Online Appendix
Felix Kersting and Nikolaus Wolf:
On the origins of national identity.
German nation-building after Napoleon

March 6, 2024

A Matching description

For our statistical analysis to be feasible, we first have to determine which children in our records belong to the same family. Our goal is to create a common family ID for all children that have the same mother and father. As is often the case with historical records, these are difficult to match as they frequently contain spelling mistakes. Therefore, we cannot use exact name matches but instead measure similarity between strings. For this purpose, we rely on the Jaro-Winkler (JW) distance (Winkler, 1990), which calculates the distance between two strings (in our case last names) in the following way:

$$sim_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases} \tag{1}$$

- $|s_i|$: Length of string i
- m: number of matching characters
- t: number of transpositions

A JW distance of 0 corresponds to an exact match and a JW distance of 1 corresponds to no similarity. The calculation itself is independent of language characteristics which makes it applicable to different languages.

Abramitzky et al. (2021) suggest some best practices for linking historical data. They are primarily concerned with how to match census data. They suggest a more conservative threshold of a JW distance less than 0.1 for two names to be considered a match. Additionally, they restrict their potential matches to people with the same place of birth and also apply a maximum age difference. Fouka (2020) uses only JW distances to match census records.

There are several properties of our dataset that make it less probable from the beginning to get false positive matches which means that we can be a bit less conservative than suggested by Abramitzky et al. (2021). First of all, we only consider potential matches in the same city. Second, our dataset comprises only a relatively short time window, from 1810 to 1821. Third, we use the mother's and father's last name for our matching approach. First names are more unreliable as their number and order is often different for the same person across different entries in our historical data. As Feigenbaum (2016) notes, last names of matched entries in the IPUMS database usually have less of a JW distance than first names. He finds that over 99% of all linked entries have a JW distance of less than 0.2. Last names also have a higher variation than first names which means that there is a relatively small probability of a married couple having exactly the same two last names as another couple. Additionally, we have the advantage that we can use two last names instead of just one in the case of census linking.

Considering the approaches cited above, we choose a JW distance threshold of 0.2. Our matching approach then goes as follows:

- Restrict the dataset to one city at a time. This means that we only consider potential matches in the same city.
- Split the data according to last name initials of mothers and fathers, similar to the blocking approach used by Abramitzky et al. (2019).
- In these subsets, we calculate the string distances between mothers' and fathers' last names separately, using last names before 1815 as the rows of our distance matrix and last names starting in 1815 as the columns. We find all instances where both the mother's last name and the father's last name from one couple match with the last names of another couple.
- We group these instances together as a family. For example, if couple a before 1815 matches couples x, y and z after 1815 and couple b before 1815 also matches x, y and z, we group all five together as the same family.

B Discussion Adjustment Aachen

In the case of Aachen, the political changes themselves led to changes in the way first names were recorded. It is possible that the French bureaucracy "francicised" many German names, although we have no direct evidence on this (Kramer, 1993, p.225). For parents' first names, this is not problematic as it is unlikely that the parents were given French names in late 18th century Aachen before the French occupation. For the children, however, this becomes more complicated because we do not know whether parents adjusted to the new rulers by choosing French first names or whether instead the French administration "francicised" German given names. To deal with this issue, we use two radically different versions of the data for Aachen:

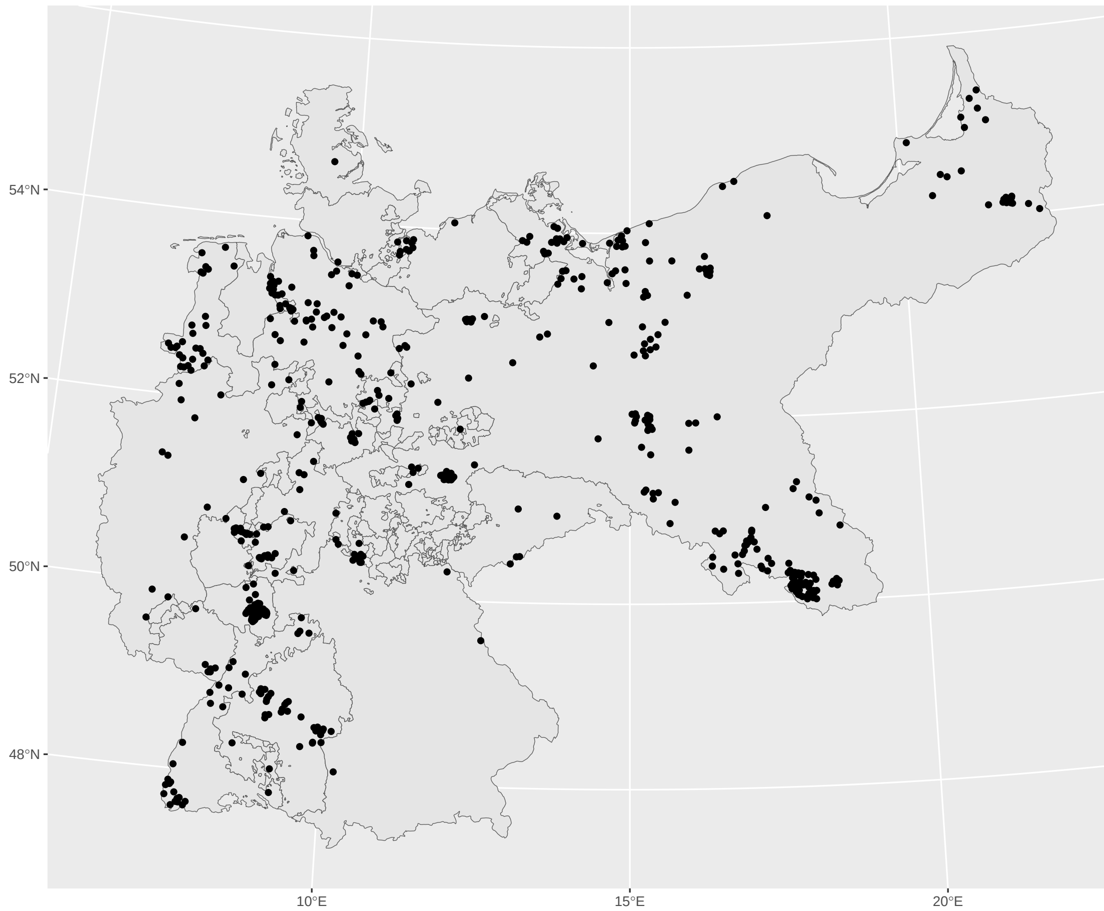
Unadjusted version We assume that all French versions of German names in the data reflect indeed the choice of parents. With this we might underestimate the share of Germanic first names for the pre-treatment period and thus potentially **overestimate** the treatment effect. Note that we do not observe a break in our data once the French administration left Aachen in January 1814.

Adjusted version As an alternative, we adopt a most "conservative" interpretation of the data for Aachen, assuming that the French administration systematically "francicised" all German first names, against the wish of the parents. Hence, we adjust the data and classify all French versions of Germanic names before 1815 as national names (e.g. Henri in 1814 would be changed to Heinrich). We lack direct evidence on this, but if anything this introduces a strong bias against us finding a treatment effect, i.e., **underestimating** the treatment effect.

Verdict As we do not observe a break in our data once the French administration left Aachen in January 1814, we prefer the un-adjusted version. We show the results for the adjusted version in Appendix C.

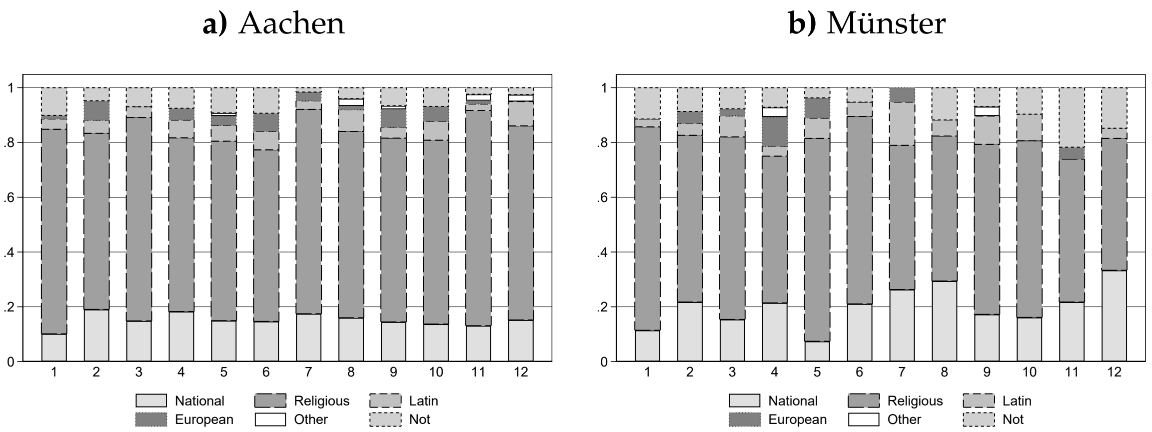
C Additional Figures and Tables

Figure C.1: Distribution of villages



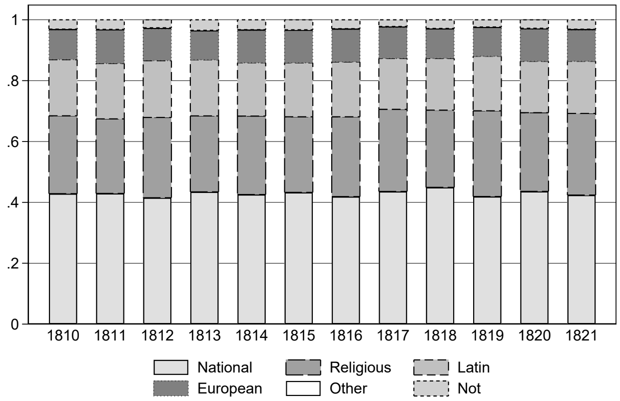
Notes: The map shows the location of the villages in our sample.

Figure C.2: Categories of first names, 1815



Notes: Figures C.2a and C.2b show the development of first name categories for Aachen and Münster, respectively, by month.

Figure C.3: Development by categories of first names in Berlin, 1810-1821



Notes: The figure shows the development of first name categories for Berlin.

Table C.1: Descriptive statistics

City	Before 1815 (year ≤ 1814)	After 1814 (year > 1814)	N per year
Aachen (adjusted)	13.69	15.63	854.56
Aachen (unadjusted)	6.24	15.63	854.56
Muenster	12.73	17.72	459.89
Frankfurt	13.30	13.81	1059.70
Hanover	19.90	23.34	595.72
Heidelberg	6.62	7.83	1442.68
Mannheim	5.95	6.62	461.58
Nuernberg	6.10	7.22	913.47
Berlin	30.44	30.81	2564.64

Notes: Share national first names (in %) by city and before/after treatment.

Table C.2: Robustness – National first names, war participation, and war decoration

	Loss Lists		Honored Soldiers		t-statistic
	N	National (in%)	N	National (in%)	
Panel 1:					
Divisionen with "non-national" leader	86.110	44.45	688	54.22	5.14***
Special first names					
... Friedrich & Wilhelm	86.110	16.26	688	14.83	-1.06
... Ernst & Hermann	86.110	4.15	688	7.00	2.91***
Panel 2:					
Korps from "old" Prussian provinces	55.080	42.70	483	57.14	6.41***
Special first names					
... Friedrich & Wilhelm	55.080	15.09	483	14.70	-0.24
... Ernst & Hermann	55.080	4.75	483	8.28	2.81***
Panel 3:					
Korps from the other provinces	69.739	41.74	417	51.32	3.91***
Special first names					
... Friedrich & Wilhelm	69.739	14.00	417	13.91	-0.05
... Ernst & Hermann	69.739	3.80	417	4.32	0.52

Notes: Panel 1: includes all divisions with leaders without national first name. Panel 2: includes the army corps from the old part of Prussia (I., II., III., V., VI. and the Garde-Korps). Panel 3: includes all corps from the provinces that became part of Prussia over the 19th century (VII., VIII., IX., X., XI., XIV., XV.). * $p<0.1$, ** $p<0.05$, *** $p<0.01$.
Sources: Verein für Computergenealogie (2014) and Königliche General-Ordens-Kommission (1878).

Table C.3: Event study regression, 1810-1821

Dep. var.: National First Name	City FE (1)
Treated × 1810	0.016* (0.081)
Treated × 1811	0.020* (0.095)
Treated × 1812	0.018 (0.159)
Treated × 1813	0.006 (0.607)
Treated × 1815	0.080** (0.015)
Treated × 1816	0.108** (0.013)
Treated × 1817	0.081** (0.045)
Treated × 1818	0.088** (0.015)
Treated × 1819	0.099** (0.031)
Treated × 1820	0.081* (0.061)
Treated × 1821	0.101*** (0.009)
Year FE	✓
City FE	✓
Observations	60168
Families	38274

Notes: * p<0.1, ** p<0.05, *** p<0.01. We only include parents without national first given name. Clustered and bootstrapped standard errors at the city level. P-values in parenthesis. Treated cities: Aachen and Münster. Control cities: Frankfurt (Main), Hanover, Heidelberg, Nuremberg, and Mannheim. Reference year: 1814.

Table C.4: Alternative control group – cities with exposure to French occupation

Dep. var.: National first name (dummy)	(1)	(2)
Treated × Post1815	0.085*	0.274*
	(0.052)	(0.050)
City FE	✓	
Family FE		✓
Observations	26137	2566
Families		1283

Notes: * p<0.1, ** p<0.05, *** p<0.01. We only include parents without national first given name. Clustered and bootstrapped standard errors at the city level. P-values in parenthesis. Treated cities: Aachen and Münster. Control cities: Heidelberg, and Mannheim. Data is collapsed in pre-and post period.

Table C.5: Names – conservative adjustment Aachen

Dep. var.: National first name (dummy)	(1)	(2)
Treated × Post1815	0.028	0.091*
	(0.104)	(0.064)
City FE	✓	
Family FE		✓
Observations	46007	4530
Families		2265

Notes: * p<0.1, ** p<0.05, *** p<0.01. We only include parents without national first given name. In this table, we use the adjustment for the names in Aachen (see Appendix B for more details). Clustered and bootstrapped standard errors at the city level. P-values in parenthesis. Treated cities: Aachen and Münster. Control cities: Frankfurt (Main), Hanover, Heidelberg, Nuremberg, and Mannheim. Data is collapsed in pre-and post period.

Table C.6: Control groups – excluding each control city once

Dep. var.: National first name (dummy)	(1)	(2)	(3)	(4)
	excluding ruler first names		including ruler first names	
Panel 1: Excluding Frankfurt				
Treated × Post1815	0.096**	0.290**	0.093**	0.230**
	(0.021)	(0.013)	(0.020)	(0.018)
City FE	✓		✓	
Family FE		✓		✓
Observations	37206	3570	37700	3822
Families		1785		1911
Panel 2: Excluding Hanover				
Treated × Post1815	0.087**	0.256**	0.092**	0.232**
	(0.027)	(0.018)	(0.027)	(0.018)
City FE	✓		✓	
Family FE		✓		✓
Observations	40729	3950	41071	4194
Families		1975		2097
Panel 3: Excluding Heidelberg				
Treated × Post1815	0.099**	0.264**	0.097**	0.229**
	(0.018)	(0.016)	(0.019)	(0.017)
City FE	✓		✓	
Family FE		✓		✓
Observations	34938	3632	35430	3842
Families		1816		1921
Panel 4: Excluding Mannheim				
Treated × Post1815	0.093**	0.262**	0.092**	0.221**
	(0.025)	(0.017)	(0.027)	(0.019)
City FE	✓		✓	
Family FE		✓		✓
Observations	42493	4266	43039	4538
Families		2133		2269
Panel 5: Excluding Nuremberg				
Treated × Post1815	0.093**	0.257**	0.092**	0.221**
	(0.025)	(0.016)	(0.026)	(0.018)
City FE	✓		✓	
Family FE		✓		✓
Observations	40216	4106	40741	4362
Families		2053		2181

Notes: * p<0.1, ** p<0.05, *** p<0.01. We include parents with national first given name. Clustered and bootstrapped standard errors at the city level. P-values in parenthesis. Treated cities: Aachen and Münster. Control cities: Frankfurt (Main), Hanover, Heidelberg, Nuremberg, and Mannheim. In each panel, we exclude one of the control cities. Data is collapsed in pre-and post period.

Table C.7: Estimator – Logit

Dep. var.: National first name (dummy)	(1)	(2)
Treated × Post1815	0.737**	0.688**
	(0.016)	(0.011)
City FE	✓	
Family FE		✓
Observations	46007	4286
Families		

Notes: * p<0.1, ** p<0.05, *** p<0.01. We only include parents without national first given name. Clustered and bootstrapped standard errors at the city level. P-values in parenthesis. Treated cities: Aachen and Münster. Control cities: Frankfurt (Main), Hanover, Heidelberg, Nuremberg, and Mannheim. Data is collapsed in pre-and post period.

Table C.8: Sample – adding parents with national identity

Dep. var.: National first name (dummy)	(1)	(2)
Treated × Post1815	0.095**	0.223**
	(0.024)	(0.011)
City FE	✓	
Family FE		✓
Observations	47281	4876
Families		2438

Notes: * p<0.1, ** p<0.05, *** p<0.01. We include parents with national first given name. Clustered and bootstrapped standard errors at the city level. P-values in parenthesis. Treated cities: Aachen and Münster. Control cities: Frankfurt (Main), Hanover, Heidelberg, Nuremberg, and Mannheim. Data is collapsed in pre-and post period.

Appendix References

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez, "Automated Linking of Historical Data," *Journal of Economic Literature*, 2021, *59* (3), 865–918.

_ **, Roy Mill, and Santiago Pérez**, "Linking individuals across historical sources: A fully automated approach," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 2019, *53* (2), 94–111.

Feigenbaum, James, "Automated census record linking: A machine learning approach," *Boston University Mimeo*, 2016.

Fouka, Vasiliki, "Backlash: The Unintended Effects of Language Prohibition in US Schools after World War I," *Review of Economic Studies*, 2020, *87* (1), 204–239.

Königliche General-Ordens-Kommission, *Königlich preussische Ordensliste 1877*, Berlin: Geheime Oberhofdruckerei, 1878.

Kramer, Johannes, "Französische Personen- und Ortsnamen im Rheinland 1794–1814," in Wolfgang Dahmen, ed., *Das Französische in den deutschsprachigen Ländern*, Tübingen: Narr, 1993, pp. 222–236.

Verein für Computergenealogie, *Verlustlisten 1870/71*, genealogy.net, 2014.

Winkler, William E, "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods. American Statistical Association*, 1990, pp. 354–359.