Mapping heterogeneous data: a case study on the French Green Infrastructure

Cécile Duchêne[1], Sébastien Mustière[1], Sandrine Gomes[1], Mathilde Kremp[1], Lucille Billon[2] and Romain Sordello[2]

1. Univ. Paris-Est, LASTIG COGIT, IGN, ENSG, F-94160 Saint-Mande, France; sandrine.gomes[at]ensg.eu, mathilde.kremp[at]ensg.eu, cecile.duchene[at]ign.fr, sebastien.mustiere[at]ign.fr

2. UMS 2006 Patrimoine Naturel, Muséum National d'Histoire Naturelle, F - 75005 Paris, France; lbillon[at]mnhn.fr , sordello[at]mnhn.fr

Abstract: To achieve a preservation and restoration of ecosystems, public environmental policies at the international level are fostering the implementation of Green Infrastructures, i.e. networks composed of areas where animal and vegetal species can live (habitat patches), and corridors to circulate between them. In France, defining existing habitat patches and corridors was ensured in a distributed manner by the Regions, the first subnational administrative level, with flexible guidelines. It resulted in very heterogeneous data in terms of level of detail, raising the question: "How to map such heterogeneous data at a supra-regional level, making them understandable while respecting the work of Regions, and with a reasonable amount of human work?". Our study focuses on habitat patches of two adjacent Regions. After making a "rough" map directly from the provided data, we explore three ways for homogenizing the map. The first method consists in generalizing the more detailed data using simple morphologic operators. The second method consists in graphically refining the less detailed data by filling the areas with a pattern taken from the more detailed data. The third method consists in drastically changing the level of abstraction of the data on both regions, while rasterizing the space. Although it would be necessary to test the resulting maps on potential users, we think the third approach is probably the only one usable.

Keywords: heterogeneity, cartography, harmonization, generalization, land use, green infrastructure

1. Introduction

The impact of human activities and infrastructures on biodiversity has become a major worldwide political concern that led to the adoption of mitigation strategies at multiple political levels – e.g. at international level the Convention on Biological Diversity (United Nations 1992) and its on-going 2011-2020 Strategic plan known as "Aichi biodiversity targets", or at European level the subsequent European Union Biodiversity Strategy to 2020 (European Commission 2011). One of the known impacts of human activity on biodiversity is *land fragmentation*: because of urban sprawl and transport infrastructures, the natural habitat of animal or vegetal species is fragmented into disconnected natural habitat patches, which prevents dissemination and genetic mixing within the species and therefore threatens their survival (see e.g. Garcia-Gonzalez *et al.* 2012, Klar *et al.* 2012). One recognized way of mitigating this phenomenon is to promote so-called *Green Infrastructures* (CBD 2017-Aichi target 15; European Commission 2013). In a nutshell, a Green Infrastructure is a network of species-friendly areas that, if dense enough, enables species to live and disseminate (Naumann *et al*. 2011, p. 1).

In line with these initiatives at international level, in 2010 the French law established the implementation of a French Green Infrastructure that has to be taken into account in urban and rural planning. This infrastructure is named "*Trame verte et Bleue*", which can literally be translated as "Green and Blue Infrastructure", green being for the ground part of the infrastructure and blue for its aquatic part. A first task to implement this decision was to build a database of existing natural or semi-natural areas to be included in the French Green Infrastructure. The responsibility of modelling and collecting the data was committed to the French Regions (first administrative subdivision of

France), both for geographical and political reasons: regions do have local specificities; and empowering the Regions with environmental planning, making this topic closer to citizens, was a choice, in line with international policies. A general framework for data modelling was established at the national level, stating that the French Green Infrastructure is composed of several subnetworks (for wet zones, woodland, littoral, etc., see Fig. 1a), and that each subnetwork is composed of "reservoirs of biodiversity" (habitat patches where species live), and corridors to circulate between them (see Fig. 1b). Regions were asked to collect data at a reference scale of 1:100k. Apart from these general guidelines, Regions were left free of their data modelling and their methods of data collection.

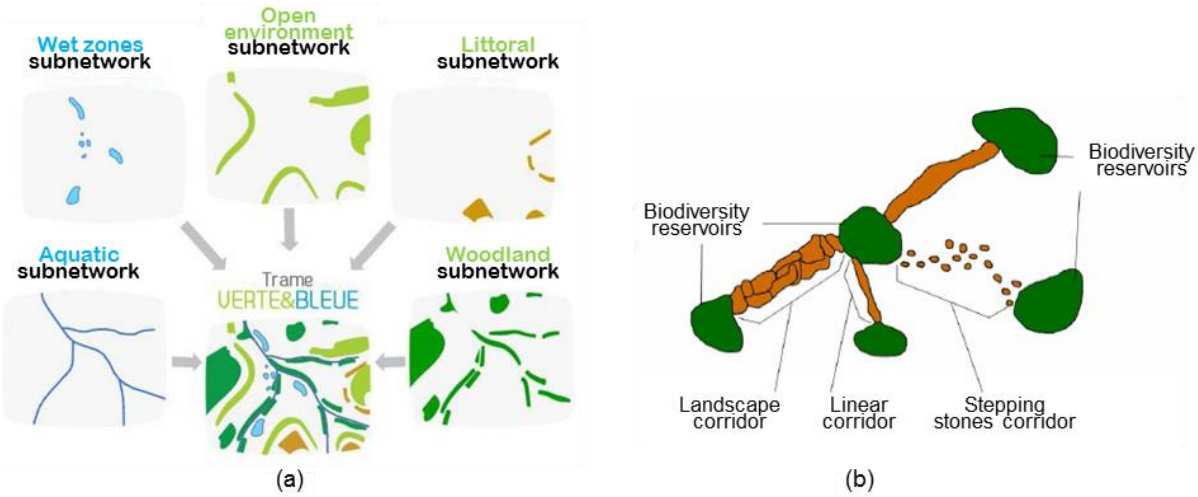

Fig. 1. (a) Green Infrastructure network and subnetworks, (b) Reservoirs of biodiversity (habitat patches), and corridors enabling species to circulate between them. Figure after Allag-Dhuisme *et al.* (2010) and SRCE Basse-Normandie (2014, p. 167).

The "Muséum National d'Histoire Naturelle" (MNHN) was mandated by the French Ministry of Environment to make a map at the national level from the data produced by the Regions. The first stage, prior to our work, consisted in standardizing the data produced by the French Regions into a common data schema. This was done by MNHN and CEREMA, a center of expertise of the Ministry of Environment (Covadis, 2014; Billon et al. 2016). The next stage, assembling these regional data, emphasized their heterogeneity, especially in terms of levels of detail (LOD). The following questions arised. How can we make a national map by assembling these very heterogeneous data? Which compromises should we look for between legibility, homogeneity (the map should be understood), and respect of the data collection work performed by the Regions (how much can they tolerate to see their data modified?)? Indeed, each Region has legally validated its data. The transformations we operate on the data for visual purpose do not intend to challenge them, but the Regions could be frightened that the resulting map eventually replaces them. This paper reports on an ongoing study conducted by the COGIT research team of IGN (the French NMA) and the MNHN that aims at exploring those questions. Until here, the study focused on two adjacent French Regions (Fig. 2 bottom-right), namely Poitou-Charentes (West) and Limousin (East), and only on reservoirs of biodiversity (corridors are not considered yet). Our aim is to analyze the data heterogeneity and the subsequent problems raised, and to explore approaches to harmonize the data in order to produce maps at coarser scales than the regional level. The target LOD is not strictly defined as an open idea is to produce a multiscale map. However, it is wished that the data are made homogeneous in order to facilitate their understanding by map readers. Note that this is needed not only at the national level. Indeed, the French administrative partitioning into Regions was redesigned in 2015 (after the data collection), resulting in 13 Regions instead of 22 for the European Territory of France: several regions were merged, including the two Regions on which our study focuses.

The issue of dealing with geographical data heterogeneity is not new and has been studied from several perspectives. A first perspective concerns data schema documentation and harmonization (see e.g. Mechouche et al. 2011 or the INSPIRE initiative). A second perspective concerns schema matching and, more particularly, special attention has been paid to the matching and conflation of land-use categorizations (e.g. Hagen-Zanker et al. 2005). A third

perspective focuses on aspects related to the harmonization of geometric LOD of the data, e.g. Stanislawski and Savino (2011) use generalization to harmonize the LOD of hydrographic networks collected at subnational level. Our case study is relatively similar, but the data are different (a set of, possibly overlapping, area patches), and the goal is definitely to create a map (i.e. a Digital Cartographic Model, Grünreich 1992) that gives the reader an idea of the spatial distribution of the Green Infrastructure on the ground, as opposed to a geographic database (Digital Landscape Model) where e.g. topology constraints could apply.

The rest of the paper is structured as follows. Section 2 briefly presents the test data and their heterogeneity. Section 3 reports on three approaches experimented to create a harmonized map, and discusses the obtained results. Finally, section 4 draws first conclusions and identifies perspectives to pursue the study.

2. Data and task difficulty analysis

The data are structured within a schema that contains three main classes of reservoirs of biodiversity, corresponding to three subnetworks of the Green Infrastructure: woodland, wetland and open environment[1] (plus littoral and rivers that have been ignored in this study). A direct mapping of the data is possible, resulting in the map of Fig. 2 (here displayed with a scale of 1:2.5M).

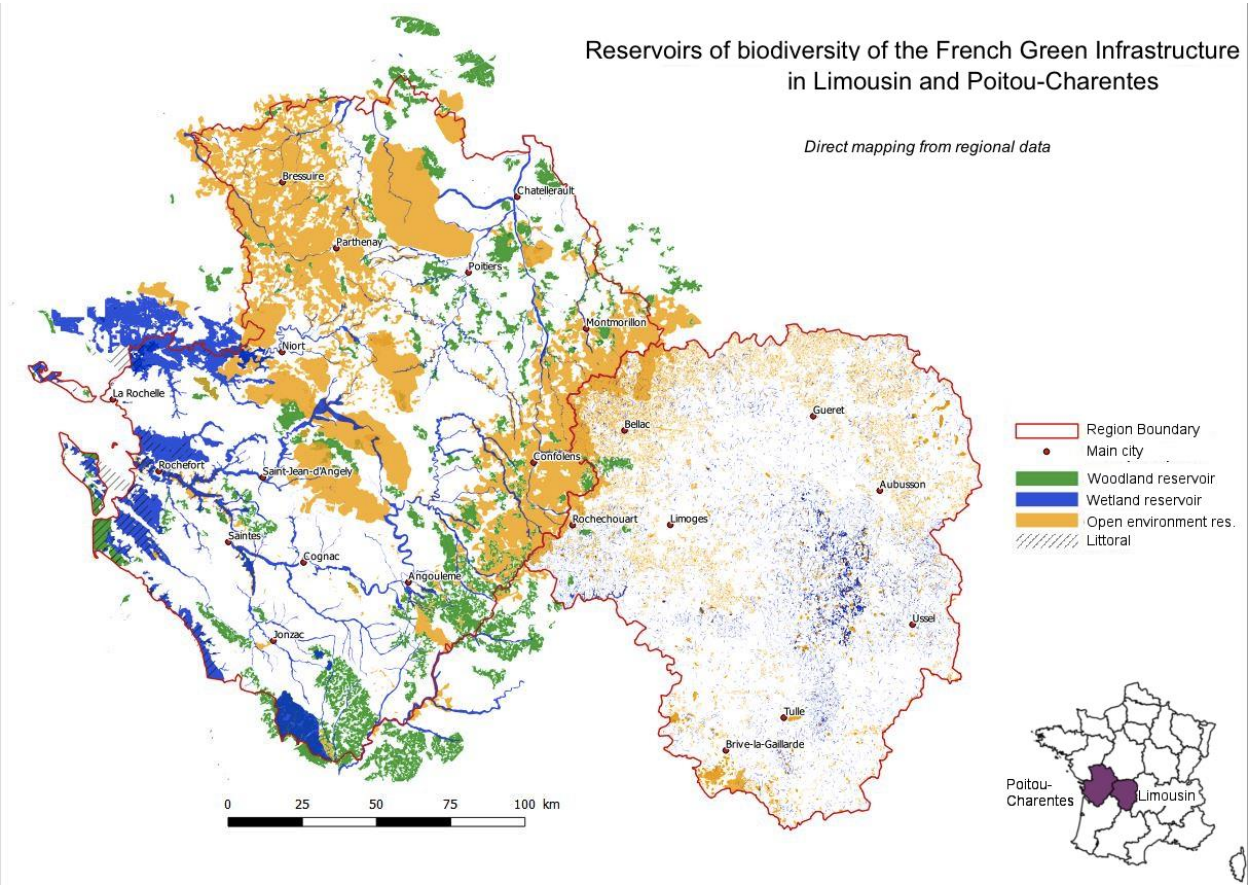

Fig. 2. Direct mapping of the reservoirs of the two regions

[1] Meadows, moorland, grass

At first sight, one may notice the differences of LOD, even resulting in a visual impression of different colors. Actually, the same colors are used for both Regions, but the reservoir layers of Limousin (East) are composed of much smaller areas than those of Poitou-Charentes (West), resulting in different local densities (see also Fig. 4b). This is due to different data capture processes, not to actual geographic differences. Limousin mainly worked on ground pixels of 50m. The cumulated area of different kinds of reservoirs are computed for each ground pixel, mainly from the land use layer of the 1m resolution vector topographic database of IGN (BD TOPO®). Pixels where the cumulated area of a given kind of reservoir is greater than a threshold are included in the corresponding reservoir layer. Poitou-Charentes worked with a mixture of ground pixels and data stemming from other sources (e.g. species surveys), but applied a morphological closure (dilatation + erosion) to the data to decrease their LOD. **LOD differences between the two regions are therefore artificial: this hypothesis was taken into account in our experimentations**.

3. Towards harmonized maps

In this section, we present three approaches that could be followed to harmonize the representations of the reservoirs at supra-regional level. We will illustrate them with simple experiments made on our two test Regions.

3.1 Generalization of the more detailed area

A first idea to create a homogeneous map is to generalize the more detailed region in order to reach the LOD of the less detailed region. In our case study, this means we aim to amalgamate sets of close and small areas into bigger areas. We follow a simple approach to do this: each reservoir layer is considered separately and a morpho-mathematical closure (dilation and erosion) is performed on it. The threshold is manually chosen to reach a visual similarity of LOD for the two regions (let us remind our hypothesis that the patterns in both Regions are actually similar). Resulting simplified layers are then superimposed.

The result of the approach is shown on Fig.3. We think this simple approach is efficient and sufficient in our case to reduce the LOD of the more detailed region and to harmonize the LOD. Graphically speaking, the result is globally legible. However, this approach of simplification may be improved. First, as the layers are handled separately with a closure, two regions belonging to two different layers can overlap after closure. Therefore the order chosen to draw the layers influences the result, as some layers may hide parts of others (Schylberg 1993 ; Harrie *et al.* 2009). Intuitively, the reservoirs with the smallest areas are better displayed on top of the others – at least in cases like ours, where we have zones in which one layer is dominant with small patches of other layers. Managing the drawing of smallest objects on top of the others was not done here due to a lack of time, which can be seen in the zone just right of the limit between the two Regions: the blue layer slightly appears by transparency but is mostly hidden by the yellow one. More tricky, in the case of very interlaced patterns (e.g. two layers initially composed of small patches that spatially alternate like white and black squares on a chess board), we can end up with one big region for each layer after closure. In such a case, using a morphological closure might not be a relevant solution and the use of typification operators (e.g. Sester 2008) could be investigated, in order to preserve the spatial patterns while decreasing the level of detail.

Besides, one limitation of the approach, if one intends to extend it to more data sources, is that this approach is only feasible if the goal is to reach the LOD of the less detailed data source: a single under-detailed source induces to create a globally under-detailed map.

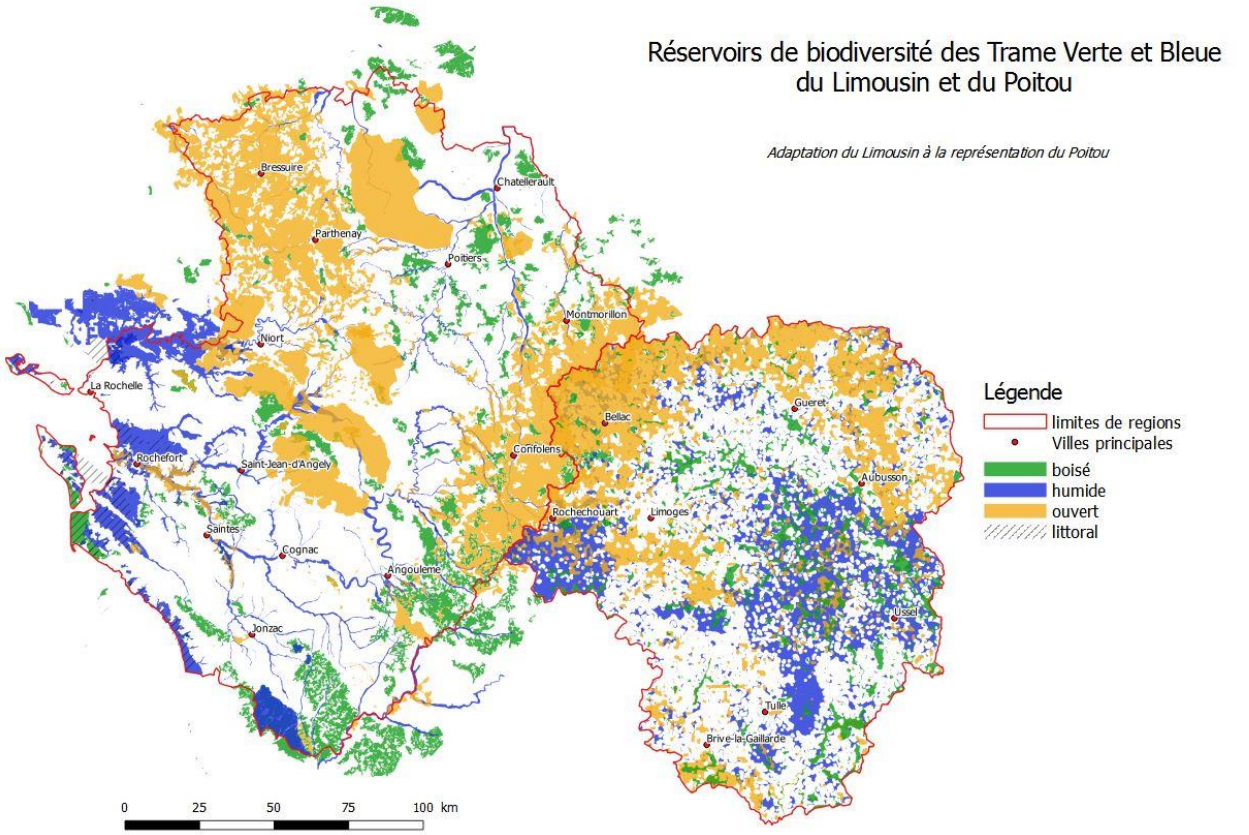

Fig. 3. Result of the "generalization" approach. Same map legend as in Fig. 2.

3.2 Graphic refinement of the less detailed area

A second approach is, oppositely, to change the representation of the less detailed region in order to graphically reach the LOD of the more detailed region. Of course, we do not aim here to "specialize" the data, i.e. to create some detailed data from less detailed sources. We only use ad hoc cartographic symbols to give the reader the impression that data have the same LOD everywhere. For this, the more detailed Region is displayed at target scale, and from this display, a small, typical rectangular image is extracted from each layer (see Fig. 4a). These images are then used as graphical patterns on the objects of the less detailed region – i.e. these objects are filled with these patterns. The result of the approach is shown on Fig. 5.

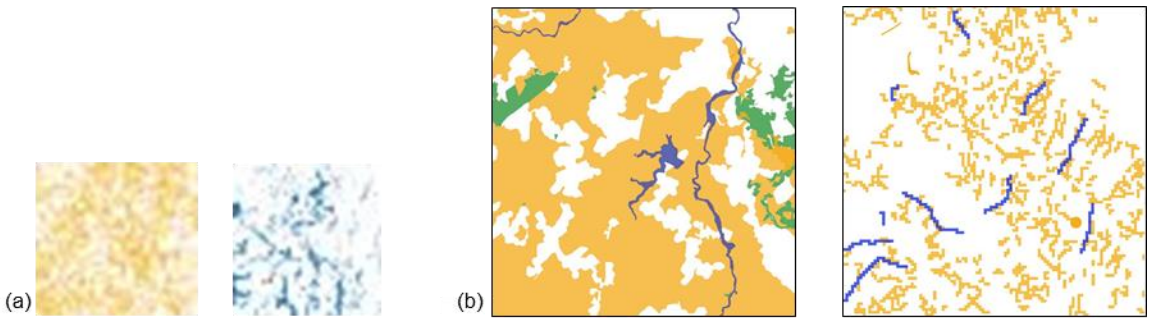

Fig. 4. (a) Some graphic patches extracted from the high resolution data to fill in the low resolution data; (b) Extracts of Poitou-Charentes (left) and Limousin (right) data on two zones of same size, magnified to show the LOD differences.

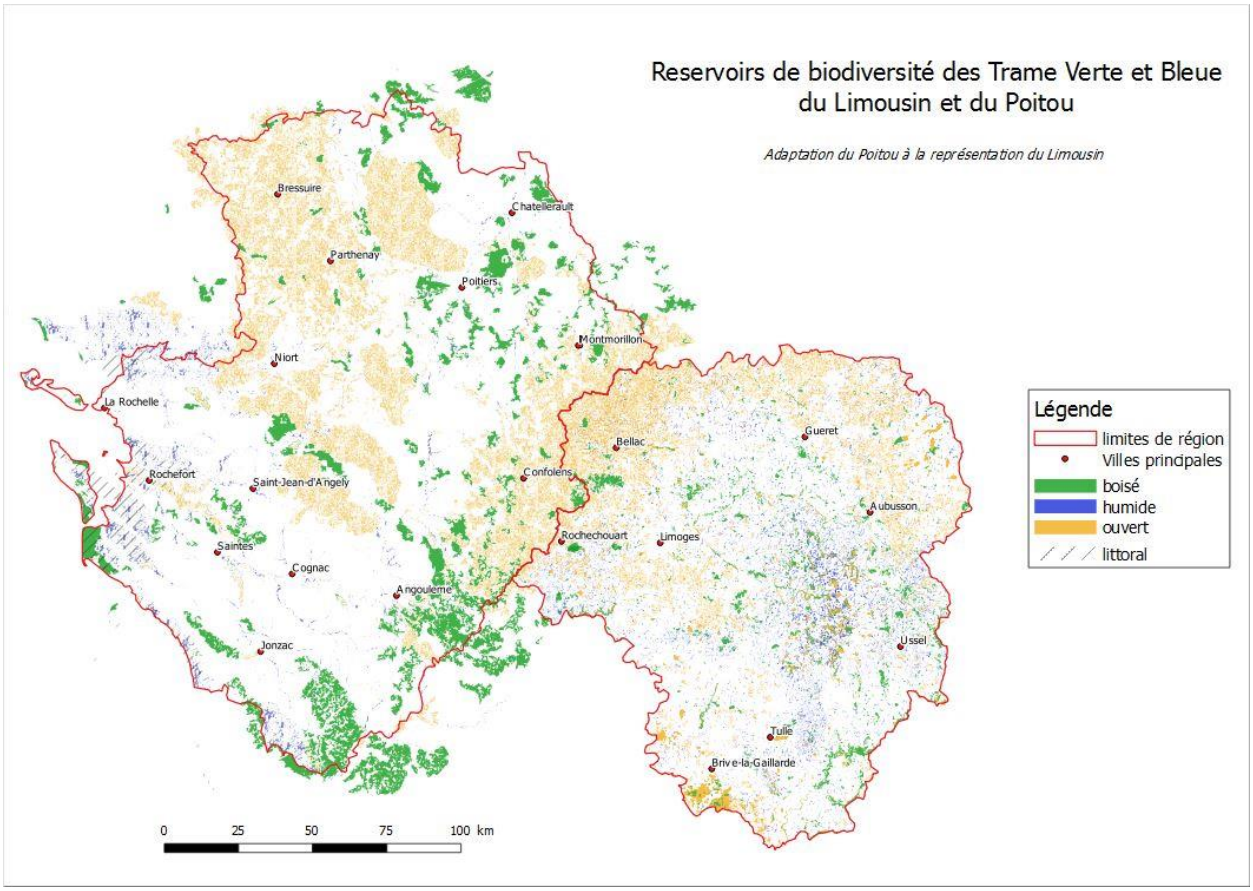

Fig. 5. Result of the "specialization" approach. Same map legend as in Fig. 2.

This is a very simple approach, not relying on any geometrical treatment. It nicely gives the impression of homogeneous data. This artificial and graphic improvement of the LOD seems efficient, even if the way readers interpret this map should be studied more in depth. However, this has a limited range of application. The target scale should not be too low, to ensure the readability of these detailed textures (for example, in our case, this is not very adapted to the blue layer, which almost disappears due to the small size of its elements). The target scale should not be too high either, to ensure that the reader does not see that the texture is artificial. It is however easy to extend this ap-

proach to a large number of regions that have to be integrated. It could also be mixed with the previous approach, if one aims at integrating over-detailed and under-detailed data sources, according to the target scale. To go further, some computer graphics approaches could be used to fill the areas with less regular textures than using everywhere the same graphic patches (Loi et al. 2013).

3.3 Abstraction by means of rasterization

A third approach consists in changing more drastically the type of representation: we do not aim anymore to make one region looking like the other one; instead, a third type of representation is defined for both regions. In this study we chose to rasterize the data, at a low resolution level. Different layers of vector data are thus transformed into a single layer of raster data. Concretely, the area is divided by means of a regular grid (size 5x5 km^2). Then each cell is associated to its majority subnetwork, with the following empirical rule: if the cumulated area of all subnetworks in a cell is less than 1/3 of the size of the cell, then the cell is considered "empty" (with no subnetwork); otherwise, the cell is considered filled with the subnetwork that has the highest cumulated area. The result of the approach is shown on Fig. 6.

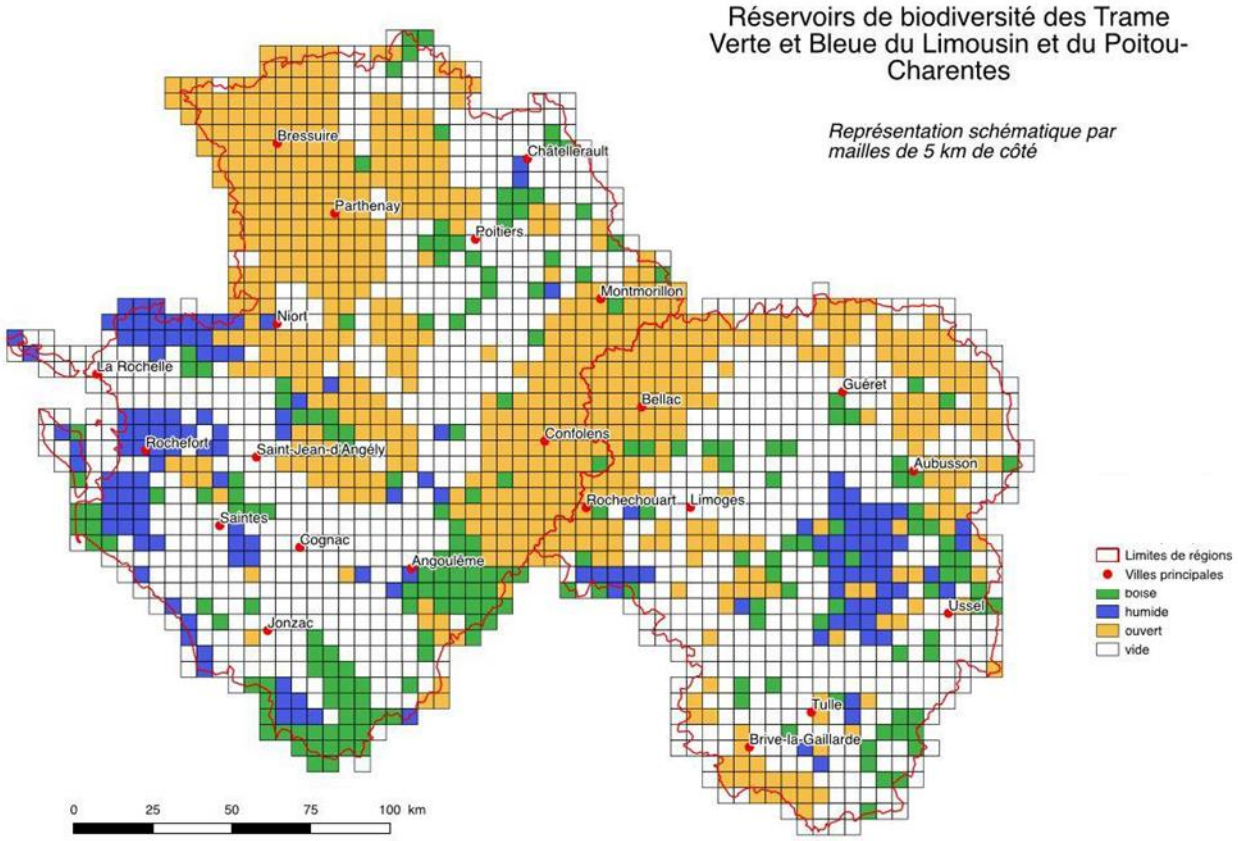

Fig. 6. Result of the "rasterisation" approach. Same map legend as in Fig. 2.

Like the previous one, the approach is relatively easy to implement and to extend to any number of data sources. It has the big advantage of limiting the risk of misunderstanding for the reader. Indeed, in the two previous approaches, the reader may not understand that one part of the data is "real data", while the other part is "artificial data" made for the sake of graphics considerations. In this "rasterization" approach, both data are modified and thus better understood as aggregated data, and not real data.

This approach is however not very well adapted if the different subnetworks have different characteristics (e.g. one subnetwork is composed of small areas and will never be the majority subnetwork) or in spatial configurations where the different subnetworks are interlaced (then the structure is lost). Another open question is how to determine relevant thresholds in this approach. The threshold used here (1/3) has been here chosen empirically, and is the same for all regions. It could be adapted to the different regions (detailed data cover less surface than simplified data).

4. Conclusions

Map generalization as well as map symbolization approaches may be used to build a harmonized map from heterogeneous data. We experimented that by means of three different approaches. These experiments, with simple tools, need to be improved now, for example with more sophisticated generalization operators to preserve geographical patterns. The resulting maps should also be evaluated and presented to map readers, in order to study what they understand from the different maps: does a map emphasize some reservoir layers, highlight or not the differences between Regions, etc.? This probably depends on the user and the context of use: a political decision maker, a kid at school, or an ecologist do not have the same needs of information on biodiversity reservoirs, neither the same interpretation of their spatial organization. Besides, our study raises two important issues.

The first issue is about the relevance of each presented approach. As already mentioned, the first two approaches may let the reader believe that the displayed data are directly the original data produced independently by the two administrative regions – which is false. Because of that, we believe that the last approach is globally more relevant. However, to implement it, we used a very basic approach (the notion of majority subnetwork in each cell), and we certainly lost a lot of information about patterns of the studied phenomena. This highlights the needs for more automated tools to produce advanced schematic maps from raw data. We need, for example, some tools to produce schematic representation like chorems (Brunet 1986, Reimer 2010). Actually, the main question behind might be that we miss a conceptual framework for levels of details for land use data.

The second issue is about the relevance of the idea of harmonizing data. Harmonizing the LOD may be (positively) thought of as hiding artificial differences between data sources with different specifications. It may also be (negatively) thought of as arbitrary hiding actual and meaningful differences between geographical configurations. In this study, we considered that the differences of level of detail were mainly artificial, i.e. the actual patterns were similar – which makes sense for the two considered Regions. But in general, this is not true. The main challenge that has to be tackled is then not to reach a complete harmonization of data, but a relevant one. Artificial, not relevant differences should be hidden while actual, relevant differences should be kept. A highly crucial and remaining issue is “What is relevant?”, however this the first question in cartography in general. More concretely speaking, this requires data mining tools to analyze the data, and process specialization approaches to adapt the treatments. When data collection processes are too different and their effects are too uncertain, external data, e.g. stemming from remote sensing, could be used to identify rough trends from one Region to another one. For example, the parameterization of our methods could be linked to the densities of the subnetworks or to the average sizes of the actual patches in each region.

Acknowledgment

The authors would like to thank the anonymous reviewers for their suggestions that helped improve the paper.

References

Allag-Dhuisme F., Amsallem J., Barthod C., Deshayes M., Graffin V., Lefeuvre C., Salles E. (coord), Barnetche C., Brouard-Masson J, Delaunay A., Garnier CC, Trouvilliez J. (2010). Choix stratégiques de nature à contribuer à la préservation et à la remise en bon état des continuités écologiques – premier document en appui à la mise en oeuvre de la Trame verte et bleue en France. Proposition issue du comité opérationnel Trame verte et bleue. MEEDDM (ed). http://trameverteetbleue.fr/sites/default/files/references_bibliographiques/guide1_comoptvb_juillet2010.pdf

BILLON L., CRIADO S., GUINARD E., LOMBARD A., SORDELLO, R. (2016). Elaboration d'une base de données nationale des composantes de la Trame Verte et Bleue à partir des données SIG des Schémas Régionaux de Cohérence Ecologique. Service du patrimoine naturel, Muséum national d'Histoire naturelle, Paris. SPN 2016 - 100: 22 p. + annexes

Brunet R., 1986, "La carte-modèle et les chorèmes", *Mappemonde*, 86(4), 2-6.

CBD (2017). Convention on Biological Diversity website. Strategic Plan for Biodiversity 2011-2020, Cities and Subnational governments. https://www.cbd.int/subnational/aichi-biodiversity-targets. Accessed February 2017.

COVADIS. (2014). Standard de données COVADIS du thème [Schéma régional de cohérence écologique]. Version 1.0. 68 pages.

European Commission (2011). COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Our life insurance, our natural capital: an EU biodiversity strategy to 2020. COM/2011/0244 final. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52011DC0244

European Commission (2013). COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Green Infrastructure (GI) – Enhancing Europe's Natural Capital. COM/2013/0249 final. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52013DC0249

Garcia-Gonzalez C., Campo D., Pola I., Garcia-Vazquez E. (2012). Rural road networks as barriers to gene flow for amphibians: Species-dependent mitigation by traffic calming, Landscape and Urban Planning, Volume 104, Issue 2, February 2012, Pages 171-180, ISSN 0169-2046, http://dx.doi.org/10.1016/j.landurbplan.2011.10.012.

Grünreich D. (1992). ATKIS—a topographic information system as a basis for GIS and digital cartography in Germany. In: Vinken R (ed) *From digital map series in geosciences to geo-information systems*. Geologisches Jahrbuch, Federal institute of geosciences and resources, Hannover, vol A(122), pp 207–216

Hagen-Zanker A., Straatman B., Uljee I. 2005. Further developments of a fuzzy set map comparison approach. International Journal of Geographical Information Science, vol.19, n.7, pp.769-785

Harrie L., Mustière S., Stuckenschmidt H. and Stigmar H. 2009. Cartographic aspects of geoportals. Presenting Spatial Information: Granularity, Relevance, and Integration. *Workshop at COSIT 2009*, Aber Wrac'h, France, September 21st, 2009.

Klar N., Herrmann M., Henning-Hahn M., Pott-Dörfer B., Hofer H., Kramer-Schadt S. (2012). Between ecological theory and planning practice: (Re-) Connecting forest patches for the wildcat in Lower Saxony, Germany. Landscape and Urban Planning, Volume 105, Issue 4, 30 April 2012, Pages 376-384, ISSN 0169-2046, http://dx.doi.org/10.1016/j.landurbplan.2012.01.007.

Loi H., Hurtut T., Vergne R., Thollot J. (2013) Discrete Texture Design Using a Programmable Approach, *SIGGRAPH Talks, 2013*

Mechouche, A., N. Abadie, E. Prouteau and S. Mustière (2011) Ontology based discovering of geographic databases content, 25th International Cartographic Conference (ICC'11), Paris, France, vol. 1, pp. 311--330, Lecture Notes in Geoinformation and Cartography, Springer, doi:10.1007/978-3-642-19143-5_18

Naumann S., McKenna D., Kaphengst T., Pieterse M. and Rayment M. (2011). Design, implementation and cost elements of Green Infrastructure projects. Final report to the European Commission, DG Environment, Contract no. 070307/2010/577182/ETU/F.1, Ecologic institute and GHK Consulting.

Reimer A. (2010). Understanding Chorematic Diagrams: Towards a Taxonomy. The Cartographic Journal 47:330-350

Schylberg, L., 1993. Computational Methods for Generalization of Cartographic Data in a Raster Environment, Doctoral Thesis, Department of Geodesy and Photogrammetry, Royal Institute of Technology, Stockholm, TRITA-FMI Report 1993:7

Sester M. (2008). Self-Organizing Maps for Density-Preserving Reduction of Objects in Cartographic Generalization. Self-Organising Maps. Applications in Geographic Information Science. P. Agarwal et A. Skupin (eds.), chap. 6, pp. 107-120.

SRCE Basse-Normandie 2014. Schéma Régional de Cohérence Ecologique de Base-Normandie, Composante 4 : Composantes de la Trame Verte et Bleue régionale, http://www.trameverteetbleuenormandie.fr/IMG/pdf/201404_SRCE_04-Composantes.pdf (Accessed May 2017).

Stanislawski L., Savino S. (2011). Pruning of Hydrographic Networks: A Comparison of Two Approaches. In Proceedings of the 14th ICA workshop on generalisation and multiple representation, jointly organised with ISPRS Commission II/2 Working group on Multiscale Representation of Spatial Data, Paris, 2011, http://generalisation.icaci.org/images/files/workshop/workshop2011/genemr2011_Stanislawski.pdf.

Touya G., Brando C. (2013). Detecting Level-of-Detail Inconsistencies in Volunteered Geographic Information Data Sets. Cartographica, vol. 48, no. 2, pp. 134-143, doi:10.3138/carto.48.2.1836.

United Nations (1992). Convention on Biological Diversity.