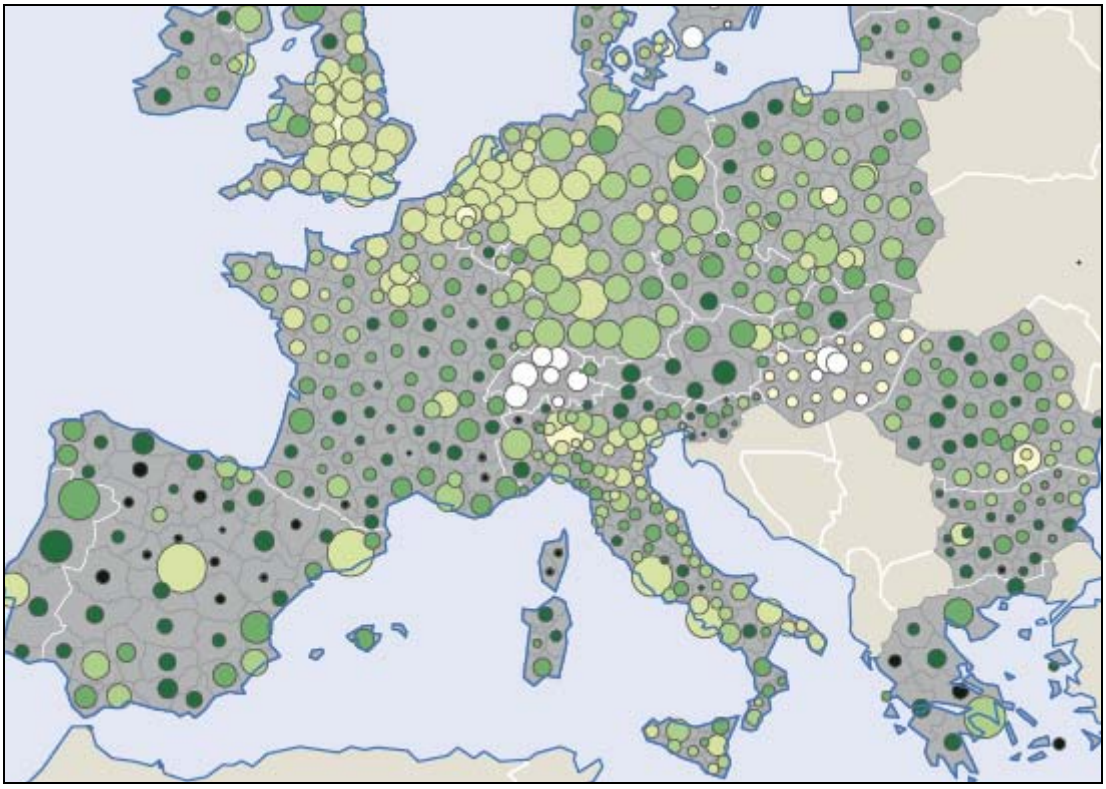
ESPON 2013 DATABASE

INCEPTION REPORT

26 September 2008



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

This inception report represents the first results of a research project conducted within the framework of the ESPON 2013 programme, partly financed through the INTERREG III ESPON 2013 programme.

The partnership behind the ESPON Programme consists of the EU Commission and the Member States of the EU25, plus Norway, Switzerland, Iceland and Liechteinstein. Each country and the Commission are represented in the ESPON Monitoring Committee.

This report does not necessarily reflect the opinion of the members of the Monitoring Committee.

Information on the ESPON Programme and projects can be found on **www.espon.eu**

The web site provides the possibility to download and examine the most recent document produced by finalised and ongoing ESPON projects.

ISBN number:
This basic report exists only in an electronic version.
Word version:

List of contributors to the inception report

UMS RIATE (FR)

Claude Grasland*

Ben Rebah Maher

Christine Zanin

Nicolas Lambert

Bernard Corminboeuf

Chloe Didelon

LIG-IMAG (FR)

Jérôme Gensel*

Bogdan Moisuc

Christine Plumejeaud

UAB (ES)

Andreas Littkopf

Juan Arevalo

IGEAT (BE)

Moritz Lennert

UMR Géographie-cités (FR)

Anne Bretagnolle

Hélène Mathian

Joël Boulier

Timothée Giraud

Marianne Guerois

TIGRIS (RO)

Octavian Groza

Université du Luxembourg (LU)

Geoffrey Caruso

* Scientific coordinators of the project

TABLE OF CONTENT

1 Introduction: from kick-off meeting to inception report

1.1 Context: A tight agenda

The lead partner (LP) of the project ESPON DB (Université Paris Diderot – Paris 7/ UMS 2414 RIATE) has been officially informed of the success of the proposal on 1 July 2008 by receiving an e-mail from the ESPON Coordination Unit, at the beginning of the summer period. The information was immediately sent to the 6 Project Partners (PP) involved in the ESPON 2013 DataBase project:

- Université Joseph Fourier Grenoble 1 (LIG), France
- Universitat Autonoma de Barcelona (UAB), Spain
- Université Libre de Bruxelles (IGEAT)
- Universitatea Alexandru Ioan Cuza, Romania
- Centre National de la Recherche (Délégation Paris A / UMR 8514 Géographie-Cités), France
- Université du Luxembourg

The ESPON Coordination Unit requested the Lead Partner to attend the KoM in Esch/Alzette on 9 July in order to launch the project from both scientific and administrative points of view. The reasons for organizing the KoM at the very beginning of the summer period were twofold:

- **From an administrative point of view** : The Inception Report and a copy of the signed Partnership Agreement had to be submitted together within 12 weeks after the approval of the project (Monitoring Committee Decision dated 13 June).
- **From a scientific point of view**, the project ESPON DB was supposed to deliver various materials, statistical and cartographical, to other ESPON projects under priority 1 that was selected during the same Call for proposals (Energy, Metropolisation, TIA, …). Without delivery of this first “mapkit tool”, the other ESPON projects would be hampered in the beginning of their activity. Moreover, the ESPON DB project should provide immediate recommendations for data collection to other ESPON projects and some of these recommendations should be written in the contract of this other ESPON project.

1.2 Consequences

i) This tight agenda introduced many difficulties, in particular because it did not take into account the fact that most administrative offices at LP and PP premises and most researchers involved in the project were not available during the summer period (15 July until end of August).

ii) The timing between the launching of the ESPON 2013 DataBase and the start of other ESPON projects is too short: other ESPON projects selected under Priority 1 were also launched in July, i.e. less than one or two weeks after the beginning of the ESPON 2013 DB project. As a result, the LP of the ESPON DB project received first requests for data and maps from the TIA project in mid-August, at a time when it was clearly impossible to send any data for legal reasons.

iii) As the contractual, financial and administrative rules of the ESPON 2013 Programme are more complex compared to the former Programme, the time required for the preparation of all official and contractual documents has increased in order to be able to launch the project on a sound basis.

We would therefore be most grateful if you could take the above mentioned points into account when assessing both scientific and contractual issues related to the launching of the project.

1.3 Request for allocation of KSS to ESPON DB Project

We would like that the possibility to include KSS in the evaluation of our project shall be once again envisaged by ESPON Monitoring Committee as the ESPON 2013 DB project addresses both scientific and political challenges. We consider that if experts from KSS had been allocated to the ESPON 2013 DB project, they would certainly has pointed and anticipated some of the difficulties we have been obliged to find during the first two months of the project and proposed useful solutions.

We consider that the ESPON 2013 DB project is very similar to other priority 1 project in terms of scientific contents and political challenges. The problems that are submitted to the project ESPON 2013 DB are indeed related to advanced research of very high level in the scientific fields of Computer Science, Database design, cartography, GIS, territorial planning. It is therefore a great frustration for researcher involved in the project to have no KSS support, able to help us to improve our work. It is also related to difficult political questions, for example in the case of map generalisation (islands, zoom on small countries…). The allocation of two KSS experts, one from scientific side and one from political is necessary for an optimal development of the next steps of the project.

2 Overview of the project

First description of the project, taking into account the objectives envisaged, the ESPON space and the geography to be covered by the ESPON 2013 Database

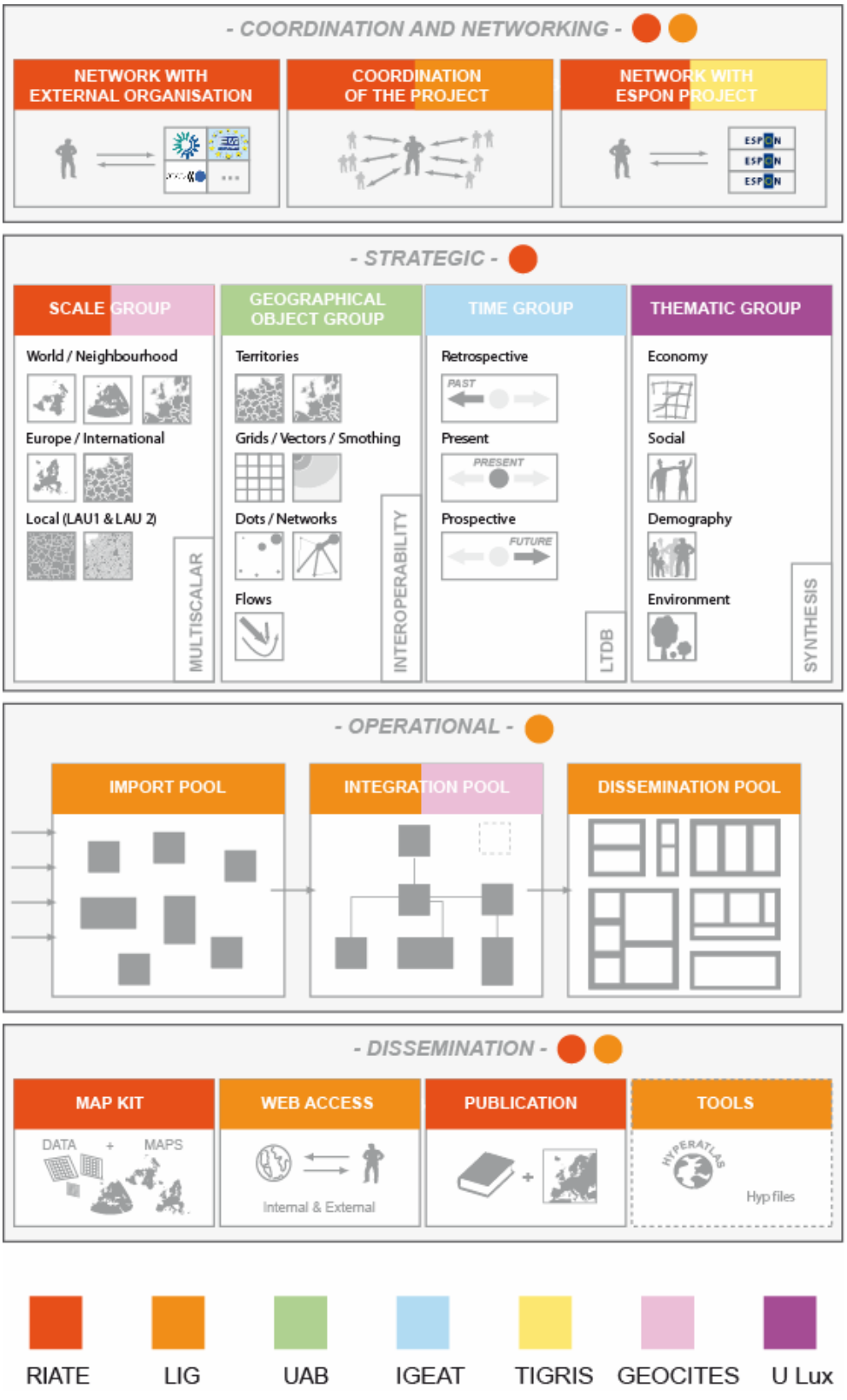
The work plan of the ESPON DB project that was selected during the tender procedure was constrained by the legal obligation to follow a precise template with maximum number of pages. This template introduced some difficulties for the presentation of ESPON DB project because two of our main field of activities should be presented separately from the core activities of the project according to the templates.

Networking and coordination (Activity A): in other ESPON projects from priority 1 (Energy, TIA, Metropolisation …) or priority 2 this activity is mainly related to internal organization of the project and some flows of data inputs and outputs. But in the ESPON DB project, this activity of networking is a central task as our project is collecting data from a wide range of external organization (EEA, Eurostat …) and is in contact with all other ESPON projects. We have therefore decided to present this part of our project as a central activity.

Diffusion (Activity D) : one more time, in other ESPON projects, the diffusion of results is a specific task that is related to the promotion of valorization of results that has been achieved. It takes place normally during the final period of activity. It is not the case for ESPON DB project where diffusion of data and maps to other ESPON projects is a prior task since the very beginning of the period of activity (e.g. Map kit tool and database to be delivered and update regularly). Moreover, one of the tasks of the ESPON DB project is to support ESPON CU in the general promotion of the results of the ESPON program. As in the case of networking and coordination, this task should be considered as a main field of activity for ESPON DB and not as a final output.

This inception report provide therefore the opportunity to propose a clearer view of the activities of the ESPON DB project than the answer to the tender where the template was not adapted to the specificities of our activities. If we admit that networking & coordination or diffusion are core part of our project and not peripheral activities, we can define our work plan for the next 30 months as a set of 4 main field of activities divided in 14 Work packages, as presented in Figure 1.This logical and hierarchical presentation of the project has been précised in a PPT presentation (see. Annex) that make more easy the understanding of the rational of the project both for ESPON CU (KO meeting in Esch/Alzette the 9th July) and the Project Partners (First Project meeting in Paris the 13th and 14th September).

Figure 1: Division of the project in activities and work-packages



In our answer to the tender, we had precisely indicated the contribution of each PP to the different activities and WP in an anonymous way (*1=RIATE, 2=LIG, 3=UAB, 4=IGEAT, 5=TIGRIS, 6=Géographie-cités, 7=UAB*) and we had also mentioned which precise tasks should be dedicated to 5 external experts (quoted 8 to 12) for which precise budgets has been allocated to RIATE (*experts 8 to 11*) and LIG (*expert 12*).

Table 1: Allocation of workforce to activities and project partners (%)

	1	2	3	4	5	6	7	8	9	10	11	12	
	RIATE	LIG	UAB	IGEAT	TIGRIS	GEOC	UL	EXP1	EXP2	EXP3	EXP4	EXP5	TOTAL
ACTIVITY A : Coordination and networking													
A.1 Networking with external org.	**3,1**	0,8	2,3	0,8	2,3	0,8	1,5	1,2			1,2		**13,8**
A.2 Coordination of the project	**4,6**	**3,8**											**8,5**
A.3 Networking with ESPON projects	**2,3**	0,8	0,8	0,8	0,8	0,8	1,5	1,2					**8,8**
Sub-Total	***10,0***	***5,4***	***3,1***	***1,5***	***3,1***	***1,5***	***3,1***	***2,3***			***1,2***		***31,2***
ACTIVITY B : Strategic Reflection on ESPON Database and Territorial Cohesion													
B.1 Scale Group	**2,3**		0,8	0,8	0,8	0,8	0,8			1,2	1,2		**8,5**
B.2 Geographical objects group	0,8		**3,1**	0,8	0,8	2,3			1,2				**8,8**
B.3 Time group	0,8		0,8	**2,3**	0,8	1,5			1,2				**7,3**
B.4 Thematic/political group	0,8		0,8	0,8	1,5		**3,1**						**6,9**
Sub-Total	***4,6***	***0,0***	***5,4***	***4,6***	***3,8***	***4,6***	***3,8***		***2,3***	***1,2***	***1,2***		***31,5***
ACTIVITY C : Operational Organization and Development of Espon 2013 Database													
C.1 Import pool	1,5	**3,1**	0,8	0,8								0,8	**6,9**
C.2 Integration pool		**7,7**	0,8			0,8						1,9	**11,2**
C.3 Export pool		**3,8**	0,8	0,8								0,8	**6,2**
Sub-Total	***1,5***	***14,6***	***2,3***	***1,5***	***0,0***	***0,8***	***0,0***					***3,5***	***24,2***
ACTIVITY D : Dissemination													
D.1 Mapkit Tool	**2,3**	1,5								1,2			**5,0**
D.2 Web access to ESPON DB		**3,1**											**3,1**
D.3 Support to ESPON CU	**1,9**		0,8		0,8	0,8	0,8						**5,0**
Sub-Total	***4,2***	***4,6***	***0,8***		***0,8***	***0,8***	***0,8***			***1,2***			***13,1***
TOTAL	**20,4**	**24,6**	**11,5**	**7,7**	**7,7**	**7,7**	**7,7**	**2,3**	**2,3**	**2,3**	**2,3**	**3,5**	**100,0**

Note : the team responsible of the coordination of a particular Activity or WP is indicated in red.

Of course, we can not fully anticipate in a very detailed way every event that will happen during the 30 months of the projects. And we can neither imagine that each PP or Expert will work in an isolated way on its tasks without interactions with the other. But it is certainly better for each of the team involved in the project to have a clear awareness of what are its main mission and what are its responsibilities of leader or associate in the different activities and work packages. From this point of view, the allocation of funds has been directly related to our evaluation of the resources that are expected to be necessary to fulfil the objectives of each activity and each work packages (Table 1).

Each activity (A,B,C,D) is considered as permanent during the 30 months of the project and the different interim reports of the ESPON DB project will propose summaries of main results obtained in the different activities and work-packages. But of course, it is not possible to consider the project as a static structure where all parts are making work independently from each other. That is the reason why we have introduced another level of organization called "**challenge**".

A **challenge** can be defined as a common objective of the partners of the project that implies the connection of different activities and work-packages. A challenge is decided in a general assembly of the project and should be precisely defined in terms of responsibilities (who is the coordinator), partnership (which researchers are involved), agenda (what are the estimated delay of realization) and concrete outputs (what are the deliverables expected). In all cases, a challenge should stick to the *objectives envisaged, the ESPON space and the geography to be covered by the ESPON 2013 Database*

The different challenges should be organized in a logical order as some tasks are interlinked and cannot be started without achievement of previous results. In most cases, a challenge implies a logical set of tasks to be realized in the field of activity, following the steps A->B->C->D. The most difficult step in the realization of any challenge is the connection between activity B and C.

The activity B "Strategic" (coordinated by RIATE) is dedicated to the research of innovative solutions for the production of new data, new maps, new indicators … interesting for the ESPON program as a whole. This activity will typically propose exciting solutions for the integration of data at different scales, based on different geometries, at different period of time, for different political use. The outputs will be innovative maps or datasets related to manual indicated how to proceed in order to reproduce this methods in different situations.

The activity C "Operational" (Coordinated by LIG) will receive the inputs from the strategic group and try to include them in a computer application where expert knowledge can be reproduced as much as possible in a standardized and automatic way. As a typical example, the methods of estimation of missing value that are proposed by an expert in activity B for a particular example should be transformed into automatic procedures that could be applied to a wider set of data.

The obvious difficulty of this transition B->C is the fact that human expert are using very complex rules to solve problems and that it is not necessarily easy to transpose this knowledge of human experts into computer applications. It is the reason why the difficulty of challenge to be solved should be carefully graduated in terms of level of increasing complexity

3 The ESPON Database Strategy

Presentation of the ESPON Database Strategy, including an analysis of the current situation of the ESPON 2006 Database and immediate needs for updating datasets (mainly in relation to EUROSTAT data, which should cover as far as possible the entire ESPON 2013 space and its regions). It should also comprehend concrete ideas in relation to data/information flow within ESPON 2013, in particular internal and external networking activities and actions. (cf point V, 2.)

This section summarizes the result of the first Project meeting held in Paris the 14-15 September where all PP was represented as far as our project officer Sandra di Biaggio from ESPON coordination unit. This meeting was prepared by a short survey sent to each PP where it was asked to indicate precisely the contributions that each PP was likely to deliver for each activity or WP and in which delay. The main output of the meeting was the definition of a set of challenges to be engaged according to precise agenda of realization.

3.1 Challenge 1 : Delivery of basic datasets derived from EUROSTAT and EEA at NUTS2 and NUTS3 levels according to NUTS2003 and NUTS2006 divisions.

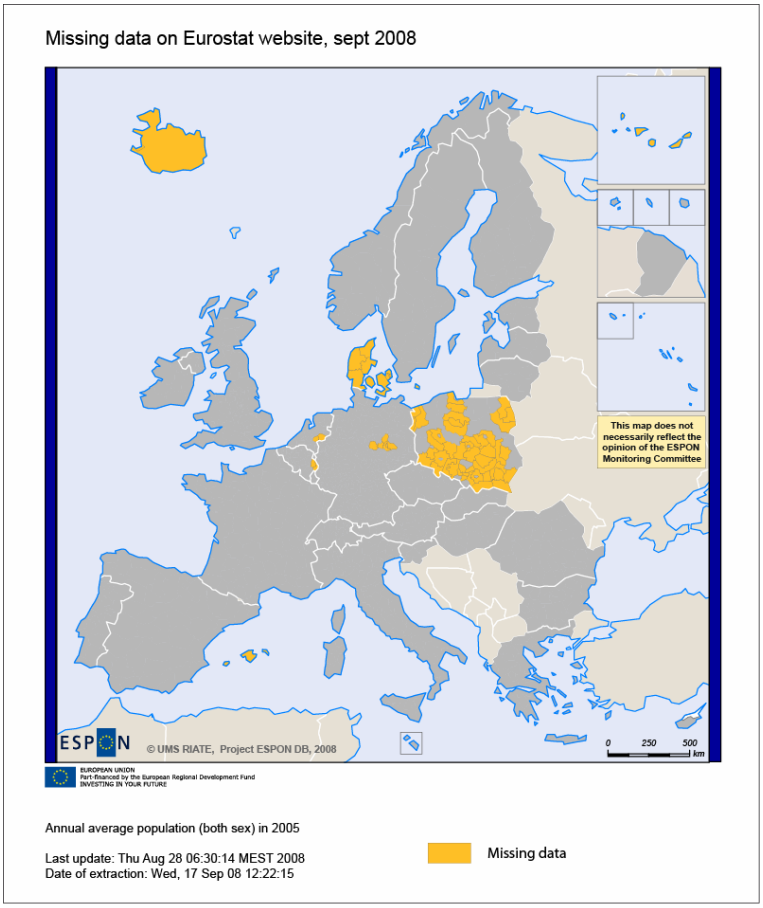
Objectives

The production of harmonized datasets covering all the ESPON space (31 countries) at NUTS 2 or NUTS 3 level has been recognized as the first challenge to be solved with an absolute priority as it is a condition of continuity with previous work realized in ESPON 2006 program. It is obvious that the new ESPON 2013 project needs immediately basic information at this level like area, population, GDP, employment, which will be used as reference for more sophisticated analysis where this project will produce more precise information in their specific fields. Moreover, the map kit tool that will be sent to this projects (see. Section 4) should not be limited to purely geometric information and should involved this basic data sets as starting point and model for more elaborated data collections. Finally, we should be able in a short delay to connect the new information elaborated by ESPON 2013 Program with former datasets elaborated by ESPON 2006 Program in order to produce time series of indicator, with the objective to support projects on the monitoring of European territory.

Analysis of the situation

Eurostat proposes currently on its website datasets at NUTS2 and NUTS3 levels using the new NUTS2006 territorial division. The problem is that most statistical data, even the more basic data like population are actually not complete for all countries. In particular, population is not available in countries where important changes has occurred in the territorial delimitation of regions and the data elaborated in the previous NUTS2003 division are no more available.

Figure 2: Availability of population in NUTS2006 divisions at NUTS3 level in Eurostat website the 17/09/2008



We have also planned an appointment with Eurostat in order to examine the delay for delivery of complete datasets on basic indicators but one more time we ignore the delay to obtain complete coverage of the 31 countries of the ESPON territory.

Proposal of two steps delivery of data

The only way to provide immediately ESPON projects with basic indicators is to produce a provisional dataset based on NUTS2003 divisions (September) and to wait the second

delivery (November) to propose equivalent data according to NUTS2006 divisions. This solution has many advantages:

RIATE has download most basic social, economic and demographic data available in NUTS2003 division in 2007, just before the moment where EUROSTAT change its website and adopted the new NUTS2006 division. At this moment, the regional datasets was generally complete for all countries with most recent information on years 2005 or 2006.

UAB has produced recently for EEA an application that convert Corine Land Cover 1990 and 2000 into NUTS2003 units NUTS3 level. This data are therefore immediately available for use and can be combined with socio-economic data collected by RIATE.

ESPON 2006 database has produced original data using NUTS1999 but also NUTS2003 divisions, which can be easily connected to basic indicator. In particular, measure of accessibility, typologies, etc.

It is obvious that the new ESPON 2013 program will use the NUTS2006 divisions as reference for the future works but it is not a problem to use the former NUTS2003 division as reference for an intermediate period.

Moreover, the ESPON 2013 database will develop tools for conversion of data between the different geometries of regional divisions (NUTS2006, NUTS2003, NUTS1999, NUTS1995 …) but will in every case store the data in their original geometries. In other words, it is not a loose of time, on the contrary, to elaborate a dataset at NUTS2003 level and it will be a useful step for the elaboration of long term time series.

Deliverables and timetable

Sept. 2008: Table of basic indicators at NUTS2 and NUTS3 level according to NUTS2003 divisions.
Oct. 2008 : Meeting with Eurostat
Nov.2008: Table of Basic indicators at NUTS2 and NUTS3 level according to NUTS2006 divisions.

Teams involved

LIG: coordinator, data model.
RIATE: harmonization of geometries
UL & IGEAT: networking with Eurostat
UAB: conversion of CLC data toward NUTS2003 and NUTS2006

3.2 Challenge 2: Harmonization of time series for basic socio-economic indicator at regional level for the period 1995-2006.

Objectives

Based on the result of challenge 1, we propose to elaborate a methodology for the harmonization of time series covering ESPON territory at regional level for the period 1995-2006 on the basis of simple indicators of regional policy (population, GDP, unemployment, age structure). The problem is not to cover immediately a great number of indicators but to define a methodology that could be implemented in the ESPON 2013 DB and reproduced by different ESPON projects.

Analysis of the situation

The problem will be to store and combine changing regional divisions (NUTS1995, NUTS1999, NUTS2003, and NUTS2006) and to propose solutions for their combination in different ways. The objective is not necessary to produce an harmonized dataset (based on NUTS 2006 divisions) but to propose methods that make possible the analysis of the territorial changes. It can include mixture of territorial division of different periods, smoothing methods, etc.

Deliverables and timetable

Dec. 2008 (Espon Seminar): presentation of different options.

Feb. 2008 (FIR): Summary of results obtained – Proposal for next steps.

Teams involved

IGEAT: coordinator, data elaboration

UL: networking with Eurostat (historical data)

RIATE: Cartography and spatial analysis

LIG: importation of data to ESPON DB

3.3 Challenge 3: Harmonization of data at World / Neighbourhood and European / regional levels.

Objectives

Based on the results of ESPON 2006 Program, we propose to examine in a systematic way how to combine datasets at world/neighborhood levels (where basic territorial units are the states) and datasets at European/Regional levels (where basic territorial units are NUTS2 or NUTS3 units). The interest of such connection is to enlarge the scales of analysis from spatial point of view (situation of ESPON territory in the world, situation of eastern and southern neighbouring countries) but also from historical point of view as time series at state level are generally more easy to obtain on long period (1960-Present) than regional time series (1995-Present).

Analysis of the situation

We can firstly use the experience gained with ESPON 2006 program. In particular the world database elaborated by project Europe in the World (ESPON 3.4.1) could be updated and eventually improved (modification of the WUTS Sytem ?). But we suggest to introduce a more general connexion with UN statistical system through an expert team specialized in the integration of world databases. The most difficult question will be the connection between the WUTS system and the NUTS system through the level of states. Indeed, a figure like the population of Germany in 2000 is not necessary the same according to a world database of UN (where Germany is a basic unit of level WUTS5) or according to Eurostat (where Germany is a macro-unit of level NUTS0). The compatibility of world database should also be analyzed for environmental data at grid level.

Deliverables and timetable

Dec. 2008 (Espon Seminar) : presentation of preliminary results.

Feb. 2008 (FIR): Summary of results obtained – Proposal for next steps.

Teams involved

RIATE: coordinator, data elaboration

UL: Compatibility with Eurostat data

UAB compatibility with EEA data

EXPERT 4: Connexion with UN statistical system

LIG: importation of data to ESPON DB

3.4 Challenge 4: Harmonization of data at European / regional and National / local levels.

Objectives

The ESPON 2006 program has revealed that many questions related to territorial cohesion can not be fully explored at NUTS2 or NUTS3 levels and need further investigations at more local levels LAU1 and LAU2 (former NUTS4 and NUTS5). Case studies providing zoom on specific territories at local level (rural areas, cross border areas, intra-urban differentiation, …) will be more and more requested in ESPON 2013 program for project of priority 2 and, in certain cases, for project of priority 1. It is therefore of utmost importance to be able to collect such type of data in ESPON 2013 Database and to develop a long term strategy.

Analysis of the situation

Eurostat has announced several time the publication of basic indicators at LAU1/LAU2 levels with related geometry, but this publication has been always delayed. ESPON has co-financed with the "mountain study" a CD-Rom realized by NORDREGIO that provide some local data at this level for basic indicators in 1990 and 2000 but the coverage of the territory is not complete and some difficulties appears with the geometry that was based on administrative situation of 1997 when data are related to 1990 and 2000. Using this CD-ROM as starting point, we propose to analyse how to develop data collection at this scale on the basis of national data provider. Test will be realized on the national statistical offices of countries involved in the project. Expert 1 will test the solutions for accessing and neighbouring countries.

Deliverables and timetable

Dec. 2008 (Espon Seminar) : presentation of preliminary results.

Feb. 2008 (FIR) : Summary of results obtained – Proposal for next steps.

Teams involved

TIGRIS : coordinator, data elaboration, test on Romania

RIATE : Geometry and cartography

Geographie-cités : Test on France

UL : Test on Luxembourg

UAB : Test on Spain

IGEAT : Test on Belgium

EXPERT 1 : Test on candidate countries

LIG : importation of data to ESPON DB

3.5 Challenge 5: Combining socio-economic data measured for administrative zoning (Nuts level) and environmental data defined on a regular grid (like Corine Land cover or any spatiomap)

Objectives

The ESPON 2006 program has mainly used socio-economic data from Eurostat, but has also sometime used environmental data, in particular the Corine Land Cover database. In most case, the environmental data was transposed to NUTS division by GIS application in order to insure a compatibility with the rest of indicators. This solution introduces some problems revealed by the MAUP study (ESPON 3.4.3) and it seems better to let open a wider range of solution for data harmonization. For example the operation grid->NUTS could be completed by a reverse application NUTS-> Grid where socio economic data collected by administrative units are transposed to regular grid or to other divisions more relevant for the analysis of environmental phenomena (e.g. Water Basin).

Analysis of the situation

Many problems described in this challenge are actually solved by EEA, in particular through the topic center ECT-LUSI which is coordinated by UAB and where LIG, RIATE and Géographie-cités are partner. Therefore, the problem is not to duplicate the work realized by EEA but to introduce a flow of data exchange between ESPON and EEA and to build common data infrastructure in order to insure full compatibility of database on each side.

Deliverables and timetable

Dec. 2008 (Espon Seminar): presentation of preliminary results.

Feb. 2008 (FIR): Summary of results obtained – Proposal for next steps.

Teams involved

UAB: coordinator, data base strategy, GIS, interface with EEA

RIATE : Geometry and cartography

LIG : importation of data to ESPON DB

3.6 Challenge 6: Constructing complex geographical objects of higher level such as cities, resulting from an aggregation of elementary objects according to a measure of relation in space (proximity, links and flows…).

Objectives

Constructing complex geographical objects of higher level such as cities, resulting from an aggregation of elementary objects according to a measure of relation in space (proximity, links and flows…). According to the previous challenge 4 and 5, in the case of cities, the elementary objects intervening in the aggregation process may come from either a grid source (for instance the built up areas), or an administrative zoning (for instance urban administrative elementary unit as LAU2)

Analysis of the situation

The aggregated objects, here the European cities, may be approached at four different levels. At each level there already exist databases linked with a specific definition. But one level may have given rise to more than one database.

Hierarchy of levels	Names	Database
1	Sub-City Districts	Urban Audit
2	Core Cities	Urban Audit
3	Morphological agglomerations	- MUAS (1.4.3 Espon Project) - UMZ (EEA)
4	Functional areas	- LUZ (Urban Audit) - FUAs (1.1.1. Espon Project, revised in the 1.4.3 Espon Project)

The 7 different databases will be collected as well as the documentations. On this basis, the work will focus at first on aligning the specifications of each data base. Some specific regions will be taken to interpret the differences between data bases defined at the same levels: for example, the LUZ and the FUAS. Then a statistical protocol will be defined and implemented for comparing two data bases (LUZ and FUAS): the aim is to integrate some metadata associated to each data base illustrating how the database describe the urban phenomena at the macro level (urbanization levels, number and

population of cities by size classes, slope of the rank-size graph etc). To the extent possible, database will be joined, city to city.

Deliverables and timetable

Dec. 2008 (Espon Seminar): presentation of preliminary results.

Feb. 2008 (FIR): Summary of results obtained – Proposal for next steps.

Teams involved

Géographie-cités: coordinator, conceptualization, interface with urban Audit

IGEAT : Socio-economic approach of urban areas

UAB: Morphological approach of urban areas.

EXPERT 1 : Urban areas in candidate countries

RIATE : Geometry and cartography

LIG : importation of data to ESPON DB

3.7 Challenge 7 : External networking (Eurostat, EEA, …)

Objectives

The project ESPON DB should develop regular contacts with ESPON coordination Unit for different activities of promotion of the program, contact with organization like Eurostat, etc.

Analysis of the situation

Each project partner is responsible for contact with specific organization and the Lead Partner (RIATE) is directly responsible from work on request addressed by ESPON coordination unit for the promotion program.

- EUROSTAT : UL, RIATE, LIG, IGEAT
- EEA : UAB, RIATE, LIG
- OECD : IGEAT
- URBAN AUDIT : Géographie-cités
- National statistical offices : TIGRIS & UL

As each of these tasks of external networking implies regular contact with ESPON coordination unit (in order to define a global strategy) we suggest that University of Luxembourg will be the coordinator of the tasks of external networking and promotion of the program. The contact with Eurostat will be a crucial milestone for this strategy of external networking. Similar procedure will be applied further with OECD, EEA , etc…

Deliverables and timetable

Dec. 2008 (Espon Seminar) : presentation of preliminary results.

Feb. 2008 (FIR): Summary of results obtained – Proposal for next steps.

Teams involved

UL : coordinator, proposal of methodology, interface with ESPON CU

RIATE : Support of ESPON CU for cartography and data

LIG : External diffusion of ESPON DB by web

All partners : Contact with External organisation

3.8 Challenge 8: Internal networking (other ESPON Projects)

Objectives

The project ESPON 2013 DB should develop regular contacts with other ESPON Projects from Priority 1, Priority 2 and Priority 3. This task is crucial as it is related to circulation of data inside the whole program.

Analysis of the situation

Each project partner is responsible for contact with specific ESPON project or specific organization and the Lead Partner (RIATE) is responsible of the general coordination of requests send by ESPON Projects. LIG is responsible on the internal and external diffusion of ESPON database on the web. Actually, we have developed a pragmatic approach and allocated contact person to each other ESPON Project :

- RIATE (C. Grasland) : contact with projects on Demography and Energy
- IGEAT (M. Lennert) : contact with project on TIA
- Géographie-cités (A. Bretagnolle) : contact with project on Metropolitan Area
- UAB (A. Littkopf) : contact with project on Climate Change
- TIGRIS : contact with project on rural development

But we are exploring a more efficient solution based on a web application where other ESPON projects will address their requests to a central board and where answer will be guarantee in a short delay.

Deliverables and timetable

Dec. 2008 (Espon Seminar) : presentation of preliminary results.

Feb. 2008 (FIR): Summary of results obtained – Proposal for next steps.

Teams involved

RIATE: coordinator, proposal of methodology, interface with ESPON CU

LIG: Diffusion of ESPON DB by web (intranet and extranet)

All partners : Contact with ESPON Projects

4 MAP-KIT TOOLS = DATA + GEOMETRIES + TEMPLATE

Presentation of a proposal for the ESPON 2013 Database design including practical considerations in relation to the ESPON map-kit tool, its compatibility with the ESPON database and needs for future updating.(cf. point V).

What is in the map kit tools (version 0) delivered on 2008 September 30th?

The Map kit tools 20013 enclose two main folders:

- The old version of map kit tools (_old_map_kit) with Espon 2006 geometries in shape files format and the last version of Espon 2006 data base delivered in April 2007.
- The new map kit tools for Espon 2013 data base named Version0 (MAP_KIT_ESPON_2013_(V0))

This new map kit is organized in two independent parts:

- A map kit for NUTS 2003
- A map kit for NUTS 2006

In each of these two map kits you will find 3 main folders:

- One for data named DATA
- One for the geometry named GEOM
- One for the layouts named TEMPLATE

4.1 The DATA folder

In this folder you will find different folders for basic indicators concerning 5 themes: area, demography, economy, labor-market and land use. Each thematic folder contains Excel and Open Office files. Every file appears two times, one file of directly useful data and one file of raw data.
The raw data files allow everyone to be able to access to the raw data whenever you could need it.
Indicators are delivered for NUTS0, NUTS1, NUTS2, NUTS2/3 and NUTS3 aggregation Europe levels and for different dates (between 1995 and 2006 depending on the indicator and the availability of the data).

4.2 The GEOM folder

In this folder you will find 3 others folders for the Europe geometries:

EUROGEOGRAPHICS geometries set with very precise layers for GIS mapping.
Data type: vector. Shape file format
Resolution: 1/1 000 000 for 2006 NUTS and 1/3 000 000 for 2003 NUTS
Geo-referencing: Coordinates in degrees (longitude, latitude) with decimal fraction and based on the ETRS89 spatial reference system.

EUROGEGRAPHICS templates set with a more generalized layer for thematic cartography.
Data type: vector. Shape file format
Resolution: 1/20 000 000
Projection: Lambert azimutal equal area.
Central meridian = 15 and latitude of origin = 50.0

RIATE templates set with generalized maps for thematic cartography.
Data type: vector. Shape file format
Projection: Lambert azimutal equal area.
Central meridian = 15 and latitude of origin = 50.0

All the templates are supplied for nuts0, nuts1, nuts2, nuts 2/3 and nuts3 aggregation Europe levels for the 2003 and 2006 nuts that are 10 templates for each set.

4.3 The TEMPLATE folder

In this folder you will find all you need for layouts, i.e. all the different elements you have to use for layouts like disclaimers or Espon logo. Also, you will find, two examples of layout, one in Adobe Illustrator format and one in PNG format.

Please note that these elements are totally provisional.

Discussion about the RIATE template

Beyond the using rights what is the purpose of the RIATE template?

The cartographic generalization is a complex process. It is the simplification of observable spatial variation to allow its representation on a map. Map generalization reduces complexity, retains spatial and attribute accuracy, accounts for map purpose and scale and provide more information or more efficient communication. The first principle of the generalization process is that the "amount of information that can be shown per unit area decreases according to geometric progression" (F. Töpfer, 1966). According to this principle, drawing a map at scales smaller than their source can give rise to map displays exhibiting graphic conflict, such that objects are either too small to be seen or too close to each other to be distinguishable. Furthermore, scale reduction will often require important features to be exaggerated in size, sometimes leading to overlapping features. Cartographic map generalisation is the process by which any graphic conflict that arises during scaling is resolved[1].

The second principle is that the way geographical objects (or spatial units) are processed depends always on their spatial context[2]. This context influence highly our way of reading maps and is influenced in reverse by the scale. From the identification of individual elements to an apprehension of the whole space all the knowing world must be recognized by any reader of the map. That's why the generalization process must respect three types of relations that an object can have with his environment: being part of a significant group, being in a particular area and being in relation with 'same level' surrounding objects.

These two principles are required for a "graphically and politically" correct generalized map. The generalization proposed by Riate follows these principles and allows high level of thematic information with a minimum of disturbance which can result from the design of the spatial units.

[1] Julio Cesar Lima d'Alge, Map Generalization, ONU presentation, Image processing division, 1998

[2] S. Mustière and B. Moulin "What is spatial context in cartographic generalisation ?" Computer Science Departement and Geomatic Research Centre, Laval University, Québec, Canada. Symposium on Geospatial Theory. Ottawa 2002.

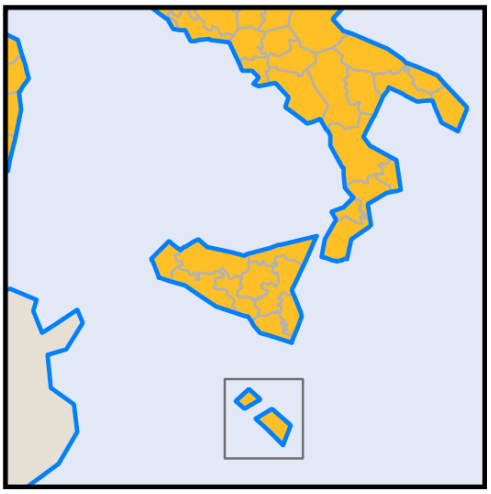
Let's us focus on two points on the map of the European regions: Norway and Malta. These two countries raise various problems during the simplification of their outlines.

MALTA

Malta is a European country but a very small island and within a certain scale it becomes an invisible point. The generalization principles allow us to emphasize the size of the country with regard to the general scale of the map. So the thematic information represented on the country remains visible.

Eurogeographics template with a resolution of 1/20 000 000

RIATE Template V1 and 2



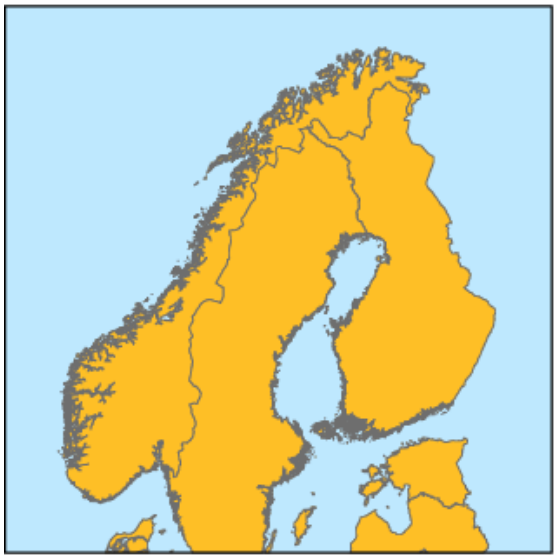
LIECHTENSTEIN

We are actually exploring solutions for the case of Liechtenstein. Solution used for Malta can not be transposed as this country is not located on maritime areas.

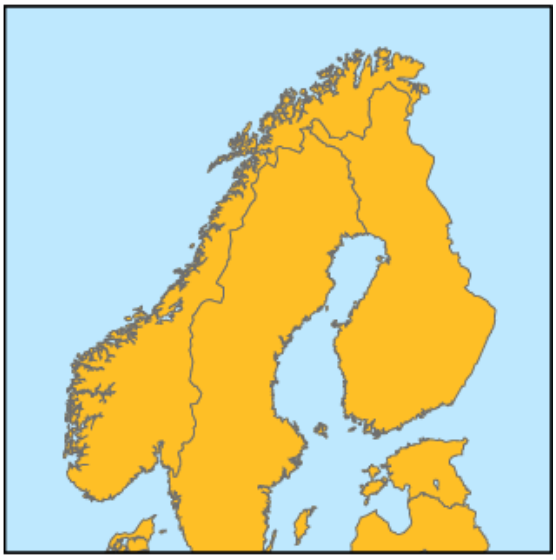
NORWAY

Norway presents outlines extremely chopped with numerous islands very close to the continent. Changes in scales can blur the outlines of country and return illegible the islands. That's why is absolutely necessary to generalize outlines while preserving the main details and characteristic aspects of the country (like islands, fjords and deltas). We can notice than solution proposed by RIATE (template V2) is very similar to the one proposed by BBR for maps presented under German Presidency.

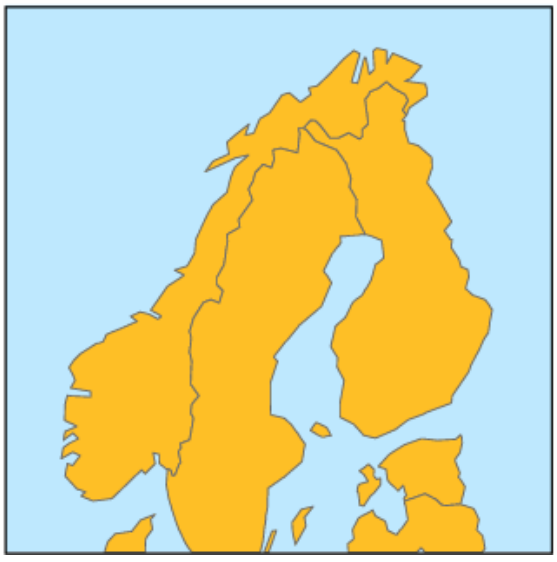
Eurogeographics template with a resolution of 1/1 000 000



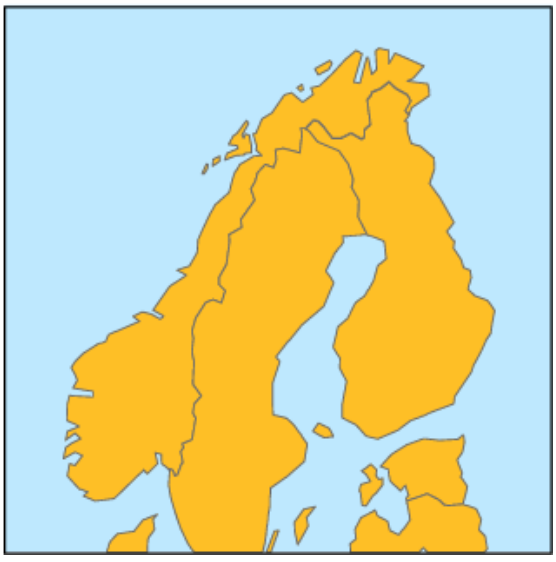
Eurogeographics template with a resolution of 1/20 000 000



RIATE template V1



RIATE template V2



5 ESPON DATABASE PROTOTYPE

First presentation of an ESPON 2013 Database "Prototype". A first visualization on how the ESPON 2013 Database could look like in terms of structure and functionalities should also be included in the Inception Report, including as far as possible some practical examples for a limited number of datasets.

After the ESPON CU has arranged a new website server contract, the database will also physically be located at this server. To be able to implement this, the project will provide, as soon as possible, a more detailed technical description of the requirements for hosting the database. Furthermore, the project will describe, in the inception report, a procedure with a time table to keep the database on the ESPON server up to date.

5.1 Espon Database 2013 challenges

First and foremost, it is important to clarify that, although the name of the application refers to a "database", the ESPON 2013 Database will be a complex Spatio-Temporal Information System (or STIS), relying on several software components: databases (for data and metadata storage), ontologies (for data and metadata structuring) and programming routines (implementing procedures in charge of data harmonization, verification and estimation).

Thus, the ESPON Database 2013 needs to provide an adequate solution to several challenges. These challenges stem from design issues, usability issues and performance issues related to the implementation of such a complex STIS.

Design challenges

The design challenges of the Espon 2013 DB are linked to the need for creating a sustainable data model. This aim will be fulfilled by: a) designing a data model capable of storing most types of data provided by statistical organisms at present and b) by allowing enough flexibility inside the model in order to allow further extensions (storage of new types of data) without having to completely redesign the application.

Enlarging the range of spatial scales with more global and more local levels: the application will take into account data both at world and local levels, extending the existing NUTS territorial partition.

Combining heterogeneous data sources: the Espon 2013 DB will deal not only with vector and thematic data recorded on territorial partitions, but also with environmental data recorded on grids and with statistical data recorded at an individual level.

Obtaining complete medium and long-term time series: the application aims not only at storing data imported from external sources, but also at implementing

data harmonization algorithms based on the knowledge of thematic experts in order to reconstitute complete past temporal series and to allow predictions about future trends and evolutions.

Exploration of new thematic fields: besides accepting the definiton of new themes and associated new indicators, the ESPON 2013 DB will offer an adequate way for storing and processing non-exhaustive (partial) datasets resulted from zooms, surveys and case studies.

Usability challenges

The Espon 2013 DB will store and manipulate a large quantity of data, coming from different sources, paradigms and formats. In order to deal with such a heterogeneity, the application will have to rely on a complex data model with advanced data harmonization capabilities. The complexity of the application will probably make it difficult to manipulate. In order to alleviate the work of the user, the application will have to: a) integrate tools in order to simplify the maintenance and updating work of experts and b) provide a simplified data repository to be accessible for non-expert users.

Performance challenges

The Espon 2013 DB will store and manipulate a large quantity of data. It will contain detailed temporal territorial data, detailed temporal grid data and some individual data. It will also store in some cases redundant indicator values (for multiple sources or for multiple publication dates). The potentially very large dataset, combined with a complex data model (required in order to integrate all types of data in one database) will most probably result in very resource consuming (in terms of CPU time and memory size) queries over the main database of the application. A way ton overcome this issue is described below.

Overall application structure and components

The Espon 2013 DB application will contain three major software components, namely the import pool, the integration pool and the export pool (see figure X). The application will be implemented in the Java Entreprise Edition programming language, which allows an excellent maintainability and evolutivity of the code, while providing the advanced functionalities required for our application (database connectivity, transaction management, security, etc.). The Java open source software development community is also very active, making possible the reuse of existing powerful software tools (for the import and export pools).

5.2 Components of the application

The import pool describes the component in charge of the data acquisition, from external sources (relational databases or files in various data exchange formats) into the main database of the application and into the two related ontologies.

The import pool is to fulfill certain requirements in order to provide rich and sound data input for the application:

It will accept a wide selection of input formats, ranging from proprietary GIS exchange formats (ESRI shape files, MapInfo mif/mid files, etc.) text-based exchange formats (like WKT, the well-known text format), proprietary database tabbed exchange formats (like Microsoft Excel and Access sheets) and XML based exchange formats.

It will allow data input from sources with different paradigms: geometric and thematic data recorded on territorial partitions, continuous data recorder on grids or TINs (triangular irregular networks), thematic data as well as flows recorded on networks of dots and last, but not least, data recorded on an individual level.

It will rely on an intelligent data acquisition process, i.e. while importing data sets, some verifications will also be performed. These verifications should assess with a certain degree of confidence the quality of the dataset to be imported: the internal coherence of the new dataset and the overall coherence with the data already stored in the database. Verifications will also allow detect changes in the universe described by the application, like changes in territorial divisions (splitting and merging of territorial units, creation or obliteration of territorial units, etc.), changes in the thematic universe (semantic changes for indicators, the creation of new indicators, etc.). These changes will be registered in the ontologies (the spatial or, respectively, thematic ontology) before being inserted into the databases.

It will propose the importation of both the data and the related metadata (in compliance with the INSPIRE directives) into the application databases and, if necessary, into the application ontologies (this is typically the case when a change occurs in the semantics (definition) of an indicator: it is easier to detect by analyzing the metadata description than by analyzing data).

The integration pool describes the main part of the application, allowing for data structuring (ontologies), data storage (databases) and data harmonization (programming routines) in the application.

The ontologies will play a structuring role for the universe of the application. Their main function is to provide a complete “dictionary” of the universe, describing all the entities of the universe of the application (spatial entities: territorial units, cities, river catchments, etc. and thematic entities, i.e. indicators), and the most relevant relations between them. The ontologies will be implemented using the OWL (ontology Web language), which is currently the de facto standard for ontologies, due to its high expressive power and its automated reasoning potential.

The spatial ontology (see Geo Onto in figure X) will provide a temporal list of synonyms for each entity name (official names, codes and any other type of identifiers linked to statistical and political naming conventions). Each entity in the spatial ontology will be considered from a historical evolution point of view: all the changes in the lifecycle of a spatial entity should be recorded, from its inception to its disappearance. Changes in the state of an entity will be recorded, under the form of splitting, merging, renaming, etc. The spatial ontology will also handle three types of relations between spatial entities: vertical relations (i.e. hierarchical and spatial inclusion relations), horizontal relations (i.e. spatial

neighboring relations) and longitudinal relations (i.e. causal historical relations). Changes in the vertical and horizontal relations will also be stored in the ontology.

The thematic ontology (see Stat Onto in figure X) will provide an exhaustive dictionary of themes, subthemes and indicators, together with precise descriptions of their semantics. Their evolution over time will also be taken into account. The indicators present in the thematic ontology will twofold: raw indicators (that cannot be produced in any way inside the application and are to be imported from external sources) and composite indicators (typically ratios, aggregated and synthetic indicators that can be calculated from other indicators known in the application). Some relations between these indicators should be also represented:

Subsumption relations will represent the semantic inclusion between themes, subthemes and indicators

Calculus relations allow defining composite indicators by means of aggregations or more complex calculi, starting from other indicators (which may be, in turn, raw or composite).

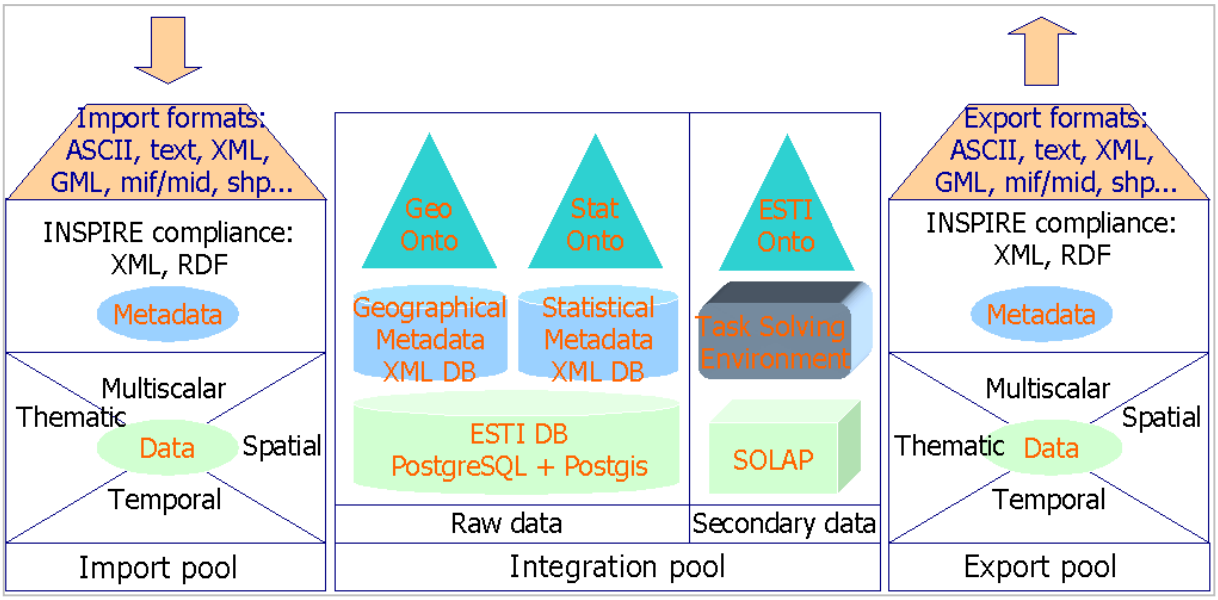


Figure X. Overall architecture of the Espon 2013 DB application.

The ESTI DB is the main database of the application, in charge of the long term storage of the raw data. The main purpose of this database is to allow the best assessment possible about the origin and the quality of the stored datasets. In order to achieve this, for all stored data, the ESTI DB will also keep a rich set of metadata. The database will also allow data redundancies in order to insure better data traceability (lineageability) and coherence:

Multiple indicator values will be allowed for the same spatial entity and the same period of time, as there might be multiple data sources providing different values for the same indicator (Eurostat, national sources, etc.)

Even given one spatial entity, one period of time and one data source, redundancies will be allowed, since statistical organisms tend to re-evaluate (and

re-estimate) permanently indicator values. For instance, Eurostat publishes indicator values for the European regions, but their values are revised on a yearly basis. Studies presenting composite indicators based on Eurostat publications might be incoherent due to the use of simple indicators with different publication years. In order to avoid such caveats, the ESTI DB will allow storing multiple indicator values for the same source, same spatial entity and same indicator, depending on year of publication.

The ESTI database will be implemented using the PostgresSQL open source database management system (or DBMS) and its spatial extension, PostGIS. PostgresSQL is a constantly evolving open source DBMS with excellent performances (compared to the market leading proprietary solutions like Oracle). It offers excellent connectivity with Java applications and it respects the OGC standards (OpenGIS Consortium, the international standardization authority for spatial data and applications). The development community around PostrgresSQL/PostGIS is very active, bug fixes for new versions are provided very rapidly and there is a plethora of free or open source software tools compatible with this DBMS. The choice of an open source DBMS is mainly motivated by its guarantee of maintainability and adaptability: open source software can be freely modified and adapted to the specific needs of any application, which is not the case of proprietary software.

A preliminary data model for the EST DB has already been presented in the answer to the tender. This data model deals with the problems of constantly evolving territorial data (changing names, codes, geometries and hierarchies). In the next phase of the project this model will be extended in order to integrate environmental grid data and constructed point spatial entities (for instance, cities) located inside networks (for instance, roads).

It is important to note that, although the use of rich metadata, of precise data lineage information and of redundancies based on publication date will allow the ESTI DB to insure better data coherence and a more precise data quality assessment, it will also increase the size of the database and it will complexify its structure. This will, in turn, make the ESI DB more difficult to use in a shared environment due to usability issues (non-expert users will probably have difficulties in querying the ESTI DB) and to performance issues (larger dataset with more complex queries will pose more strains on the hardware resources).

In order to overcome these issues, a **simplified database** will be set up in the more advanced stages of the project (see the SOLAP component in figure X). This database can be seen as a repository for a subset of harmonized data exported from the ESTI DB, more useable by non-expert users due to a smaller dataset size and a simpler data schema. Depending on the evolution of the ESTI DB, this database will be implemented either as a traditional relational database, or as a data warehouse, allowing for faster response to queries and on-line analytical processing (OLAP). The database will be implemented in the PostgresSQL/PostGIS DBMS, with the possible addition of an open-source data warehouse server (the choice of the specific software and the desirability of this solution is to be decided in more advanced stages of the project).

The **harmonization method repository** (ESTI Onto in figure X) is a structured set of programming procedures (implemented in Java, based on the JDBC enterprise suite for database connectivity) for data harmonization. These methods will implement the methodology elaborated by the thematic experts in the strategic thinking phase of workpackage B. They will have access to the ESTI

DB in order to allow harmonizing existing raw data. The results of such methods are to be re-injected in the ESTI DB, injected in the simplified database, or exported directly into exchange formats and made available for other applications or projects.

The export pool describes the component in charge of data exportation and dissemination towards external databases and applications, and providing rich data output in different formats. It will allow providing a wide selection of output formats, ranging from proprietary GIS exchange formats (ESRI shape files, MapInfo mif/mid files, etc.) text-based exchange formats (like WKT, the well-known text format), proprietary database tabbed exchange formats (like Microsoft Excel and Access sheets) and XML based exchange formats. It will allow the export of both the data and of the related metadata (in conformity with the INSPIRE directives) for external applications and projects.

Physical hosting of the application and management of server resources

In the early phases of the project, hosting of the application will be ensured by the partner LIG, wich will also be responsible of the data acquisition tasks. In a more advanced stage of the project, once the simplified database will be set up, it will be transferred to the Espon CU in order to be made available online, in order to provide a direct access to data for other Espon projects and external partners. The main database of the application (requiring more database administration expertise, together with constant testing and prototyping) will be hosted by LIG until the end of the project, when the whole application will be transferred to the Espon CU.

5.3 Definition of challenge 9

Analysis of the situation

The Espon 2013 DB application will be an operationalization of the expert knowledge coming from the strategic reflection for the challenges mentioned in the previous section. The main objective is to implement the data acquisition and harmonization tasks usually performed manually by thematic experts so that these tasks can be automatically performed by the computer. Of course, the strategic thinking must reach a high level of maturity and clarity in order for such tasks to be completely transferable to a machine. Given this strong constraint, operationalization of expert knowledge will proceed in an incremental manner.

Deliverables and timetable

Dec. 2008 (Espon Seminar): presentation of our strategy (in terms of data models and ontologies) for answering challenges 1 and 2 (see section 3).

Feb. 2008 (FIR): demonstration of a first prototype including an operational answer to the challenges 1 and 2. Presentation of our strategy (in terms of data models and ontologies) for answering challenges 3 and 4 (see section 5). Challenges 5 and 6 will be tackled in the next period (from March 2008 to Feb 2009).

Teams involved

LIG: coordinator, proposal of data model and methodology, i

RIATE : provide expert knowledge and data for prototyping

ULB : provide expert knowledge on harmonization methodology

Expert 5: consulting on data models and software architecture

6 CONCLUSION : from Inception report to First Interim report

Work plan until the 1st Interim Report.

The general elements of the work plan until the first interim report has been clearly defined in **section 3** (ESPON database strategy) where 8 major challenges are identified with precise allocation of responsibilities to partner, precise objective and agenda for the deliverables to be produced.

Concerning the specific tasks related to mapkit tools, the **section 4** provide some complements on the work to be engage (also described in Challenge 1).

Concerning the computer application (challenge 9) – that depends from all other challenges – the **section 5** provide detailed explanation on the work plan until first interim report.

In the short term the milestones of the agenda are the following one :

- October 2008 : contact with Eurostat / networking with other Espon project / Signature of expert contracts
- November 2008 : preparation of new MapKit tool
- December : Presentation / Discussion of results at ESPON Seminar – Feed backs from ESPON MC and ESPON CU
- January : Project meeting (including experts).
- February : Delivery of FIR

7 ANNEXES

7.1 Annexe 1 : The 10 Commandments for data collection[3]

In this section, we detail the good practices for data collection according to our experience and the recommendation of national and international statistical institutes like for example Eurostat (see box 3)

EUROSTAT defines the quality of statistics with reference to seven criteria:
1. **Relevance:** an inquiry is relevant if it meets users' needs. The identification of users and their expectations is therefore necessary.
2. **Accuracy**: accuracy is defined as the closeness between the estimated value and the (unknown) true value.
3. **Timeliness and punctuality in disseminating results**: most users want up-to-date data which are published frequently and on time at pre-established dates.
4. **Accessibility and clarity of the information**: statistical data are more valuable when they can be easily accessed by users, when they are available in standard formats, and when they are adequately documented.
5. **Comparability**: statistics for a given characteristic have the greatest usefulness when they enable reliable comparisons of values through space and time. The comparability component emphasizes the comparison of the same statistics between countries in order to evaluate the meaning of aggregated statistics at the European level.
6. **Coherence**: when originating from a single source, statistics are coherent in that elementary concepts can be combined reliably in more complex ways. When originating from different sources, and in particular from statistical surveys of different frequencies, statistics are coherent as long as they rely on common definitions, classifications and methodological standards.
7. **Completeness:** domains for which statistics are available should reflect the needs and priorities expressed by the users of the European Statistical System.

Source : http://forum.europa.eu.int/irc/dsis/coded/info/data/coded/en/gl011043.htm

Box 1. Definition of quality of statistics by Eurostat

Previous experiments led by many European research projects have clearly shown that the **quality** of the database is much more important than the **quantity** of information. It has been proved many times at European level that a small database perfectly integrated is much more useful and efficient than a wide database which is a simple compilation of heterogeneous indicators built without applying any rules of quality control.

[3] This annex is an extract from the "ESPON HANDBOOK FOR DATA COLLECTION, HARMONISATION AND QUALITY CONTROL" realized by ESPON Project 3.2 at the end of the ESPON 2006 Programme. It provides useful recommendations for data collection in ESPON 2013 Program.

Derived from the practice of ESPON 2002-2006, we can propose 10 "commandments" which define the general rules to be followed in order to build an objective, efficient, harmonized and evolutive database.

AN OBJECTIVE DATABASE

Commandment 1: "You shall always use the most primary sources of information"

Commandment 2 : "If you can not use primary sources, you shall indicate precisely the path leading to the data collection"

The precise identification of the **initial sources** of information (census data, survey…) is an absolute necessity for the quality control of information. In any case, it should be possible to identify **the path** of elaboration of any figure of the ESPON database, from tertiary or secondary sources to primary (initial) sources of information. Each transformation or modification of primary sources should be clearly identified and registered when secondary sources are introduced in the ESPON database.

For instance, most data used by ESPON are secondary sources, because they were download on Eurostat website but are derived from initial sources which are national census elaborated in each member state. In this case, a precise identification of sources would imply the storage of the references to each national statistical institute in charge of the production of census data and the precise time of these censuses. But it is not necessary because this work of documentation has most of the time been perfectly done by EUROSTAT. Then, in this very precise case, it is sufficient to indicate the reference of EUROSTAT where all those precisions are available. But in other cases, databases are tertiary or quaternary sources.

For example, some data about European regions can be download on the website of the French Observatoire des Territoires[4] but these data are derived from Eurostat data which are derived from national sources … and are therefore tertiary sources. This data are correctly documented but should not be used by a TPG which should use original sources.

[4] http://www.territoires.gouv.fr/

In some cases, data are not documented at all, as the famous data of states of the world proposed by CIA under the name of "*The World Factbook*[5]". This publication of the american government proposes an extraordinary database with 200 to 300 indicators describing all states of the world (at least, the one which are recognised as such by United States …). But the origin of data are never precisely described and only administration of the United State is mentioned as data provider. The real producers from first level (states) or second level (United Nations) are not indicated. It means that the user of this data has no possibility to check and verify the figures: either he believes what is said by the US Government, either he does not. This is a typical violation of our 1st and 2nd commandment (see Box 4).

The World Factbook is prepared by the Central Intelligence Agency for the use of US Government officials, and the style, format, coverage, and content are designed to meet their specific requirements. Information is provided by Antarctic Information Program (National Science Foundation), Armed Forces Medical Intelligence Center (Department of Defense), Bureau of the Census (Department of Commerce), Bureau of Labor Statistics (Department of Labor), Central Intelligence Agency, Council of Managers of National Antarctic Programs, Defense Intelligence Agency (Department of Defense), Department of Energy, Department of State, Fish and Wildlife Service (Department of the Interior), Maritime Administration (Department of Transportation), National Geospatial-Intelligence Agency (Department of Defense), Naval Facilities Engineering Command (Department of Defense), Office of Insular Affairs (Department of the Interior), Office of Naval Intelligence (Department of Defense), US Board on Geographic Names (Department of the Interior), US Transportation Command (Department of Defense), Oil & Gas Journal, and other public and private sources
Source : https://www.cia.gov/cia/publications/factbook/docs/contributor_copyright.html

Box 2. Example of criticable sources : the CIA World Factbook

AN EFFICIENT DATABASE

Commandment 3: "You shall always collect the raw count variables rather than ratio or other indexes derived from their combination"

Commandment 4: "When you are not able to provide raw count variables, you shall indicate the weight to be used for aggregation".

In many cases, the indicators used for territorial planning are a mathematical combination (addition, substraction, division, product…) of raw count variables which are not directly useful but are, in practice, the kernel information from which all indicators are directly or indirectly derived. A good database structure should absolutely store those kernel

[5] https://www.cia.gov/cia/publications/factbook/index.html

variables (real information) and not necessarily store the derived indicators (virtual information) which can be automatically computed when request.

In a short term perspective, this principle may seem heavy to apply. For instance, if one want to use a variable like the median age of population in NUTS3 regions, one has (1) to store all the age structure of those NUTS 3 regions and (2) to store the formula of median age computation in a SGBD. Apparently, this solution is time and human resources consuming, but, eventually it produces a very important gain of time, ressources and quality in a long term perspective because:

- The spatial aggregation or disaggregation of data is then much easier. In the case of the median age, it is impossible to estimate the values at NUTS 2 level if we have stored only the median age at NUTS 3 level, even if the result is weighted by population (it would be possible with mean age of population, but not with median). But it is very easy to aggregate all age structure (which are count variables) from NUTS 3 to NUTS 2 and then to apply the formula of median age computation which has been stored for NUTS 3 and remain available at NUTS 2.
- Many indicators are derived from the same kernel variables: which means that with a limited number of good kernel variables, it is possible to produce a very wide set (probably infinite) of indicators and derivated variables. Storing kernel variables can favour the production of new indicators which will not be possible if this kernel information had not been stored. Imagine for example that a TPG has produced the indicators Z1=(A/B) and another TPG the indicator Z2 = (C-D)/E, the strategy of kernel indicators (storage of A,B,C,D,E) make possible the construction of many other indexes like (C-D)/B or A/E which would not have been possible if we had only stored the indexes Z1 and Z2.

- Statistical tests are generally not available or biased if the kernel information is not available in the database structure: in the very simple example of GDP/inh., a good statistical evaluation of heterogeneity can not be made by a simple comparison of regional ratios but implies a direct examination of the unequal repartition of the raw count values of population and wealth. Generally speaking, the use of ratio is very dangerous in statistical analysis because results are non weighted and subject to random variations in small areas. Keeping the initial count variables from which ratio are derived is a necessary condition for a correction of those biases.

A HARMONIZED DATABASE

Commandment 5: "You shall always explain precisely your procedure for time harmonization".

Commandment 6: "You shall always explain precisely your procedure for territorial harmonization".

Commandment 7: "You shall always explain precisely your procedure for thematic harmonization".

Time harmonization

In the European context, especially if we take into account the enlargement of databases from 15 to 27 countries, it is not possible to use primary sources without modifications and harmonization. An obvious example is related to the census year and date which are different in most European states. Thus, if we want to evaluate the regional distribution of population at 1st January 1990 or 2000 for all European regions, we will be necessarily obliged to introduce estimations for many states which have a different census time. Those estimations are not a problem as far as the estimation procedure is clearly indicated in the database structure.

An ideal situation would be the storage of data and formula used in order to produce harmonized information derived from kernel information. For instance, if we want to evaluate the population of France regions in 1980 starting from the census variables of 1975 and 1982, we should store (1) the regional populations of France in 1975 and 1982 (primary kernel information) and (2) the precise formula used for the estimation (linear, exponential, …) of population in 1980 from population in 1975 and 1982.

It is necessary to keep in mind that those rules applies only to raw count variables and are not necessary for indicators which are derivated from combination of raw count variables variables. If we take the example of a regional index of median age of population in France in 1980, the database should indicate that it is the result of a formula applied to age structure in 1980, which is derived from a formula of interpolation derived from the age structure in 1975 and 1982 …

Territorial harmonization

Another crucial problem is related to the harmonization of European territorial divisions, which is not only a technical problem but also a political problem, as far as the attribution of structural funds is related to particular values (or thresholds) established at a particular level of territorial division. Practices of "gerrymandering" or simple evolution of regional divisions are responsible for a very difficult problem in the establishment of long term time-series at regional level. Users of regional

statisticis in Europe unfortunately have to face with gaps, discontinuities and breakdowns in regional time series, which leads to a very pernicious lack of efficiency in the production of study on trends at European level. It has been many times underlined (especially in ESPON project 3.2) that it is impossible to make good territorial prospective (prospective trends) if it is impossible to compare present situations to past ones (retrospective trends)

It is thus necessary to distinguish between "official maps at a given time t" which are required to use the official delimitation of the period t, and "trend maps on a long period [t1, tn]" which can not use official delimitations because of regular changes in the official delineation of political and administrative boundaries. If the structure of the ESPON database is not able to take into account the evolution of territorial division, then one has to face the dilemma presented in Figure 5: either a great diversity of indicators with high spatial resolution, but only for a very short period of time, is available, or longer time series, but only for a limited number of indicators and with low spatial resolution are present.

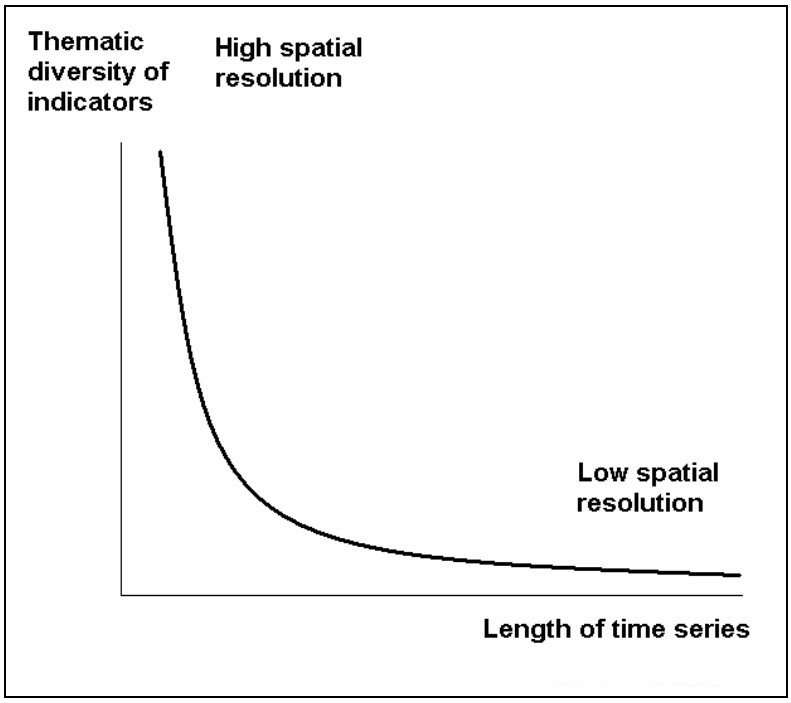


Figure 1 The dilemma of time vs thematic harmonization

The structure of the ESPON database used in the period 2002-2006 was typically subject to this dilemma. And the choice was made to have a high spatial resolution (NUTS 3) and a great diversity of indicators but with very short length of time series (generally less than 5 years). An alternative structure of database can be proposed for ESPON 2013 called "Long Term Data Base" which can allow to avoid the dilemma and to produce at the same time long term time series for a limited number of indicators.

Thematic harmonization

In addition to the problem related to temporal and spatial harmonization of European databases, it is necessary to focus also on the problems linked to the harmonization of definitions used in European states. Possible biases may be introduced by differences in the definition of variables and/or collection of information procedures.

A classical example of those biases is related to the measure of the infant mortality rate. At first sight, this index is based on a very precise definition (ratio between death between 0-1 year and number of birth) but in fact many problems of harmonization have been revealed by demographers. In certain states, for instance, children who die in their 1st day of life were not recognised as "birth" (and thus, neither as "dead between 0 and 1"). Accordingly, they were classified as "birth dead" and not taken into account in the computation of the infant mortality rate, which introduces an important reduction of the ratio, when compared to other states.

In the case of ESPON, one has to be very careful about those possible biases and it is important to store in the database the possible biases related to a lack of harmonization in the definition, as well as statistical system practices when those biases are established by experts. According to the ESPON Guidance Papers, biases related to data issues should be pointed out and explained by the TPG in all ESPON final reports in a specific section on data issues. For many crucial subjects (unemployement, R & D, accessibility, …) this precise criticism of indexes is, in a sense, more important than the value of indexes stored in the database and represent the real added value of the ESPON Programmeme. We have illustrated this point through several examples in Chapter 3.1.

For all commandment 8, 9 and 10, it is important to mention the existence of an ESPON template on a database agreement to be applied each time that an ESPON project needs to use a significant part of a database. This agreement is available under request to the ESPON Coordination Unit.

AN EVOLUTIVE DATABASE

Commandment 8: "You shall propose indicators that can be further collected by others than yourself".

Commandment 9: "You shall propose indicators that can be further update in the future"

Commandment 10: "You shall propose indicators that can be further collected in new countries"

According to the previous principles, we intend to propose the design and development of an open and evolutive database which will be probably limited in a first step, but can be further developped geographically, historically and thematically in order to produce a cumulative knowledge base on European Spatial Developement.

An open database means that every researcher involved in the ESPON Programmeme will be able to contribute to its development (as a data provider or as an expert) and that, conversely, all researchers involved in the ESPON Programmeme will have the right to use this database for the purpose of the program. This interactive open access of all ESPON research members to the ESPON database may be technically complicated (problems of security). However, to our opinion, it is above all a guarantee of quality of the results since all indexes involved in the ESPON database will be subject to the collective evaluation by a community of more than 200 researchers of all the European Union and accessing or neighbouring countries.

Reseach teams working for the ESPON Programmeme should therefore avoid the use of "private" databases which are their own property and that they are not ready to deliver to the community of other ESPON researchers. Moreover, they should always contact the ESPON coordination unit whenever they discover a database of particular interest that could be integrated into the ESPON database.

To illustrate, when the TPG ESPON 3.4.1. Europe in the World has decided to use an harmonized historical database on population and GDP of states of the World published by OECD and achieved by the independent researcher Angus Madison, the decision has been taken to negociate the use of this database for all ESPON members and not only for the lead partner or research teams of the ESPON project 3.4.1. As a result, future TPG of ESPON 2013 which will produce research on Europe in the World or historical evolution of European economy will benefit from the integration of this data in ESPON database.

A more difficult question is related to indicators which are derived from mathematical model or specific databases belonging to research units. A good example is provided by the indicators of accessibility produced by ESPON project 1.2.1 *Transport*. What is at stake here is how such indicators could be updated in the future if there are some evolution in the transportation network, or if the regional delimitations changes, or if the territorial scope of ESPON is enlarged (inclusion of Turkey or Balkanic countries), etc. It is obvious that what should be stored in the ESPON database is not only the regional scores of accessibility but also the transportation network and the program/algorithm used for the computation of accessibility.

In ESPON 2013, the focus should not be put on the **quantity** of indicator produced by TPG's but rather on their **quality** which includes the possibility to update values regularly in the past and the future, and to enlarge geographically their coverage when needed.

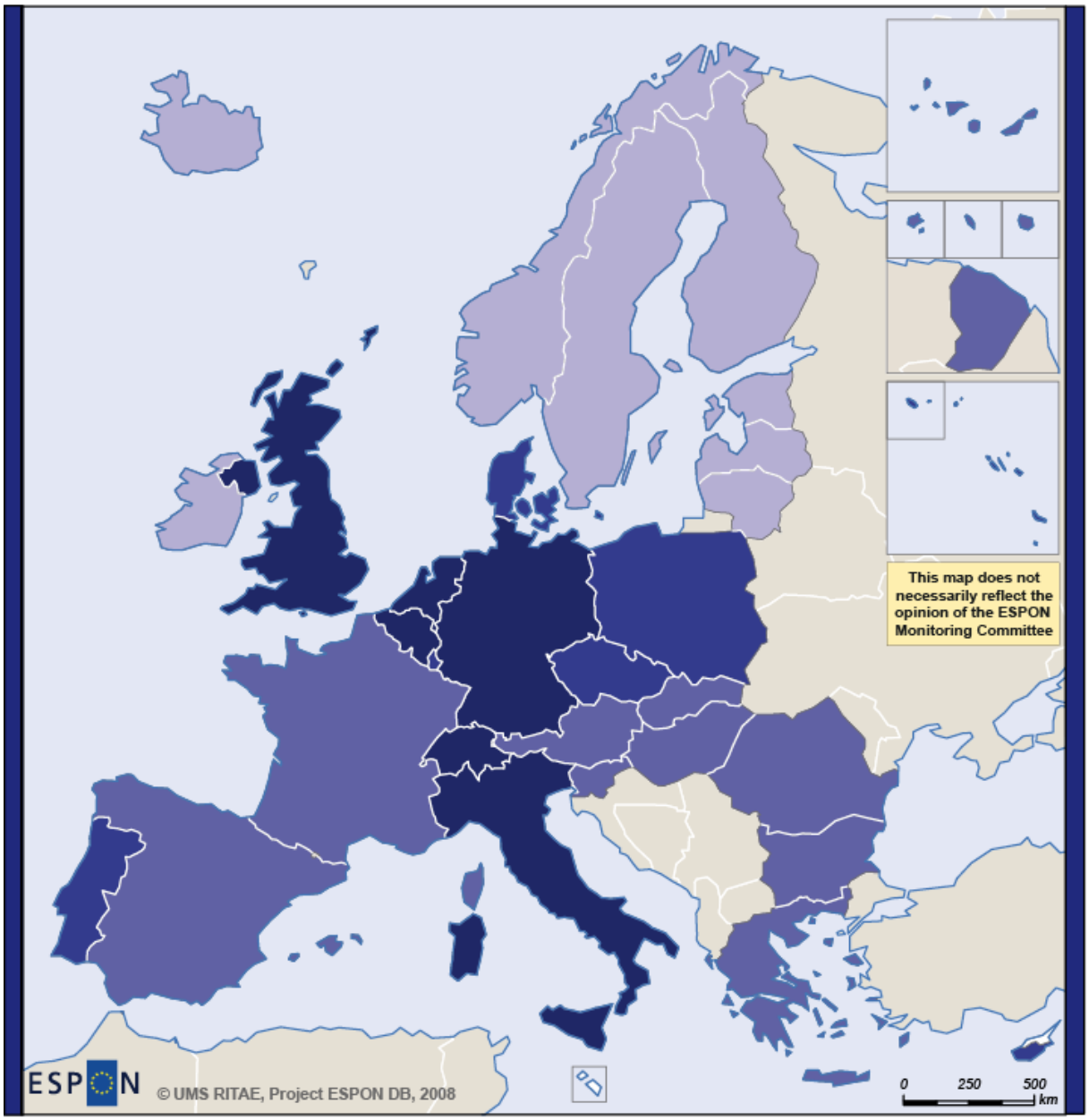
7.2 Annexe 2 : Sample of maps realized with the Mapkit tool

The ESPON Coordination Unit has expressed the wish to add a set of maps realized with the new map toolkit in order to be able to present it to ESPON Monitoring Committee. This concern in particular the discussion about map generalization proposed by RIATE.

We propose therefore :

- 5 maps of population density in 2003 at different NUTS level
- 1 map of population combined with GDP/capita in 2003
- 1 map of population combined with natural area/inhabitant in 2000.

Population density, 2003



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

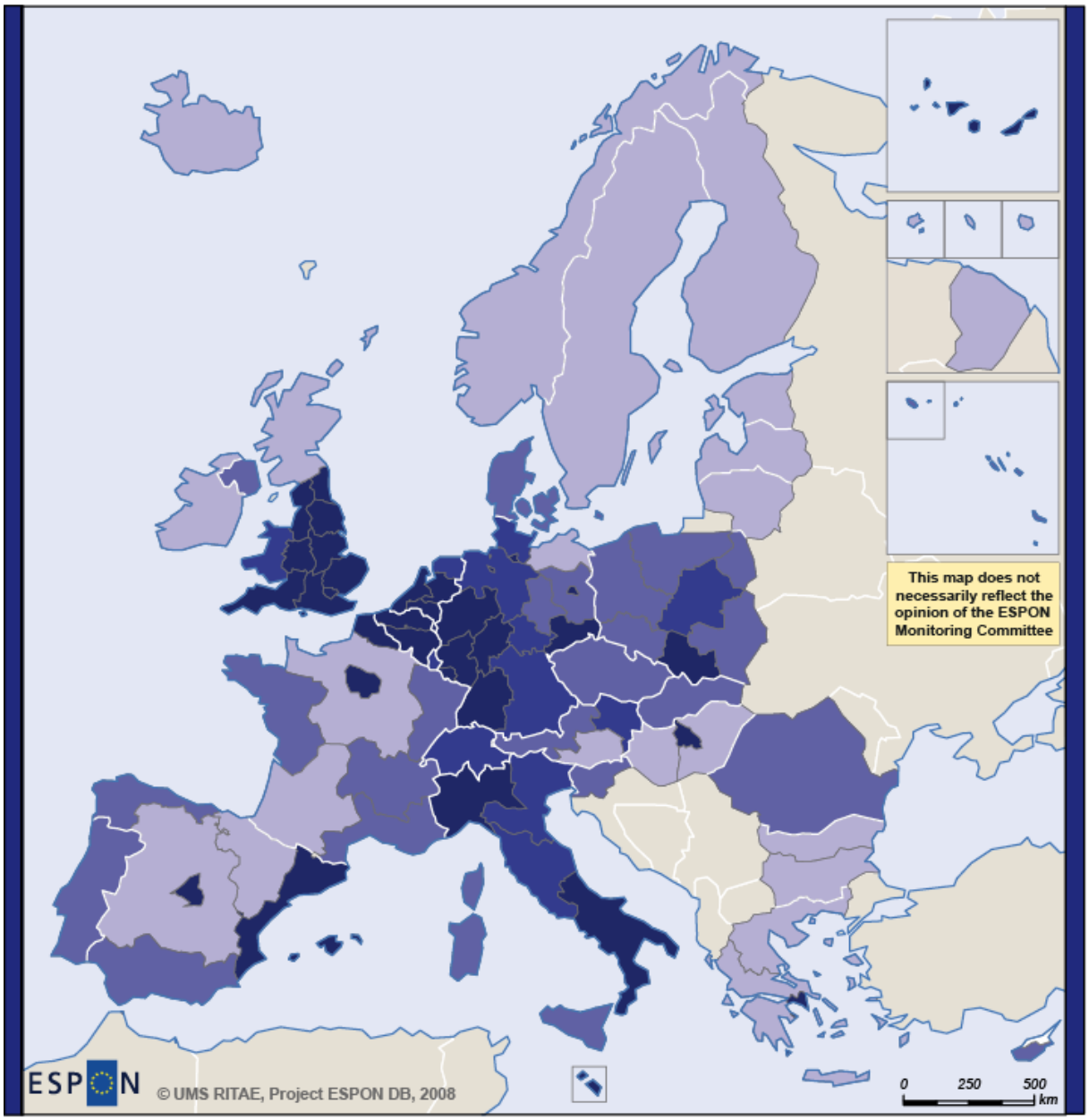
Population density
(inh per km²)

- 2
- 70
- 110
- 170
- 391

Source:
ESPON 2013 Database
Source for administrative boundaries:
UMS RIATE
Origin of data:
© European Communities, 1995-2008, 2007

NUTS 0 - 2003

Population density, 2003



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

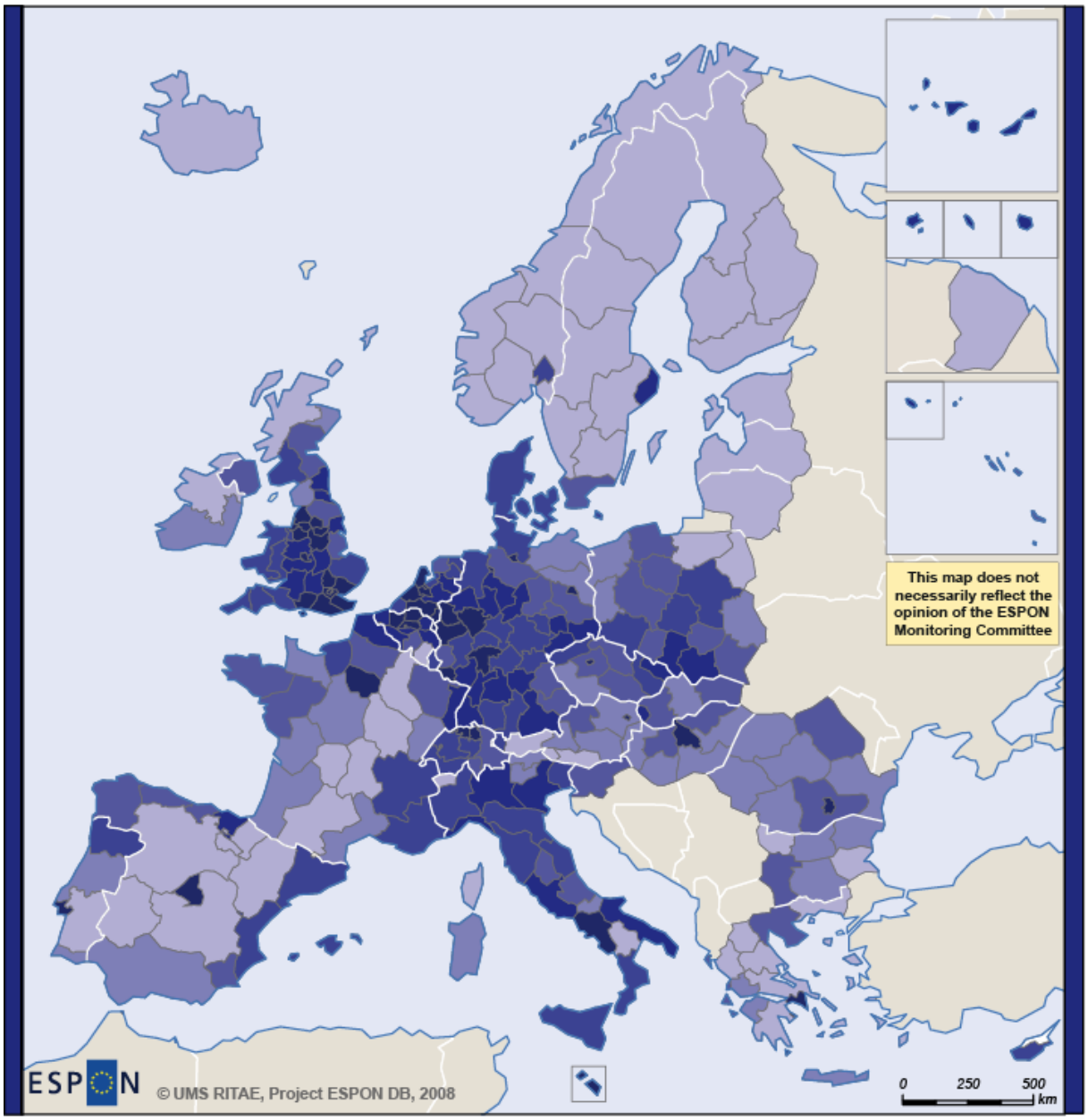
Population density
(inh per km²)

- 2
- 85
- 140
- 190
- 6 186

Source:
ESPON 2013 Database
Source for administrative boundaries:
UMS RIATE
Origin of data:
© European Communities, 1995-2008, 2007

NUTS 1 - 2003

Population density, 2003



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

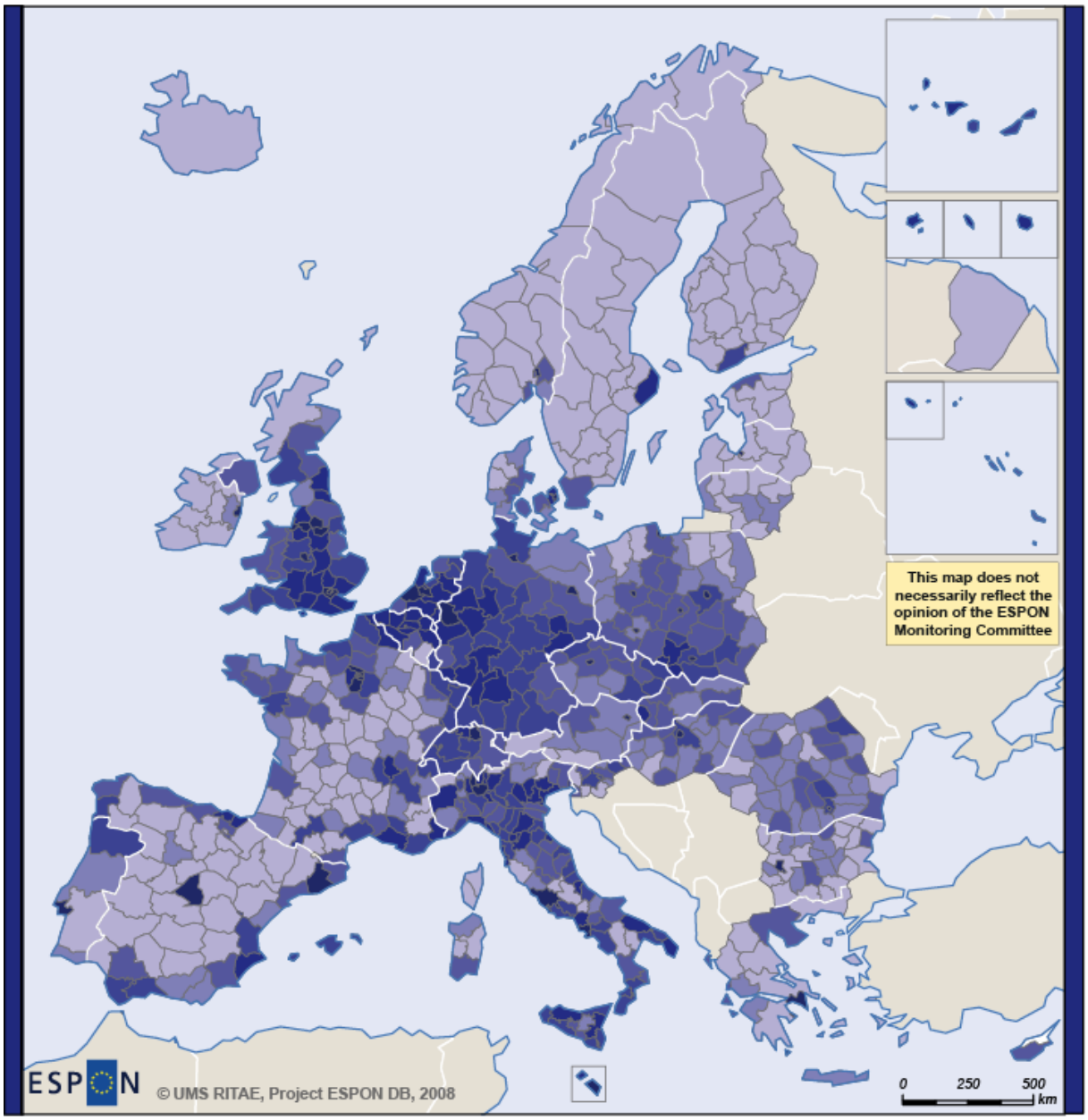
Population density
(inh per km²)

2
60
90
125
205
390
9 073

Source:
ESPON 2013 Database
Source for administrative boundaries:
UMS RIATE
Origin of data:
© European Communities, 1995-2008, 2007

NUTS 2 - 2003

Population density, 2003



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

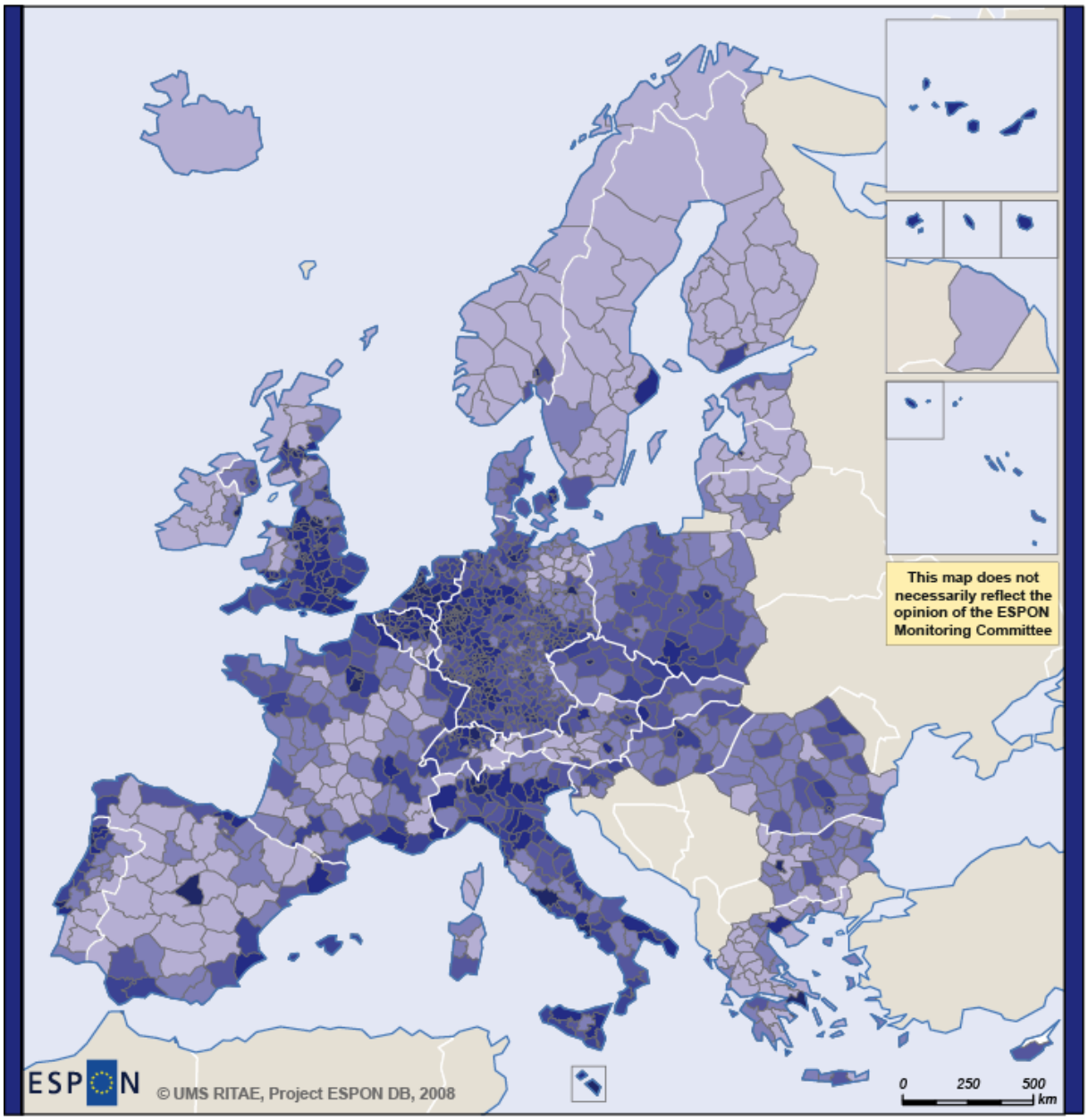
Population density
(inh per km²)

1.5
60
90
145
245
640
20 548

Source:
ESPON 2013 Database
Source for administrative boundaries:
UMS RIATE
Origin of data:
© European Communities, 1995-2008, 2007

NUTS 2/3 - 2003

Population density, 2003



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

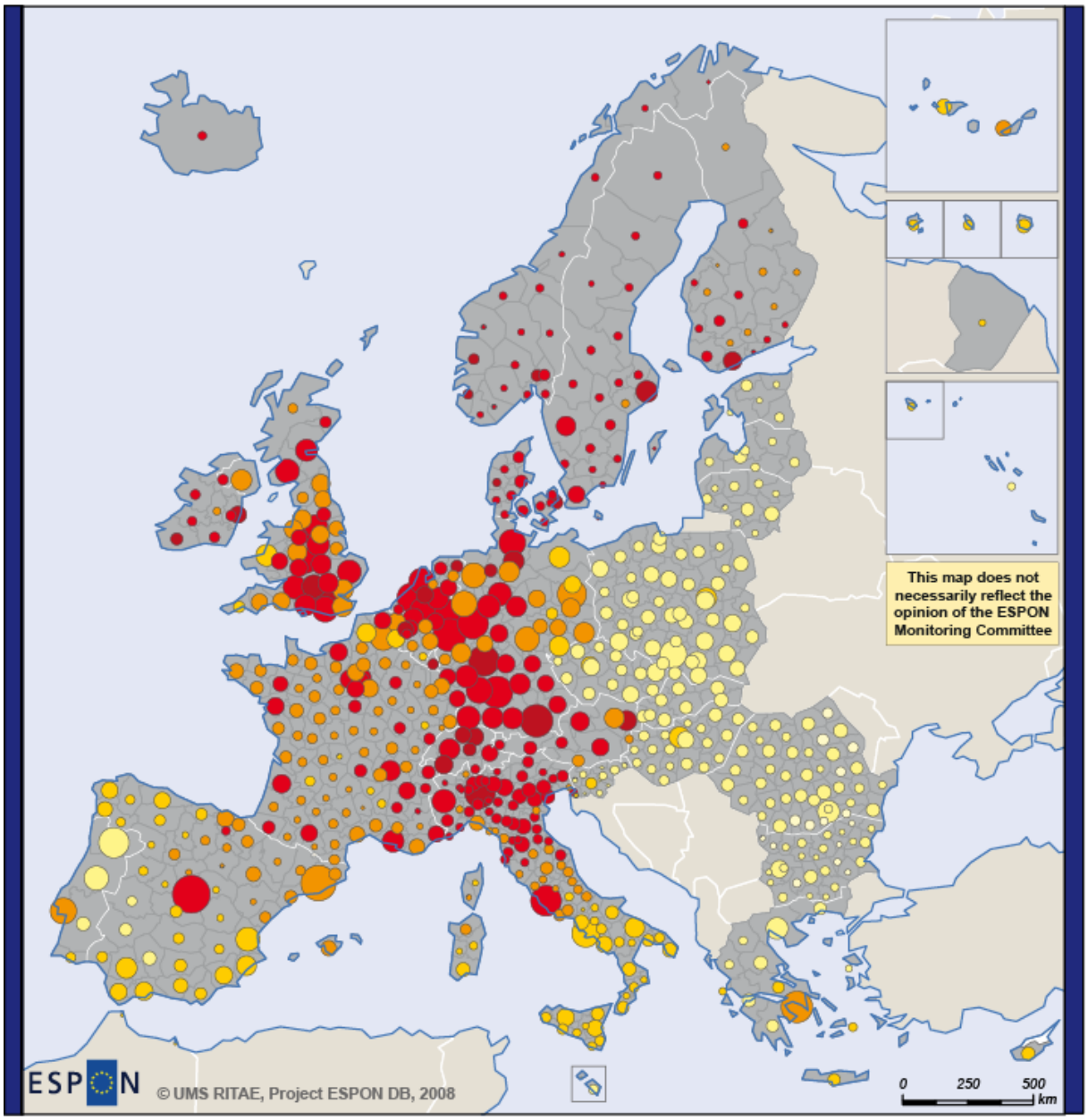
Population density
(inh per km²)

- 1.5
- 50
- 90
- 140
- 235
- 645
- 20 548

Source:
ESPON 2013 Database
Source for administrative boundaries:
UMS RIATE
Origin of data:
© European Communities, 1995-2008, 2007

NUTS 3 - 2003

GDP per capita, 2003



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

GDP per inh
(euro/inh)

- 1 290
- 1 660
- 12 670
- 18 170
- 23 675
- 34 680
- 72 221

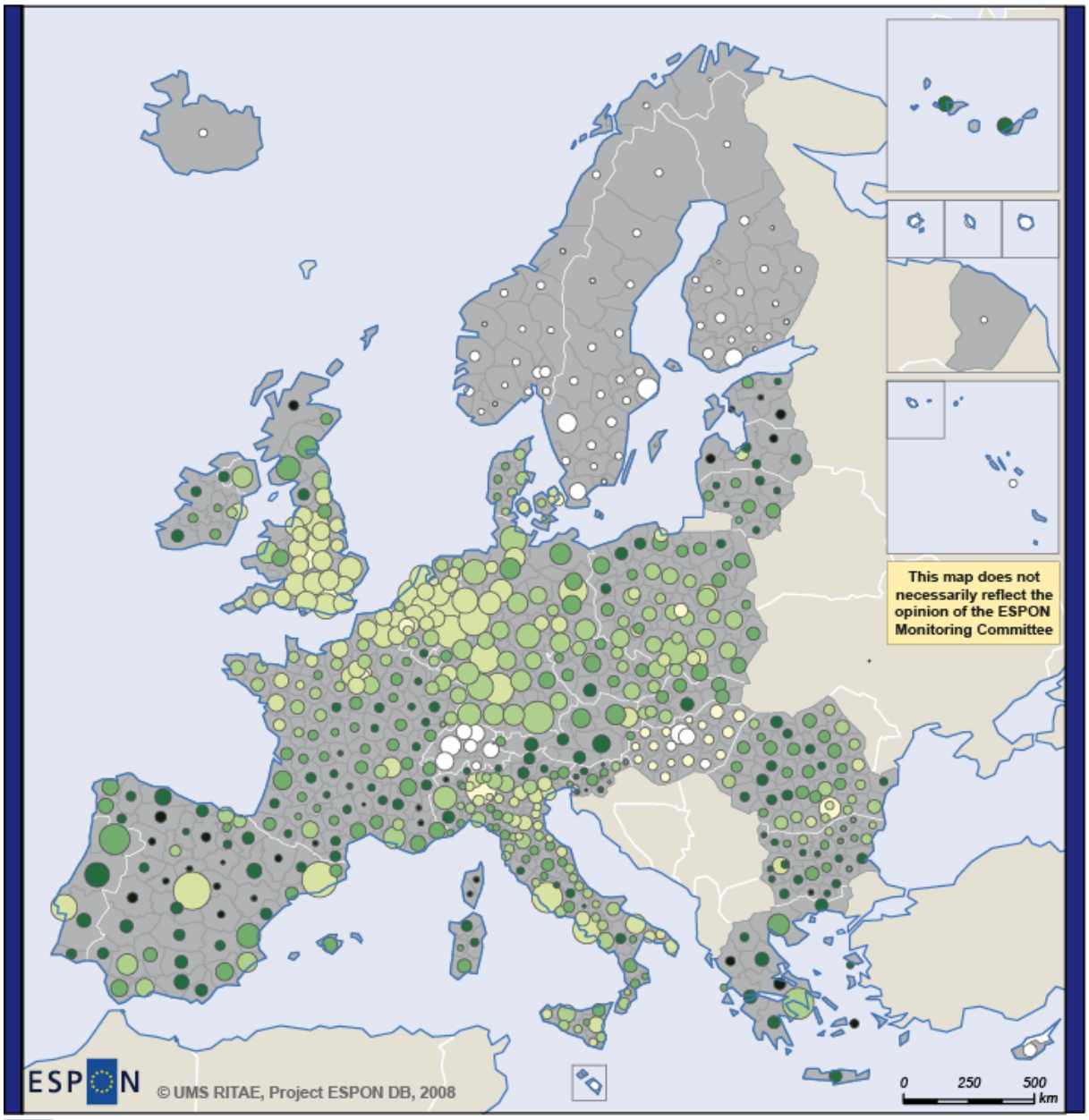
Total population
(thousands inh)

- 5 640
- 3 000
- 1 000
- 500

Source:
ESPON 2013 Database
Source for administrative boundaries:
UMS RIATE
Origin of data:
© European Communities, 1995-2008, 2007 and 2008
ESPON Database - April 2007

NUTS 2/3 - 2003

Semi-natural and natural land per capita, 2000



EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

Total semi-natural and natural land, 2000 *(ha/ thousands inh)*

- 0
- 3
- 111
- 283
- 662
- 2 000
- 9 635

Total population, 2003 *(thousands inh)*

- 5 640
- 3 000
- 1 000
- 500
- No data

Source:
ESPON 2013 Database
Source for administrative boundaries:
UMS RIATE
Origin of data:
© European Communities, 1995-2008, 2007 and 2008
EEA, 2007

NUTS 2/3 - 2003