From "Free Information" to Its (Geo)referencing and Analysis: The "Costs" of Open Source

# From "Free Information" to Its (Geo)referencing and Analisys: The 'Costs" of Open Source

TAMARA BELLONE, FRANCESCO FIERMONTE, CHIARA PORPORATO*

SUMMARY: *1. Introduction – 2. From Data to the Geo-referred Information – 3. Geocoding – 4. Geo-Data Mining – 5. Geographic Knowledge Discovery and Participatory Mapping – 6. Conclusion*

## 1. INTRODUCTION

In principle, within a "participatory decision-making"[1] all actors into play can "collaborate" because they share – preferably jointly – the information required to properly manage data flows and processes called for the contingent situation. To ensure to "communicate", students should also use the same codes for the same conventional signs, which are coded and universally recognizable, also using identical or at least easily recognizable "instruments" for a "standardized format" – through simple computing processes.

Mapping – also in its ultimate form, the digital one – is not an exception to the "rules" described above. The signs that appear on a paper must be clearly understood and shared to ensure an absolutely immediate reading without misunderstandings. Even before caring for the "visual" aspect it is essential to step back. It is essential, therefore, to find and use "official" data sets widely meta-documented and fit for purpose. In this context, during recent years in Europe, several projects and directives have been involved to specify and formalize not just the "representation" – a topic that remains open and of not immediate solution – but the methodology required to "share" catalogs of data on a large scale, in order to:

(a) tackling unnecessary and harmful phenomena of duplication - which also generate problems of "certification" and "official nature" of sources, among other things;

* T. Bellone is associate professor at Technical University of Turin (Politecnico di Torino). She is active in the field of Data Processing in Geomatics; F. Fiermonte works in the District of Architecture (Politecnico di Torino) at LARTU - Laboratory of Territorial and Urban Research; C. Porporato has a Ph. D. in "Drafting and Surveying for the Protection on the Territorial and Building Heritage" at the Politecnico di Torino.

[1] T. BELLONE, A. CITTADINO, F. FIERMONTE, *Participatory Mapping. Information Sharing on the Web*, in "Cartographic Challenges. Movement, Participation, Risk", Bergamo, 2009.

(b) promoting processes of "derivation" and sharing, for example, for a given scale, a certain amount of data layers should not exist to represent the administrative units (country, regions, provinces and municipalities in the Italian case), but only one from which you can reach the others, even through the information sharing through the display of services such as WMS - Web Map Service, WFS - Web Feature Service, WCS - Web Coverage Service and so on, officially recognized and certificated;
(c) ensuring the free use of the information, encouraging public participation and access to justice (in court proceedings in relation to the environment)[2].

For example, we can mention the INSPIRE project[3] and the SEIS[4] Directive, to which we refer for completeness.

The problem of the availability of "official data" should be resolved or at least the path taken takes out to be correct. As far as this subject is concerned, we highlight the efforts taken up by the *Regione Piemonte* with the project *Dati Piemonte* (public data are of everyone) that provides a valuable informative wealth. However, the sources not always fully satisfy the requirements of analysis and research. Increasingly, in fact, we ourselves need to create geo-referred/geo-related data also starting from textual information available online. It is known that any set of information, with full address, can be "transferred" to a map description crossing road databases with the address books of interest. The process is called "indirect georeferencing". However, how easy is to deal with it? What are the problems and the "costs" that one must endured in "open source" environments and "open source" tools? This paper will attempt to analyze the problems inherent in such operations, highlighting the key points and the critical issues identified.

## 2. From Data to the Geo-referred Information

To look up information it may be sufficient to start from a search engine (for example Google, Bing or other) or, even better, making reference to

[2] Aarhus Convention, *Citizens Right to Access Environmental Information*, http://ec.europa.eu/environment/seis/citizens_rights.htm.

[3] Acronym that stands for Infrastructure for Spatial Information in Europe. See European Community, *SEIS - Shared Environmental Information System*, http://ec.europa.eu/environment/seis.

[4] Acronym that stands for Shared Environmental Information System.

official records. In the first analysis all the feedback in PDF format can be left out, especially if protected by a password, as it avoids making copies of records to be processed. Therefore, if possible, it is better to use directly the accessible information published in spreadsheets (OpenOffice Calc), databases (OpenOffice Base) or text files, preferably the structured ones (CSV - Comma Separated Value).

About the previously cited shared formats, we found the ability to download – in addition to the already mentioned georeferred data – CSV information (for example, among others, the "List of shopping malls") on the Dati.Piemonte.it site[5]. In order to correctly decode such information, however, additional software tools such as, for example, Google Refine[6] are required. If this is not enough, it is possible to use a simple text editor (with a good computer skill), procedures that is clearly described on the Dati.Piemonte site. Similar situations, although they require free add-in type, however, extend the pre-processing time and lead to higher costs (man/days) concerning the georeferencing process.

## 3. GEOCODING

The information, in order to be properly managed and represented in its correct component spatial reference, shall be added to addresses, preferably if it is normalized. As it is known, a hypothetical toponym, such as "Verdi", should be correctly ascribed to get a truthful control for localization. Giuseppe Verdi, the composer, is obviously different from Mario Verdi, hypothetical hero of the Italian Risorgimento. After having been normalized, namings should be divided, if necessary, in basic components: street, preposition, street name, area, etc. in order to allow a "(semi-)automatic process" from instruments of "massive geocoding" such as Batch Geocoder[7]. This online tool, which is "free for limited numbers of records", returns the coordinates (WGS84 geocentrics), expressed in latitude-longitude, beginning from the addresses and from their civic numbering. We are facing, even in this case, with processing simplifications. The coordinates, in fact, are returned by interpolating the house numbering on single road arcs – between intersection and intersection – and with a range of house numbers on the left and right (left-right/from-to). This is obviously a process which, in ad-

[5] REGIONE PIEMONTE, *Dati Piemonte*, http://www.dati.piemonte.it/.

[6] See https://code.google.com/p/google-refine/.

[7] See http://www.BatchGeocode.com.

dition to all the typical limitations of the starting databases (update, reliability, quality, length of road between two intersections, etc...) also suffers from inherent problems of "understanding", such as accented letters or abbreviations of place names, that, if not managed properly to the source, may render ineffective a large part of the work.

Broadly speaking, any GIS - Geographic Information System type software application "is able", starting from the coordinates of the points just obtained, to return graphically the "correct" location of the points that appear then to be geo-referenced in all respects. Unfortunately the situation is not optimal as it seems and does not allow to be certain of the results because a lot of the steps are performed by hand or using procedures offered sight unseen and with no means to intercept and handle the error. On a limited number of recognizable objects (by their area, shape or size) we can verify the correct correspondence using a certified document (such as "technical map" or "orthophotomap"). However, it is a question of long and not always feasible checks. A school building, for example, could not be so "evident" or "recognizable" onto the map or seen from above. Generally, the unidentified locations are then assigned to "coordinates zero" and therefore it is necessary to investigate and to work out adequately what is left "pending", if possible.

## 4. GEO-DATA MINING

Why is it so important to think about DM - Data Mining[8], especially referring to the "geographic information"? Maybe, just for a simple consideration: the amount and the volume of (geographic) data are grown a lot in the last decade and they are increasing day after day in a massive way[9]. No possibilities, therefore, exist to use the "data" correctly without converting them into "structured information" or simple creating "information from

[8] For a description about the concept, see Wikipedia at http://en.wikipedia.org/wiki/Data_mining.

[9] "Geographic data collection devices linked to LATs - Location-Aware Technologies, such as the global positioning system, allow field researchers to collect unprecedented amounts of data. Other LATs such as cell phones, in-vehicle navigation systems and wireless Internet clients, can capture data on individual movement patterns", see J.P. WILSON, A.S. FOTHERINGHAM (eds.), *Geographic Data Mining and Knowledge Discovery*, in "Handbook of Geographic Information Science", Blackwell Publishing, 2008, http://www.geog.utah.edu/ ~ hmiller/papers/Handbook_GIS.pdf.

data"[10]. For this reason, if thinking about a specific "geographic location" (and all that is close to it) is relevant, considering its historical background is also strategic. Such analisys can measure or explain the land consumption, or simply show where should be the "optimal place" for a new kindergarten. So, what is DM? We can answer in a simple way: DM is not a Google (Advanced) Search. DM is, instead, the process of analyzing data to identify "patterns" or "relationships". To do this task, it is important to use almost two softwares, a spreadsheet and a statistical analysis package, for example OpenOffice Calc and Orange[11] to "join" some of these data together. A GIS software (an open source one is better, of course) is also required to perform all the (geo)spatial analisys of interest (Quantum GIS, for example). Data mining is also known as KDD - Knowledge Discover in Databases[12]. And knowledge means "all that is interesting" and not known *a priori*.

How many "data" exists? How many of them is it possible to convert into "information", much better if reusable and upgradable in time? From KDD to GKD - Geographic Knowledge Discovery, and its core sector, the GDM - Geographic Data Mining, the way is not so easy. In fact, we have to notice that "Spatial" concerns any phenomenon where the objects can be embedded within some "formal space" that generates implicit relationships among the objects[13]. So, the matter is: "how discover knowledge from databases"? How use data mining to create and manage "information"? Using a sentence by Wilson and Fotheringham, "data mining involves the application of techniques for distilling data into information or facts implied by the data" if we have to:

- understand the world that surrounds us (according to literature, we can call it the "background knowledge");

[10] K.-H. ANDERS, *Data Mining for Automated GIS Data Collection*, in Fritsch D., Spiller R. (eds.), "Photogrammetric Week '01", Wichmann Verlag, Heidelberg, 2001, pp. 263-272.

[11] See http://orange.biolab.si/ and T. CURK, J. DEMŠAR, Q. XU, G. LEBAN, U. PETROVIČ, I. BRATKO, G. SHAULSKY, B. ZUPAN, *Microarray data mining with visual programming*, in "Bioinformatics", vol. 21, 2005, n. 3, pp. 396-398.

[12] M. ESTER, H.-P. KRIEGEL, J. SANDER, *Spatial Data Mining: A Database Approach*, in "Proceedings of the 5th International Symposium on Large Spatial Databases - SSD '97", Berlin, Springer-Verlag, 1997, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.18.4661&rep=rep1&type=pdf; U.M. FAYYAD, G. PIATETSKY-SHAPIRO, P. SMYTH, R. UTHURUSAMY, *Advances in Knowledge Discovery and Data Mining*, Menlo Park, AAAI/MIT Press, 1996, pp. 1-34, http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf.

[13] J.P. WILSON, A.S. FOTHERINGHAM (eds.), *op. cit.*

- verify, organize[14] and clean the data (that always must be followed by their own metadocumentation) before starting any operations and analisys;
- according to our need, choose the appropriate "data mining type" to perform the analysis or the research (such as the correct classes, associations, rules, clusters, outliers and trends discussed in more detail below).
- "build (and consolidate) the Knowledge" using all the input data (and their spatial relationship).

Of course, the KDD process is not a linear one, with a sequenced starting and ending point: "analyst will re-sequence and even revisit steps based on the sought knowledge and the nature of the information uncovered within the process"[15].

Data mining should reveal, among the others, the patterns of interest and to perform this we have to consider several steps to "organize" the input. So, broadly speaking, we have to classify the data, analysing their association and make prediction (if possible) about their correlations and effects, perform "cluster analysis"[16] and "outlier analysis"[17].

## 5. Geographic Knowledge Discovery and Participatory Mapping[19]

GDK - Geographic Knowledge Discovery is the process of extracting information and knowledge from massive geo-referenced databases[20]. How-

[14] Geographic data must share the same GCS.

[15] J.P. Wilson, A.S. Fotheringham (eds.), *op. cit.*

[16] Techniques for classifying data objects into similar groups.

[17] Outliers are data objects that appear inconsistent with respect to the remainder of the database[18]. While in many cases these can be anomalies or noise, sometimes these represent rare or unusual events to be investigated further. For example, outlier analysis has been used in detecting credit fraud, determining voting irregularities or severe weather prediction. See S. Shekhar, P. Zhang, Y. Huang, R. Vatsavai, *Trends in spatial data mining*, in Kargupta H., Joshi A., Sivakumar K., Yesha Y. (eds.), "Data Mining: Next Generation: Challenges and Future Directions", AAAI/MIT Press, 2003.

[19] M.F. Goodchild, *Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0*, in "International Journal of Spatial Data Infrastructures Research", vol. 2, 2007, pp. 24-32, http://www.geoinformatics.cn/wp-content/uploads/citizensasvoluntarysensors.pdf.

[20] SDWs - Spatial Data Warehouses are data warehouses that also include both "spatial and aspatial data", J.P. Wilson, A.S. Fotheringham (eds.), *op. cit.*

ever, to do this (for the nature of geographic entities, relationships and data) the "standard of the KDD techniques is not sufficient" to explain the complexity of real world and to reflect it into related schemes to build structured information[21]. We have to think also at the spatial relationship and at proximity rules. These can be defined and understood using terms such as location (where is it?), distance (how far is it?), direction (on which direction?) and/or topology to manage data integrity, shared geometry, adjacent and connected features. Last, but not least, we are compelled to think about the relations among data, spatial relationship and timeline series.

The goal is to arrive into a new geographic visualization, that someone calls GVis - Geographic Visualization, that is "the integration of scientific visualization with traditional cartography"[22].

Probably, participatory mapping is not the last, neither the least, nor the first argument of this short paper but, without raising and sharing the "information", it could be not even start. The accessibility to the official data, such as the lists of street maps (certified toponyms), for examples, and procedures apt to the verification and standardization of the addresses, restricting the preliminary operations of cleaning and recognition, would allow to obtain, in less time, a better quality output. The use of "open source" instruments, of course, limiting the costs of basic software, allows to share instruments and resources and to invest in training on alternative procedures and on freely available tools.

In the last years, in spite of all the efforts made for greater sharing and participation, from a point of view of the ordinary user, the view has not substantially changed.

"Data sharing" (or "participatory mapping" as you prefer) in order to function, needs (at least) of three levels + 1 of interoperability[23]

- technical (communication through shared "interfaces");
- semantic (standards related to content and meta data documentation)

[21] For example, we could think about "the complexity of spatial objects and relationships as well as their transformations over time, the heterogeneous and sometimes ill-structured nature of geo-referenced data, and the nature of geographic knowledge" J.P. WILSON, A.S. FOTHERINGHAM (eds.), *op. cit.*; see also S. SHEKHAR, P. ZHANG, Y. HUANG, R. VATSAVAI, *op. cit.*

[22] J.P. WILSON, A.S. FOTHERINGHAM (eds.), *op. cit.*

[23] M. SALVEMINI, *La direttiva INSPIRE: punti fermi, priorità, impegni per i governi nazionali*, 2004, http://151.100.2.84/wordpress/wp-content/uploads/2008/02/salvemini-la-direttiva-inspireasita-2004.pdf.

- institutional (collaboration and sharing aimed at overcoming the institutional barriers);
- training (activities designed to increase the geographical culture);

In this latter respect, the extension of ECDL[24] certification in GIS environment (Geodesy, Cartography, commercial or open source software and tools) and the strengthening of training offer, that such certification requires, are extremely important.

## 6. Conclusion

In conclusion, to facilitate the consolidation of a "participatory mapping", leading to the birth of an "ACTIVEsharing" (in Italian "partecipAttiva", Active_and_Participatory) mapping, it is highly desirable to succeed in:

- "standardizing" the mapping component in the access pages, for example by sharing the same prefix or suffix;
- making cartographic inventories, related resources and applications running on the Internet "more visible" or "not hidden" to search engines[25];
- publishing lists of WebGIS, WMS, etc. services, taking care of including the main functionalities of each one (e.g., downloading data: yes/no);
- increasing availability of downloadable datasets (including meta documentation) locally;
- guiding the user to a "conscious" use of data, even through the use of distance learning instruments of e-Learning (e.g., Moodle);
- promoting and supporting ECDL GIS certification;
- allowing users to have the possibility to consult the Web pages and the services offered by it in the language of the state of the country of belonging as well as in English.

[24] The european GIS certification (Geographic Information System) is a program designed to demonstrate the professional knowledge on the use of GIS systems and of their main components and functions. It is addressed to all those who are called, in their professional area, to work with systems (GIS) that correlate the phenomena or variables to the territorial dimension, creating thematic maps and analytical reports in various formats, http://www.ecdlgis.it.

[25] Appropriate bookmarks drowned in the code (meta tags) make resource easily recognizable.

We want to conclude our paper with an Italian example about a possible use of georeferencing. An Italian research team from the Politecnico of Torino is purposing to georeference the information for teaching[26]. This new approach could increase the appeal of some scientific disciplines. Moreover, it could be a user-friendly tool, for the new students generation, the BIT generation, in addition to the traditional ones commonly used for studying. In particular, the group realized two study cases: about the biography of Sir Isaac Newton and Ernesto Schiapparelli. The information about these two scientists, using a timeline, was georeferenced on maps and visualized on Google Earth. Going into details, these researchers developed a tool that is able not only to geocode the information, translating it into KML format (and view it into Google Earth) but also to create an html page with CSS style language[27] file in order to browse all the data using a compatible hypertext client software.

[26] A.C. SPARAVIGNA, R. MARAZZATO, *Georeferenced Lives*, 2012, http://arxiv.org/abs/1203.0500v1.

[27] See http://en.wikipedia.org/wiki/Cascading_Style_Sheets.