Eric Grosso*

Integration of historical geographic data into current georeferenced frameworks: A user-centred approach

Keywords: data integration; historical geographic data; old maps; georeferencing.

Summary: Historical data can be seen as having two spheres of interest to users. The first being that historical data, mainly old maps, are often visually attractive and represent cultural and artistic information. The second is that historical data contains invaluable information which is often unmapped or not represented in current maps or data. Historical data is thus particularly of interest to various individuals and professional groups, from historians to ecologists and from archaeologists to glaciologists. In order to exploit, analyse and process historical data, these actors have identified a need to integrate it into a recent georeferenced framework. A georeferencing process is thus essential.
In this context and based on the fact that historical geographic data will be increasingly available in a vector format, this paper deals with methods to integrate vectorised historical geographic data into current georeferenced frameworks and proposes a user-centred approach in order to take into account both users knowledge and users constraints in this context. This paper introduces the idea than several georeferencing processing methods are possible according to users needs. An example is given through the integration of a Cassini map sheet into the current topological component of the French geographical reference framework.

Introduction

Over the last few years the creation and diffusion of geographical information have considerably increased. In order to benefit from this information as much as possible, either to carry out better analysis or to study the temporal evolution of georeferenced data, users have expressed a strong need to couple their data with other data provided both by producers or other users. The need to integrate historical data is particularly expressed as it often contains invaluable information which is often unmapped or not represented in current maps or data. Historical data is of particular interest to ecologists (study of forest evolutions, comparison of ground occupation on various dates, study of climate evolution, etc.) (see e.g. Sanderson and Brown 2007), archaeologists, historians, and also to research scientists who work in the field of simulation (research of evolution rules based on historical data). To enable users to couple this data, a data integration process is needed (Parent and Spaccapietra 2000, Sheth and Larson 1990). This integration can be done by integrating all user data into a common frame of reference which is usually the most detailed database available to the institutional geographical data producers. Consequently, the goal of these actors is first to digitise historical data, then to integrate it into a recent georeferenced framework. Finally, this data can be vectorised thus enabling a better exploitation, analysis and processing of the historical data and giving a more meaningful result. Contrary to the digitalisation and georeferencing processes which are often used, the vectorisation process is rarely used. This is in part due to the fact that users are not necessarily familiar with the effect of the georeferencing process on vectorised data. In this context, this paper focuses on the integration of vectorised historical geographic

* PhD Student, Institut Géographique National, COGIT Laboratory, 73, avenue de Paris, 94365 Saint-Mandé cedex [eric.grosso@ign.fr]

data into current georeferenced frameworks and proposes a user-centred approach in order to take into account both users knowledge and users constraints.

Historical geographic data: from an archive process to analysis

Over the last few decades, the digitalisation of historical data has been used to archive, to protect and to preserve historical geographic data such as old maps, archaeological plans, monographs, charts, etc. In parallel, this phenomenon has consequently contributed to the wider diffusion of this data.

This diffusion in turn also increased its utilisation, mainly by ecologists, glaciologists, archaeologists, historians and more generally by research scientists. Although it is possible to use data in its actual state or digitalised, users realise that this data becomes more useful if its content is directly integrated into a recent frame of reference, thus enabling that data to be manipulated using a Geographical Information System (GIS). Indeed, the process of integration of historical data into a recent frame of reference increases the possibility of analysis with current data. An integration process is thus essential. Currently, the method used to integrate this data consists of georeferencing the digitalised image onto the ancient document.

Several projects already propose tools adapted to this data to help users deal with problems arising from the historical data integration process. For example, the Old Maps Online project aims to help with online publishing and georeferencing of scanned historical maps (Pridal and Zabicka 2008) based on the observation that the ability to use old maps is critically under-utilised in the current Internet environment. Another example is MapAnalyst (Jenny et al. 2007), a tool to visualise planimetric accuracy of historical maps.

Finally, to go one step further, users are able to obtain even better results thanks to the vectorisation process. It is noted that a growing number of projects choose to use this solution, rather than using the simpler method of georeferencing raster data, as it offers the most possibilities for complex data analysis (Dupouey et al 2007, Noizet 2009). In this last context, this paper describes firstly the specificities of vector data, then the classical geoprocessing method, and finally proposes an improved georeferencing method which is illustrated through an example.

From raster data to vector data

A map sheet is generally digitalised in one or more raster data. This data contains all themes of a map in a single layer. From this single raster layer, the vectorisation process provides a specific data layer for each theme (see Fig. 1). Therefore, users can manipulate data separately and have a specific focus on a particular theme.

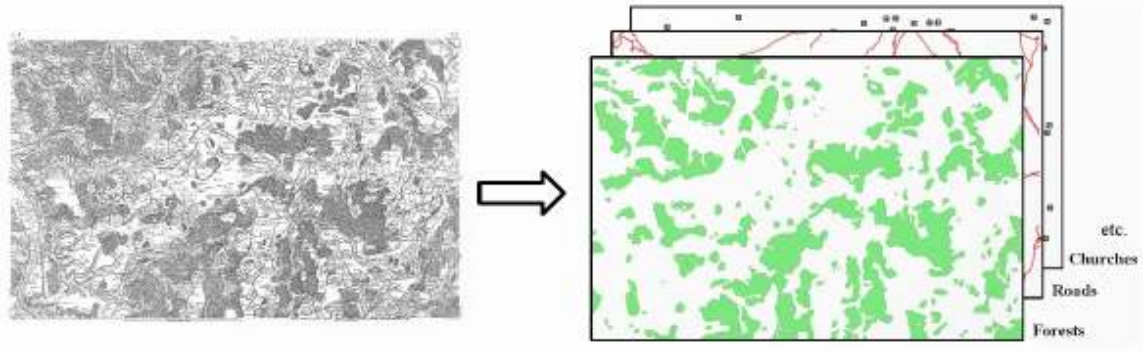

Figure 1: Different vector layers digitalised from one raster map.

Vector data is not only made up of geometric locations but also includes associated attribute information such as e.g. the name of a town, the nature of a religious place or the type of a road (e.g. path, gravel road, gravel road with trees, road in construction, etc.).
To enable the user to correctly enrich that data it is important to note that a good knowledge of historical data is needed in the vectorisation process. This is true particularly in the case of map legends. This information can be found in books whose authors study specific data (see e.g. for Cassini maps, Pelletier 1990), or in the specifications of a map (see e.g Institut Géographique National 1950) or in books which compare different historical data (Costa and Robert 2009).

Georeferencing process

To georeference data, users firstly need to define at least three control points. Control points are traditionally described as a couple of planimetric coordinates which enable to compute a spatial transformation from a source frame (here the historical data frame) to a target frame (here the recent frame of reference). A couple of control points loosely consists of a source control point and a target control point.
Once control points have been selected, a georeferencing process (comparable to a global spatial adjustment) needs to be carried out. Users have a choice among several possible spatial transformations to achieve this task: affine transformations, Helmert transformations (four or seven parameters), transformations based on a gravitating model (Langlois 1994), triangulation and rubbert sheeting (White 1985), second (or higher) order polynomial transformations, thin-plate spline method, etc. Users also have a choice among several possible resolution methods to mathematically solve the problems related to spatial transformations.

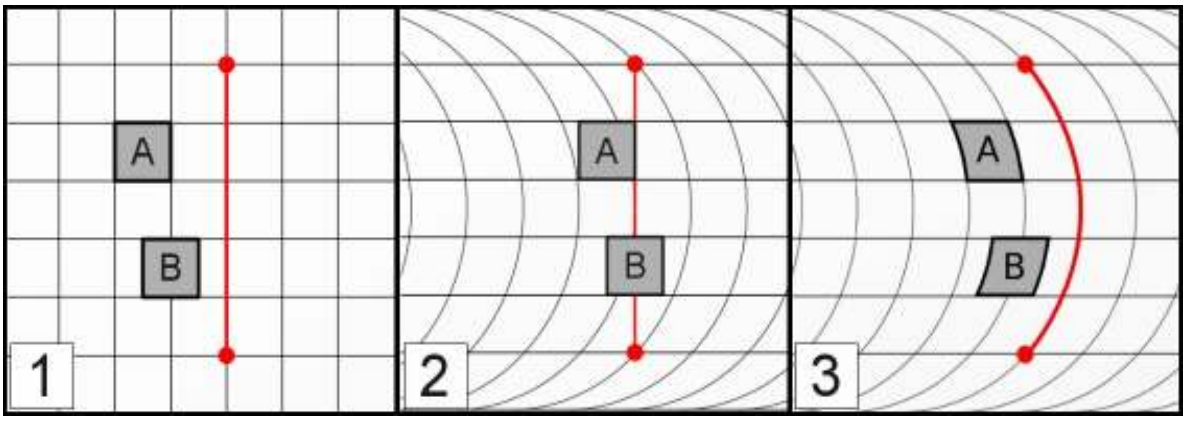

Figure 2: Possible problems of georeferencing vector data.

Finally, we outline the method of how to apply the spatial transformation to vectorised data. Indeed, some spatial transformations uniformly applied (as applied to raster data) to all the coordinates of vectorised data can provide some incorrect or inaccurate results. Figure 2 shows this problem through a fictive example.
Figure 2-1 shows the vectorized data before the georeferencing process. Features A and B can be assimilated to buildings, and the red segment to a road. Buildings are made up with four points, one in each corner, and the road is made up only with two points (one at each extremity). A transformation grid is computed as seen in the background of figures 2-2 and 2-3. In figure 2-2, the transformation applied to buildings is to move the centroids using the grid and to build the same

building around these centroids with maintaining buildings shapes. The transformation applied to the road is to move each point using the grid (if each point of the building had been displaced using the grid – the building A would have touched the road). As a result, the topological relations between objects changed. It is thus very complex to have a quick and objective analysis of the old map in comparison with the current one. To rectify this, the transformation has to be applied to resampled vectorised data as shown in the figure 2-3. However, in this last case, it is difficult to apply an analysis process (e.g. orientation analysis, surface analysis, etc.) to the transformed data.
Consequently, our approach takes into account the problems of georeferencing vectorised data, as outlined above.

A user-centred approach to georeferencing historical vector data

Global approach

We begin with two observations. Firstly users are not necessarily familiar with the georeferencing process, notably with spatial transformations. Secondly, users sometimes need to use georeferenced data for other tasks than those which simply consist of overlapping several layers, e.g. to analyse the orientation of geographical features.
Our approach would therefore propose the "most adapted" transformation regarding user needs and the most adapted mathematical way to solve the given problem (even if a least squares adjustment is traditionally used to solve the problem). "Most adapted" means here that the transformation has to satisfy different kinds of constraints. The objective is to minimize distortions and to take into account some possible user constraints.
Distortions have to be quantified to know how a transformation can accurately map all control points. This can be done by computing the Root Mean Square (RMS) error based on the residual errors (a residual error is the distance between the target control point and the associated transformed source point). This indicator gives a good assessment of the consistency and the accuracy of a transformation between the different control points. Nevertheless, even if the RMS error is low, some residual errors can be particularly significant (e.g. due to misplaced control points). In this case, a couple of control points can be removed to improve the transformation.
To take into account some possible user constraints, the transformation has to minimize as much as possible length, angular or surface distortions. Information is in this case added to the system, for example "an affine transformation implies that straight lines remain straight, parallel lines remain parallel, rectangles may become parallelograms". Boutoura and Livieratos (2006) provide useful and detailed information about the effects of many different spatial transformations in an historical map context.

Local approach

To go further than the classical georeferencing process, noting that the georeferencing process is generally based on a global transformation without consideration of local distortions, we propose a further solution of local spatial adjustment.
Control points can be used in a more refined way in order to improve the spatial adjustment process. Indeed, control points can theoretically be weighted. However in practice control points are commonly weighted identically even though this does not adequately reflect reality. This choice is explained by the fact that the weighting of control points is a complex task requiring very good

knowledge of the quality and accuracy of historical data. This required knowledge is generally gained using georeferencing: we therefore enter a vicious circle.
To partly solve this problem, a global spatial transformation is computed based on a set of control points built on top of different kinds of objects (e.g. road intersections, churches, etc.). Thanks to the computation of intersections between the different vector layers and the set of control points, it is possible to group the control points by categories (to follow our example, road, church, etc.).
Then by using the computed residual errors, we determine what group has the smallest residual errors and thus, order the groups of control points by quality, instead of by weighting them.
Based on the "best" control points, a global transformation can firstly be computed. This transformation is applied to all vector layers. Three solutions are now possible: a solution based on data matching as proposed by Sester et al. (2007), a sequential approach based on the control points or a combination of these two last approaches.
The data matching approach aims at adjusting source geometries in order to have the same geometries as the target objects. To be automated, this approach involves that the semantic correspondences between objects have to be known and then the types of objects must be identified. Consequently, if users do not create the correspondences, this approach could be difficult to implement.
The sequential approach consists of computing local transformations based on the control points. Using firstly the control points which provide the best accuracy to spatial transformations (following the approach of qualification of these points described above), local transformations can be computed and thus be applied to all the vector layers in the concerned local areas. Then the second group of control points in terms of quality enables to compute other local transformations. But contrary to the first step, these local transformations are applied to all the vector layers except the one linked to the first group of control points. Then we iterate the process until the last group of control points. This method can be quite inefficient if the number of control points is small. Indeed, some areas can not be concerned by any deformations, due to the fact that they do not contain control points.
Finally, the combination of these two approaches uses firstly the control points to link the objects. Therefore, links between objects are created and enable the data matching approach to be used.
These solutions are not in opposition. On the contrary, they can be complementary. Indeed, according to users needs, several georeferencing methods are possible, contrary to the raster georeferencing process which generally uses one way to georeference data for a given map.

Example: Integration of a Cassini map sheet into a recent database

To illustrate the approach described above, the goal here is to integrate the vectorized layers of the Cassini map sheet number 48 (covered area: Vezelay and Cosne) into a frame of reference, the BDTOPO®. First, let us introduce this data.

Data description

The Cassini map has been created following a demand by Louis XV to produce a precise map of France. Four generations of the Cassini family went on to produce the 180 maps covering all of France at a scale of 1:86,400. Edited from 1752 through 1815, these maps were more accurate and detailed than any previous maps. The Cassini map is made up of different themes such as roads, bridges, rivers, streams, towns, villages, water mills, wind mills, mountains, forests (with some

types of woods such as pine forests) and some types of land-cover (e.g. briar, moor, meadows or vineyards). Finally, the Cassini map constitutes the first main French toponymic survey.
The BDTOPO® is the topological component of the French geographical reference framework produced by the French National Mapping Agency. This database has a metric precision and is composed of many different themes such as road network, hydrographic network, train network, electrical network, buildings, administrative areas, activities areas, land-cover, toponyms, etc.

Vectorisation process of the Cassini Map

The vectorisation process has been done manually using the free Open Source GIS Quantum GIS[1]. The following themes have been vectorised:

- Road network: gravel roads with trees, unpaved roads, country roads, track roads,
- Hydrographic network: rivers, streams, canals, ponds,
- Religious buildings: churches, abbeys – monasteries and nunneries –, priory churches for monks, priory churches for nuns, commanderies,
- Isolated buildings: country houses, court houses, post offices, water mills, wooden wind mills, rock wind mills, guard rooms, towers, milking sheds,
- Towns and villages,
- Forest.

Toponyms have not been directly vectorised. Hence, the toponyms linked to a building, a town, a village or a forest have been "vectorised" as an attribute in the associated attribute information of these themes. The toponyms which have no direct relations with human activity have not yet been referenced in the data.
In terms of timing, the production of a vectorised data set fitting to a single Cassini map sheet (covering nearly 110 kilometers by 80 kilometers) takes between one month and one month and a half.
The results of this vectorisation process are illustrated below with an extract of the Cassini map sheet number 48. The first figure (see Fig. 4) shows the original map. The second one (see Fig. 5) shows the overlapping between the original map and the vectorised data. Finally, the third one (see Fig. 6) shows the vectorised data only. The following legend is used to represent the vectorised data (see Fig. 3):

Figure 3: Legend used to represent the vectorised data.

[1] http://www.qgis.org/ - Quantum GIS website

Figure 4: An extract of the Cassini map sheet number 48.

Figure 5: Overlapping between raster and vector data.

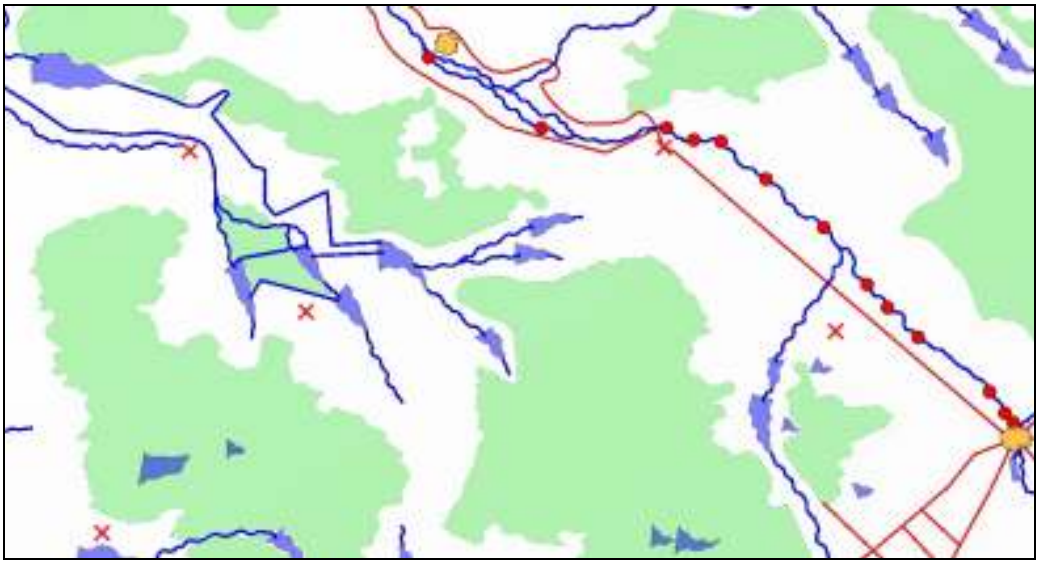

Figure 6: Vector data after the vectorisation process

Control points selection

The Cassini map has been built on top of a triangulation network of which the summits are the churches, priory churches and abbeys, or more precisely the top of the bell towers (Pelletier 1990). Consequently, a natural selection of control points is based on churches. In Cassini maps, different symbols are used to represent churches, priory churches and abbeys as illustrated below in Fig. 7. All these symbols have in common a white circle which represents the summit of each bell tower. Consequently, the source of the control point is defined as being at the center of that circle.

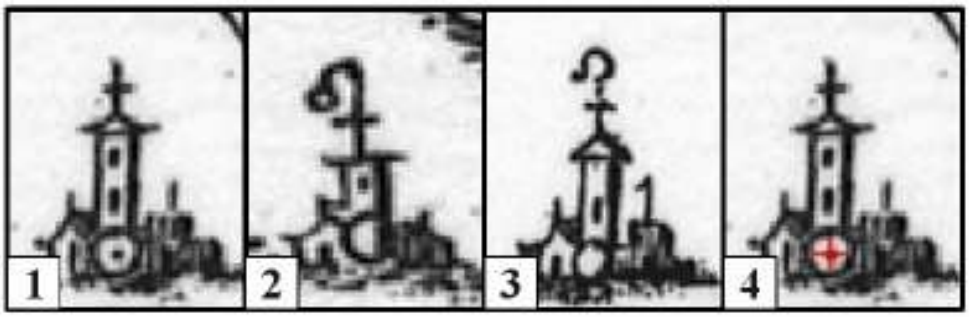

Figure 7: Different symbols for religious buildings – 1. Church, 2. Priory church, 3. Abbey – and the control point (4).

The equivalent of these religious buildings in BDTOPO® are the noteworthy buildings of which the "nature" is equal to "Church". These buildings have a polygonal geometry, not a punctual geometry. To fit the geometric type, centroids are automatically computed. However, this raises an accuracy problem as centroids do not exactly fit with the planimetric position of the bell tower summits.

Following this approach, nearly two hundred couples of control points have been created based on religious buildings, as outlined in the figure below (see Fig. 8). In addition, control points based on hydrographic and road networks are added, following the methodology described above.

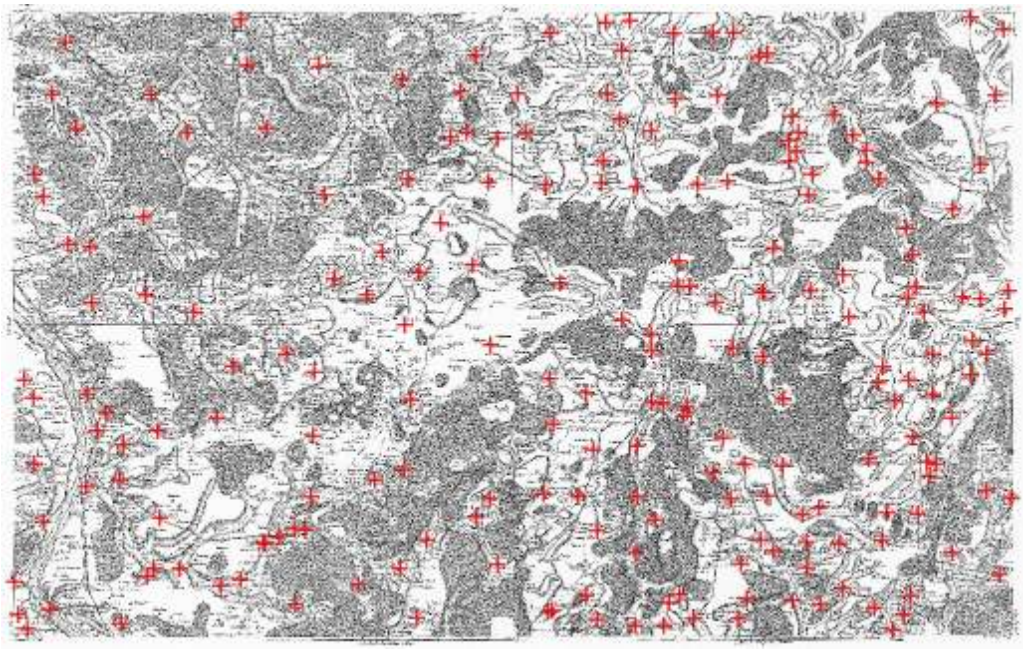

Figure 8: Control points based on main religious buildings.

Geoprocessing

The system tried different solutions to transform and to solve the transformation problem. Finally, the "most adapted" transformation found here is an affine transformation whose parameters are computed using the Structured Total Least Norm method (Felus 2006). The residual errors are from 24 meters to 350 meters, up to 350 meters in some rare cases, with a mean of 204,38 meters and a standard deviation of 112 meters. This transformation is applied to all vectorised data.
Moreover, some local adjustments are computed for some vectorised data using additional control points (based on hydrographic and road networks) or following some rules such as "a water mill is close to a river".

Results

The first result is that there is a real improvement, in terms of visualization, in comparing historical and current vector data. Indeed, in the case of raster data, it is often quite difficult to compare one specific theme. Using only vector data enables to only overlap a single historical theme with a single historical theme of each data (historical and current) and even to adapt the current data to the historical one. Fig. 9 shows a comparison example between the Cassini road network and major BDTOPO® roads. A major road is defined here as a road with a width superior to 5 meters. It is quick and easy to see that historical roads fit pretty well to the current major roads. Another example is given in Fig. 10 with the comparison between Cassini hydrographic network – rivers, canals and ponds – and BDTOPO® hydrographic network.
The second result is directly based on our approach, using here a global and a local transformation. In the example given in Fig. 11, we compare the Cassini water mills and BDTOPO® water mills. The global transformation computed above is used to globally georeference the Cassini water mills. Then a local transformation is applied based on the rule "a water mill is close to a river". Cassini water mills are thus adjusted to fit with the hydrographic network of the BDTOPO®. As a result, it is visible that there were three times more water mills in the studied area than today. Finally, using this process and more recent maps than those of Cassini, it would perhaps be possible to improve the current database with the adding of old water mills (in ruins or not).

Figure 9: Comparison between Cassini road network (in black) and major BDTOPO® roads (in orange and red).

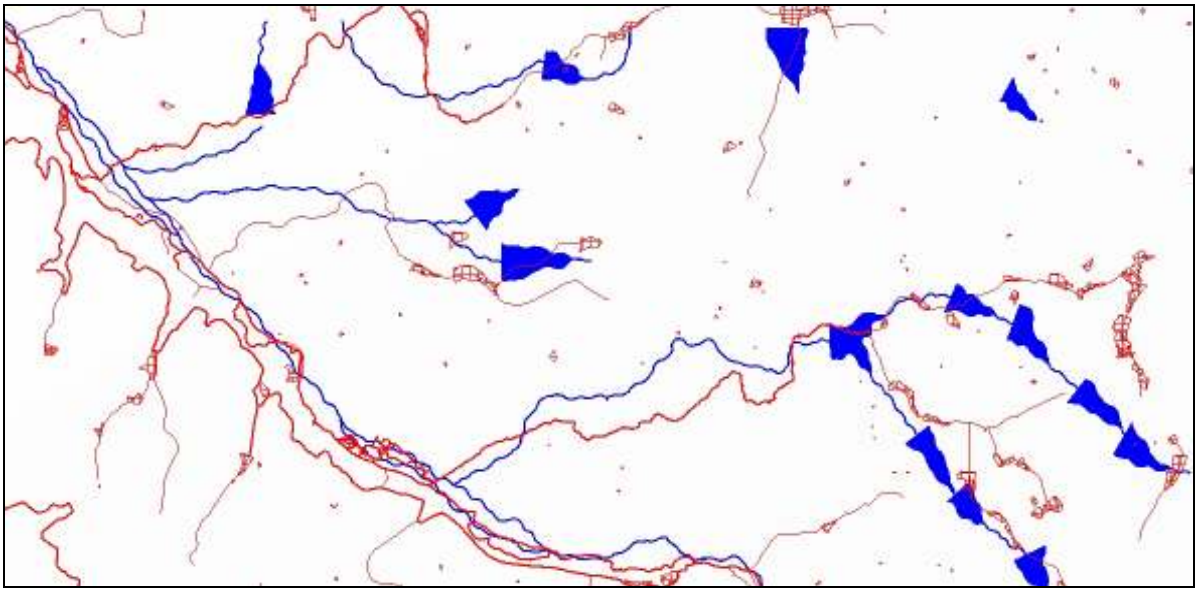

Figure 10: Comparison between Cassini hydrographic network – rivers, canals and ponds – (in blue) and BDTOPO® hydrographic network (in orange and red).

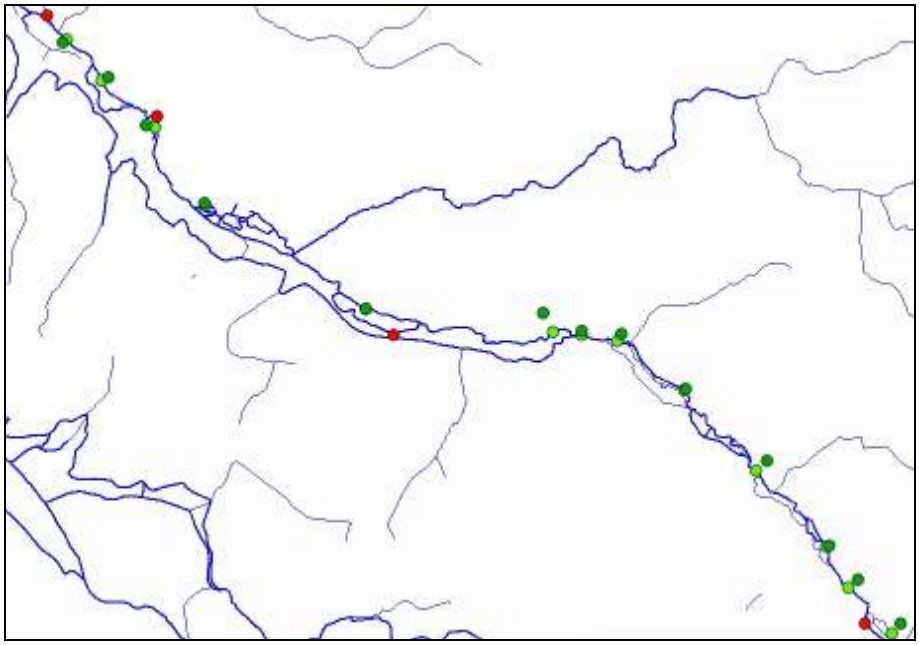

Figure 11: Comparison between Cassini water mills (water mills after global spatial transformation in dark green, water mills after global and local transformations (in green) and BDTOPO® water windmills (in red).

Conclusion

This paper introduced a method to deal with the integration of historical data into a current reference framework, based on a georeferencing process. This process is flexible and can be seen both as a global spatial transformation and as a combination of a global and local spatial transformations. This process has been illustrated through the example of the integration of a Cassini map sheet into the topological component of the French geographical reference framework. However, this method is currently partially automated. Therefore, the goal is to fully automate the process,

to design this process through a web services architecture, which will enable the provision of an online application through a web interface.

References

Boutoura, C. and E. Livieratos (2006). Some fundamentals for the study of the geometry of early maps by comparative methods e-Perimetron 1 (1): 60-70. Viewed January 2010, http://www.e-perimetron.org/Vol_1_1/Boutoura_Livieratos/1_1_Boutoura_Livieratos.pdf

Costa, L. and S. Robert (2009). Guide de lecture des cartes anciennes. Editions errance, Paris.

Dupouey, J.-L., J. Bachacou, R. Cosserat, S. Aberdam, D. Vallauri, G. Chappart and M.-A. Corvisier De Villèle (2007). Vers la réalisation d'une carte géoréférencée des forêts anciennes de France. Revue du Comité Français de Cartographie (CFC), Vol. 191, 85-98.

Felus, Y. A. (2006). On Linear Transformations of Spatial Data Using the Structured Total Least Norm Principle. In: Cartography and Geographic Information Science, Vol. 33 (3), 195-205.

Institut Géographique National (1950). La Nouvelle Carte de France au 20:000 - son utilité, son exécution -. Institut Géographique National, Ministère des travaux publics, des transports et du tourisme.

Jenny, B., W. Adrian and H. Lorenz (2007). Visualising the Planimetric Accuracy of Historical Maps with MapAnalyst. Cartographica, 42-1, 89-94.

Langlois, P. (1994). Une transformation élastique du plan basée sur un modèle d'interaction spatiale, applications en géomatique. Journées de la Recherche sur les SIG, GDR 1041 Cassini, INSA Lyon. 241-250.

Noizet, H. (2009). Les plans d'îlots Vasserot, support d'un système de l'information géographique historique de Paris. In: EAV, La revue de l'école nationale supérieure d'architecture de Versailles, 14, 86-95.

Parent, C. and S. Spaccapietra (2000). Database Integration: The Key to Data Interoperability. In: Papazoglou M., Spaccapietra S. and Tari Z. (eds), the MIT Press.,Advances in Object-Oriented Data Modeling, MIT Press, 221-253.

Pelletier, M. (1990). La carte de Cassini, l'extraordinaire aventure de la carte de France. Presses de l'école nationale des Ponts et Chaussées.

Pridal, P. and P. Zabicka (2008). Tiles as an approach to on-line publishing of scanned old maps, vedute and other historical documents. In: e-Perimetron, Vol. 3 (1), ISSN: 1790-3769, 10-21.

Sanderson, E. W. and M. Brown (2007). Mannahatta: An ecological first look at the Manhattan landscape prior to Henry Hudson. Northeastern Naturalist, vol. 14, 4, 545-570.

Sester, M., G. von Gösseln and B. Kieler (2007). Identification and adjustment of corresponding objects in data sets of different origin. In: Proceedings of the 10th AGILE Conference, Aalborg, Denmark.

Sheth, A. and J. Larson (1990). Federated database systems for managing distributed, heterogeneous and autonomous databases. In: ACM Computing Surveys (22:3), 183-236.

White, M.S. and P. Griffin (1985). Piecewise linear rubber-sheet map transformations. The American Cartographer, vol. 12 (3), 123-131.