ETH zürich

Political territories from the 16th century: Automated construction and enrichment of geo-spatial data

Ramona Roller

PhD candidate at Chair of Systems Design, ETH Zurich (Prof. Frank Schweitzer)

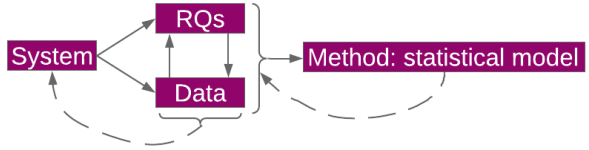
Territories of the Holy Roman Empire (HRE) around 1400

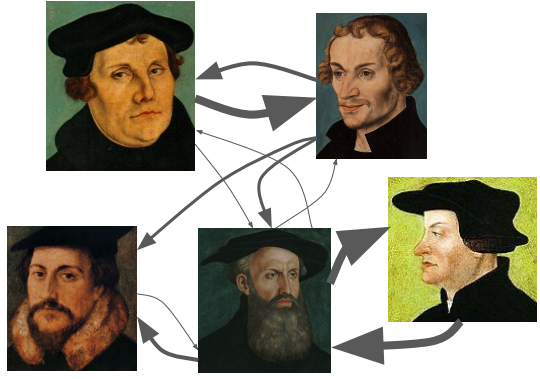
Wikipedia

- Digitised
- Structured
- Enriched
- Applied

Background and research focus

- **Background**
  - BSc Psychology with Cognitive Science
  - MSc Computational Science
  - PhD candidate in data-driven modeling of socio-historic systems
- **My research**
  - **System**: Letter correspondence of reformers
  - **Data**: 20,000 letters, 3,000 people, sending- and (receiving) dates + locations
  - **Method**: Social network analysis
    - Who writes how many letters to whom
    - From where to where and when
    - No letter content
    - ➔ **Insights about Reformation with quantitative methods**





Motivation

Research question

How does the religious ideology of reformers affect their letter correspondence network?

- No information on religious ideology of reformers in letter correspondence data
- Holy Roman Empire (HRE) was comprised of lots of territories (patchwork rug)
- Territory ruler determined religion
- Reformers moved between territories
- Test assumption that reformers moved to places that support their religious ideology
- **Match reformers to territories ➔ assign religious ideology to reformer**

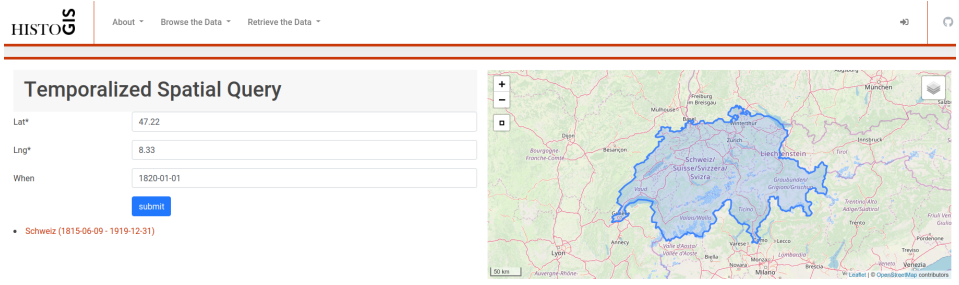
Motivation

Research question

How does the religious ideology of reformers affect their letter correspondence network?

- Match reformers to territories ➔ assign religious ideology to reformer

HistoGis: Where was point X,Y at time Z



Assign religious ideology to reformers within a certain period

- **Steps**
  - **Temporalised spatial query**: Where was reformer (point X,Y) at time Z ➔ territory A
  - **Content query**: Which religious ideology did territory A support at time Z ➔ Lutheran
  - **Inference**: Reformer was Lutheran at time Z
- **Problems**
  - No geo-political vectorised maps from 16th century available
  - No structured data on territory-religion match available

Assign religious ideology to reformers within a certain period

- **Approach: Construct geo-spatial dataset of territories**

- **Desired Output**
  - GeoDataFrame of reformers who are matched to activity places, territories, and religious ideologies (potentially over time)

reformer	placeGeo	territoryGeo	territoryName	activityStart	activityEnd	religion
Luther	Point(...)	Polygon(...)	Kfm. Sachsen	1530	1540	lutheran
Calvin	Point(...)	Polygon(...)	Genf	1530	1540	calvinistic

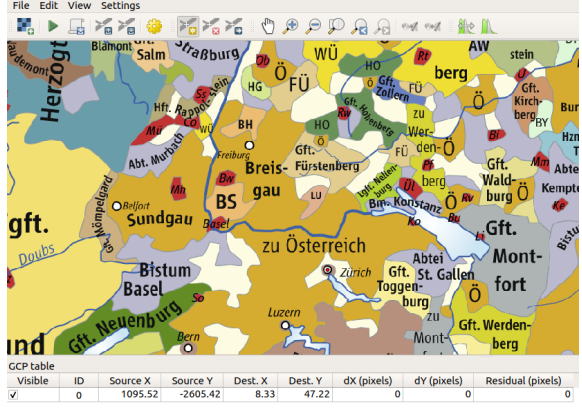
Workflow: Vectorise raster map

Vectorise raster map: Input

- Raster file of geo-political map
- Territories separated by colour

Vectorise raster map: Scan, pre-process, geo-reference

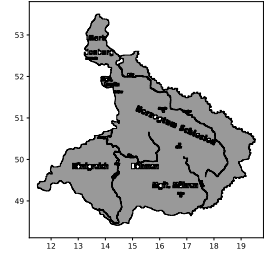
1. Scan geo-political map
2. Pre-process scan
   - Crop legend
3. Geo-reference scan
   - Look-up geo-coordinates of locations

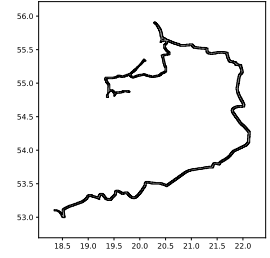


Vectorise raster map: Polygonise

1. Scan geo-political map
2. Pre-process scan
3. Geo-reference scan
4. Polygonise map
   - Extract polygons based on pixel colour

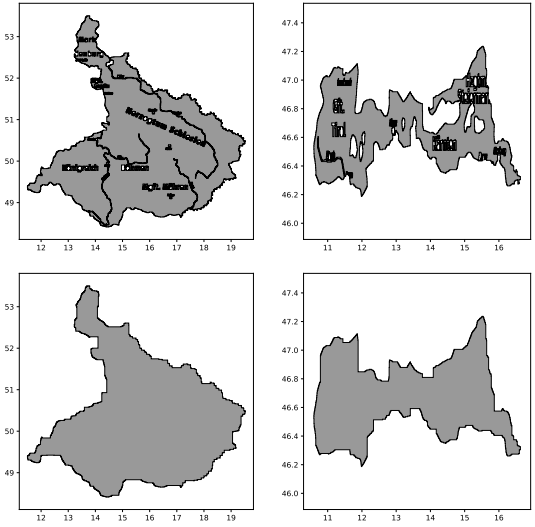
	red	geometry
390155	240.0	POLYGON ((13.70047412479831 53.50353094605324,...
292477	247.0	POLYGON ((17.12797152035539 53.55143751701807,...
552917	206.0	POLYGON ((8.992564741574409 46.46126501422215,...
102464	247.0	POLYGON ((20.5424034874669 55.61577521132106, ...
507497	213.0	POLYGON ((15.5122135359187 47.23648043528951, ...





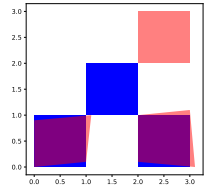
Vectorise raster map: Fill

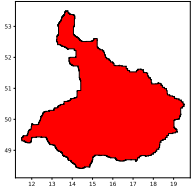
1. Scan geo-political map
2. Pre-process scan
3. Geo-reference scan
4. Polygonise map
5. Fill holes in polygons
   - Buffer function
   - Measure accuracy by randomly sampling points in bounded rectangle and categorising them as
     true positives, *false positives*,
     true negatives, *false negatives*

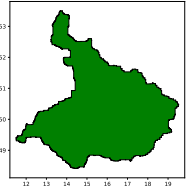


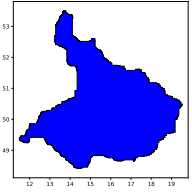
Vectorise raster map: Prepare filtering

1. Scan geo-political map
2. Pre-process scan
3. Geo-reference scan
4. Polygonise map
5. Fill holes in polygons
6. Prepare filtering of non-territory polygons
   1. Filter by area
      - Hardcoded rule: Discard all polygons with $area < threshold$
   2. Match polygons of different colours
      - **If** $area(poly_R) \approx area(poly_B)$ **and** $0.95 * area(poly\_R) \leq area(poly_R \cap poly_B)$: **Match**









	red	green	blue	geometry
0	4.0	3.0	4.0	POLYGON ((20.12430977722094 54.98863464595235,...
1	4.0	3.0	4.0	POLYGON ((18.24288808104642 54.400690365944, 1...
2	16.0	17.0	17.0	POLYGON ((12.25021156735244 52.69782952526276,...
3	4.0	3.0	4.0	POLYGON ((20.79500177076869 52.43216581356172,...
4	61.0	97.0	169.0	POLYGON ((11.56645414535705 54.40940065156943,...

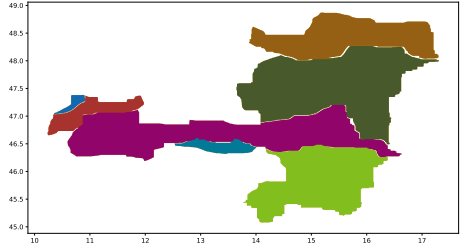
Vectorise raster map: Filter

1. Scan geo-political map
2. Pre-process scan
3. Geo-reference scan
4. Polygonise map
5. Fill holes in polygons
6. Prepare filtering of non-territory polygons
7. Filter non-territory polygons by colour

- Rivers
- Labels
- Borders
- Sea

Vectorise raster map: Merge

1. Scan geo-political map
2. Pre-process scan
3. Geo-reference scan
4. Polygonise map
5. Fill holes in polygons
6. Prepare filtering of non-territory polygons
7. Filter non-territory polygons by colour
8. Merge polygons belonging to the same territory
   - **If** same colour **and** shared border **or** intersection: **Merge**
   - Does not work for enclaves

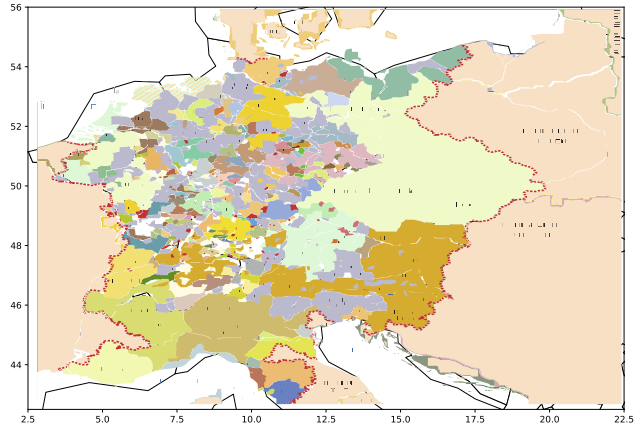


Vectorise raster map: Output

	red	green	blue	originalId	geometry
0	247.0	224.0	196.0	7912	POLYGON ((14.55843725939442 45.33328302513392,...
1	247.0	224.0	196.0	7898, 7907, 7908, 7909	POLYGON ((19.98494521 52.50620324, 19.98494521...
2	240.0	250.0	201.0	7911, 7432, 7883, 7788, 7888, 7860	POLYGON ((19.0181035 49.94102412339576, 19.018...
3	247.0	224.0	196.0	7779, 7910, 7533, 7444, 7897	POLYGON ((4.593870498229246 46.77048015408438,...
4	213.0	172.0	48.0	7904, 7905, 7653, 7852, 7797, 7896, 7900	POLYGON ((13.765801267 45.4465167383, 13.79193...

- 6,562 territory-polygons
- For comparison: 611,854 polygons after polygonisation step

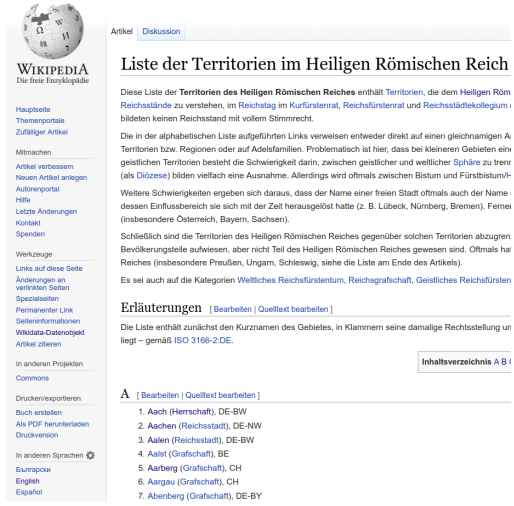
Vectorise raster map: Output



Workflow: Crawl Wikipedia

Crawl Wikipedia: Get territory names

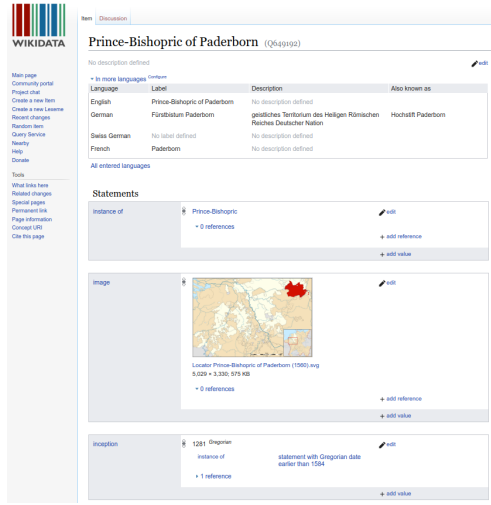
1. Crawl Wikipedia's list of HRE territories
   - Parse Wikipedia's page on 'List of territories in the Holy Roman Empire'
2. Extract HRE territory name from html



	old_name	modern_name	type_of_rule	modern_country
0	Aach (Hegau)	Aach	[Herrschaft (Territorium), Herrschaft]	[DE-BW]
1	Aachen	Aachen	[Freie Reichsstadt, Reichsstadt]	[DE-NW]
2	Aalen	Aalen	[Freie Reichsstadt, Reichsstadt]	[DE-BW]

Crawl Wikipedia: Structured

1. Crawl Wikipedia's list of HRE territories
2. Extract HRE territory name from html
3. Structured crawl
   - Query `wikidata` to extract HRE territory attributes



	old_name	type_of_rule	description	instance_of	inception	dissolved	religion	geoLoc
215	Landgrafschaft Hessen-Kassel	['Graf#Landgraf', 'Landgrafschaft']	state of the Holy Roman Empire in 1567–1803	state in the Holy Roman Empire	+1567-01-01T00:00:00Z	+1803-01-01T00:00:00Z	Lutheranism	(9.497778, 51.319722)
158	Stift Gandersheim	['Reichsstift']	monastery	monastery	+0801-01-17T00:00:00Z	+1802-01-01T00:00:00Z	Lutheranism	(10.026097222222, 51.870397222222)
435	Kurpfalz	['Graf#Pfalzgraf', 'Pfalzgrafschaft']	historical territory of the Holy Roman Empire	countship	+1085-01-01T00:00:00Z	+1803-01-01T00:00:00Z	Calvinism	(8.01667, 49.5)

Crawl Wikipedia: Semi-structured

1. Crawl Wikipedia's list of HRE territories
2. Extract HRE territory name from html
3. Structured crawl
4. Semi-structured crawl
   - Crawl Wikipedia's infoboxes of HRE territories

Territorium im Heiligen Römischen Reich

Hochstift Paderborn

Alternativnamen	Stift Paderborn, Hochstift Paderborn, Fürstbistum Paderborn, Paderborner Land
Entstanden aus	im 14. Jahrhundert herausgebildet aus Herzogtum Sachsen
Herrschaftsform	Wahlfürstentum/Ständestaat
Herrscher/Regierung	Fürstbischof, Administrator oder in Vakanz: Domkapitel

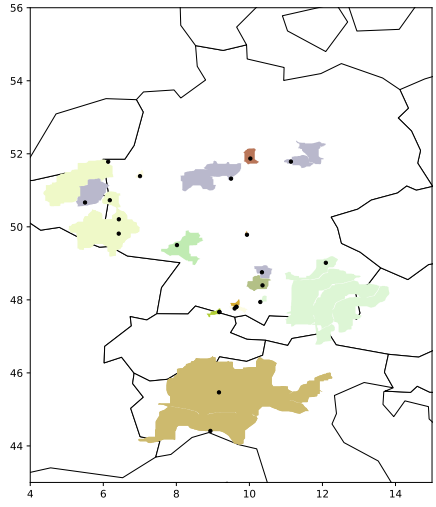
	old_name	type_of_rule	infobox_keys	inception	dissolved	religion	geoLoc
441	Pfalz-Zweibrücken	['Pfalzgrafschaft']	['name', 'wappen', 'karte', 'bildtext', 'vorlä...	1444: [[Pfalz-Simmern-Zweibrücken]], 1444:...	1793/1801: [[Erste Französische Republik|Frank...	bis 1533: [[Römisch-katholische Kirche|katholi...	(nan, nan)
16	Anhalt-Köthen	['Fürstentum']	['name', 'wappen', 'karte', 'periode', 'regier...	NaN	NaN	[[lutherisch]], ab 1525	(nan, nan)
44	Fürstentum Bayreuth	['Markgrafschaft#Markgraftum', 'Markgraftum']	['name', 'wappen', 'karte', 'regierungsform', ...	NaN	[[Königreich Bayern]] (1810)	[[katholisch]] bis 1525, seitdem [[lutherisch]]	(nan, nan)

Crawl Wikipedia: Evaluation of crawlers

1. Crawl Wikipedia's list of HRE territories
2. Extract HRE territory name from html
3. Structured crawl
   - ⊕ No hardcoded string-formatting needed
   - ⊖ No time-stamped data (change of religious ideology)
   - ⊖ Large loss: from 687 to 30 territoies
4. Semi-structured crawl
   - ⊕ Time-stamped data available (change of religious ideology)
   - ⊖ Hardcoded string-formatting needed
   - ⊖ Large loss: from 687 to 58 territoies

Workflow: Assemble territory-specific data set

Assemble territory-specific data set



1. Match polygons to territories
   - Spatial join

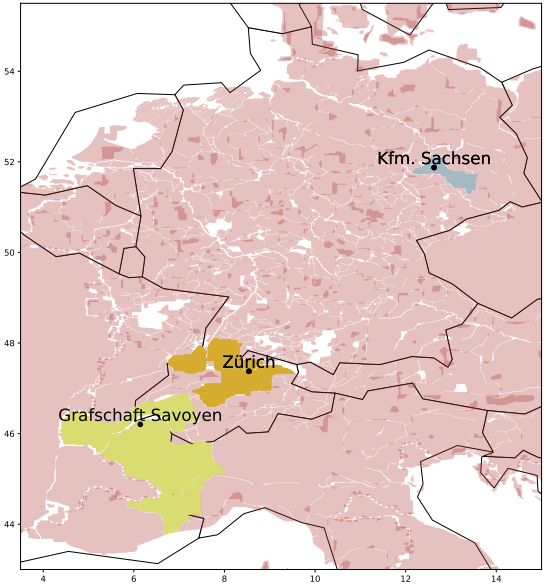
	old_name	red	green	blue	geometry	inception	dissolved	religion
6	Kurpfalz	193.0	236.0	179.0	POLYGON ((8.287031605513741 49.06999556039447,...	1085-01-01	1803-01-01	Calvinism
8	Landgrafschaft Hessen-Kassel	186.0	185.0	205.0	POLYGON ((9.384527594944018 51.39564182360063,...	1567-01-01	1803-01-01	Lutheranism

Workflow: Assign ideology to reformer

Assign religious ideology to reformer

1. Approximate place of activity
   - Location where reformer sent+received most letters
2. Match activity places to territory
   - Spatial join
3. Match religion of territory to reformer

	name	geometry	territory	religion
81	martin luther	POINT (12.6279659 51.8739831)	Kfm. Sachsen	lutheran
85	philipp melanchthon	POINT (12.6279659 51.8739831)	Kfm. Sachsen	lutheran
31	heinrich bullinger	POINT (8.541694 47.3768866)	Zürich	reformed
160	ulrich zwingli	POINT (8.541694 47.3768866)	Zürich	reformed
33	johannes calvin	POINT (6.1431577 46.2043907)	Grafschaft Savoyen	calvinism



Workflow: Evaluation

❶ Automated vectorisation of raster map

- Generalisable approach with few manual inspections (buffer size, cut-off area, filtered colours)
- Spot test are promising

➔ Systematic validation with manually vectorised map

➔ Refine exploration of parameter space

➔ Filter first then fill holes

❷ Crawl Wikipedia

- Enrich shapefiles with non-geo-spatial attributes
- Unstable solution which requires a lot of manual work

➔ Also consider other more structured data sources

Summary

1. **Generalisability**
   - Automated vectorisation by RGB-values
   - Enrichment by Wikipedia
2. **Feedback**
   - Validation and publication
3. **Communication of reformers**
   - Effect of religious ideology on letter correspondence network (initial RQ)
4. **Migration of reformers**
   - Factors and events driving the movement of reformers
     (e.g. support for religious ideology, wars, characteristics of ruler)
5. **Confessionalisation**
   - Factors and events driving adoption of protestant ideologies (e.g. situation in neighbouring territories)