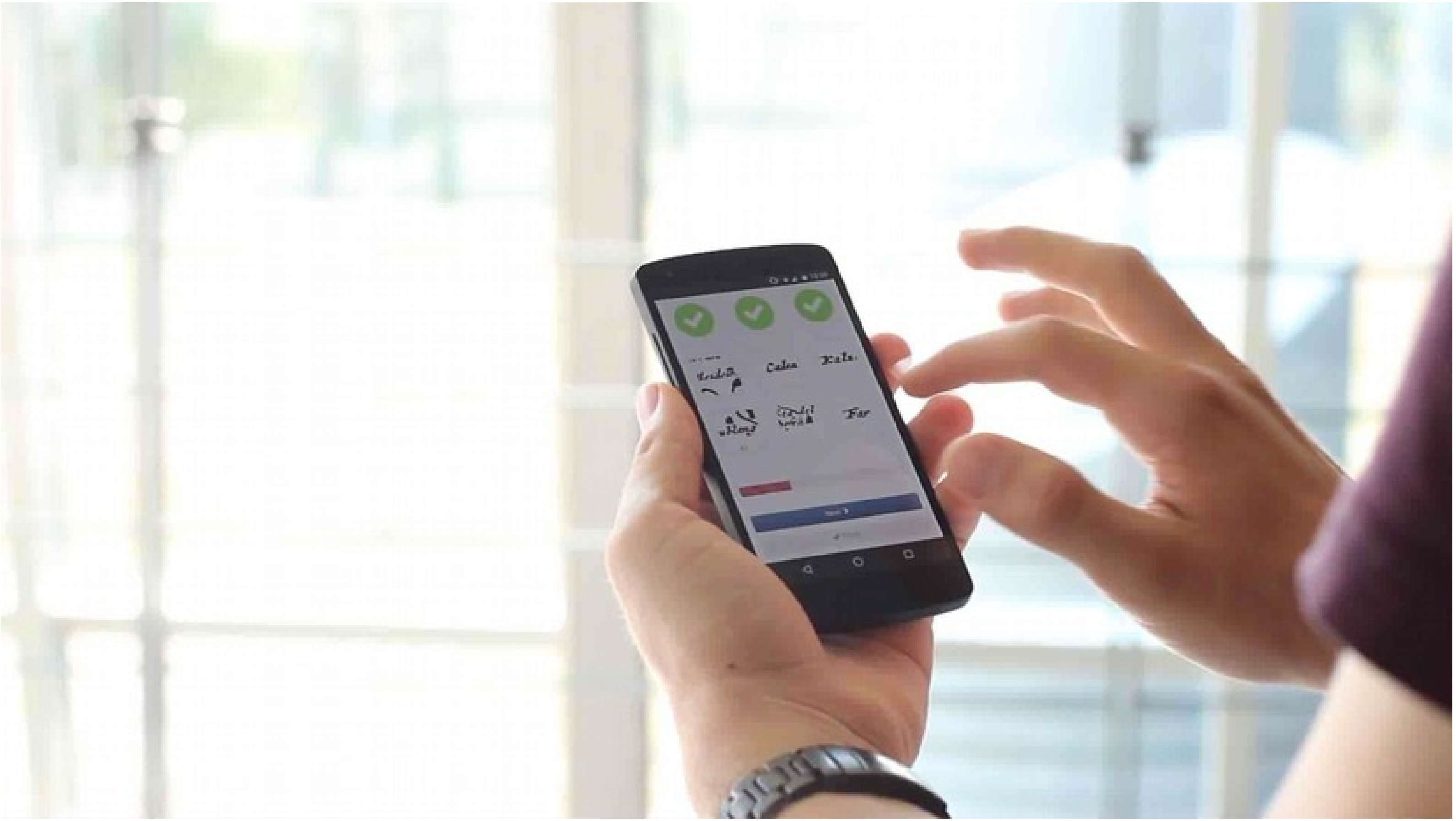
Segmentation

Clustering and Matching

Optical Character Recognition

Georeferencing

Smartphone



Open Questions & Future Work

- Develop remaining modules in extraction pipeline
  - Cluster template matches (e.g. into labels)
  - Use already collected information for OCR
  - Georeferencing, ...
- Should the pipeline really be sequential?
- Crowdsourcing! Yes, but how exactly?
- What other algorithmically-guided user interactions?