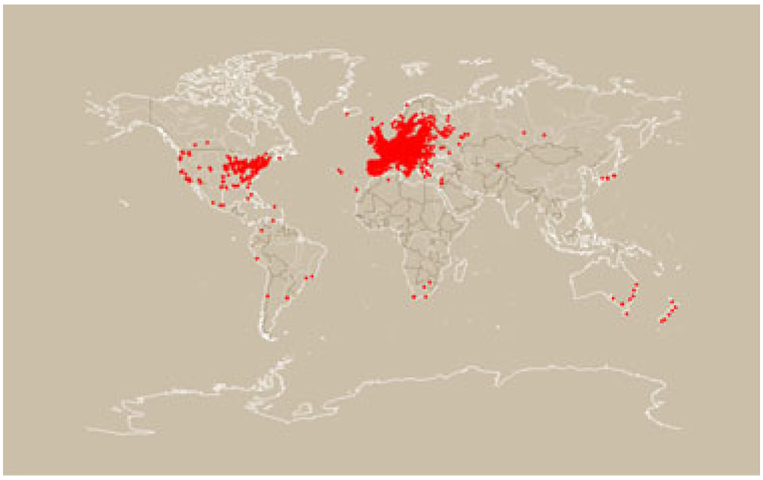
DALEK
Georeferences for the Bernstein Paper Atlas

Vlad Atanasiu[1], Claire Priol[2], Anne Tournieroux[2], Ezio Ornato[2]



(1) Commission for Scientific Visualization, Austrian Academy of Sciences, Vienna, Austria, atanasiu@alum.mit.edu, http://www.viskom.oeaw.ac.at

(2) Laboratory for Medieval Studies of Paris, National Center for Scientific Research / University Paris 1, Paris, France, erialch@gmail.com, anne.tournieroux@wanadoo.fr, ezornato@vjf.cnrs.fr, http://lamop.univ-paris1.fr/W3/

Purpose

Dalek provides georeferences for the Bernstein Paper Atlas (http://www.bernstein.oeaw.ac.at). That is placenames, geographical coordinates and region names of the databases part of the Bernstein workspace. There are approximately 14.000 records in the dataset, representing approximately 7.500 unique and identified places.

The difference between Dalek and a generalist gazetteer is the attribution of georeferences according to what the placenames represent in the various sources. For example, 'Paris' likely means 'Paris, Texas' in a database on incunabula repositories and 'Paris, France' in the case of a database on printing places of the Renaissance. More subtle cases of disambiguation involve deeper historical knowledge and decision making in respect to georeferencing.

The file has two development support tools: Wiz, for visualizing locations on a map, and Dibuk, for debugging possible errors in georeferencing. Dalek is distributed in three formats: Microsoft Excel 2003 file (convienient spreadsheet format), Unicode UTF-16 comma-delimited text file (for compatibility and long term preservation), KML files (visualization in GoogleMaps and GoogleEarth).

Description

The following describes the columns of the Dalek file.

A. *Serial* – serial number of placenames.

B. *Status* – information about the nature of the placenames.

Georeferenced placenames – placenames for which coordinates are provided.

location – locations considered punctiform (e.g.: city, abbey, castle).

region – locations considered surfaces: administrative units, geographical and historical regions (e.g.: county of York, Land of Bayern, Transylvania). Regions were normalized to the principal locations in the region (e.g.: county of York > city of York, Land of Bayern > city of München, Bohemia > Prague).

Placenames not georeferenced – placenames for which coordinates are not provided

ambiguous – placenames referring to more than one location that could not be disambiguated (e.g.: Gazzo appears once in the region Emilia Romagna, once in Lombardia and twice in Veneto).

not identified – placenames that could not be identified (e.g.: Two Waters, Dorf "Zahaeim").

C-D. *Placenames*

C. *Original placename* – placenames as given in the sources. Accents are preserved.

D. *Normalized placename* – placenames providing a unique name homonymous placenames in the sources (e.g.: Bécs, Vienne, Vienna, Wien > Wien). The local or most common name is the preferred one.

E-F. *Coordinates* – Decimal coordinates, rounded to two decimals. When more than one set of coordinates were found for a given location, only one was retained. Empty fields mean that the location is of a type that didn't allow georeferencing.

E. *Latitude*

F. *Longitude*

G-K. *Regions* – Names and codes of the placenames' regions (e.g.: Oxford > Oxfordshire).

G. *Code* – Code of the region. For EU-27 and EFTA countries this is the code level 3 of the Nomenclature of Territorial Units for Statistics 2003 (http://ec.europa.eu/eurostat/ramon/nuts/ and http://www.bfs.admin.ch/bfs/portal/fr/index/international/11/geo/analyse_regionen/11.html. For other countries the country code according to ISO 3166-1 alpha 2 was used (http://www.iso.org/iso/country_codes.htm).

H. *Country* – Name of the country as given by the two standards named above.

I. *NUTS 1 (2003)* – name of NUTS region level 1, state codes for Canada and United States.

J. *NUTS 2 (2003)* – name of NUTS region level 2.

K. *NUTS 3 (2003)* – name of NUTS region level 3.

L. *NUTS 4 (2003)* – name of NUTS region level 4.

M-AA. *Sources* – shows in which sources the placenames are present.

M. *BP* – Briquet Printed (repertory of watermarks and papers). Charles Moïse Briquet, *Les Filigranes. Dictionnaire historique des marques du papier dés leurs apparition vers jusqu'en 1600*, Paris, 1907, 4 vol. ; 2nd edition Leipzig 1923. / C. M. Briquet, *Les Filigranes. The New Briquet, Jubilee Edition*, Allan Stevenson (ed.), Amsterdam, 1968, 4 vol. / http://www.ksbm.oeaw.ac.at/_scripts/php/briquet.php

N. *BR* – Briquet Repositories. Libraries, archives and other collections were the papers described in BP were recorded by the author.

O. *GWP* – Gesamtkatalog der Wiegendrücke print places (quasi complete, extended catalog of existent incunabula). Locations were incunabula were printed. http://www.gesamtkatalogderwiegendrucke.de

P. *GWR* – GW Repositories. Collections were the incunabula in GW are preserved.

Q. *IBPH* – International Bibliography of Paper History. http://www.bernstein.oeaw.ac.at > Bibliography

R. *ISTC* – Incunabula Short-Title Catalog (quasi complete, compact catalog of existent incunabula). http://www.bl.uk/catalogues/istc/

S. *LKH* – Likhachev (repertory of watermarks and papers). J.S.G.Simmons, Bé van Ginneken-van de Kasteele (ed.), *Likhachev's watermarks*, Amsterdam, 1994.

T. *LR* – Collections were the papers described in LKH were recorded by the author.

U. *PF* – Piccard Findbuch (repertory of watermarks and papers). Gerhard Piccard, *Die Wasserzeichenkartei im Hauptstaatsarchiv Stuttgart*, Stuttgart, 1961-97, 17 vol.

V. *PO* – Piccard-Online (repertory of watermarks and papers). http://www.landesarchiv-bw.de/piccard/

W. *PR* – Collections were the papers described in PO were recorded by the author.

X. *WILC* – Watermarks in Incunabula printed in the Low Countries (repertory of watermarks and papers), National Library of the Netherlands, The Hague. http://watermark.kb.nl

Y. *WILCR* – Collections were the papers described in WILC were recorded by the author.

Z. *WZMA* – Wasserzeichen des Mittelalters [Watermarks of the Middle Ages] (repertory of watermarks and papers), Commission for Writing and Book in the Middle Ages, Austrian Academy of Sciences, Vienna. http://www.ksbm.oeaw.ac.at/wz/wzma.htm

AA. *WZMAR* – Collections were the papers described in WZMA were recorded by the author.

AB. *BIR* – Biraben (list of places where plague epidemics occurred between 1347-1600). Jean-Noël Biraben, *Les Hommes et la peste en France et dans les pays européens et méditerranéens*, Paris/La Haye, Mouton/École des hautes études en sciences sociales, 1975-6. A year by year list of places of occurrence of plague is given in the Yersinia dataset (look under Atlas in http://www.bernstein.oeaw.ac.at).

AC. *Remarks* – Various remarks, mostly about how ambiguities were solved or why some placenames are considered as ambiguous.

Quantities

The following tables provide a quantitative insight in Dalek as of 23.06.2008. Updates to the files might modify the quantities.

General overview

Notice that almost one fifth of the records given by the sources could not be georeferenced and that approximately one third represent homonyms. The fifth column in the next table represents percentages of unique values in respect to the total amount of records (first column).

	Records		Unique values		
Locations	11 378	*82 %*	7 462	*78 %*	*66 %*
Regions	741	*5*	424	*4*	*57*
Ambiguous	765	*6*	738	*8*	*96*
Not identified	991	7	975	*10*	*98*

Places of paper use

The column PF&PO highlights the difference in the content of these two sources, 40% of the records being unique to one or the other source. For the acronyms see the section describing the Dalek fields. This and the following tables give numbers of unique values in dataset.

	Total	BP	LKH	PF&PO	PF	PO	WILC	WZMA	
Locations	4 424	1 713	162	3 309	1 974	2 950	21	77	
	100	*38*	*3*	*74*	*44*	*66*	*0,4*	*1*	*%*
	▬	■	▮	▬	■	▬	▮	▮	
Regions	186	104	20	109	35	96	1	10	
Ambiguous	536	194	5	355	153	261	1	7	
Not identified	849	211	12	639	219	496	1	0	

Places of paper preservation

	Total	BR	LR	PR	WILCR	WZMAR	
Locations	364	199	33	98	143	16	
	100	*54*	*9*	*9*	*39*	*4*	*%*
	▬	■	▮	▮	■	▮	
Regions	4	1	3	1	0	0	
Ambiguous	1	0	0	0	1	0	
Not identified	0	0	0	0	0	0	

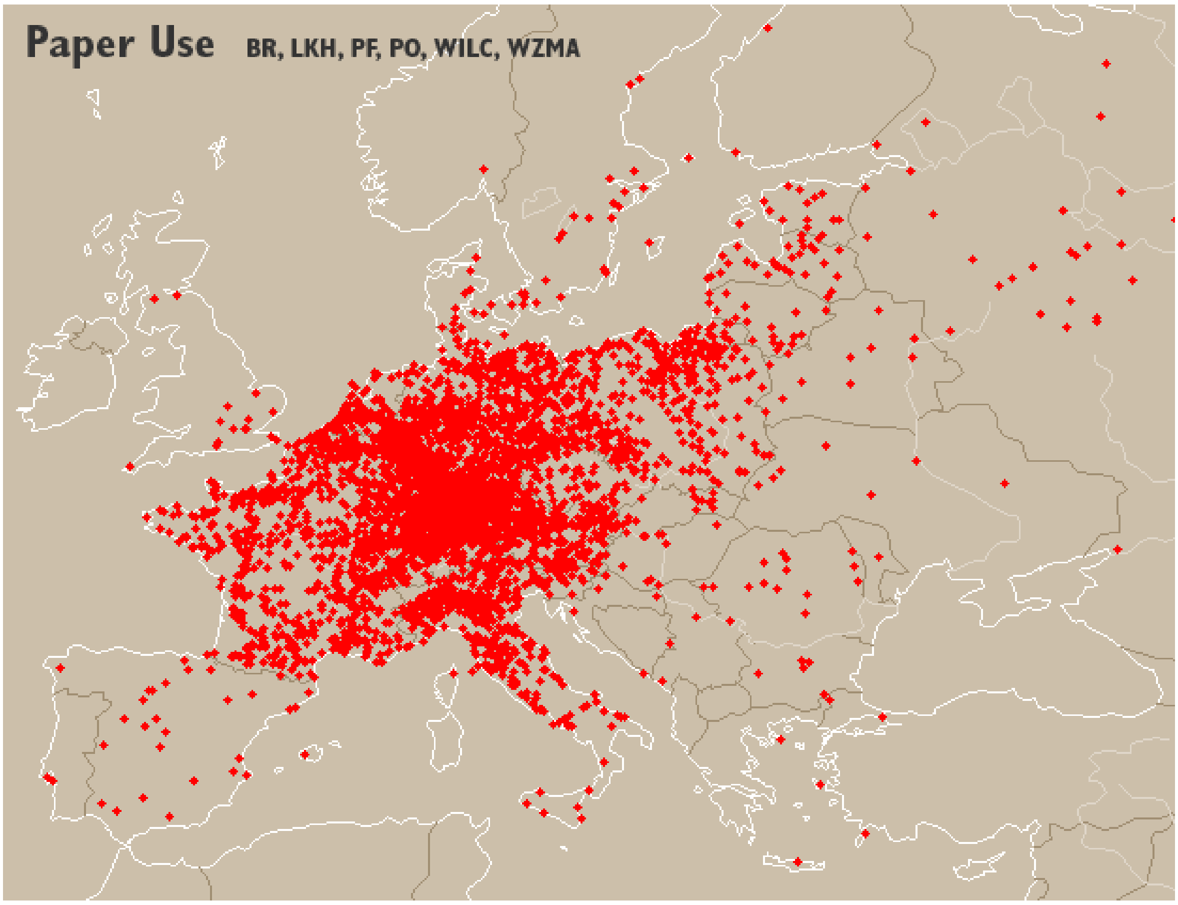
Incunabula printing and preservation places

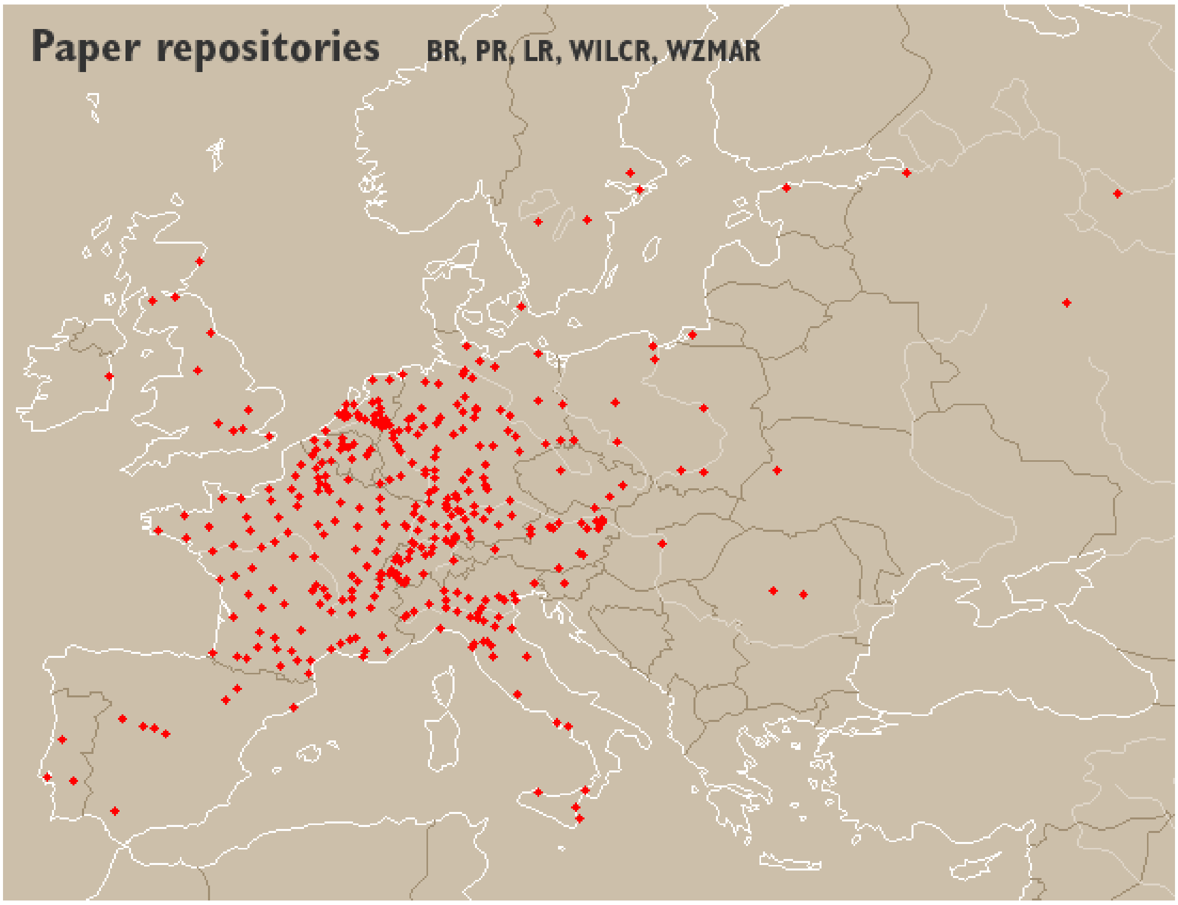
	Total	GWP	ISTC	GWR
Locations	286	275	241	1 946
Regions	4	1	4	1
Ambiguous	1	0	1	4
Not identified	1	0	1	2

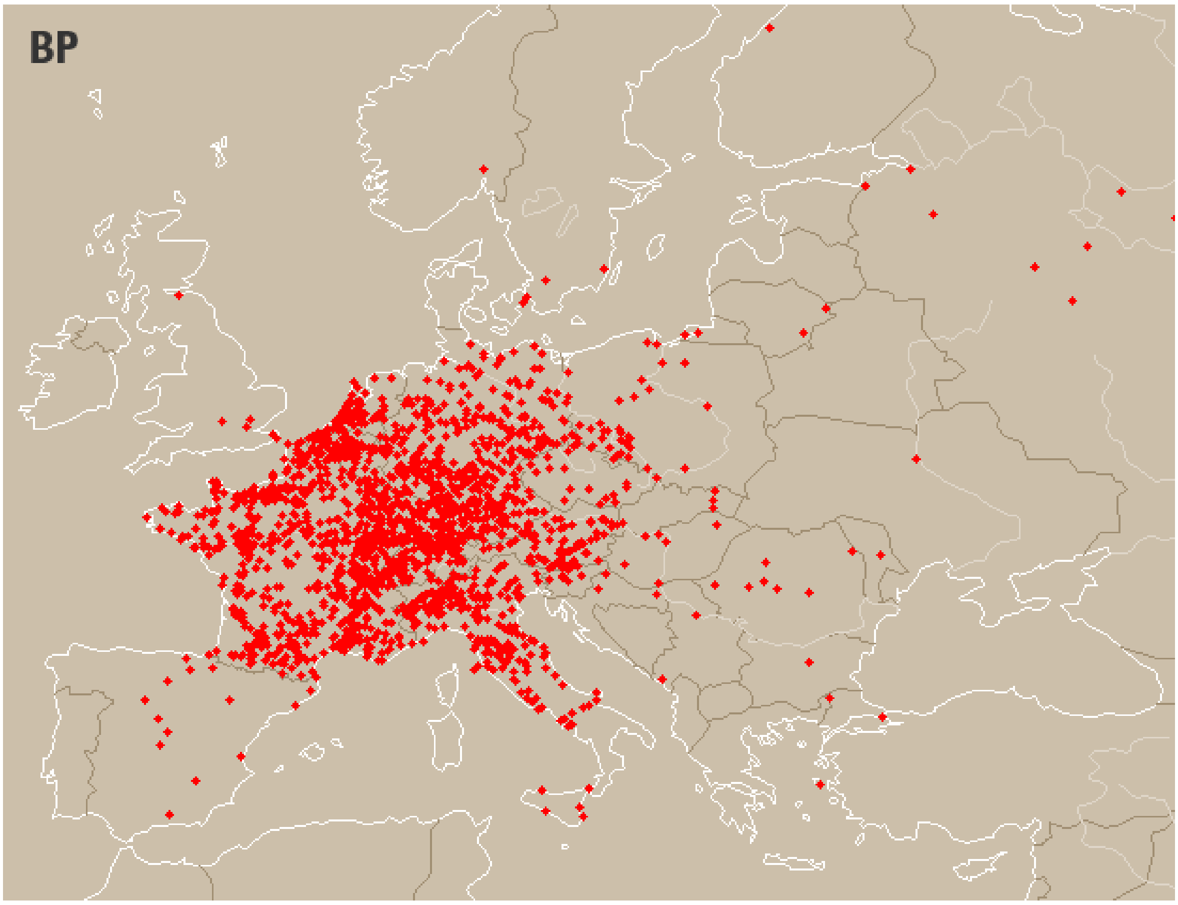
Contextual sources: bibliography and plague

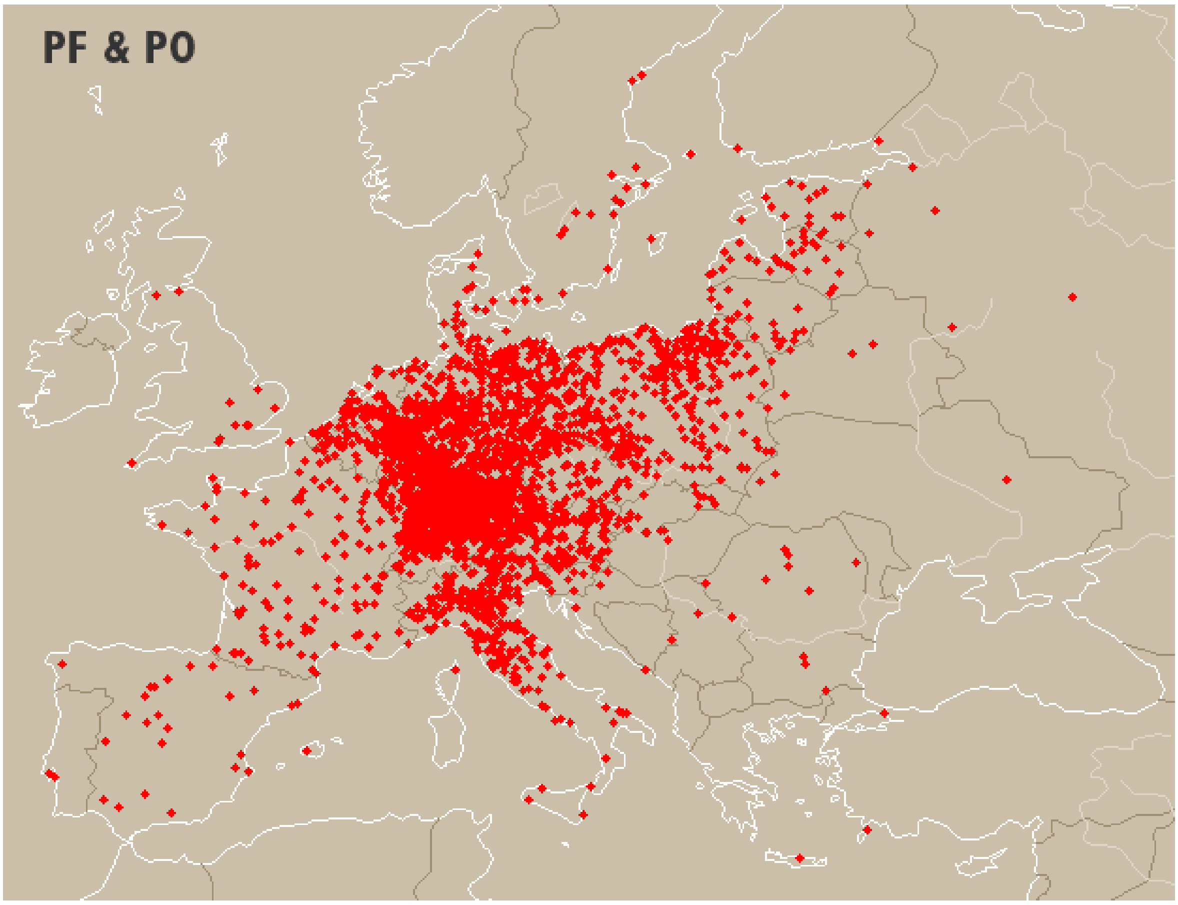
	IBPH	BIR
Locations	2 271	821
Regions	208	169
Ambiguous	184	36
Not identified	55	68

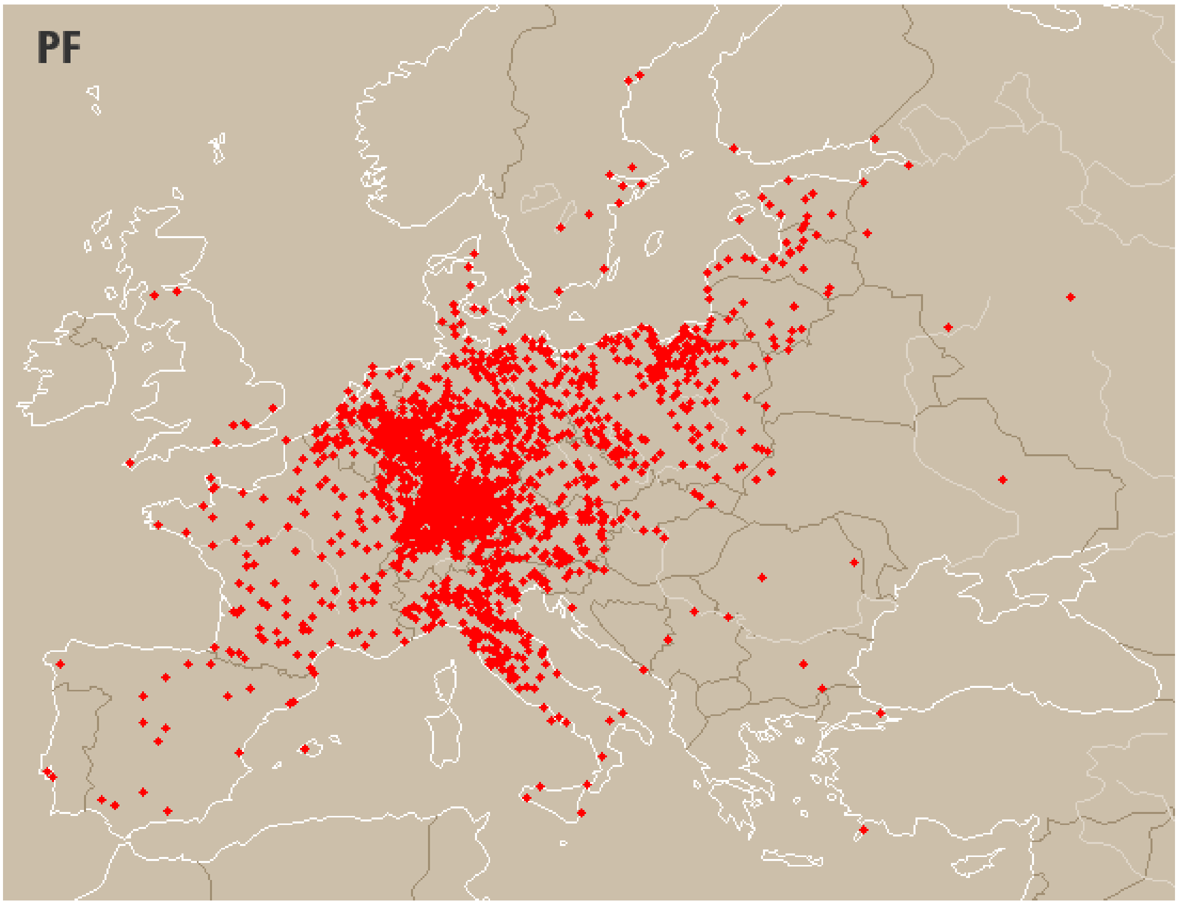
Visualizations

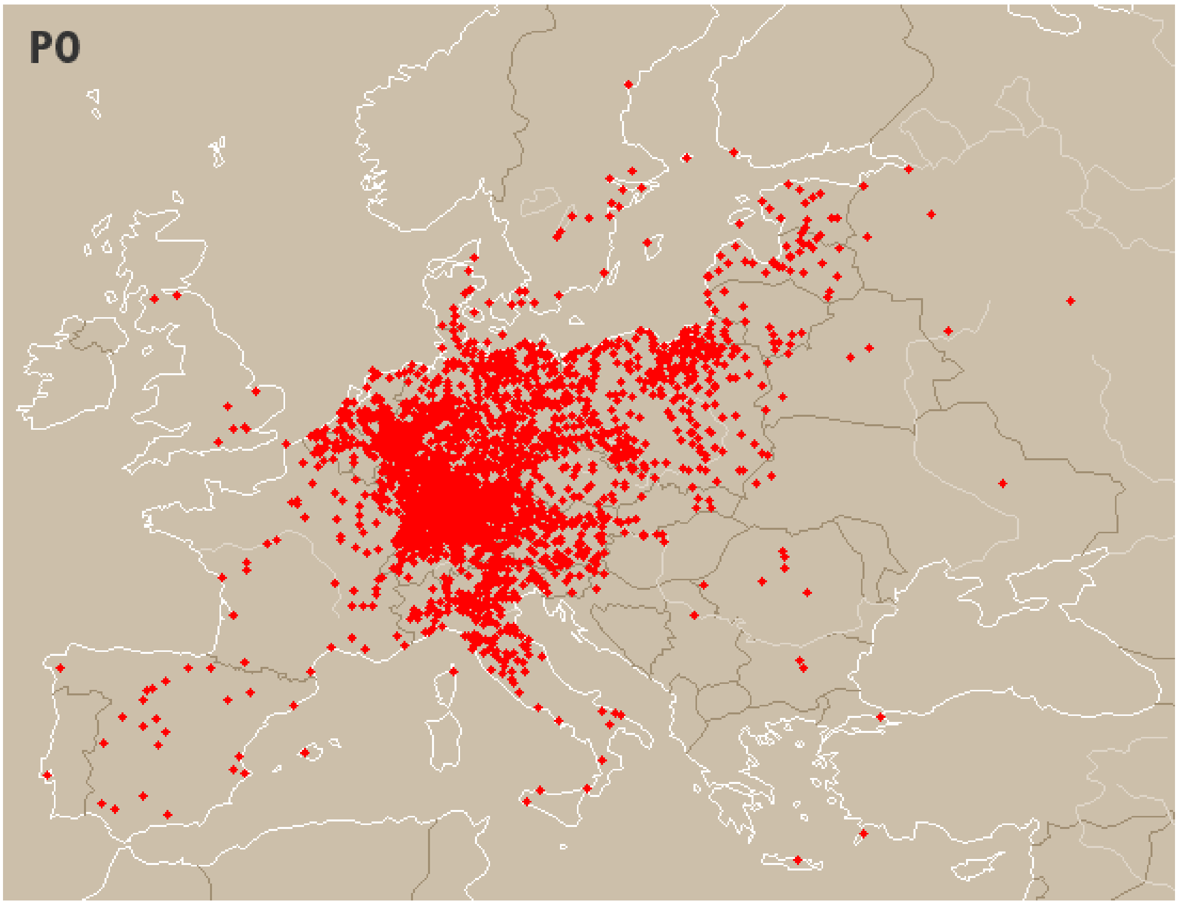


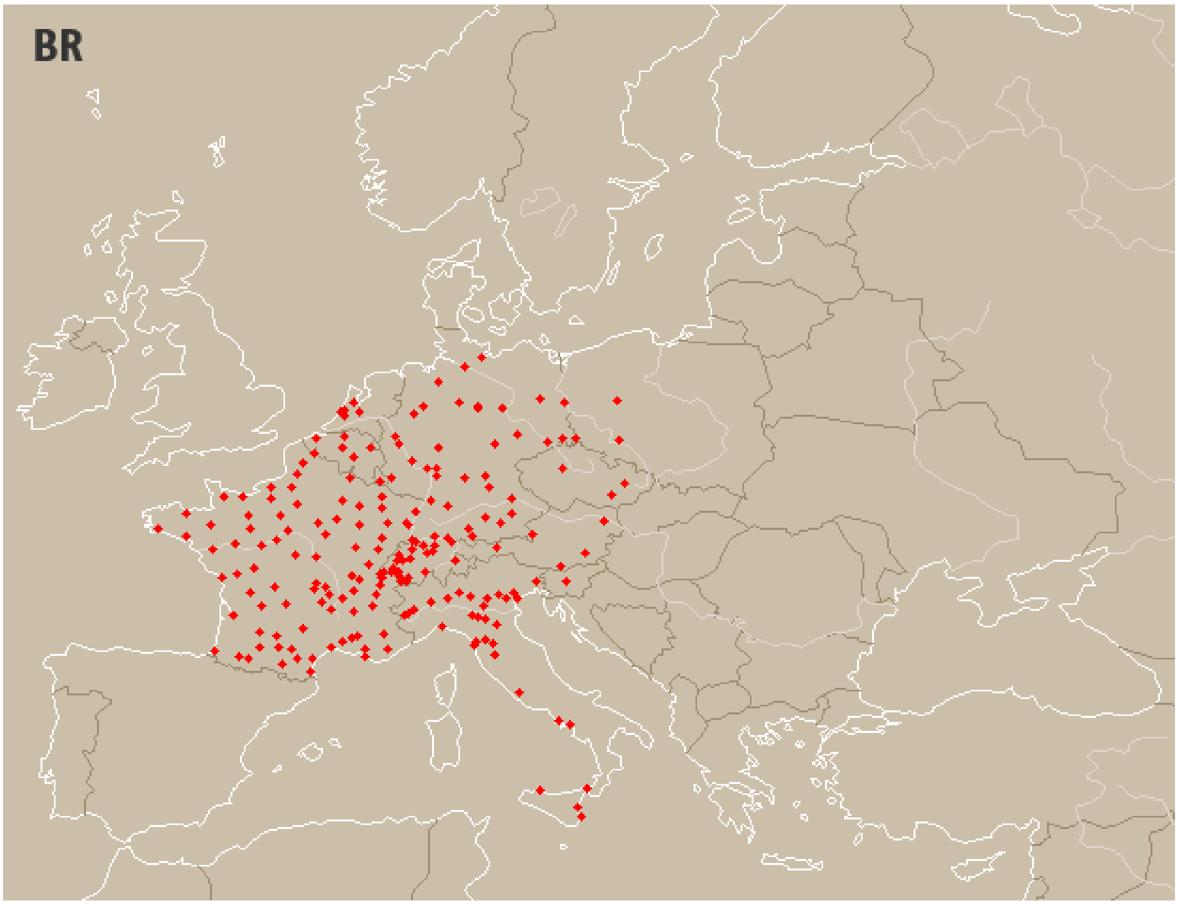


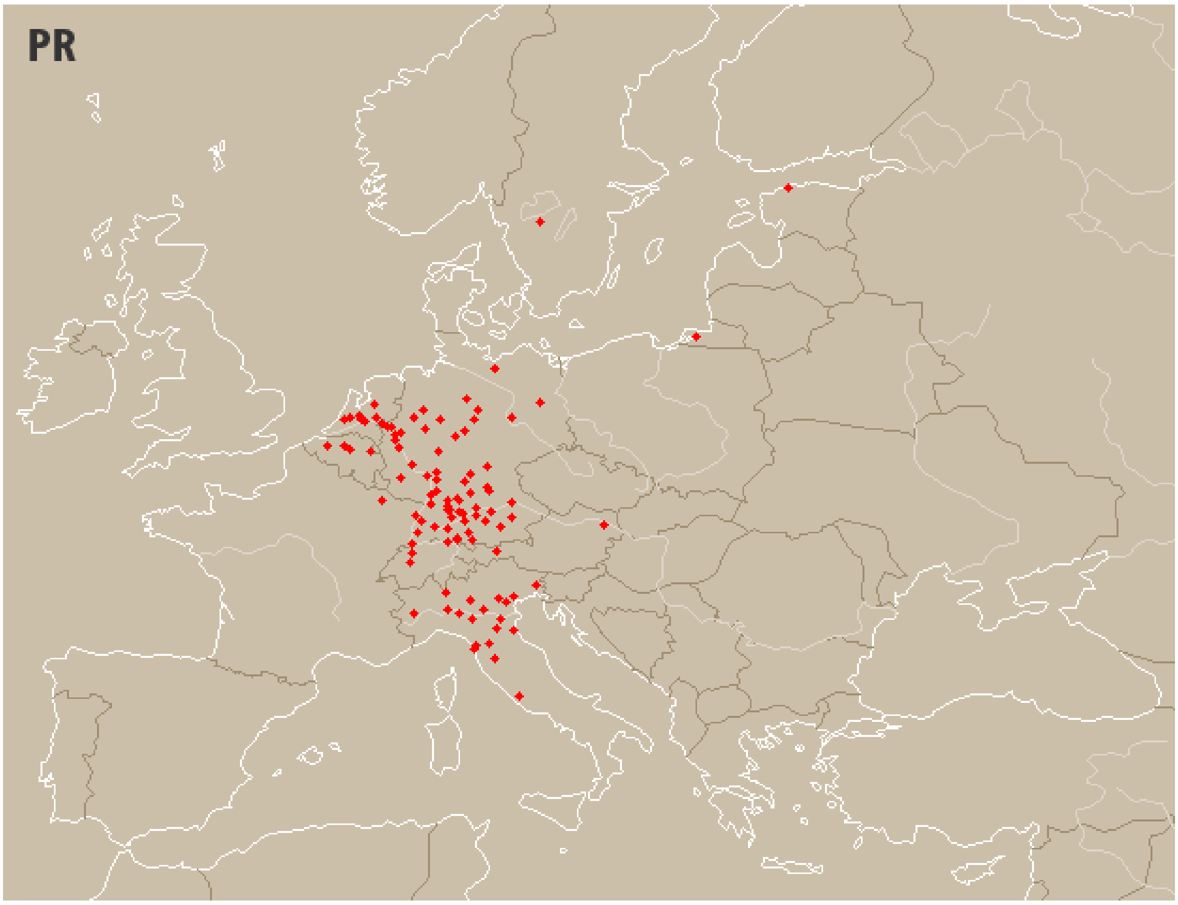

BP

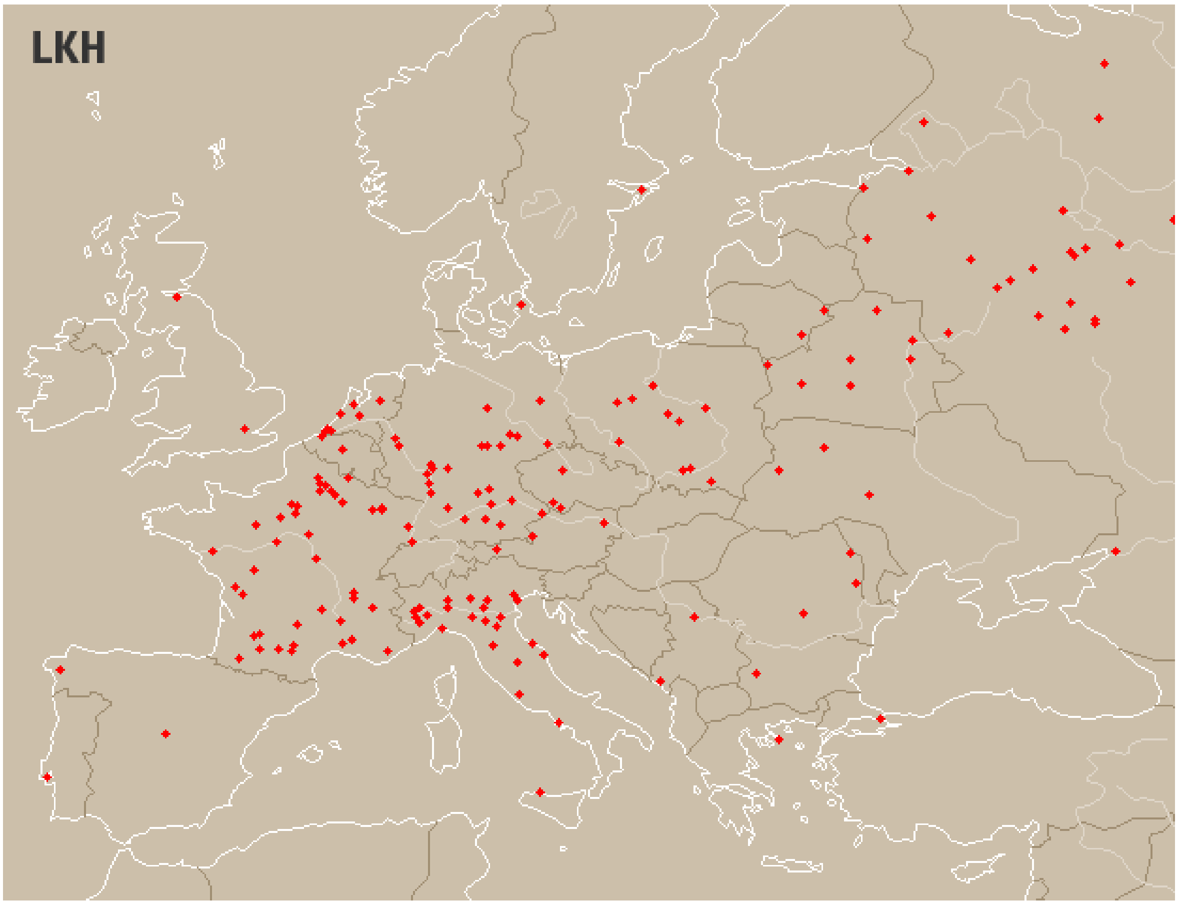

PF & PO

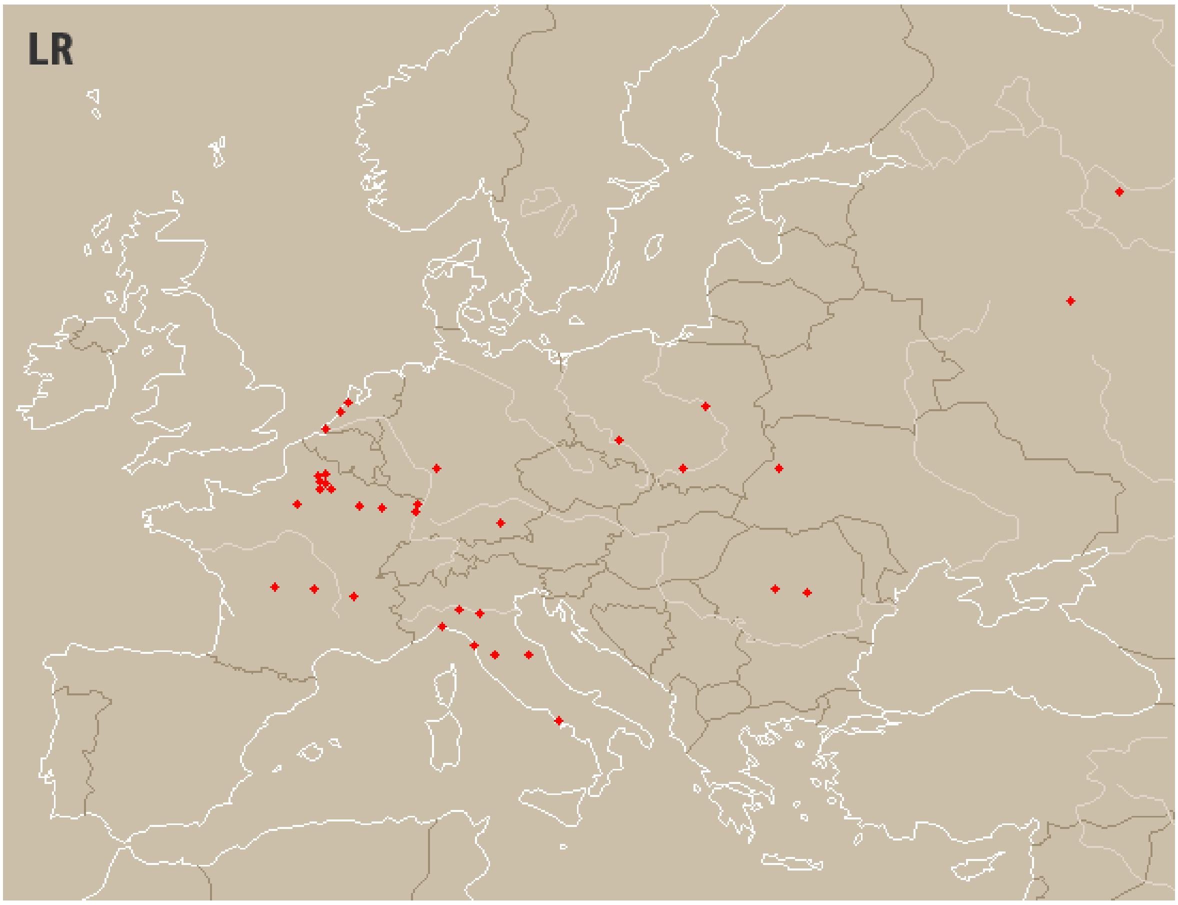

PF

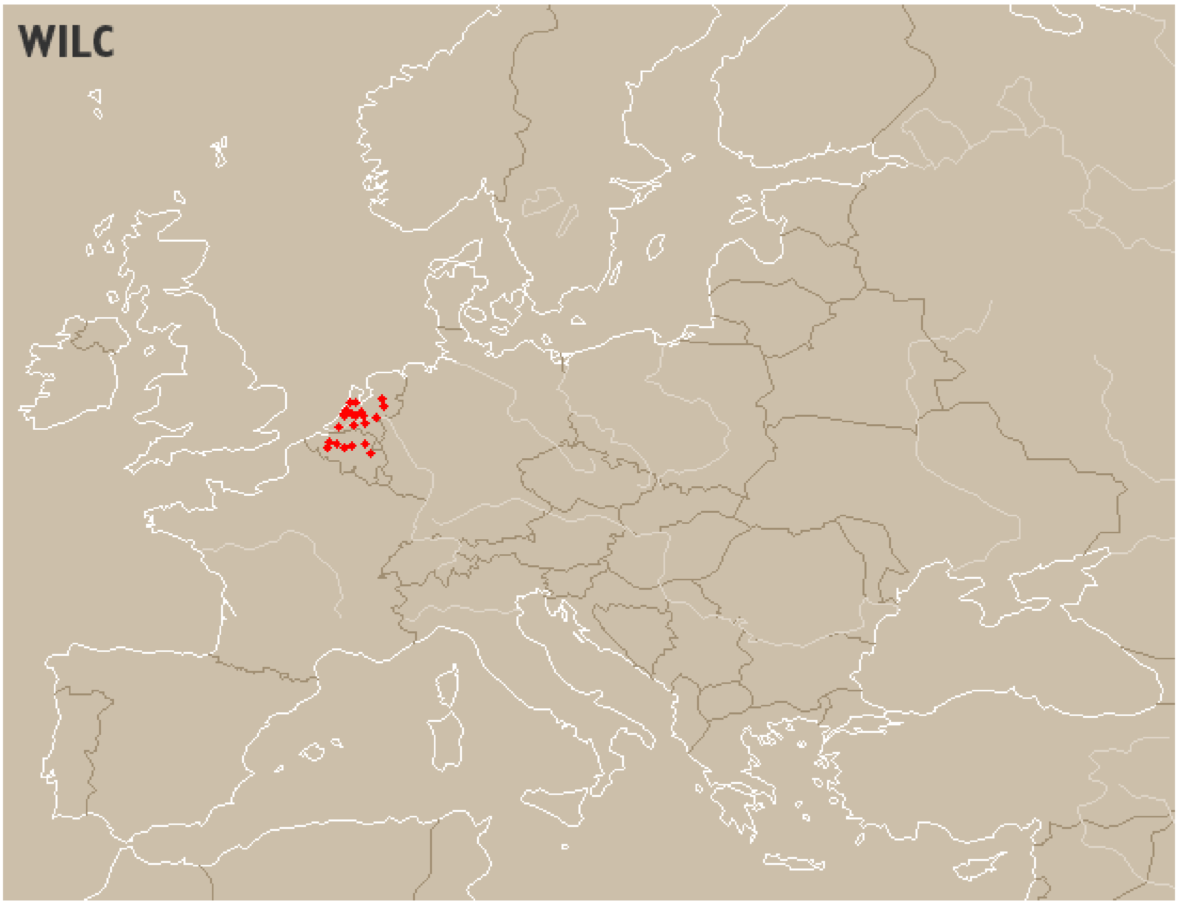

PO

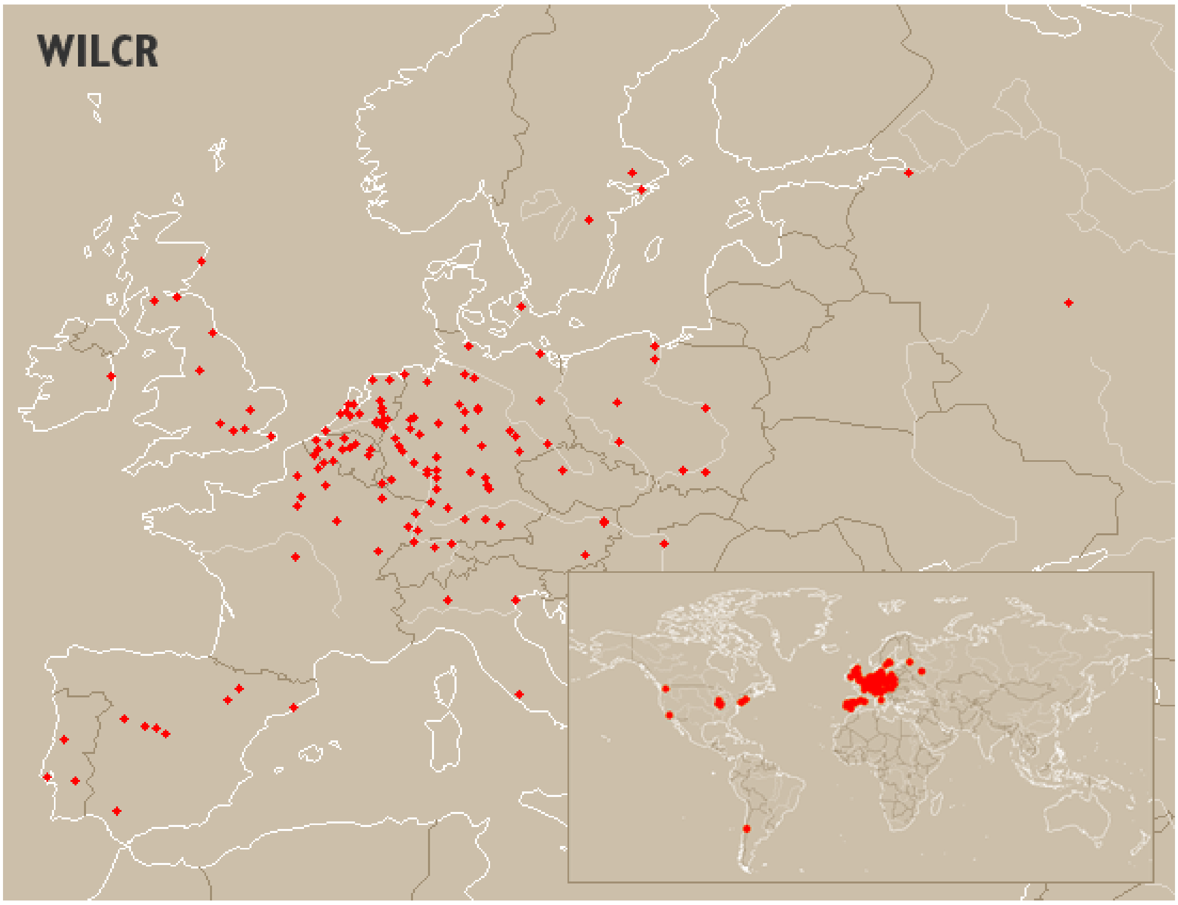

BR

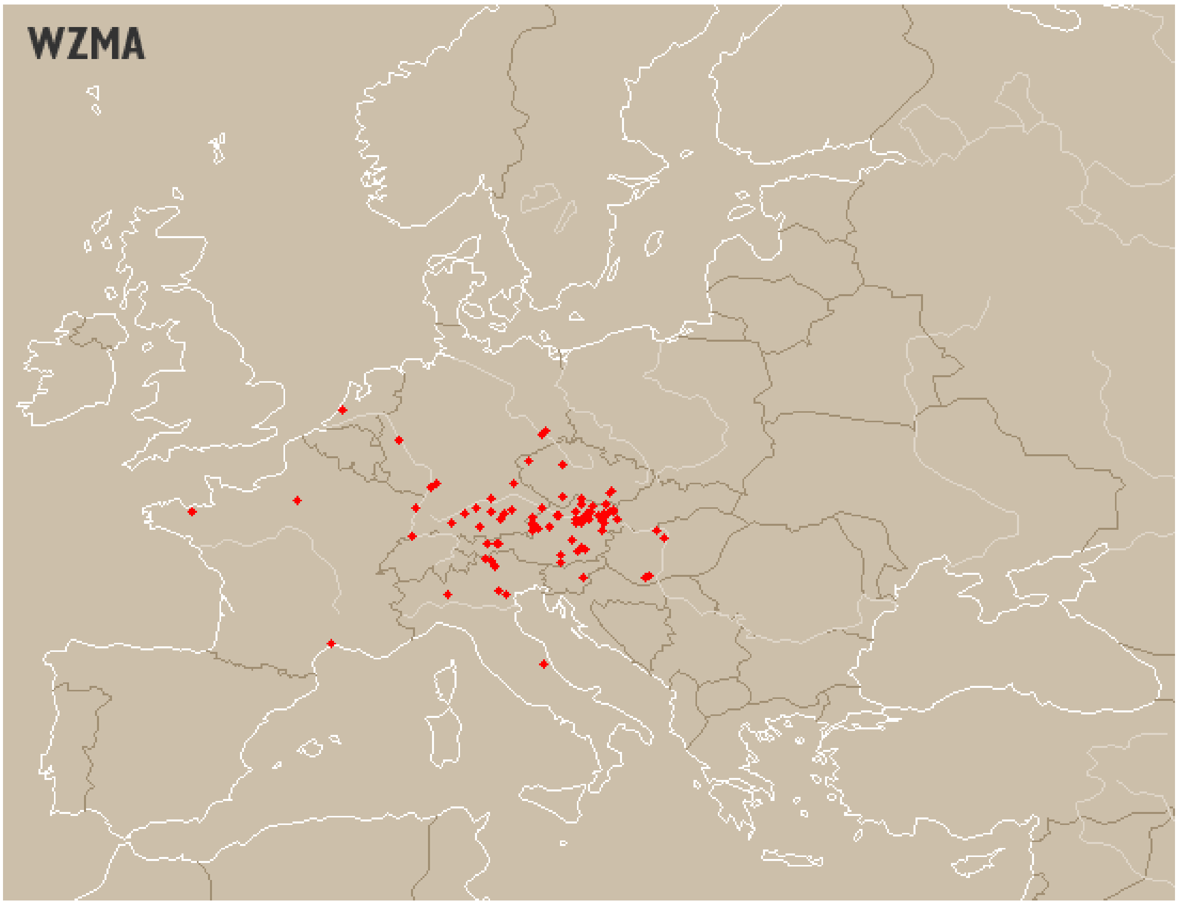

PR

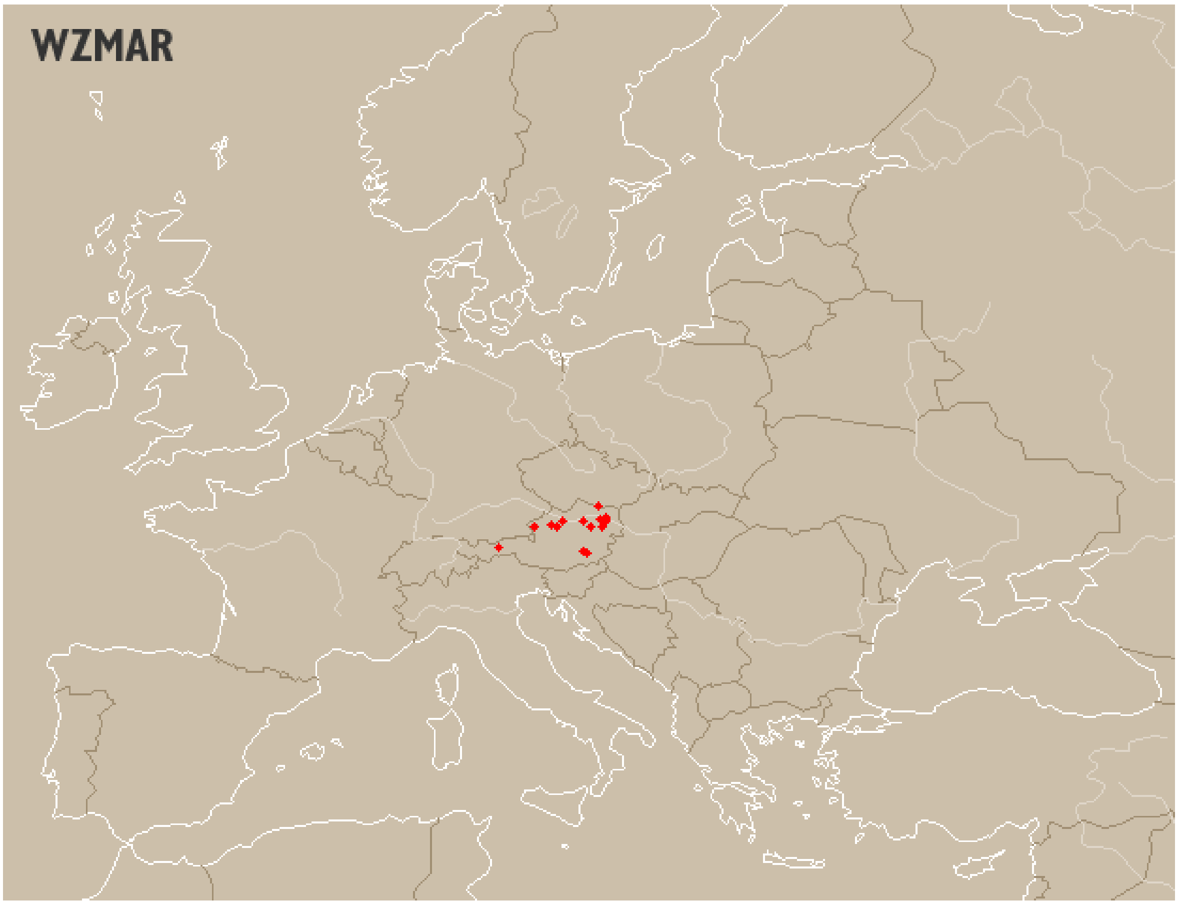

LKH

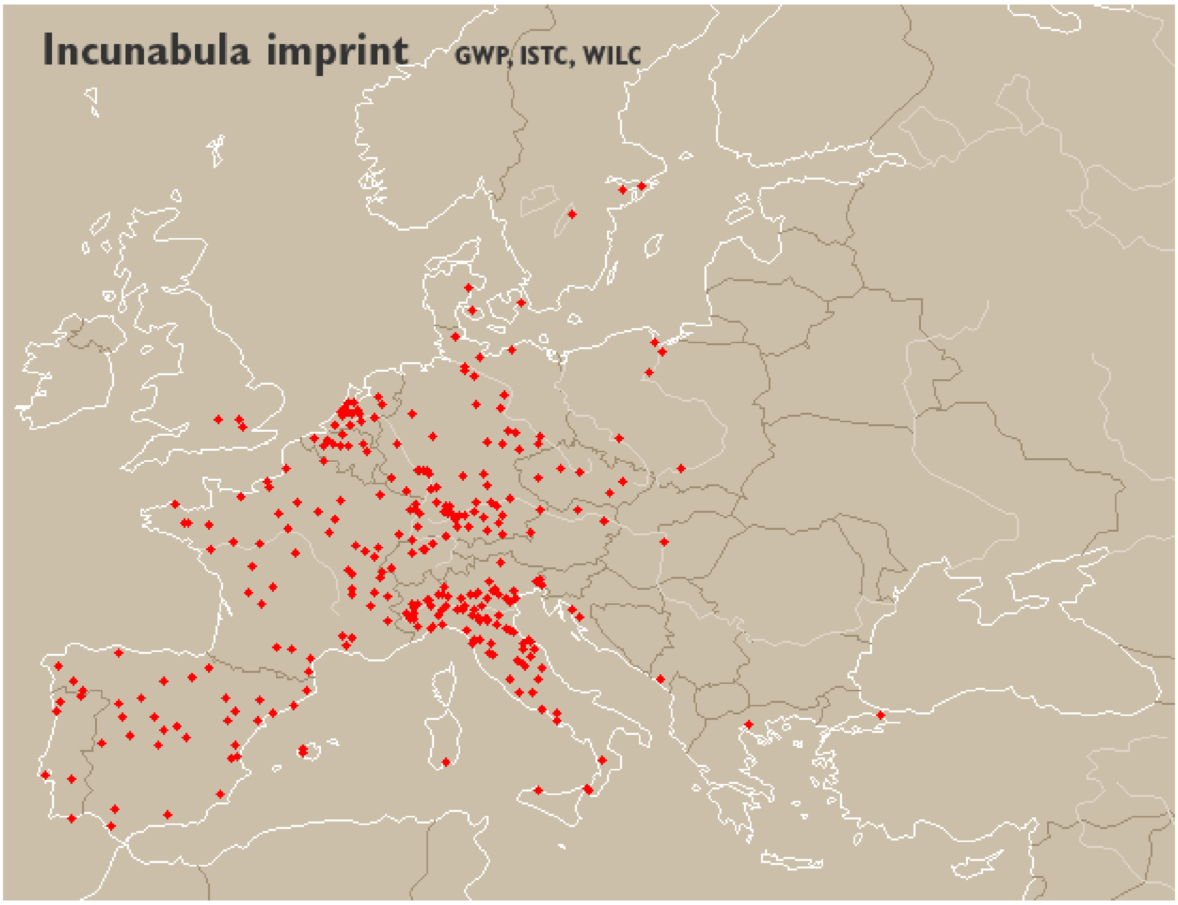

LR

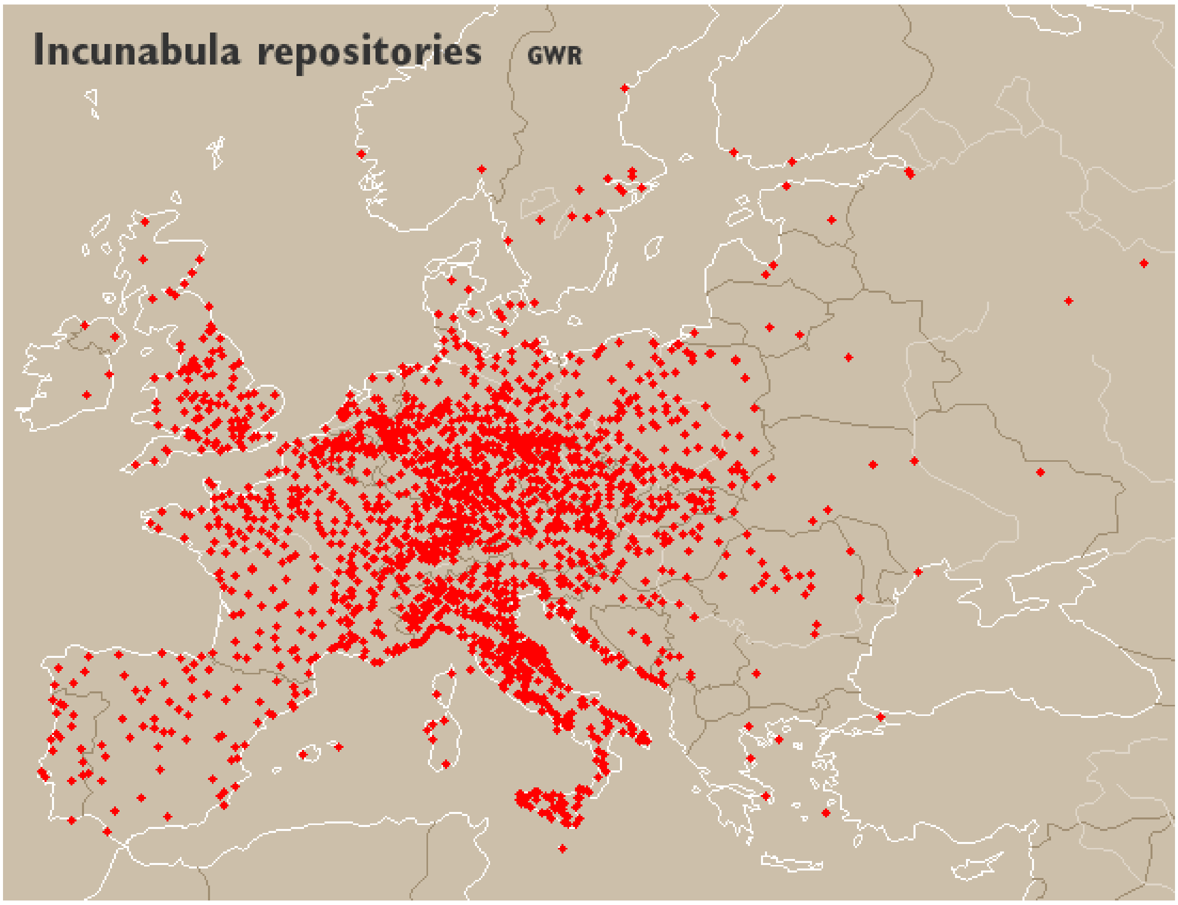

WILC

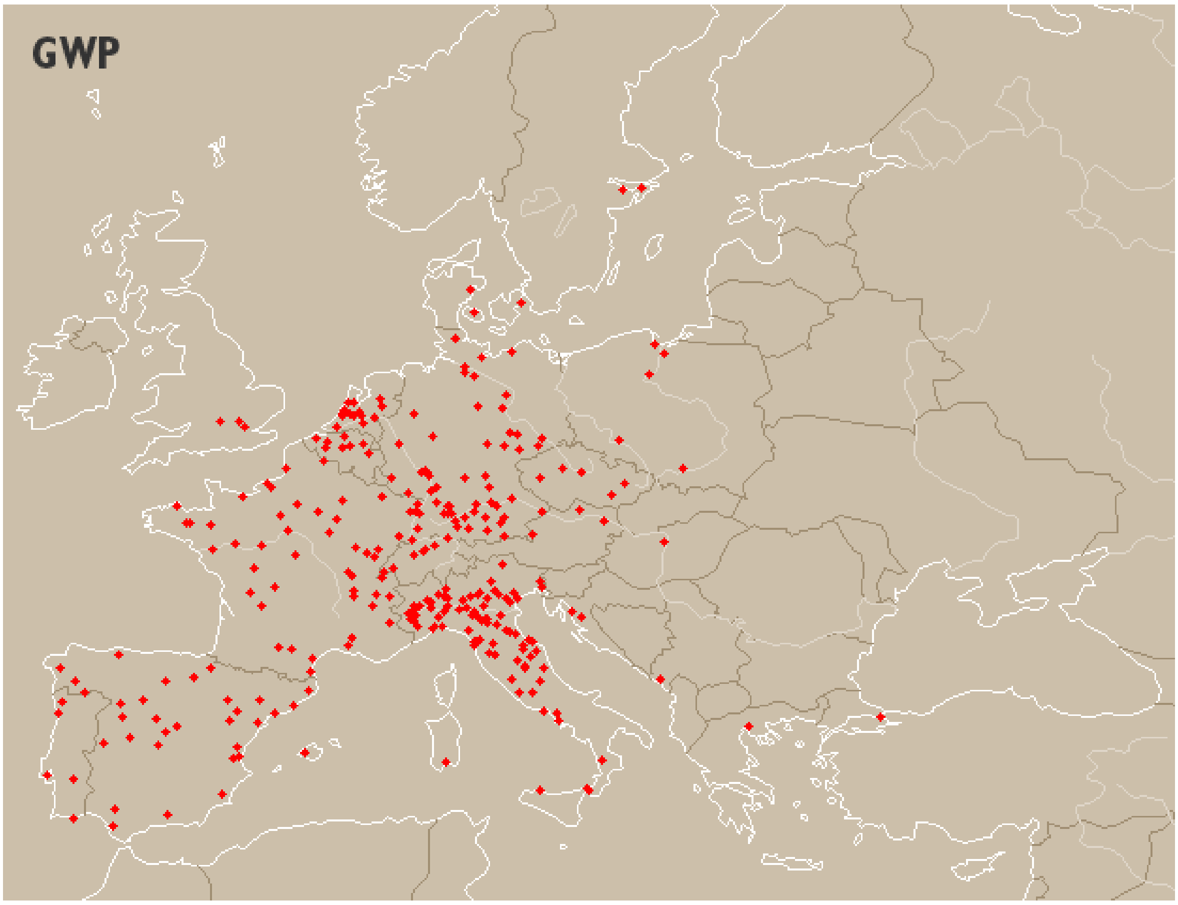

WILCR

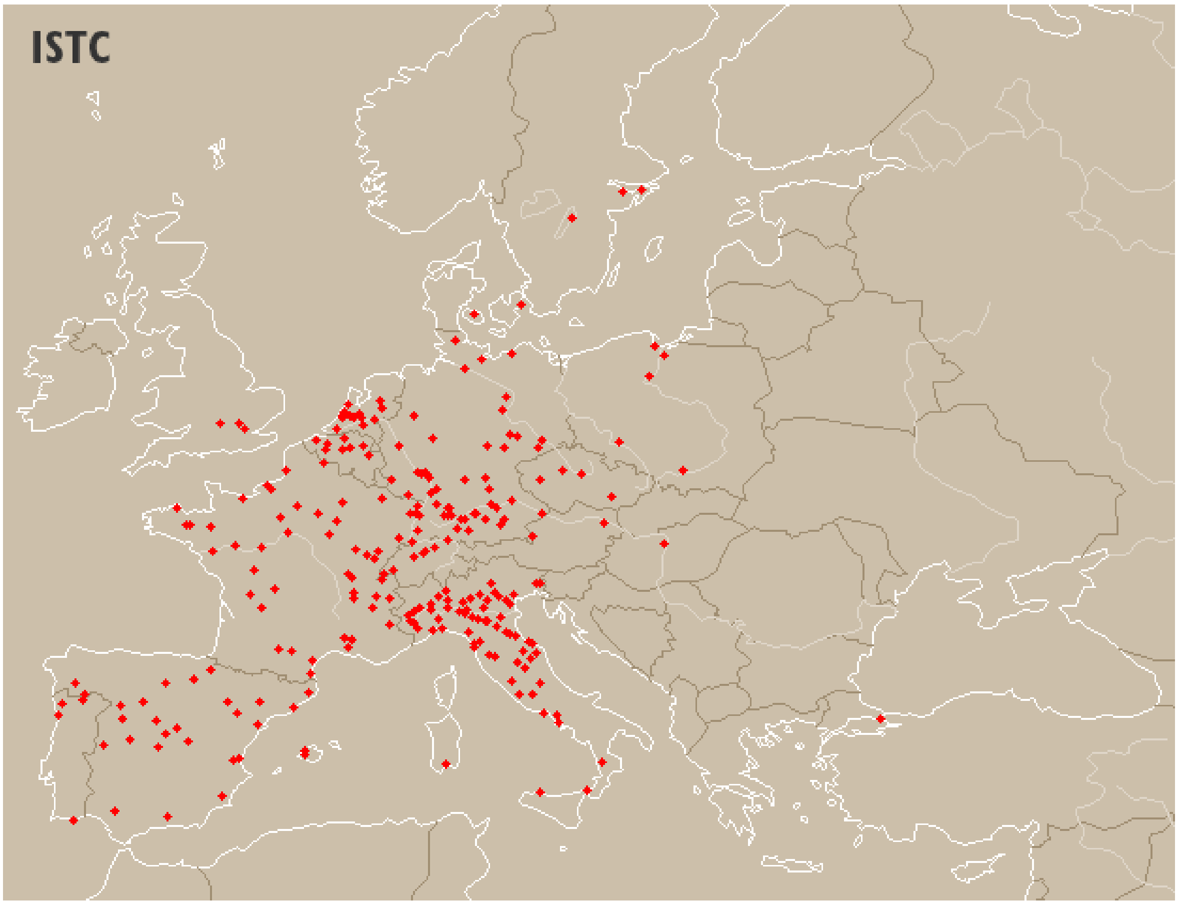

WZMA

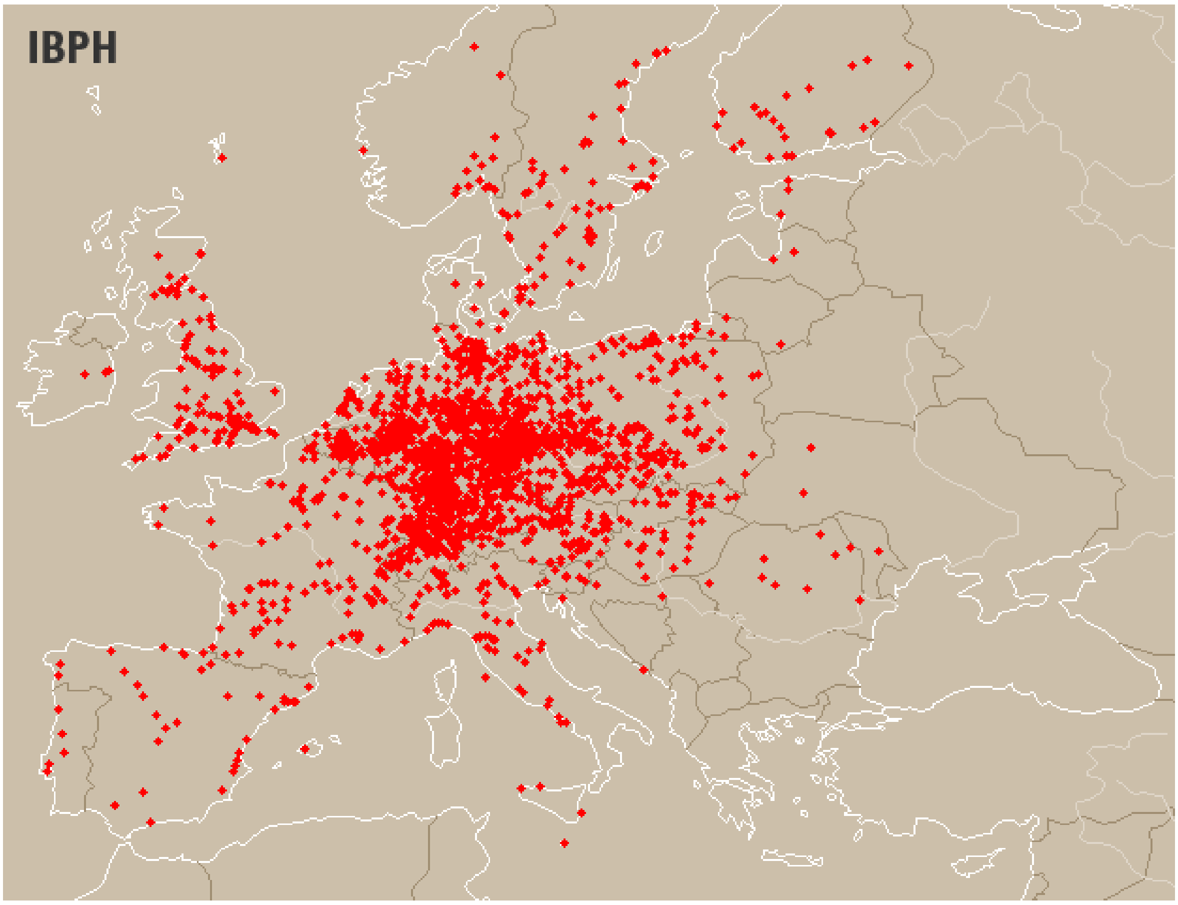

WZMAR

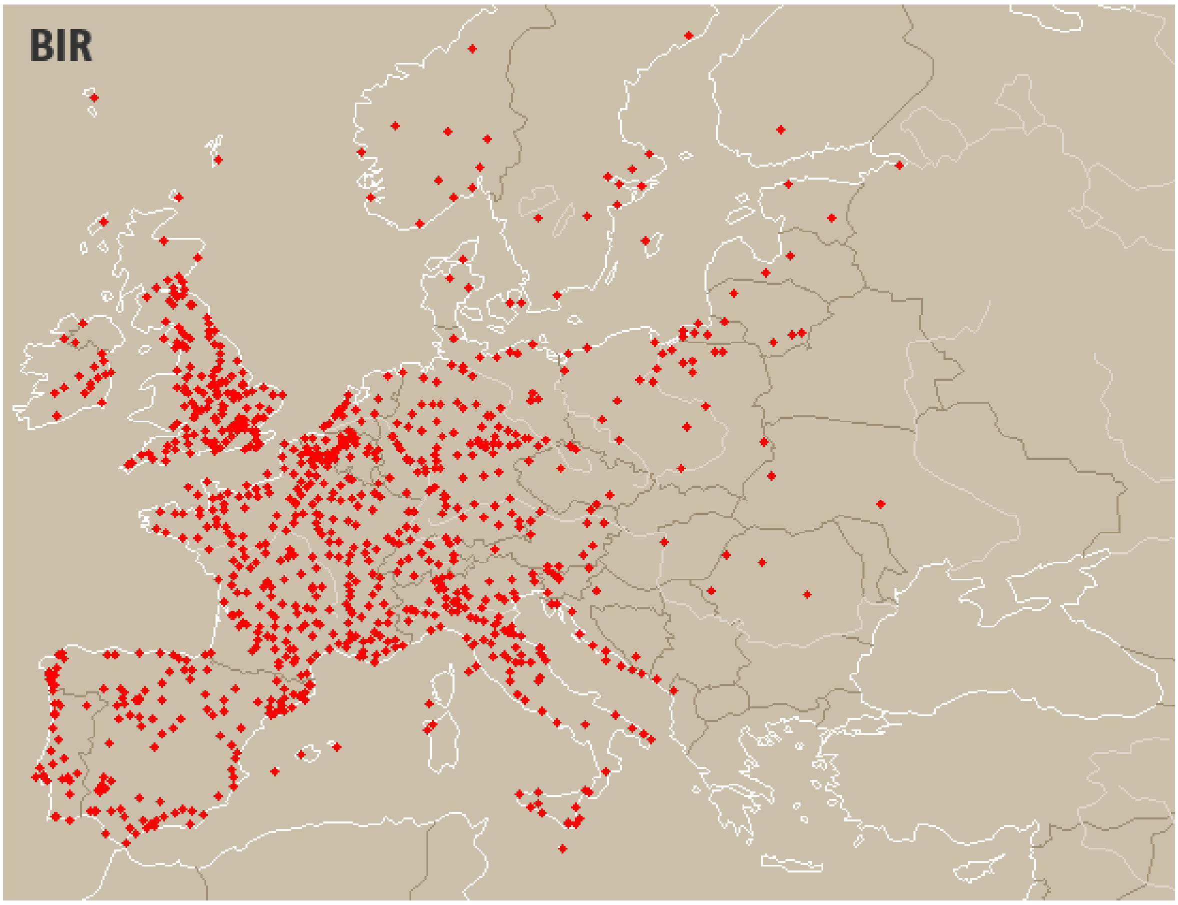

Incunabula imprint
GWP, ISTC, WILC


Incunabula repositories
GWR


GWP


ISTC


IBPH


BIR

Method

The following details the process by which the Dalek file was created.

1. *Extract place names* – Place names were received as digital files from the database owners (PO), downloaded from their website (WILC) or typed in the computer the information found in printed sources (BP).

2. *Identify locations* – The location of place names given in the sources was identified using online atlases (see point 5). When necessary several sources were consulted to minimize the risk of excluding homonym places.

3. *Disambiguation* – In many cases the same place name is given to distinct location ('Paris' in France and 'Paris' in Texas, USA). Following rules have been followed to decide to which location to attribute the place name:

a. *Additional data in the source* – Sometimes the source provides clues to the location of the place name, such as mentioning a nearby city or river.
b. *Geographical likelihood* – A bigger city is more likely to be meant than a smaller city.
c. *Historical likelihood* – Paris, Texas, USA is not a likely location of a place name of a book printed during the time of the Renaissance.

4. *Reattribution* – Although some placenames are identified, they are either too narrowly defined or too wide, so that reattribution has been operated.

a. *Inclusion* – Streets and neighborhoods are attributed the name of the respective cities (e.g.: 'Bread Street' and 'Lambeth' are normalized to 'London'), and abbeys and castles the names of the villages or cities close to which they are located.
b. *Focusing* – Names of regions are normalized to the principal settlement in the area (e.g.: 'Alsace' > 'Strasbourg'). For cross-border regions the attribution might be arbitrary ('Flanders' is partly in present day Belgium, France and Netherlands).

5. *Find coordinates* – Once it was established which location the source means, its geographical coordinates had to be found. All coordinates were converted to decimal degrees and rounded to two decimals, giving a precision of about 1 km. The main sources of information are GeoNames (http://www.geonames.org) and Wikipedia (http://en.wikipedia.org/wiki/Main_Page), various other online services being also used, such as Getty Thesaurus of Geographic Names (http://www.getty.edu/research/conducting_research/vocabularies/tgn/) and Falling Rain Genomics (http://www.fallingrain.com/world/).

6. *Attribute administrative units* – Providing the administrative unit to which a place belongs allows performing geographical statistics by clustering point-locations into areas. The overview thus gained helps the historian in his investigative work of the past. This is the reason why along with the coordinates' information on the administrative units was provided. Because administrative units are dynamical historical entities – created, abolished, their boundaries changed – it was decided to refer to the present state of administrative units. Statistical units (NUTS) where chosen over administrative units (AU) because there were available to the creators of this dataset both as nomenclature and vector boundaries for almost the entire area covered by the dataset (Europe).

Limitations

While great care has been taken to georeference the placenames, errors might subsist, some which are even not stemming from the authors of the file. Here is a list of error sources identified during the development of Dalek. Most of the errors could be identified using the Dibuk software.

— *Historical sources issues*

1. error on the placename in the original document

2. error in reading the document by the modern-time author of the source used by Dalek

— *Geographical sources issues*

3. error in the gazetteers on coordinates, names, accents, regions of belonging (e.g.: non-ASCII characters are stripped of their accents in GeoNames and the NUTS)

4. incorrect boundaries due to digitization errors or faulty generalizations (e.g.: in Matlab there are two discontinuities in the boundary of the Faroe Islands)

5. reorganization of regional boundaries (e.g.: breakup of Yougoslavia, reorganization of NUTS every few years)

6. several closely situated coordinates for the same placename (e.g.: when one source geolocates Paris with the coordinates of the Eiffel Tower and another with those of the Notre Dame – which one to follow?)

7. coordinates with various levels of precision (e.g.: 45.1 and 45.11 considered as two distinct locations)

8. system transformation errors (e.g.: coordinates and datum)

— *Development issues*

9. typing errors

10. mixing of systems (e.g.: decimal/sexagesimal coordinates, geographical datum)

11. correct coordinates but attribution to wrong region (e.g.: human error or map scale with insufficient detail)

12. variants not identified as such (e.g.: 'Dubrovnik' and 'Ragusa' are the Serbo-Croatian and Italian names of the same place, which is not self-evident without the necessary background knowledge)

Credit

The concept of this dataset originates with Ezio Ornato, with input from Vlad Atanasiu; georeferencing was done by Claire Priol and Anne Tournieroux; development software for visualization and debugging written by Vlad Atanasiu. The work has been undertaken in 2007-8 at the Laboratory for Medieval Studies of Paris, National Center for Scientific Research / University Paris 1, Paris, France, within the project "Bernstein – The Memory of Papers" (http://www.bernstein.oeaw.ac.at), co-funded by the European Commission, under the programme eContentPlus (ECP-2005-CULT-038097/Bernstein).