eContentPlus

European Commission
Information Society and Media Directorate-General

ECP·2005·CULT·038097/Bernstein
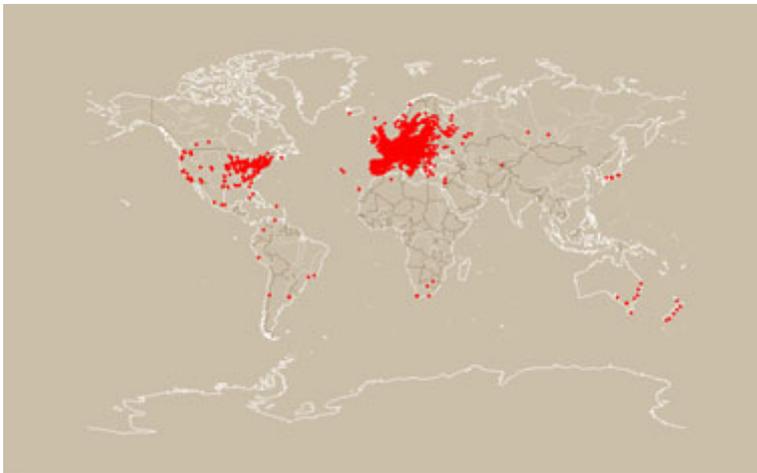http://www.bernstein.oeaw.ac.at

Deliverable no. 27, ref. 2.6, Geographical & chronological metadata

Places georeferenced in the Bernstein databases

Deliverable number	D2.7
Dissemination level	Public
Delivery date	30 January 2009
Status	Final
Author(s)	The Bernstein Consortium

1. Summary

The Bernstein integrated databases contain about 120.000 records describing papers and watermarks of which 90% many provide information about where and when the items were used – for example in which place and in which year a book was printed on a paper with a specific watermark. Provided coordinates are added to the records, maps of the distribution in time and space of papers and watermarks can be produced. Thus new perspectives on paper history emerge, as well as new ways to interact and evaluate the data. The impact is further increased by contextualization, when the data on paper is used in the context of other digital datasets, on book history, economy or political history, to name just a few. This is the background of the work on geographical and chronological metadata in project Bernstein.

A first batch of metadata was delivered in July 2007, as part of deliverable nr. 10, ref 2.2, GIS data batch no. 1 (http://www.bernstein.oeaw.ac.at/twiki/bin/viewfile/Main/GisServices?filename=Deliverable10_D2.2_GIS_Data_v1_2007.07.23. pdf). The dataset is called Dalek and downloadable from (http://www.bernstein.oeaw.ac.at/twiki/bin/view/Main/ GisServices > Datasets > Dalek > dataset). The present deliverable contains additions and upgrades to the previous. The format, amount of the data and the methodology are presented in deliverable nr. 10 and the documentation accompaigning the georeferences dataset (http://www.bernstein.oeaw.ac.at/twiki/bin/view/Main/ GisServices > Datasets > Dalek > documentation). Here are presented the novelties: new datasets, new capabilities and corrections, scientific spin–off–s and plans for the future.

A note on automation

There are numerous services online that provide coordinates for placenames, so why not simply download these lists for Bernstein? The reason is because of the homonyms. Many placenames refer to different locations and the knowledge which one is implied by a specific occurrence in a specific document necessitates – given the technological means available to the Bernstein Consortium – human intervention. To give an example geonames.org puts at fingertips length the coordinates N 48° 51', E 2° 20' in France, N 33° 39', W 95° 33' in Texas, USA plus 147 other populated places for "Paris". It doesn't say however which one is the right one, since it doesn't know if I'm reading an article about Wim Wenders or the *The Hunchback of Notre Dame*.

2. Novelties

The novelties concern new georeferenced datasets, new capabilities of how the data can be used, and corrections to the previously delivered data.

1. Datasets

Georeferencing of the four Bernstein databases, four major paper and watermarks repertories and four contextual datasets has been achieved. Through this work approximately *250.000 records* with geographical information or *80% of the total records* in the cumulated sources are georeferenced, representing *7.846 unique locations*.

A. Bernstein databases:	1. International Database of Watermarks and Paper used for Prints and Drawings (use places [U] and repositories [R])
	2. Piccard Online [U&R]
	3. Watermarks in Incunabula printed in the Low Countries [U&R]
	4. Watermarks of the Middle-Ages [U&R]
B. Repertories:	5. Briquet [U&R]
	6. Likhacev [U&R]
	7. Piccard Findbuch [U]
	8. Watermarks in Incunabula printed in España
C. Contextual datasets:	9. Biraben (medieval plague epidemies)
	10. Gesamtkatalog der Wiegendrücke [U]
	11. Incunabula Short-Title Catalog [U]
	12. International Bibliography of Paper History

The bibliographical references for the above datasets are given in the Dalek documentation; for the repertories see also deliverable nr. 26, ref. 2.7, Repertories concordances (http://www.bernstein.oeaw.ac.at/twiki/bin/view/Main/ DocumentsArchive > Deliverables).

2. Capabilities

1. *Fuzzy regions*: The new dataset makes possible to represent regions that have a fluctuating boundary history or no precisely known boundary trace. Every European state is a potential example in this regard. The solution was to georeference regions by a single coordinates pair used to put the regions name on the map, as is the case for mountain chains in cartography. A "The Alps" label clearly situate them without however defining their border.

2. *Multiple placenames*: The ability to encode in a standardized way records attributed to multiple locations has been improved. It is now possible to include also such records in the statistics and mapping of Bernstein data.

3. *Uncertain localizations*: In the new georeferences dataset the information on an uncertain localization as given by some records is maintained and usable in the statistics and maps.

The following example shows cases of how fuzzy regions, multiple placenames and uncertain localizations are encoded.

Original	*Normalized*
Basel und Klosterneuburg	Basel & Klosterneuburg
Solothurn oder Luzern	Solothurn | Luzern
Wien od. Klosterneuburg	Wien | Klosterneuburg
Wien oder Klosterneuburg (?)	Wien ? | Klosterneuburg ?
Wien/Klosterneuburg (?)	Wien ? | Klosterneuburg ?
Wien/Niederösterreich (?)	Wien ? | Niederösterreich ?
Österreich/Bayern	Österreich | Bayern
Wien?	Wien ?
Graz ?	Graz ?
Gmunden (?)	Gmunden ?
Ober- oder Niederösterreich (?)	Oberösterreich ? | Niederösterreich ?
Tirol und Steiermark (?)	Tirol ? & Steiermark ?
Straßburg (?) (Kärnten)	Straßburg ?
Süddeutschland (Dillingen ?)	Süddeutschland | Dillingen ?
Südostdeutschland (od. Böhmen (?)	Südostdeutschland ? | Böhmen ?

3. Corrections

Various types of mistyping and erroneous georeferences where fixed in the previous Dalek file. A software, Dibuk, was developed to assist debugging, allowing automatic identification of the following possible errors:

- Variants: same placenames and same coordinates
- Focus: same placenames with slightly different coordinates
- Homonyms: same placenames with sizeable different coordinates
- Proximity: different placenames, but very close coordinates
- Outliers: coordinates outside a given region
- Precision: rounding coordinates to a given number of decimals

Dibuk is downloadable from and documented on the page http://www.bernstein.oeaw.ac.at/twiki/bin/view/Main/GisServices > Tools.

A peek inside the Dalek file

N°	Status	Original Placenames	Normalized placename	Latitude	Longitude	Code	Country	NUTS 1	NUTS 2	NUTS 3	NUTS 4	Custom	BRQ-U	BRQ-R	GW-U	GW-R	IBPH	ISTC	LKH-U	LKH
6 775	locality	Luxemburg	Luxembourg	49.61	6.13	LU000	Luxembourg (Grand-Duch	Luxembourg (Grand-l	Luxembourg (Grand-	Luxembourg (Gra			x	x		x	x			
6 776	locality	Luxemburg	Luxembourg	49.61	6.13	LU000	Luxembourg (Grand-Duch	Luxembourg (Grand-l	Luxembourg (Grand-	Luxembourg (Gra			x							
13 662	locality	Luxemburg, BN	Luxembourg	49.61	6.13	LU000	Luxembourg (Grand-Duch	Luxembourg (Grand-l	Luxembourg (Grand-	Luxembourg (Gra										
6 777	region	Luxemburg <Provinz>	Luxembourg (Grand-Duché)	49.82	6.13	LU000	Luxembourg (Grand-Duch	Luxembourg (Grand-l	Luxembourg (Grand-	Luxembourg (Gra							x			
6 778	locality	Luxeuil	Luxeuil	47.82	6.38	FR433	France	Est	Franche-Comté	Haute-Saône			x	x		x				
6 705	locality	Lucerna	Luzern	47.08	8.27	CH061	Schweiz/Suisse/Svizzera	Schweiz/Suisse/Sviz	Zentralschweiz	Luzern			x							
6 706	locality	Lucerne	Luzern	47.08	8.27	CH061	Schweiz/Suisse/Svizzera	Schweiz/Suisse/Sviz	Zentralschweiz	Luzern			x							
6 780	locality	Luzern	Luzern	47.08	8.27	CH061	Schweiz/Suisse/Svizzera	Schweiz/Suisse/Sviz	Zentralschweiz	Luzern			x	x		x	x			
6 781	region	Luzern <Kanton>	Luzern	47.08	8.17	CH061	Schweiz/Suisse/Svizzera	Schweiz/Suisse/Sviz	Zentralschweiz	Luzern							x			

This image shows part of the georeferences file Dalek incorporated into the Bernstein databases. Note the normalization of names undertaken, the georeferences – coordinates and region names –, as well as the specification in which source the placename is to be found (from column M on). Also visible is how the same placename can refer to different entities across sources or inside the same source: in IBPH "Luxembourg" means either the city or the duchy. This is the sort of disambiguation that had to be performed in order to accommodate the Bernstein databases.

3. Chronological data

Chronological data are mentions of centuries, years, months and days to which a record pertains. Work on this type of data consisted in harmonizing the data format across datasets and expressing centuries in year ranges, so that the integrated search software can access and process this type of data. There was no need for producing an independent dataset like for the georeferences.

4. Spin-off-s

Having started with a simple need to geolocate places in the Bernstein databases, this work has lead to an increased accumulation of expertise in historical georeferencing. After several conferences on Geographical Information Systems, it became clear that a need exists for comprehensive research and training of parties interested in spatial representation of historical data. The writing of a *Handbook of Historical Georeferencing* is thus considered, as well as expanding the acquired expertise in projects beyond Bernstein.

5. References

Georeferences file Dalek
Dalek documentation
Georeferences debugging software Dibuk

> http://www.bernstein.oeaw.ac.at/twiki/bin/view/Main/ GisServices > Datasets | Tools

Deliverable nr. 10, ref 2.2, GIS data batch no. 1
Deliverable no. 27, ref. 2.6, Geographical & chronological metadata
Deliverable no. 28, ref. 4.5, Historical paper cartography tool
Deliverable nr. 26, ref. 2.7, Repertories concordances

> http://www.bernstein.oeaw.ac.at/twiki/bin/view/Main/ DocumentsArchive > Deliverables