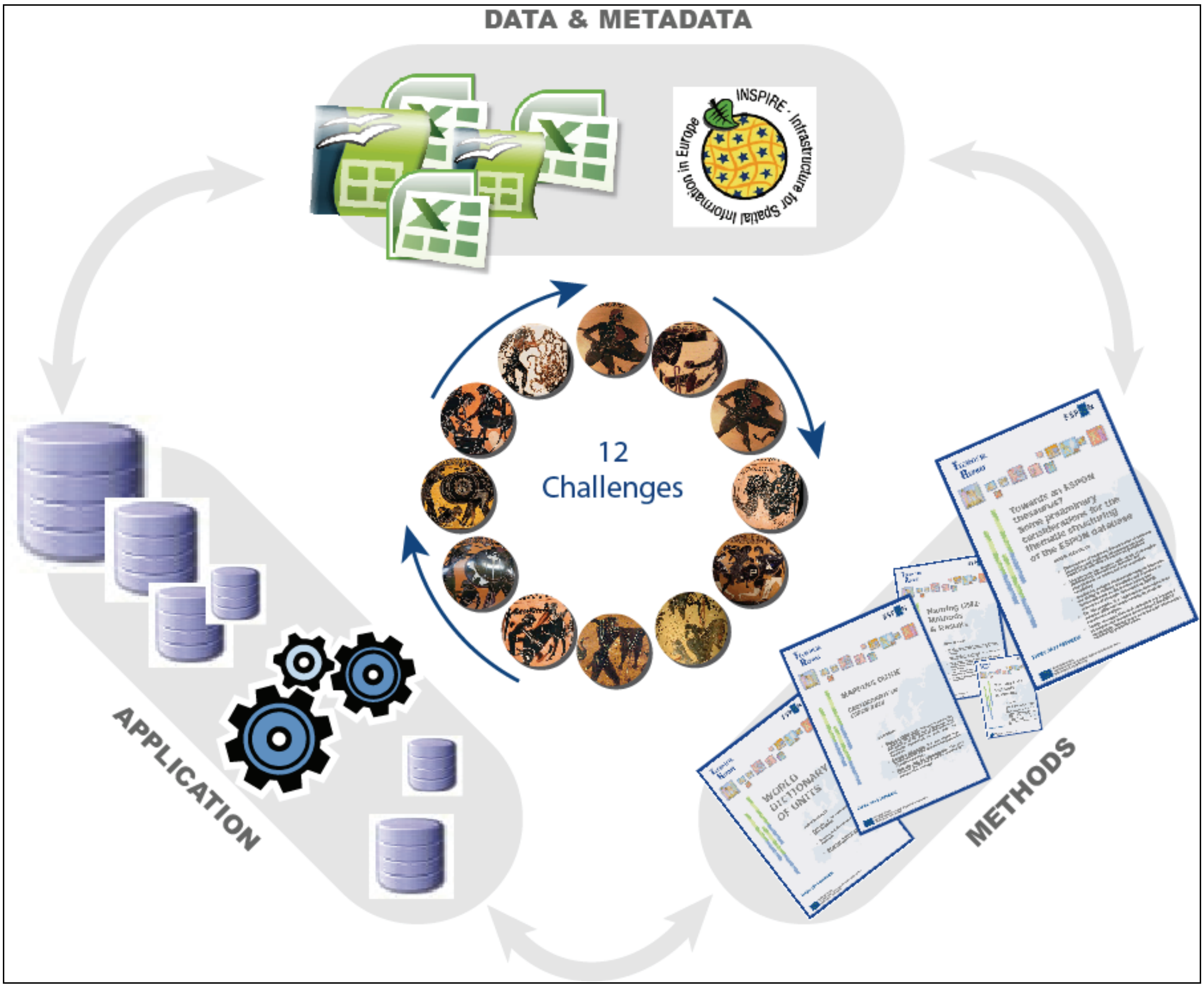
ESPON 2013 DATABASE

SECOND INTERIM REPORT

2010 February 26



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

This first interim 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 first interim report

UMS RIATE (FR)
Claude Grasland*
Ben Rebah Maher
Ronan Ysebaert
Christine Zanin
Nicolas Lambert
Bernard Corminboeuf
Isabelle Salmon

LIG (FR)
Jérôme Gensel*
Bogdan Moisuc
Christine Plumejeaud
Marlène Villanova-Oliver
Anton Telechev
Benoît Le Rubrus

UAB (ES)
Andreas Littkopf
Juan Arevalo
Roger Milego
Maria-José Ramos

IGEAT (BE)
Moritz Lennert
Didier Peeters

UMR Géographie-cités (FR)
Anne Bretagnolle
Hélène Mathian
Timothée Giraud
Marianne Guerois

TIGRIS (RO)
Octavian Groza
Alexandru Rusu

Université du Luxembourg (LU)
Geoffrey Caruso
Nuno Madeira

National University of Ireland (IE)**
Martin Charlton
Paul Harris

National Technical University of Athens (GR)**
Minas Angelidis

Umeå University (SE)**
Einar Holm
Magnus Strömgren

UNEP/GRID (CH)**
Hy Dao
Andrea De Bono

* Scientific coordinators of the project
** Experts

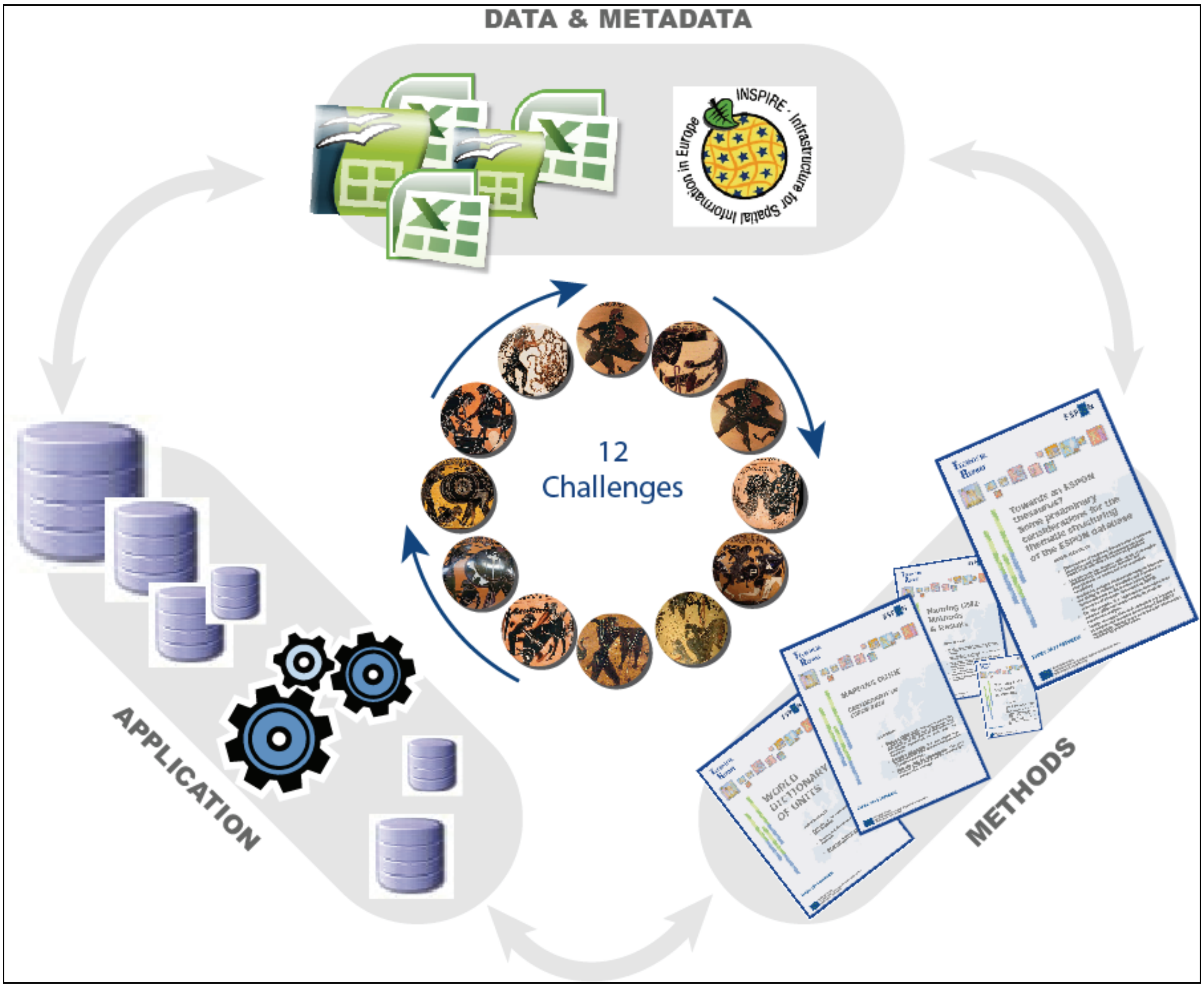
TABLE OF CONTENTS

1 Introduction

1.1 Overview of the project

The Figure 1 presented at the ESPON meeting in Malmö proposes a synthetic view of the division of work inside ESPON DB project and progress made since the First Interim Report

Figure 1 : Overview of ESPON DB Project



The 12 challenges have been the core of the project since the beginning. They provide a simple and efficient division of work between partners and experts, each of them being responsible for one challenge, eventually possibly in association with others. But challenges will have to be integrated in a more synthetic way in the second part of the project, which is illustrated by the three work areas defined as Methods, Application, Data and Metadata.

Data and metadata. The amount of data present in the ESPON database is the most obvious output of a project called "Database". It is also the easiest way to evaluate progress made at ESPON level because it includes both basic data collected by ESPON DB project itself, and other data collected by all ESPON projects. But it is important, in our opinion, to insist on the fact that ***metadata are probably more important than data itself***. More precisely, it is not useful to enlarge the ESPON Database if data are not very accurately described, (definition, quality, property copyrights). We acknowledge that the elaboration of such metadata was not an easy task, both for the ESPON DB project and for other ESPON projects and we apologized for that at the Malmö meeting. But we are convinced that, without this collective effort, the sustainability of the ESPON program will not be ensured on the long run.

Methods, presented in the form of standalone booklets called ***Technical Reports*,** are the methodological supports of data and metadata and represent the second major contribution of the ESPON DB project. In the 12 challenges, we have explored a great number of options that could enlarge the scope of data collected and used in the ESPON project. This knowledge was produced by the ESPON DB project itself with many inputs from other ESPON projects dealing with specific geographical objects (e.g. FOCI for urban and local data; Climate Change and RERISK for Grid Data; DEMIFER or EDORA for time series at NUTS2 or NUTS3 levels; the priority 2 projects for local data). Technical Reports focus on questions that are regularly asked in ESPON projects and try to summarize collective knowledge. Some Technical Reports provide clear solutions. Some identify shortcomings or dead-ends. Others focus on questions of cartography, in particular the mapping guide that has been made available on the ESPON website[1].

Applications are different computer programs elaborated by project partners for data management, data query or data control. It is important to understand that ESPON database is not made of a single application doing everything, but of a set of interlinked applications with different purposes in the data integration process. Many misunderstandings appeared in the beginning of the project in relation with this issue and many efforts were made to clarify the vocabulary. A basic distinction has to be made between ***an interface for query*** that will be made available on the ESPON website in March 2010 and an ***application for data management.*** The second one is the interface "back office" but it also fulfills more general objectives of data integration. These two major applications are designed and implemented by the computer science research team LIG, but it is important to note that other partners and experts of the project contribute

[1] http://www.espon.eu/main/Menu_ScientificTools/MappingGuide/

to this work. In particular, the UAB team has contributed to the elaboration of the *metadata editor* with LIG. It has also developed the ***OLAP program for NUTS to GRID*** conversion. The UL team has adapted a ***specific program of text mining*** for the elaboration of ESPON Thesaurus. The experts of NCG have developed ***application for outlier detection in R language***. Finally, the expert team UNEP-GRID is building a specific program for the benchmarking of data at State level provided by UN and Eurostats, etc.

Even if a wide set of ambitious options have been explored during the first period of the ESPON DB project, it is true that during a certain period of time our project has been working more profitably on the building of solid foundations than on the delivery of final results. (Figure 2)

Figure 2 : The ESPON DB Project at the beginning period

1.2 Organisation of the Second Interim report

As in the case of the First Interim Report (FIR), the aim of this Second Interim Report (SIR) is to produce a short report where only major information is reported. Every technical development is put in annex in the form of technical reports.

The review of progress made by challenges (Part 2) is the core part of the report that provides synthetic information on the work done since the FIR. A first group of challenges is related to the production of specific datasets or specific expertise on different types of geographical objects: collection of basic data at regional level (2.1), harmonization of time series (2.2), enlargement of regional data towards global (2.3) or local (2.4) levels, combination of social and environmental data (2.5), and collection of urban data (2.6). A second group of challenges is more closely related to data integration process in order to build an integrated data model that can be implemented as a computer application (2.7). The involvement of the expert teams is related to the specific description of new challenges: spatial analysis tools for quality control (2.8), collection of data on neighbouring countries (2.9) and exploration of individual data and surveys (2.10).

The work plan until the final report (Part 3) is a description of tasks that will be achieved during the last period of activity of the ESPON DB 2013 project. It is organized by challenge, as in part 2, in order to facilitate the evaluation of work achieved. At the project midpoint we have decided to stop the exploration of innovative ideas and to focus mainly on the consolidation of results achieved so far. The technical reports will be updated and a final version will be delivered within the final report.

The perspectives and needs for further improvements (Part 4) are tasks that ***will not*** be achieved during the actual ESPON DB 2013 project but that have been identified as important for our successors during the 2011-2013 period. This is not an exhaustive list and it has to be completed by the ESPON Coordination Unit, other ESPON Projects and stakeholders (EEA, Eurostat, OECD)…

The Draft Technical Reports (DTR), annexed to the Second Interim Report**,** are a full part of the present SIR but are also considered as non final versions, or "work in progress". Each challenge is improving this document and it is only with the Final Report that they will be considered as definitively achieved.

1.3 Coordinator’s message

The coordinators of the ESPON DB 2013 project, Claude Grasland (UMS RIATE) and Jérôme Gensel (LIG) take the opportunity of the present SIR to address a message to the ESPON Coordination Unit and the ESPON Monitoring Committee.

Many progresses have been made in ESPON 2013 concerning data flow

The ESPON DB 2013 Project, in partnership with other projects from Priority 1 (TIPTAP, EDORA, DEMIFER, FOCI, RERISK) and Priority 3 (Demography, Accessibility, Lisbon Indicators, Typology, …), has elaborated a substantial database on European regions and cities, with very important added value for policymakers working on territorial cohesion. This database, that will be available on the ESPON website in March 2010 through an innovative computer application, will play a major role in the promotion of ESPON network and ensure a wider diffusion of results presented in the form of papers. At the same time, ESPON has developed stronger partnerships with data providers (Eurostat, EEA, National Statistical Agencies,) and data users (DG REGIO or DG AGRI).

The ESPON 2013 Program as a whole is starting to be recognized as an important player in the field of databases at the European scale. The contribution of ESPON DB 2013 Project to this recognition has been crucial on several points:

A very strict definition of rules concerning metadata and quality check: this goal has been extremely time consuming (as INSPIRE directive and ISO norms were not directly applicable to many data used in ESPON). Even if it was a difficult constraint for our project, as for the other ESPON projects, the strict codification of metadata is absolutely crucial for ESPON external recognition.

The integration of various types of geographical objects : even if regional data (NUTS2 and NUTS3) remain actually dominant in the ESPON Database project, this one has been designed in order to open the door for data elaborated at upper and lower scales (World by states, local units) and for data using different geometries (cities, networks, …).

The attempt to enlarge time series towards past and future: as spatial planning is necessarily dynamic and prospective, we cannot limit our investigation to a short term period. But it has been demonstrated many times that it is impossible to enlarge future previsions (t+20 years) without an equivalent gain of information on past trends (t-20 years).

… but many difficulties have also been encountered …

The first set of difficulties that we have faced with this project was related to the **ESPON agenda** and the fact that our Priority 3 projects started at the same time than other Priority 1 projects (*DEMIFER, FOCI, TIPTAP, EDORA, RERISK*) and data release (*Demography, Accessibility, …*). Starting 6 months before the other projects, would have allowed delivering immediately basic data to the other

ESPON projects and elaborating our metadata model or mapkit tool, avoiding the use of an intermediate version that was imperfect and had to be modified several times. Therefore, starting earlier would have been better for all the parts involved.

The second set of difficulties was related to some difficulties of communication with the **ESPON Coordination Unit**, in particular concerning the website (which was not available at the delivery time of our computer application) and the meeting with EUROSTAT (which was delayed many times). There were also some misunderstandings concerning the contribution of UMS RIATE to the design of ESPON posters for the Prague's meeting…

The third and most serious set of difficulties has been related to reporting and financial control. We know that the rules of the ESPON program are what they are, and that they could not possibly be changed before a new phase after 2013. But we also know that the European Commission has insisted in 2008, after the crisis, on the necessity to make the rules of control easier and to avoid unnecessary administrative burdens. Our feeling as coordinators is that the situation of ESPON is actually very critical on this question of financial control, with the danger of blocking the achievement of the ESPON DB 2013 Project. The coordinators of the project have indeed observed that more and more work time, normally devoted to the productive part of the project, is transferred to the management of administrative burdens related to "every-six-month-reports". And this burden is not limited to the coordination team but also spread all over the project partners, with the only exception of expert teams (that are not submitted to the same constraints).

2 Review of the project working progress

For simplicity reasons, the presentation of progress made is presented by challenges (sections 2.1 to 2.10). But some cross-challenge activities are presented in a final section (2.11) as they are not directly related to a particular challenge but implied the contribution of several partners. This concerns in particular the activity of support to the coordination unit and the external or internal networking.

2.1 Challenge 1: Collection of basic regional data

Coordinator: RIATE

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

The data collection for basic indicators has been finished with the delivery of data of June 2009 (ESPON Seminar in Prague). It contents the following indicators (table 1), collected for the all ESPON Area in the NUTS 2006 delineation:

Indicators	Period of reference	Geographical objects	Geographical coverage
Age pyramid by 5 years age-group	2005	NUTS0 NUTS1 NUTS2	ESPON Area (31 Countries)
Unemployed persons	2000,2001,2002, 2003,2004,2005, 2006,2007	NUTS0 NUTS1 NUTS2 NUTS2/3	ESPON Area (31 countries)
Active population	2000,2001,2002, 2003,2004,2005, 2006,2007	NUTS0 NUTS1 NUTS2 NUTS2/3	ESPON Area (31 countries)
Total population	2000,2001,2002 2003,2004,2005, 2006	NUTS0 NUTS1 NUTS2 NUTS3	ESPON Area (31 countries)
GDP in Euros	2000,2001,2002 2003,2004,2005, 2006	NUTS0 NUTS1 NUTS2 NUTS3	ESPON Area (31 countries)
GDP in PPS	2000,2001,2002 2003,2004,2005, 2006	NUTS0 NUTS1 NUTS2 NUTS3	ESPON Area (31 countries)

Table 1: Data collection of ESPON DB Project in June 2009

Data comes mainly from Eurostat. Missing values have been provided either by including data from other data providers (National Statistical Institutes, ESPON 2006 Database) or by computing statistical estimations. The complete lineage of the values is described in the metadata.

The activities of challenge 1 have in fact moved to some cross-challenges activities (table 2):

i. Data check of ESPON Projects: In February 2010, eight datasets from ESPON Projects have been checked (TIPTAP, Territorial Observation 1 & 2, DEMIFER, TeDi, ESPON Climate, ESPON Typology, Lisbon Territorial Indicators). The work consists to check if projects respect data and metadata rules defined by ESPON Database Project; give some advice on how to organize and precise as well as possible data and metadata; and to synthesize the knowledge of data gathered within the ESPON 2013 Program (degree of completeness of the datasets…)

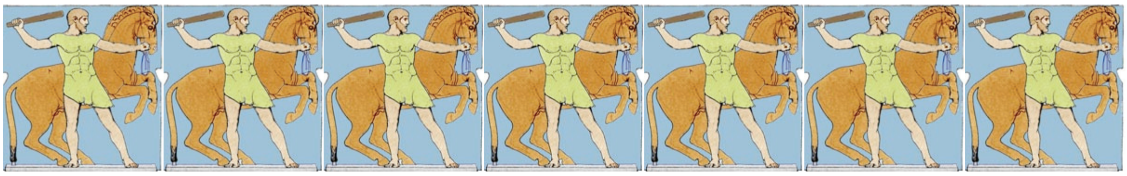
ii. Collection and harmonization of data produced within the other challenges of the ESPON Database Project. This process has namely allowed to make available data in November 2009 (presented during the ESPON Seminar in Malmö, table 2).

These activities are expected to continue until the end of the project.

Indicators	Geographical objects	Challenge involved
Corine Land Cover 2000	NUTS0 NUTS1 NUTS2 NUTS3 (version 2006)	5 (social/environmental data)
Population and area from 2000 to 2006 in Western Balkans and Turkey	"SNUTS0" "SNUTS1" "SNUTS2" "SNUTS3"	11 (Enlargement to neighbourhood)
Name, area, centroïd, and population 2001 of Urban Morphological Zones	UMZ	6 (Urban data)
Estimated population in 2005, 2010, 2015, 2020, 2025 and 2030 according to a central scenario of population projection	NUTS2 (version 2003)	Networking activities + challenge 1

Table 2: Data flow within ESPON Database Project in November 2009

2.2 Challenge 2: Harmonization of time series



Coordinator: IGEAT and RIATE

Harmonization of time series for basic socio-economic indicators at regional level for the period 1995-2006.

Background

ESPON DB 2013 is_a project that aims to improve the access to time series data. The issue of time series is a recurring necessity for ESPON projects as well as several European institutions mainly DG REGIO and ESUROSTAT. In spite of its importance, this process has not been very adequately initiated by the previous ESPON DB 2006 project.

The issue of time-series data can be assimilated fundamentally to the lack of data for a territorial unit either because the territorial unit in question has changed in the course of time, or because data are simply missing. Difficulties to build time series data can be related firstly to the lack of archived databases. Indeed, EUROSTAT, which is the principal provider of European statistics, does not archive all its database versions. It only keeps the last version of a given database. Secondly, information about historical changes of NUTS is often either missing or uncertain.

Time series approach can be organized in two main steps. Firstly, there is the collection and exploration of historical databases (New Chronos from EUROSTAT, cohesion reports from DG-REGIO…). This step aims to provide a review of continuous time-data series could be built form these data bases. Additionally, we have explored NUTS changes between 1995 and 2006. This exploration resulted in the compilation of the dictionary of NUTS changes which allows the review of territorial changes (codes, names and geometries). But the most important contribution of the dictionary is the identification of the genealogy (lineage) of NUTS which proves very useful for the harmonization of time-series data. The result of this first step will be used to build continuous time-series data. The conceptual model and its computer implementation are in progress.

State of the work: Conceptual framework and exploration of data sources

As it was planned in the previous interim report, during 2009, research has focused on the first step such as the exploration of the archived data bases (especially New Chronos) of Eurostat and territorial NUTS changes.

Besides the data available on the Eurostat internet portal, we obtained a CDRom with the Windows-only New Chronos application, i. e. the Eurostat archives. This CDRom was unsuitable for the needs of the project because of its web interface designed exclusively for data consultation and not for data exportation. The data were also stored on the CDRom in a specific file format unknown from us which led us to spend time on finding technical workarounds to finally extract and store these in a format we could handle.

The data appeared to be organized in 271 tables and 16 categories. We made an inventory of their content in order to have an idea of their completeness, i.e. the covered time span and the covered territory. The administrative division used is NUTS 1999, and all european countries that are currently EU members are represented. Data completeness depends, of course, on the type of data, the nuts level, the years considered, and, as could be expected, the completeness of these archives decreases with older data.

To provide here an exhaustive list of the content, even in a synthesised way, would be a nonsense because of the sheer number of variables and parameters. The data currently available on the Eurostat internet portal and the data included in these archives are partially the same, except that they do not use the same nuts reference system.

These data will be included in the Database system but will depend on the time series conversion tool to mix them with the current Eurostat data. Reversely, since they refer to an older nuts genealogy (1999) they might be useful in the next step of the time series harmonization challenge, but probably as a means of validation, to be compared with the values that our tool will compute for the 1999 nuts references.

Concerning historical NUTS territorial changes, the database was built by combining several sources such as the Official journal of the European Union, Eurostat, National statistical institutes and other European organisms like the DG Regio. In addition to data collection, our expertise has been based on experiences exchange with DG Regio and Eurostat.

The main conclusions of this expertise are:

Data available are very heterogeneous

Data quality is largely varying

Lack of good practice and experiences of handling territorial boundaries

Nuts changes knowledge: systemic approach of NUTS territorial changes

The benchmarking of sources and experiences has showed the complexity of NUTS territorial changes. Following Swianczny (2000) who states that: *"In order to create a truly time integrative GIS, the focus has to change from spatial to temporal and from analyzing changes between events to the analysis of the change itself"*, we propose an appropriate approach to formalize the Nuts changes. This approach will be based on an explicit description of changes.

Because the formalization of Nuts changes is complex and has to take into account several parameters (type of changes, temporal period and scalar dimension), we propose a cubic model which emphasizes the relationships between these parameters. This means that the result of territorial modifications depends on the type of changes (name, code, and geometry), the period of time and the territorial level. Thus, we used the concept of systemic approach (Figure 3).

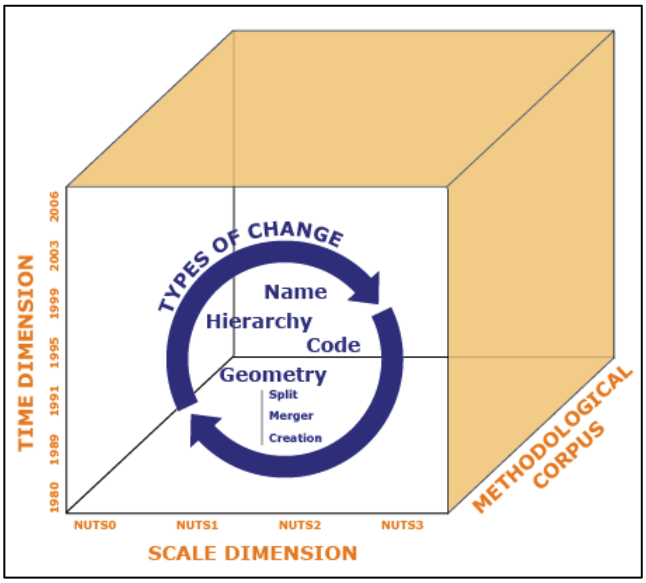


Figure 3: cube structure of NUTS changes formalization

We demonstrate our approach through the analysis of the example of Italian Nuts between 1995 and 2006:

Concerning the temporal dimension, tow orders can be distinguished:

The period determines the degree of discontinuity of the data sets. Indeed, the extension of the period increases the discontinuity because of the complexity of changes that may have occurred. In the case of Italian Nuts, if we consider the whole period (1995-2006), we can see a big discontinuity in the data sets. However, the data set will be complete between 2003 and 2006.

The building of time series data could be considered in either a prospective or retrospective territorial approach. The prospective view consists in transposing old data sets onto a recent version of Nuts (data 1995 onto Nuts 2006 for example). However, the retrospective view consists in transposing recent data sets onto old Nuts versions (data 2006 to Nuts 1995). Each approach requires a different methodology. For example, 2003 version data should be disaggregated to be integrated in Italian Nuts 1 level 1999 version. However, the 1995 version data should be aggregated.

As for the Scalar dimension, it is linked to the hierarchical structure of Nuts (Nuts 1 level is subdivided into Nuts 2 level, which is in turn subdivided into Nuts 3 level). In fact, the changes which occur in higher levels (1 and 2) have various consequences on lower territorial levels. The territorial reform of Italian Nuts 1 level in 2003, consisting in merging and changing codes of units, has caused a change of codes of Nuts 2 and Nuts 3 units. Moreover, reforms of higher Nuts levels (Nuts 1 and Nuts 2) could have more complex implications on lower levels. The creation of 5 new Nuts 2 units in Denmark in 2003, by splitting DK00, has caused very complex territorial reorganization on Nuts 3 level units.

Regarding Relationships between changes, the change of geometry is a determining factor in the time series data building process. On the whole, three types of unit spatial changes can be identified: the loss of area, the gain of area and deformation (which means territorial boundaries redistribution without loss of area). Based on these primary types of changes, we have developed a conceptual corpus to describe further types of changes (dictionary of changes). The dictionary of changes aims to answer the following questions: what happened? How did it happen? And what were the results?

For example, the Danish territorial reforms in 2003 could be described as follows:

Nuts 1 level: there are no changes

Nuts 2 level:

The Split of DK00 (change of geometry)

Official disappearance of DK00

Creation of 5 new Nuts 2 units: DK01, DK02, DK03, DK04 and DK05

Nuts 3 level:

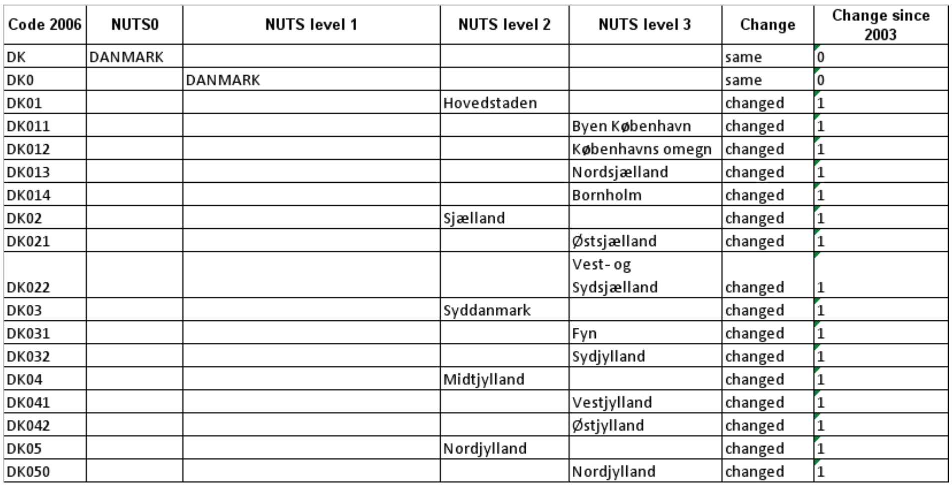
Change of code which means change of belonging to a superior unit (hierarchy): Funy DK008 (2003) and DK031 (2006), Bornholm DK007 (2003) and DK014 (2006)

Complex changes of geometry for the rest of units which have caused the disappearance of 12 units and the creation of 10 new units

This formalization, which is further explained as part of a draft technical report (annex), should not be seen as a normative approach, but rather as a descriptive one which will be improved in the next steps of the project.

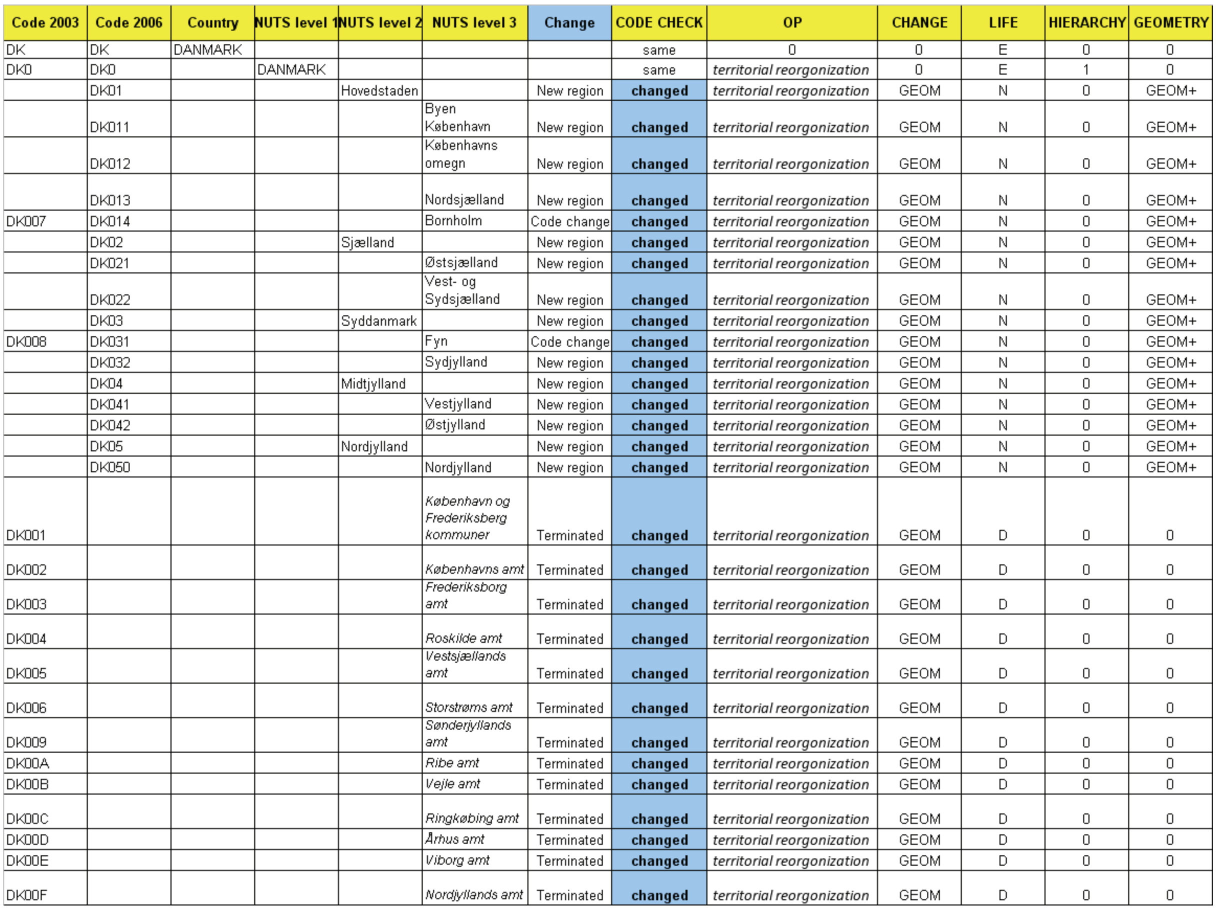
Presentation of the exploration results

The results of this exploration may be presented in different ways depending on the users' needs. The examples that we present illustrate the progress of the complexity of the issue of nuts changes formalization: location of change, identification of change and genealogy (lineage) of spatial units.



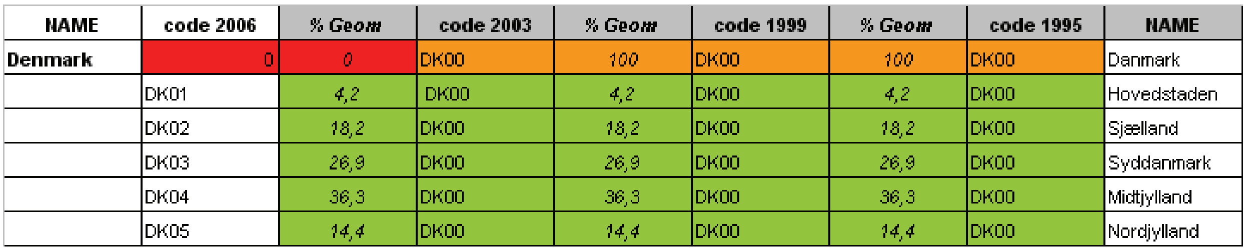
Code 2006	NUTS0	NUTS level 1	NUTS level 2	NUTS level 3	Change	Change since 2003
DK	DANMARK				same	0
DK0		DANMARK			same	0
DK01			Hovedstaden		changed	1
DK011				Byen København	changed	1
DK012				Københavns omegn	changed	1
DK013				Nordsjælland	changed	1
DK014				Bornholm	changed	1
DK02			Sjælland		changed	1
DK021				Østsjælland	changed	1
DK022				Vest- og Sydsjælland	changed	1
DK03			Syddanmark		changed	1
DK031				Fyn	changed	1
DK032				Sydjylland	changed	1
DK04			Midtjylland		changed	1
DK041				Vestjylland	changed	1
DK042				Østjylland	changed	1
DK05			Nordjylland		changed	1
DK050				Nordjylland	changed	1

Table 3: Extract of the table of changes locations: Danish nuts units between 2003 and 2006



Code 2003	Code 2006	Country	NUTS level 1	NUTS level 2	NUTS level 3	Change	CODE CHECK	OP	CHANGE	LIFE	HIERARCHY	GEOMETRY
DK	DK	DANMARK					same	0	0	E	0	0
DK0	DK0		DANMARK				same	*territorial reorgonization*	0	E	1	0
	DK01			Hovedstaden		New region	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
	DK011				Byen København	New region	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
	DK012				Københavns omegn	New region	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
	DK013				Nordsjælland	New region	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
DK007	DK014				Bornholm	Code change	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
	DK02			Sjælland		New region	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
	DK021				Østsjælland	New region	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
	DK022				Vest- og Sydsjælland	New region	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
	DK03			Syddanmark		New region	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
DK008	DK031				Fyn	Code change	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
	DK032				Sydjylland	New region	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
	DK04			Midtjylland		New region	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
	DK041				Vestjylland	New region	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
	DK042				Østjylland	New region	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
	DK05			Nordjylland		New region	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
	DK050				Nordjylland	New region	**changed**	*territorial reorgonization*	GEOM	N	0	GEOM+
DK001					*København og Frederiksberg kommuner*	Terminated	**changed**	*territorial reorgonization*	GEOM	D	0	0
DK002					*Københavns amt*	Terminated	**changed**	*territorial reorgonization*	GEOM	D	0	0
DK003					*Frederiksborg amt*	Terminated	**changed**	*territorial reorgonization*	GEOM	D	0	0
DK004					*Roskilde amt*	Terminated	**changed**	*territorial reorgonization*	GEOM	D	0	0
DK005					*Vestsjællands amt*	Terminated	**changed**	*territorial reorgonization*	GEOM	D	0	0
DK006					*Storstrøms amt*	Terminated	**changed**	*territorial reorgonization*	GEOM	D	0	0
DK009					*Sønderjyllands amt*	Terminated	**changed**	*territorial reorgonization*	GEOM	D	0	0
DK00A					*Ribe amt*	Terminated	**changed**	*territorial reorgonization*	GEOM	D	0	0
DK00B					*Vejle amt*	Terminated	**changed**	*territorial reorgonization*	GEOM	D	0	0
DK00C					*Ringkøbing amt*	Terminated	**changed**	*territorial reorgonization*	GEOM	D	0	0
DK00D					*Århus amt*	Terminated	**changed**	*territorial reorgonization*	GEOM	D	0	0
DK00E					*Viborg amt*	Terminated	**changed**	*territorial reorgonization*	GEOM	D	0	0
DK00F					*Nordjyllands amt*	Terminated	**changed**	*territorial reorgonization*	GEOM	D	0	0

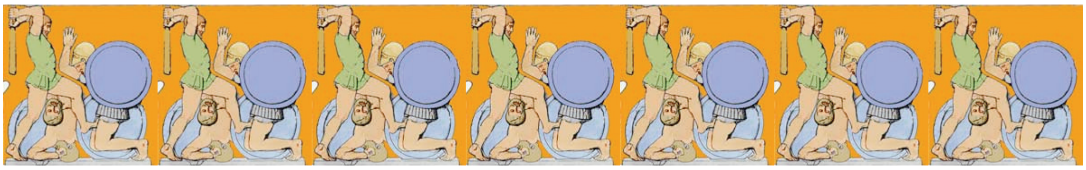
Table 4: Extract of the table of changes identification: Danish nuts units between 2003 and 2006



NAME	code 2006	*% Geom*	code 2003	*% Geom*	code 1999	*% Geom*	code 1995	NAME
Denmark	0	*0*	DK00	*100*	DK00	*100*	DK00	Danmark
	DK01	*4,2*	DK00	*4,2*	DK00	*4,2*	DK00	Hovedstaden
	DK02	*18,2*	DK00	*18,2*	DK00	*18,2*	DK00	Sjælland
	DK03	*26,9*	DK00	*26,9*	DK00	*26,9*	DK00	Syddanmark
	DK04	*36,3*	DK00	*36,3*	DK00	*36,3*	DK00	Midtjylland
	DK05	*14,4*	DK00	*14,4*	DK00	*14,4*	DK00	Nordjylland

Table 5: Extract of the table of nuts units genealogy: Danish nuts2 units between 2006 and 1995

2.3 Challenge 3: World / Regional data



Coordinator: RIATE & UNEP

Harmonization of data at World/Neighborhood and European/regional levels.

The first obvious aim of this challenge is to provide data for ESPON projects working at global scale, like the new project on "Globalisation" launched in February 2010. This challenge aims also to complete some discontinuous time series at NUTS2 or NUTS3 levels by means of disaggregation of time series available at State level. The work done by UMS RIATE and expert team UNEP on this challenge is summarized in the draft technical report "*ESPON World database*".

World Units and aggregation levels

- **Building a coherent dictionary of World Units**: The main role of UNEP/GRID in the ESPON Database 2013 project is to define a methodology for combining data available at world/neighborhood levels with those at European/regional levels, as well as to provide world data (around 200 states) for selected basic indicators, for the period 1960-present (and present-2050 in the case of demography). Several approaches can be developed to select countries; such as thresholds based on surface and/or population size, economy…Every approach has positive and negative aspects. We consider the official list of countries from main international "thematic" providers. The number of countries and the definition of "what is a country" for each provider do not correspond in several cases: for example Gibraltar is considered as a separate entity for all sources, apart from World Bank (GBR).
- **Exploring aggregation levels**: The related question of aggregation units has been also addressed. For the time being, aggregations are performed according to existing hierarchies such as (1) WUTS that have been proposed in ESPON 2006 but will be certainly improved by new project on globalization in ESPON 2013; (2) Official aggregation proposed by international organization like Chelem, Espon 2006, United Nations, World Development Indicators, UNEP GEO data portal. Grouping of countries based on thematic variables (e.g. income) are evolving through time (e.g. countries moving from low to middle income category) but the criteria that establish categories are stable through time, thus we can re-build past

hierarchies. In a next phase of the ESPON Database 2013 project, alternative hierarchies could be developed in order to better suit the ESPON needs.

Collecting World data and linking with Eurostat regional data

- **Collecting a first set of structural data:** The preliminary version of the World Database (v1.0) includes two main groups of variables: population and carbon dioxide (CO2) emissions. It will include, in a second stage (v2.0), variables on land-use and economic categories. To fill the database, we have develop methods for (1) re-computing of past data series for today separate countries. Example: the pre-1991 relative shares for the former Yugoslavia will be calculated based on data available after 1991; (2) fill gaps of data inside or in the extremes of the time series: values will be calculated by linear extrapolations and interpolations. In both cases methodologies and standards developed under the UNEP GEO Data Portal process were applied.
- **Linking World data with Eurostat regional data with a "Gap Tracker Tool":** Our goal has been to design a methodological tool (named "Gap Tracker") for explaining the differences between global databases and Eurostat data. For this purpose, two sample datasets were prepared (1) Europe in the ESPON database (EIE) with data at state level mainly derived from Eurostats; (2) Europe in the World database (EIW) with data obtained at state level from global organization like UN. In order to increase compatibility between EIE and EIW datasets, a systematic process was set up for analyzing of the differences. The Gap tracker tool is currently under testing and will be further improved

Providing geometries and aggregation levels for an ESPON World Mapkit

We have identified two relevant sources for cartography: (1) very precise geometries of World like Eurogeographics extension for World (compatible with EU); (2) generalized map of World like FAO GAUL (Admin 0 to Admin2-3).

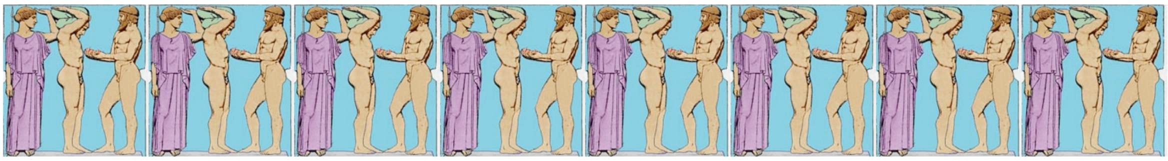
UMS RIATE has started the elaboration of a world mapkit similar to the one used in ESPON 2006 by project ESPON 3.4.1 Europe in the World. The task has mainly focused on the compilation of basic spatial units ("pieces of the World" that could be compatible with the different data provider at world scale in terms

Networking

According to the agreement between ESPON DB 2013 and FP7 EuroBroadMap, some data will be exchanged by both project and the same codification and geometries will be used when it is possible. The project FP7 has achieved in

December 2010 several matrixes of flows between countries at world scale related to Trade, Migration, FDI and Diplomatic Relation. This data could be made available to ESPON projects working on globalisation.

2.4 Challenge 4: Regional / Local data



Coordinator: TIGRIS

Harmonization of data at regional (European) and National local levels

In accordance with the proposals set out in the FIR, the TIGRIS team had to develop a sum of strategies able to spot and to collect information at LAU2/ LAU1 scale, information mobilized in order to fulfill an adequate local database. Our six objectives retained in our activity were declined depending on the search for equilibrium between data harmonization and collection opportunities.

First explorations

The first stage in our work was an apparently simple one which means that we had to finalize a sample database for at least two countries in the ESPON space, such as Romania and Bulgaria. It was a testing situation for TIGRIS ability to draw the strategies and the frames in order to obtain further local databases. Due to geometry metamorphosis and some technical difficulties, populating this sample database proved to be quite a challenge. That's why we passed on to *plan B* and we tried to properly integrate the information available for two other countries and an option was made for Czech Republic and Slovakia. Again, despite our intention and despite the resilience of the spatial LAU2 geometry (unlike Romania and Bulgaria, in the ex-Czechoslovakia few administrative reforms altered the local spatial frame), completely populating a database for the two new countries was an illusion, exceeding our possibilities to harvest all the indicators and to prioritize them.

A second step in TIGRIS work in Challenge 4 consisted of somehow a fuzzy collection of indicators describing the local level for as many countries as possible. LAU2 and LAU 1 indicators for Norway, the Baltic States, partially Italy, Austria, Luxembourg or Belgium were stored in a temporary database and shall be integrated in a GISCO based database frame, after a solid matching between LAU2 codes offered by the NSI and the base map coding system. Exploring different information sources and different indicators formats was quite a time consuming job, apparently without immediate outputs but with a strong formative dimension, able to indicate some good practices in the process of data collection.

The third stage of our work focused on the construction of the available list of LAU2 indicators for the countries included in the ESPON space. As much as possible, the list was completed. However, the work on this objective largely depends on a secondary task: translating all the indicators into English and placing Internet links for a further attempt to collect them. One major issue concerning this stage of work (as specified in FIR) is the actualization of the available information, a tricky task when new indicators become available or when old indicators may be ported to some other stocking platforms. One way or another, the second and the third step were strongly connected as a working approach and taking the both steps emphasized the problem of the indicators collection priority, as signalized in the TR.

Some of the problems encountered during the work on the fourth objective intersected some spatial analysis issues also shared by other teams (the change in the administrative limits or the discontinuities in the time series). Deriving and tracing the history of modifications in the LAU 2 unit's geometry was possible thanks to the access to the GISCO and MUNIS files, permitting us to trace modifications since the early '90 to 2006. A *just in time* geometry correction strategy should be envisaged in order to quickly respond to different demands concerning studies at LAU 2 scale. Largely depending on the quality of the basemap geometries, this exercise should be prudently regarded.

To our knowledge, except the 2001 LAU 2 population variable, there is no other *indicator* chronologically harmonized for the ESPON space. The financial availability of the information throughout the NSI should also be considered and it is not a quite a relaxing issue. That's why populating the database with only one indicator, declined at LAU 2 scale for all the involved countries, was a task with a double layered output: checking one of the objectives exposed in the FIR and framing the structure of one of our deliverables (an MS ACCESS based database for 31 countries). At the same time, we have managed to integrate in one indicator demographic and spatial information by the bias of the population potential at different bandwidth (25 and 50 km), using Euclidean distances. Being a derived indicator (such as the density one) it is quite difficult to consider this task as a database filing one.

SIRE database exploration: a potential entry for local data knowledge

Finally, recovering the SIRE Database represented our last challenge. While the inner structure of the SIRE files is a particular one (a hierarchical mix of spatial attributes from NUTS 0 to ex-NUTS 5 with different chronological and semantically marks), our expectations concerning the integration of the SIRE information in a proper LAU 2 database are partially fulfilled. Creating a list of coherent equivalencies between the SIRE coding system and the GISCO (our main support for geometries and database interrogation) was complicated by the different coding system (1991 as reference year vs. 2001).

Up to this moment our work has shown that dealing with more than 100 000 LAU 2 units, in different contexts, involves the obligation to sacrifice something, either the speed collection of data, in the name of some unstable indicator's quality, or the quantity of data, maybe sometimes of prerequisite importance. Ideally, we would like to use *neither/ nor* instead of *either/or*, but for the moment we prefer to cope with the immediate ESPON LAU 2 reality, rather than with the next one.

2.5 Challenge 5: Social / Environmental data

Coordinator: UAB (ETC-LUSI)

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 and Background

The aim of the challenge 5, leaded by the UAB, is to define a suitable methodology for integrating and making comparable data coming from statistical sources (e.g. EUROSTAT) and measured by administrative units, together with environmental data stored by natural unit or regular grid structure (e.g. Corine Land Cover).

The MAUP study results and recommendations, the bibliography research on existing methodologies and our experience at the UAB, as European Topic Centre on Land Use and Spatial Information, led us to the conclusion that the best way to downscale socioeconomic data and make them comparable with other kind of data, is using a regular grid structure based on the 1 km European Reference Grid, in which each cell takes a figure of the indicator or variable.

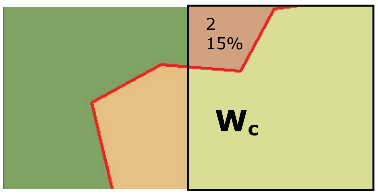
This document, based on the technical report of the Challenge 5 (November 2009), presents the results of the work that has been carried on by the UAB team since the First Interim Report of the Espon 2013 Database Project (March 2009) Methodology

Depending on the nature of each indicator or variable, a different kind of integration procedure must be applied. In this regard, we have defined and tested with different data the following three integration methods:

Maximum area criteria: the cell takes the value of the unit which covers most of the cell area. It should be a good option for uncountable variables.

Proportional calculation: the cell takes a calculated value depending on the values of the units falling inside and their share within the cell. This method seems very appropriate for countable variables.

Proportional and weighted calculation: the cell takes also a proportionally calculated value, but this value is weighted for each cell, according to an external variable (e.g. population). This method can be applied to improve the territorial distribution of a socioeconomic indicator.



Cell value = Wc Σ (Vi * Sharei)
Vi = Value of unit i
Sharei = Share of unit i within the cell
Wc = weight assigned to cell c

In the example: $W_c * (V_1 * 0.85 + V_2 * 0.15)$

Figure 4: Schema of the proportional and weighted calculation

The next table specifies some of the variables used and which integration method has been applied.

Integration methodology	Data source
Maximum area criteria	Urban Morphological Zones 2000, EEA
Proportional calculation	Unemployment rate total, 2001, Eurostat
Proportional and weighted calculation	GDP in euro per inhabitant, 2002, Eurostat
	Weighted by: JRC's population density 2001 grid

Table 6: Integration method and data selected.

In order to facilitate the testing processes an ESPONDB toolbox within ArcCatalog has been develop for each methodology described before. The next figure shows the general schema of the processes of the integration methods.

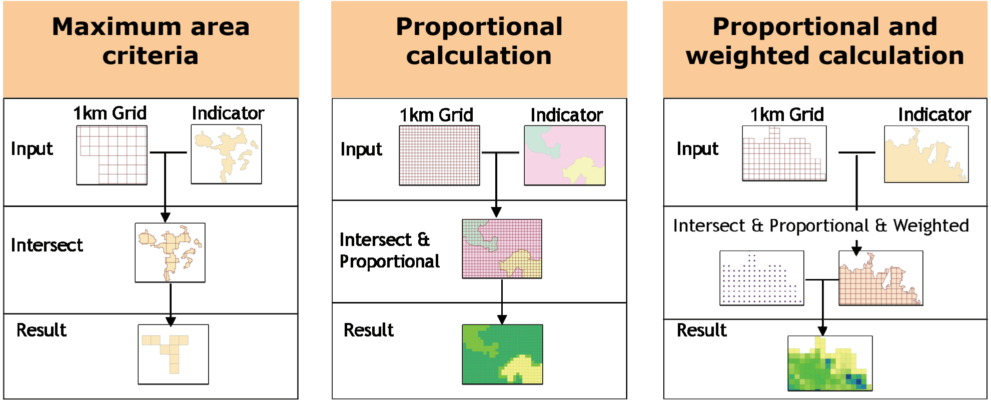


Figure 5: General schema of the integration processes.

The integration of ESPON socio-economic data and environmental data will be based on the building and distribution of ESPON OLAP (On-Line Analytical Processing) cubes using the most updated data.

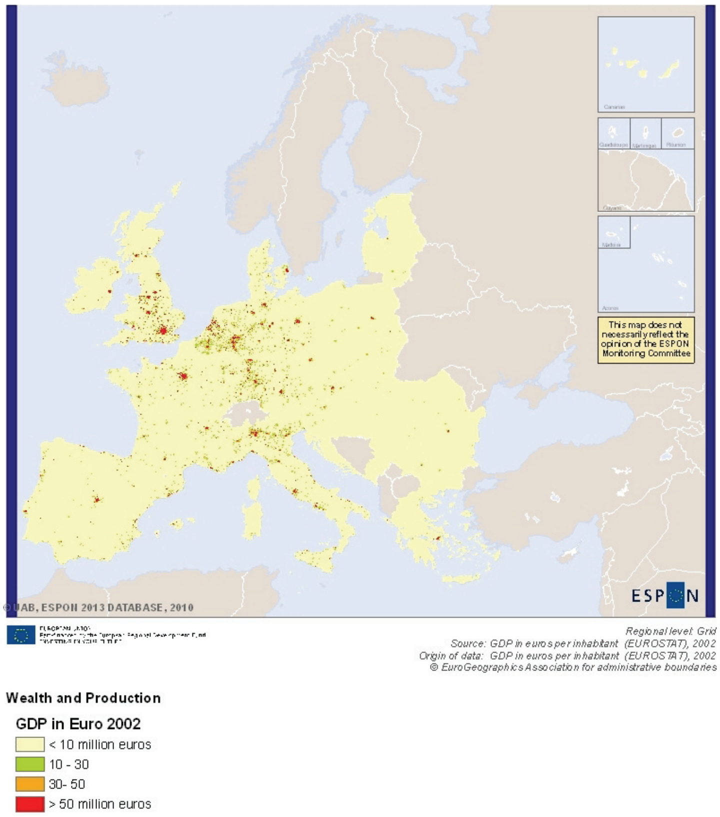
Results

The "proportional and weighted" aggregation method is the most representative one to appreciate the usefulness of these methodologies. In the result presented

[2] http://www.eea.europa.eu/data-and-maps/data/population-density-disaggregated-with-corine-land-cover-2000-1

here the GDP in euro per inhabitant 2002 (Eurostat) has been downscaled and weighted by population (JRC's population density 2001 grid dataset) 1.

The GDP is concentrated in the biggest urban areas, where most of the people are living and somehow higher in the grid cells belonging to the richest regions in Europe. Consequently, this method of redistributing and weighting data by grid cells is useful to be somehow independent of the administrative (arbitrary) divisions (figure 6, Distribution of GDP in Euro 2002 by grid). This case is highlighted for example in the south-west of Ireland, where the Nuts3 region (IE025) is very big, but the richness is concentrated mainly around the Cork city (Figure 7 Sample of the South-West of Ireland).


This map does not necessarily reflect the opinion of the ESPON Monitoring Committee
© UAB, ESPON 2013 DATABASE, 2010
ESPON
Wealth and Production
GDP in Euro 2002
< 10 million euros
10 - 30
30- 50
> 50 million euros

Regional level: Grid
Source: GDP in euros per inhabitant (EUROSTAT), 2002
Origin of data: GDP in euros per inhabitant (EUROSTAT), 2002
© EuroGeographics Association for administrative boundaries

Figure 6: Distribution of GDP in Euro 2002

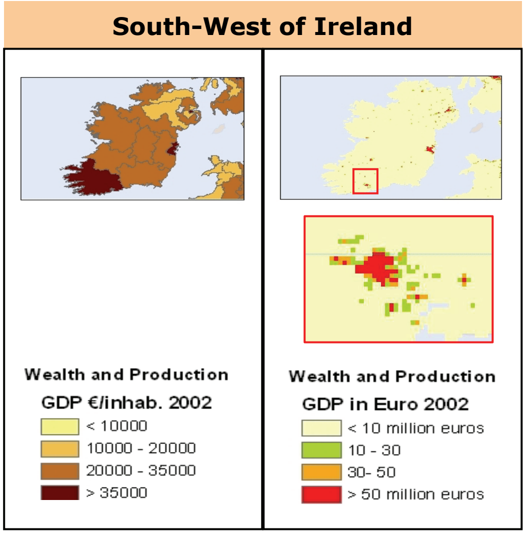


Figure 7: Sample of the South-West of Ireland.

Conclusions

The next points sum up the main conclusions about the outcomes of the work done:

- Disaggregating socioeconomic data by a regular grid is the best solution in order to downscale such information reported by administrative areas.

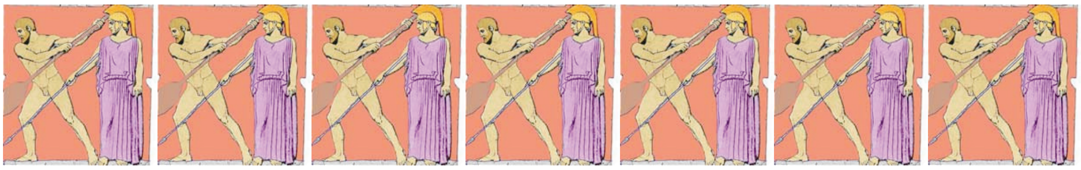
- The 1 km European Reference Grid is a good option to undertake the disaggregation due to have an European coverage and follow Inspire specifications.

-The "proportional and weighted" aggregation method is the one that gives better results, plus some added value to the downscaling.

- Different methods are independent from the source data format and can be applied to vector and raster format.

-This methodology allows the integration of socio-economic in an OLAP cube, which facilitates the comparison and analysis of such data together with land cover data, for example.

2.6 Challenge 6: Urban data



Coordinator: Géographie-cités

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…).

Since the First Interim Report, we have followed three main directions.

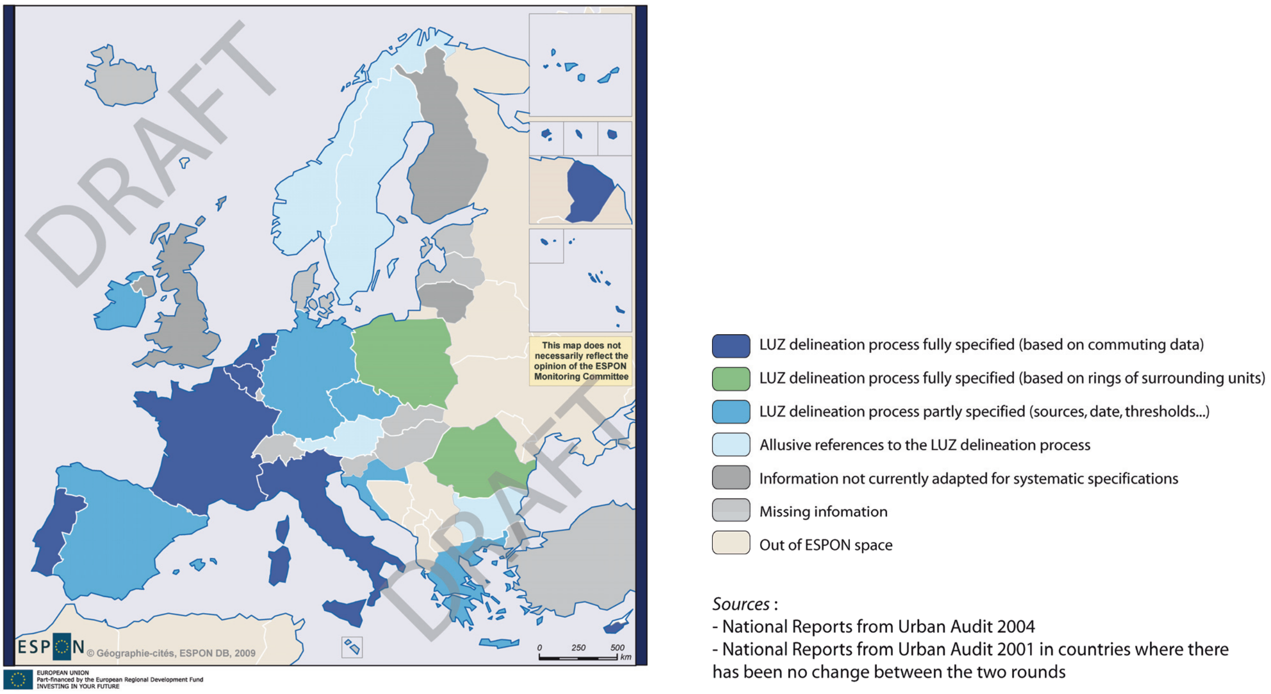
Metadata conceptual framework

Two different levels have been followed. At a general level, we have taken part in many meetings about the metadata framework of the ESPON 2013 project (Grenoble, Barcelona and Paris, see Activity Report January/June 2009). Our participation consisted principally on the expertise on how cities and cities databases could be integrated in Espon metadata profile and in Espon database. In October 2009, we have integrated UMZ metadata in the ESPON DB model and we have delivered the UMZ shape files to the teams in charge of the database implementation.

Considering the specific topic of cities, we have worked further on the semantic expertise that was announced on the previous FIR. To remind the context, we had gathered 9 urban databases, described in Table 1 on the FIR (UA core cities and LUZ 2001 and 2004, MUAS, UMZ, Proxy LUZ and FUA). Concerning FUA (Espon 1.1.1 and 1.4.3), the databases and documentations still remain incomplete so that we have not been able to pursue the semantic expertise. For Urban Audit databases, the different exchanges that were engaged with UA and FOCI team (especially the June Luxembourg meeting) have clarified a point that was not clear for us: the 2001 delineations are updated as soon as a country decides to change its rules (for example Spain, passing from proxy-functional delineation to more actually functional ones) and the previous data are not stored by UA. So it appears that the semantic expertise has to be lead on the 2004 UA reference year, and not on the 2001 one. We have also improved our LUZ metadata using the documentation gathered by Theodora Brandmueller from the Urban Audit team, and corrected some aspects of our first typology (presented in Prague June meeting). This is the reason why a second typology of LUZ delineations has been made, and presented in Malmö December meeting (Table 7 and Figure 8).

	Data on commuting flows	References of data	Quoted criteria and thresholds	Pre-existent zones / previous national zoning	References of pre-existent zones
Poland		Ring	🟨		
Romania		Ring	🟨		
Belgium	🟨	1991 census	🟨	Région urbaine	🟨
France	🟨	1999 census	🟨	Aire urbaine	🟨
Italy	🟨	1991 census	🟨	Sistemi Locali	🟨
The Netherlands	🟨	-	🟨	Stadsgewesten	🟨
Cyprus	🟨	2001 census	🟨		
Portugal	🟨	2001 census	🟨		
Greece	🟨	2001 census	-		
Spain	🟨	2000/2001 census	-		
Czech Republic	🟨	1991/2001? Census	-		
Croatia	🟨	-	🟨		
Ireland	🟨	-	🟨		
Germany	🟨	-		Planning Region?	-
Sweden	-	-		Local Labour Market Area	-
Norway	-	-		-	
Austria	🟨				
Bulgaria	🟨				

Table 7: Larger Urban Zone delineations (Urban Audit)
Source: National Reports (UA2004 and UA2001)


DRAFT
This map does not necessarily reflect the opinion of the ESPON Monitoring Committee
ESPON
© Géographie-cités, ESPON DB, 2009
EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE
0 250 500 km
LUZ delineation process fully specified (based on commuting data)
LUZ delineation process fully specified (based on rings of surrounding units)
LUZ delineation process partly specified (sources, date, thresholds...)
Allusive references to the LUZ delineation process
Information not currently adapted for systematic specifications
Missing infomation
Out of ESPON space

Sources :
- National Reports from Urban Audit 2004
- National Reports from Urban Audit 2001 in countries where there has been no change between the two rounds

Figure 8: A typology of LUZ delineations in Europe

Concerning the different other databases (MUA, Proxy LUZ, UMZ), the rules used to build urban objects have been extracted and compared. Some intermediate results have been presented at the RIATE February meeting (comparison MUA/UMZ). We are now preparing a technical report which will integrate all these metadata expertise (see above "Work to be done").

A new version of UMZ database

As announced in the FIR, we have prepared a new version of the UMZ database (CLC2000), creating a geometric attribute (centroid, the method is described in Technical Report "Naming UMZ"), adding population from the V.4.1 density grid of the EEA (Gallego 2007) and giving one or several names according to a methodology fully described in the Technical Report (Figure 9). The database is now ready to be used in urban studies, as showed on the rank-size graph usually built by urban planners and researchers (Figure 10). We have also integrated the different observations made by the ESPON-CU in order to improve the Technical Report "Naming UMZ" (see the new Figure 1 of the Technical Report and associated explanations). A further work is mentioned at the end of this report, on specific countries (Great Britain, Ireland and Portugal) and will be described in more details in next section ("Work to be done").

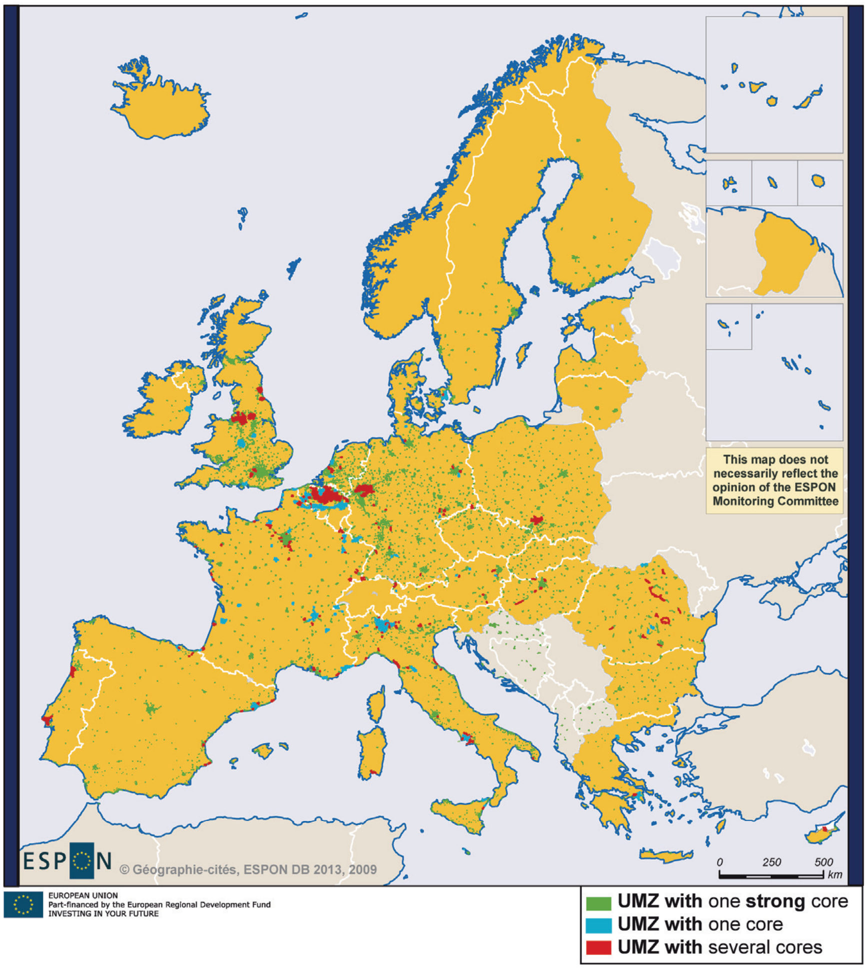

This map does not necessarily reflect the opinion of the ESPON Monitoring Committee
ESPON
© Géographie-cités, ESPON DB 2013, 2009
0
250
500
km
EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE
🟩 UMZ with one strong core
🟦 UMZ with one core
🟥 UMZ with several cores

Figure 9: UMZ typology according to naming results ("one strong core", "several cores" and "one core")

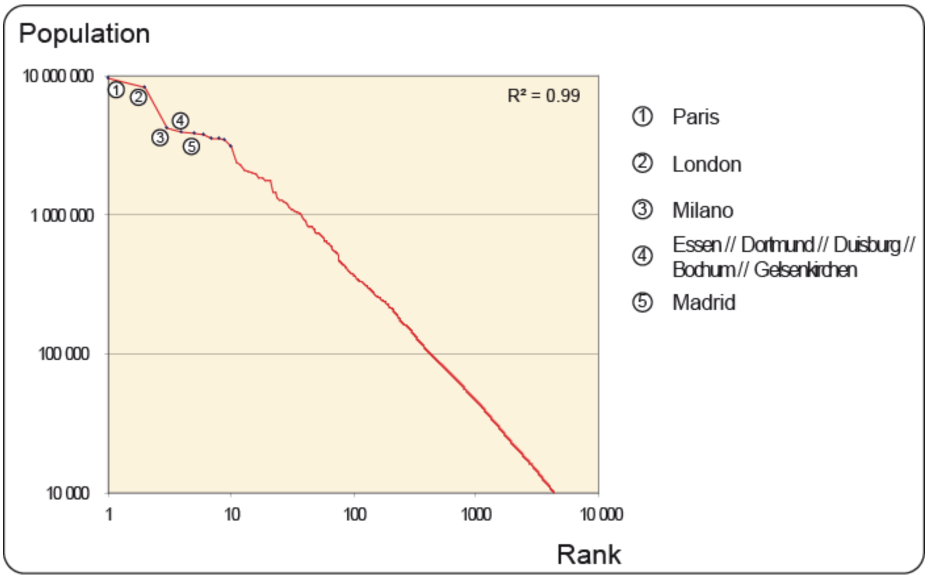


Figure 10: Rank-size graph and names of the main UMZ (CLC2000)
Source: EEA (CLC2000)

Semantic expertise and database comparison

In order to improve our knowledge about the use of UMZ for urban studies (validation process), we have defined a comparison protocol that takes into account semantic as well as geometric differences between UMZ and national urban databases (countries where are defined morphological agglomerations). Until now, the expertise has been developed by comparing UMZ to French and Danish morphological urban areas, and the work has been engaged for Sweden, in collaboration with Challenge 12. The first results, obtained for France and Denmark, show that the compatibility between databases is high. There is only an average difference of about 5% for urban populations (see table 8). Furthermore there is no systematic under or over estimation of urbanization by UMZ as compared to national databases (UMZ seems more extensive than urban areas are in Denmark, but less extensive in France). At a local level, the main differences are observed for some French urban areas and are related to specific types of settlement patterns (industrial or coastal conurbations). For Sweden, some preliminary results displayed a similar range of differences.

Population size	Denmark	France
More than 1 million (inh.)	3,4	-10,9
500 – 1000 000	1,3	-23,2
100 - 500 000	-1,0	-9,5
50 – 100 000	2,4	-7,9
20 – 50 000	3,6	-9,9
10 – 20 000	1,6	-12,3
5 – 10 000	8,9	-6,6
Less than 5 000	1,3	-2,1
Total	6,2	-6

Table 8: Deviation between UMZ and urban areas population (%)

Source: EEA (CLC2000), INSEE-RGP1999, Statistics Denmark-2001

2.7 Challenge 7, 8 and 9: data integration and retrieval process in the Espon database

Coordinator: LIG, RIATE, UAB and University of Luxembourg

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…).

This part illustrates the ESPON DB application working progress. It was built by merging challenges 8 and 9. Thesaurus issue was, also, added.

The data integration and retrieval process implementation and development can be divided into four main issues which are much linked: thesaurus, metadata profile, ESPON database model and ESPON database interface.

2.7.1 Espon thesaurus: first implementation

Metadata issues emerged among the partnership has a crucial element to link data organization for data communication and sharing. The first discussions on metadata were quite intensive and frequently involved some difficulties in tuning the different positions. Previous meetings revealed some inconsistencies with this regard that urged proper clarification. The project meeting held on April 23-24, 2009, in Barcelona, constituted a follow-up those meetings to deepen the discussions on metadata models and standards. Project partners were invited to review concepts, models and profiles. In addition, it allowed the definition of short- and mid-term strategies to further advance on the technological implementation and eventually to establish a work plan with task distribution for the following months.

The UL was given the task to make some preliminary considerations regarding the construction of corporate thesaurus. For this purpose, a draft technical report has been produced containing some reflections on the importance of such initiatives to structure geographical databases into themes and sub-themes that could facilitate information retrieval by end-users. This activity involved an intensive desk research to gain background information on the subject and provide useful guidance to other project partners. To this end, UL carried out a literature review of international guidelines on how to construct corporate thesaurus to structure knowledge and facilitate information retrieval. This overview was important to demonstrate that thesauri-based structures are organized by standardized relationships, such as equivalence, associative, and hierarchical. Besides, comprehensive examples of online thesauri (e.g. ILO, UNESCO, and OECD) have been described to ease understanding.

Within this scope, we argued that qualitative and quantitative text analysis applications may be very supportive to ensure the thematic structuring of the ESPON 2013 DB and further advance on the harmonization of concepts developed by ESPON. Such potentialities have been initially applied in some ESPON scientific reports to determine occurrence and co-occurrence of keywords that could be considered when defining themes and sub-themes. This methodological approach has been then applied to many other reports and eventually progresses presented in Paris, 1-2 October 2009, during the 2nd General Meeting of the ESPON 2013 DB Project. During the same event, our colleagues from RIATE presented the latest developments on the user interface prototype for data warehouse.

Because data should be accessible through the Internet, a communication protocol and query language needed to be specified. This urged UL to develop a short-term solution to classify indicators, determine naming conventions and harmonize coding schemes. For this purpose, a first proposal has been submitted for discussion among project partners during a technical meeting held at UNEP, in Geneva, 9-10 November 2009.

Based on constructive comments and suggestions, UL carried out a comprehensive exercise to overcome this problem. As a first approach, we assembled a list of first-level themes defined by international database classifications. This is meaningful because most of these databases, such as UNEP, EEA, or Eurostat, have provided and will continue to provide raw data on environmental and socio-economic issues to develop ESPON indicators and indices. The usefulness of such approach constitutes an opportunity to harmonize terminology used by some of the most prominent statistical agencies and therefore enable policy-makers, practitioners, and researchers to adopt a common language of understanding.

With this regard, each word (or expression) used as a first-level theme has been listed, evaluated in terms of similarity, and ultimately aggregated into similar themes. The following step involved data preparation to identity patterns. To this end, themes have been transformed into a binary valued matrix to ease the interpretation of results and eventually enhance the visual perception of similarities. In order to capture other potential patterns, we decided to include the ESPON 2006 DB structure of first-level themes and identify specific features that could validate or refute our clusters analysis.

The structure embedded in the matrix facilitated the definition of preliminary themes. This was very symptomatic after applying generalized association plots, or GAP (Chen, 2002; Wu et al., 2008), an open source tool that offers the possibility to identify proximities between subjects (i.e. word(s) that define a theme) and variables (i.e. database classifications). The preliminary results have provided substantial information on how to comprehend our data collection. It became clear, for instance, that certain themes are more representative to some databases while others are less visible. Words such as "Agriculture", "Population", "Transport", or "Energy" are exceptionally transversal. To a certain extent, this result justifies the need for adopting such words as first-level themes within the ESPON 2013 DB. When we exclude the previous ESPON structure the association matrix slightly changes its appearance. This showed that some themes gain more visibility while others express a reverse tendency.

Nevertheless, the primary group has been kept very alike. Similarly, we have identified a less prominent group, mostly clustered on environmental issues, but totally disconnected from the above mentioned cluster. Themes such as "Tourism", "Land Use", "Climate", "Resources" or "Health" lose their importance if not included in the same matrix as ESPON 2006 DB.

This analysis, which is further explained as part of a draft technical report, should not be seen as a normative approach, but rather a descriptive one. However, we have to point out that the choice of themes itself is very crucial for the success of the ESPON 2013 Program. Indeed, one could ask if this theme or that were emphasized more, or if an attempt was made to add one theme or another. Taking into consideration the limits of this methodological approach, we believe that our preliminary results should be seen as images of the future or, alternatively, as elements that correspond to the needs of a particular moment.

Against this background, those themes that have not been mentioned in this first proposal should be considered as less interesting, although this assumption should not be taken as granted. It is widely known that the current and future dynamics of the EU policy agenda will shape the research demand of the ESPON 2013 Program. This is of extremely importance, not only for the program itself but also to the database. For the moment, it is not feasible to address all the relevant political, environmental or social issues, even if we consider different approaches to conjecture about the degree to which EU priorities will develop and gain more or less visibility.

Another problem that emerged along this proposal concerns naming conventions and coding schemes. Using as a reference the latest list of ESPON indicators, we noticed that naming conventions vary according to the criterion defined by each research team. We argue that consistent definitions for commonly used terms would improve the harmonization of naming conventions and therefore the risk of having identical indicators with different names. Given that, this is a very difficult matter to resolve, mainly because we are dealing with textual information. Moreover, naming conventions should not be seen as a way to replace metadata. This will require additional efforts, such as the development of a glossary or handbook to assist in clarifying terms that could potentially be used to label ESPON indicators.

Persuade dataset suppliers to adopt common standards of coding schemes is a permanent challenge. As explained above, we noticed that some of the applied research projects under Priority 1 and 2 of the ESPON 2013 Program have defined their own rationale to label indicators. Despite the usefulness of such exercises internally, the degree of ambiguity is increasing when applying different methods to label indicators. This is often the case among well-popularized indicators, such as unemployment.

Within the ESPON 2013 DB project the above mentioned situation is becoming increasingly problematic to further progress on user interface prototype. Indeed, if no harmonization is employed the capacity to deduce information from codes becomes rather difficult. To a certain extent, coding schemes are not used to express the content of data but rather an attempt to homogenize codes for indicators, indices and other measures. Nevertheless, some information needs to be provided and, most importantly, it needs to be arranged in a consistent, clear way to avoid potential misunderstandings. With this regard, UL has introduced an innovative coding scheme to label indicators. Such method suggests a minimum number of characters that should be used to assemble relevant information about data. The process of structuring has been organized in a sequence of five fields where each one of them describes some of the specificities embedded in each indicator. This experimental procedure has been then applied on approximately 140 ESPON indicators delivered up to date.

2.7.2 Data and metadata models implementation

We will divide our presentation of our activities since the FIR in two categories: conceptual work and implementation work. Since the FIR, our conceptual efforts have been directed in three main directions:

1. The elaboration of a complete vector data and metadata profile for the ESPON 2013 DB
2. The extension of the ESPON DB model in order to receive data and metadata compatible with the defined profile
3. The definition of two ontologies for the ESPON 2013 Database:
    a. A temporal ontology of territorial units (to be implemented by RIATE)
    b. A thematic ontology (to be implemented by UL and RIATE)

We describe these activities in the sections below.

A data and metadata profile for vector data

Metadata profile

The definition of a data/metadata profile is an essential task in order to ensure both the long term compatibility of the ESPON Database with the other players of the statistical data scene (global, European or national institutes) by complying with the existing standards and rules.

The major standards and directives that were taken into account were:

- The ISO 19115 standard on geographic metadata;
- The metadata rules encompassed in the INSPIRE directive
- The SDMX standard for statistical data and metadata

This task has shown to be quite difficult to deal with and took longer than initially expected because there is a lack of consensus among existing standards and, also, the existing standards and formats are not exhaustive for the purposes of the ESPON 2013 Database. The ISO 19115 standard and the INSPIRE directive are fairly similar and, although they do not match in terms of the exact details required as metadata for a geographic dataset, conceptually they are easy to harmonize. The drawback of both standards is that they are aimed only at geographic data and, as such, provide very poor means to describe statistical data from a thematic point of view and from the data quality point of view. Typically, statistical datasets are not homogenous from a lineage viewpoint, the values may have different sources and, potentially, different levels of quality or confidence. This crucial issue has to be tackled by the ESPON DB Project. Although the SDMX standard is devoted to statistical data and metadata, it only establishes an open exchange format for statistical metadata: it gives the means to represent any statistical metadata but without stating which metadata is critical for exchange and compatibility purposes. As a consequence, the ESPON DB data/metadata profile adds precise data quality description at the value level (lineage information, etc.) that allows users to have a deep understanding of the origin of the data and of the successive transformations that the data have been subject to (error correction, estimation, etc.). More precisely, this profile describes metadata about indicators measured or collected on vector spatial units (NUTS, cities, etc.), or, in short, vector metadata.

Another important issue to consider in the elaboration of the data/metadata profile is user friendliness and usability. Our objective was to demand data providers (ESPON projects, etc.) the least amount of effort as possible when filling up some of the metadata fields, so that they can fill more accurately other critical metadata fields. As a consequence, all the metadata fields that can be inferred from the data themselves (e.g. spatial coverage) are not required by the ESPON DB metadata profile (to be filled by providers) and are to be inferred by the ESPON DB instead and made available only in the output metadata.

The ESPON DB metadata profile contains three types of information:

1. General Dataset information (identification of the dataset and of the producer)
2. Indicator metadata (thematic description and methodology of each indicator in a dataset)
3. Value metadata (detailed description of data subsets in terms of lineage and copyright constraints)

A detailed description of the metadata fields is given in the technical annex concerning the metadata editor.

A new model for the ESPON DB with complete metadata support

In order to keep the ESPON DB in phase with the ESPON DB data/metadata profile, the model of the database has been modified accordingly. The database is used for storing both the data and the metadata, as a matter of fact, from a database point of view; we could simply consider metadata as data with a higher level of abstraction. The ESPON DB model describes the following categories of information:

- Indicator metadata
- Contact metadata
- Lineage and copyright metadata
- Dataset metadata
- Indicator values
- Geometrical data

The data and metadata model is also conceptually ready to receive other types of vector statistical data than NUTS, e.g. cities or world data. However, in order for the ESPON database to seamlessly integrate world, NUTS and local data, just making a flexible database model is not enough. Temporal geographic ontologies describing the horizontal, vertical and transversal (temporal) relations between geographical units need to be created. The purpose of ontologies is described in more detail in the next section.

2.7.3 Definition of ontology needs for the ESPON 2013 DB

From the beginning of the project, we have anticipated that we would need to create and use ontologies in order to build a sound ESPON Database. The term 'ontology' is quite polysemic, but within the scope of the ESPON Database we refer to ontologies as dictionaries (or, more precisely, structured vocabularies) of features (indicators, territorial units, etc.) that contain verified, complete and coherent information about a certain field. The role of ontologies inside the ESPON Database is threefold:

1. To provide an efficient way to enforce data consistency and to avoid data redundancy: new data is typically tested for consistency with existing ontological information in order to detect errors or potentially duplicated entries
2. To simplify the task of filling metadata: for instance, providers can simply choose among the terms in the ontology (in the metadata Web editor) instead of providing all the information about them each time
3. To allow richer metadata output for the community (users get complete and sound metadata with their data)

Since the FIR, our work on the data/metadata profile, editor and on the model of the ESPON Database also allowed us to define more clearly the needs, possibilities and priorities for ontologies within the ESPON Database. We choose to develop two main ontologies in the near future:

The thematic (or indicator) ontology is a full dictionary of indicators that are to be stored in the ESPON Database and of the relations between them. The team from the University of Luxembourg has undertaken the development of this ontology. The main challenges in building this ontology are:

1. Creating a standardized codification system for all the indicators, so that every indicator has a unique, non-ambiguous code
2. Creating a classification (with either simple of multiple classification) hierarchy, doubled with aggregation/inclusion relations between indicators
3. Linking each indicator present in the dictionary to one or more elements of a thesaurus seen as keywords. The thesaurus can then be used for the exploration and querying of the database.

The spatial ontology is a full dictionary of NUTS territorial units and the changes they have been subject to (from NUTS0 to NUTS3 level). This ontology is crucial for most computer assisted data harmonization process. The RIATE team undertakes most of this task.

For a complete integration of world, NUTS and local data in the ESPON database, two other spatial ontologies need to be implemented. Each of these ontologies units (the global and the local spatial ontology) need to give a complete list of the spatial units and to describe their temporal evolution and the relations between them. Once the global and local ontologies completed, they need to be connected to the existing NUTS ontology in order to make the database integration complete. However, the creation of these ontologies requires considerable effort. This is explained in more detail in the section describing needed improvements for the second phase of the ESPON database project.

During this period, our implementation work has focused mainly on three activities: i) implementation of the data and metadata excel templates, ii) implementation of the first version of ESPON database and Web interface, and iii) implementation of the second version of ESPON database and Web interface.

The data and metadata templates

The data and metadata templates, implemented as formatted Excel spreadsheets have been mainly developed by RIATE. The template matches perfectly the conceptual data and metadata profile. It contains additional comments, examples and formatting in order to increase usability and readability.

2.7.4 The first version of the database and Web interface

The first version of the ESPON database and Web interface has been described in detail in the FIR. This database was based on the well-known open-source DBMS PostgreSQL and its spatial extension, PostGIS. The Web application for data download was based on Java technologies and uses a framework for Web application development (Java Server Faces). Since the FIR, more work on this version has been carried out until the end of June for debugging and minor adjustments in order to make the application completely compliant with the requirements of the CU. The main strong points are:

1. Independent, lightweight Web application easy to deploy and requiring minimal resources from the client machine. The idea behind the chosen architecture is that, basically, any computer can be used for accessing the application, provided that the Web browser used is recent enough.
2. Dynamic, easy to understand query interface, with dynamic display of the query criteria for data retrieval. Each query criterion (spatial extent, time, indicators, etc.) is displayed as a list of choices that correspond to what is available in the database. As soon as the user chooses one or more items in one of the lists (e.g. some indicators), the content of the other lists is updated, taking into account the choices as a partial query (e.g., the list of years presents only the years for which some data are available for the chosen indicators). This allows users to be informed on the fly about data availability while composing their queries, instead of blindly querying the ESPON database.

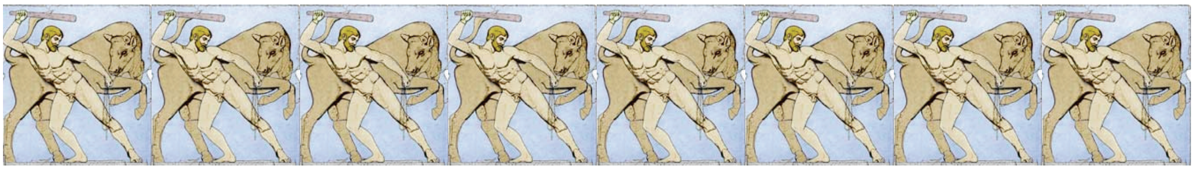
Prototyping the first version of the Web interface and of the ESPON database also allowed us to point out some weaknesses in our approach and to perceive some improvements to be addressed in the second version. The main weaknesses were:

1. Incomplete metadata support, due to insufficient maturity of the ESPON metadata profile. Thus, the development of the first database version started right at the beginning of the project when the ESPON data and metadata profile had not been completed. The metadata support was incomplete both in the database model and in the Web interface.
2. Insufficient data/metadata exploration capabilities in the interface. While, on the one hand, the use of dynamic criteria lists allow the users to have an idea of what is available in the database, on the other hand, scrolling through the lists when too many choices are available (which will be more and more the case since the ESPON DB will receive more and more data) appears to be cumbersome for the users.

The second version of the ESPON DB Web interface and database aims at solving these issues and providing a complete solution for metadata exploration. The positive features present in the first version were maintained. The Web interface for the ESPON database is based on a dynamic Java Web application, developed with the Struts framework. The database itself is based on the same PostgreSQL/PostGIS DBMS as the first version. The main novelties of the second interface version are:

1. A unique Web application that assists users in all the tasks related to the ESPON database (metadata editing, data and metadata upload, data discovery, exploration and downloading). The complete application flow is described in the technical annex on the ESPON DB Web application.
2. The possibility to display and query every metadata content existing in the ESPON database. Each metadata field described in the ESPON metadata profile can be used as a search criterion within the ESPON database. The users also have the possibility to view the complete metadata files of the datasets before actually downloading the datasets to their computers.
3. Two versions of the Web interface are available, the first proposes the most common query criteria, while the second version (accessible via an “advanced search” link) displays and handles all the metadata attributes as search criteria.
4. The display of data completeness at different scales (completeness of the whole dataset and completeness of the data at each level of detail).

2.8 Challenge 10: Spatial analysis for quality control



Coordinator: NCG and RIATE

Objectives: To develop spatial analysis and data mining methods in order to identify exceptional values

Work since the first interim report has been concerned with the identification and evaluation of suitable techniques for the detection of exceptional values in the ESPON 2013 Database. These techniques should eventually be implemented in the database. This work has also considered the impact of the Modifiable Areal Unit Problem (MAUP) on the detection of exceptional values.

Our second interim report examines how mathematical, statistical and spatial analysis tools can be applied to the database in order to find 'logical input errors' and 'statistical outliers'. In both cases, 'exceptional values' can arise but it is not always clear if such values relate to input errors or true values that are statistically-outlying. In this respect, reliably determining the nature of an exceptional value is important, especially as input errors should be treated differently to statistical outliers. For example, input errors are usually corrected or removed, whilst suspected outliers are usually flagged for further scrutiny.

The outcome of the report is a targeted review of existing outlier-detection tools in the field of statistics, data mining and spatial analysis, and an examination of how they can assist in the detection of errors/outliers in the database for improved quality control. The methodological review has a clear focus on spatial analysis with respect to outlier-detection; and is complemented by worked examples on an ESPON-type test data set, where chosen techniques are demonstrated. Worked examples are coded using open-source software so that the applied techniques are easily transferable. The list of techniques that are applied should not be considered as exhaustive, but form a cross-section of useful techniques which are appropriate for the ESPON 2013 Database.

Our report considers a number of different types of exceptional values that may arise in the database, and how to identify them. We demonstrate the utility of chosen detection techniques using subsets of one over-arching example test data set (primarily at the NUTS 3 level). One such data subset includes a number of (logical) input errors that have been deliberately introduced. Detection techniques are implemented within the R[3] statistical computing system, in a series of six worked examples. The corresponding R scripts are presented in an appendix to the report.

[3] R is an open source environment for statistical computing and graphics. (http://www.r-project.org/)

Data for the ESPON 2013 database consists largely of spatial data; that is, data which refer to the measured attributes of the spatial regions used in data collection, manipulation, analysis and reporting. These regions, various levels in the NUTS[4] and LAU[5] hierarchies, refer to locations on the earth's surface within the administrative and political boundaries of the EU. Spatial data have some interesting properties, and these may be used to aid the detection of values which are, in some sense, unusual.

There are several different classes of exceptional value which we often encounter. These include (a) input errors arising from logical inconsistencies in the data; (b) observations which are unusual when the spatial characteristics of the data are ignored; and (c) observations which are unusual when the spatial characteristics of the data are taken into account.

Identified input errors may be corrected or removed (the metadata may have to include this, or a report produced on the action taken). Observations which are unusual in a statistical sense may be flagged for further investigation, as either outliers or potential input errors. Any observation which is thus flagged may be left alone, replaced, or removed – again there may be some need to report the action(s) which have been taken.

Logical input errors arise when there is a mismatch between the correct data values for a region and those which are supplied. We term these 'logical' errors because their identification can often be made using a set of logical, often mathematical, deductions. A logical input error may occur for a number of reasons, for example when: (a) the wrong NUTS code has been allocated to data for input to the database; (b) wrong values have been input by an operator; (c) data have been displaced within a column or swapped between columns; (d) a data item is presented which has one or more 'impossible' values (for example, a data value of -2, 4.5, 11 or B for a land use class variable that can only take positive integers from 1-9). Such a list is not exhaustive, and should grow as different input error-types become apparent (i.e. at this stage, we are not expected to foresee all input error possibilities).

We can also use statistical techniques to detect input errors; for example an unemployment rate of 27% that has been mistyped by transposing the digits to 72 would appear as a statistically-outlying value in the upper tail of the distribution.

Aspatial statistical outliers may exist as outliers in a univariate or a multivariate sense. For univariate data a simple graphical display such as the boxplot forms an ideal starting point. The components of the boxplot (the medians and quartiles of the distribution) lead to the definition of inner and outer fences; data values which are outside of the outer fences should be examined and possibly flagged as outlying. If the data for the variable in question is markedly skewed (which is often the case with social data: income is a good example), then the boxplot can be adjusted in the light of the skew. If two variables are considered jointly, then a data item might only appear as unusual

[4] NUTS: Nomenclature of Territorial Units for Statistics

(http://ec.europa.eu/eurostat/ramon/nuts/basicnuts_regions_en.html)

[5] LAU: Local Administrative Units (http://ec.europa.eu/eurostat/ramon/nuts/lau_en.html)

when both indicators are viewed together: here the bagplot, an extension of the boxplot is a helpful tool.

Values which are unusual when several variables or indicators are considered simultaneously are multivariate outliers. A useful technique here is to examine those observations which have a large squared Mahalanobis Distance; the Mahalanobis Distance is a scale invariant measure of the distance between objects, taking into account the covariance structure of the data. A second technique is to apply a Principal Component Analysis (PCA) to reduce the dimensionality of the data set, where in the resultant transformed space, outliers may be more readily identified. The advantage offered by the several PCA approaches that we test is that computation of the PCA transform is rapid, and is therefore suitable for complex, high dimensional data sets.

Spatial statistical outliers can be identified when the spatial structure of the data is additionally taken into account. Observations which may not be identified as unusual when the spatial component is ignored may be spatially outlying (i.e. unusual with respect to its neighbours). To ignore the spatial element in such cases would give rise to false negatives. Here Hawkins' spatial outlier test provides a convenient statistical technique for detecting outliers in univariate spatial data.

Statistical prediction models may also help in detecting outliers where the model's prediction errors are unusually large. Whilst it is necessary to specify a suitable model for this type of analysis, examination of the prediction errors can be a powerful tool in outlier detection. Observations can be outliers in (a) the variable being predicted (a dependent variable) or (b) a variable that is informing (or explaining) this prediction (an independent variable). Useful predictors include multiple linear regression, attribute space local regression, and geographically weighted regression; all of which are applied. These predictors may also be useful if the imputation or prediction of missing data is required in the database.

Spatial clusters may be formed when a group of observations identified as outliers may actually be spatially clustered with a substantive reason for their 'unusualness' (i.e. false positives are to be avoided as well). In this respect, various nonstationary modelling techniques are applied that identify local (or regional) changes in the spatial process according to some key moment or relationship. Here, geographically weighted summary statistics; geographically weighted regression and Anselin's local version of Moran's I (for spatial autocorrelation) are the chosen techniques.

NCG has already reported on the effects of the **MAUP** for ESPON 2006. The spatial structure of the reporting units is an important consideration. It has been known since the early 1930s that the values of statistics for spatial units are not only conditioned on the size of those units but also their configuration. In our report we demonstrate this by applying outlier detection techniques to spatial units of different sizes in the NUTS hierarchy.

Another database project is concerned with the effects of the reporting units being used for the ESPON database changing over time.

2.9 Challenge 11: Enlargement to neighbourhood

Coordinator: RIATE & NTUA

Objectives: To collect data at regional and local level for neighbouring countries of ESPON territory

An expertise on data availability in Western Balkans and Turkey (Candidate Countries / CC and Potential Candidate Countries / PCC) was the main issue developed concerning Espon DB enlargement to European neighbourhood. The work done since the FIR could be summarized as follows:

Compatibility of spatial administrative divisions with the EU NUTS classification

Turkey, Croatia and FYROM have already adopted this classification. For the rest Western Balkans countries, the criteria of the population weight (formal criterion) and the administrative capacity (informal criterion) are fulfilled in the majority of the existing administrative divisions (regions, districts etc) of these countries, therefore respective ”similar NUTS” divisions could be used for the work on data without considerable problems.

Data required and existing data

The data required are mainly referred to the following aspects: (a) Demographic and social: Population, households, dwellings etc per appropriate categories. (b) Economic aspects, employment: Active population, employment / unemployment, GDP etc. (c) Environmental aspects. We give in the Annex 2 of the Technical Report the Table 1 of the existing data per CC / PCC, per group of themes and per census / survey in which are based.

Data availability and quality at level NUTS 0 / country level

In general, it is very satisfactory for all CC / PCC. Most of the data are provided by Eurostat, additional data are provided by the National Statistical Offices (NSO). In more detail, the following issues are covered satisfactorily: General Economic Background, Employment, Innovation and Research, Economic Reform, Social Cohesion, Environment**,** Population and social conditions, Industry, trade and services, Agriculture, Forestry and fisheries, External trade, Transport, Environment and Energy, Science and technology.

Data availability and quality at NUTS2 level

It is in general very satisfactory for Croatia, FYROM and Turkey -which have adopted the EU NUTS classification. *Eurostat* provides data on: Agriculture, demography, economic accounts, science and technology, tourism, labour market. Some additional data for specific topics are provided by the *NSO*.

It is less satisfactory for the other Western Balkans countries. Relevant data are provided by the NSO and other sources.

Data availability and quality at NUTS3 level

*For Croatia, FYROM and Turkey, it is very satisfactory for the following sections (data provided by Eurostat)***:** demography, economic accounts, tourism, labour market. Some additional data for specific topics are provided by the NSO of these countries.

For the other Western Balkans countries, data are provided only by the NSO and other sources. For *demography and labour market*, it is good only for some of them while for the rest it is nearly acceptable. For the *rest issues*, there are important differences according to the country. Concisely, availability is more satisfactory for Serbia, much less satisfactory for the rest countries.

Inclusion of the Western Balkans and Turkey in the scope of the ESPON Database

Taking into account that necessary reliable data at the appropriate NUTS level or "similar NUTS" level exist for the CC / PCC except Kosovo (under UN Security Council Resolution 1244), all *these countries should remain in the scope of the ESPON Database;* few data for Kosovo should be included at the moment in the Database.

Western Balkans and Turkey NUTS geometries

The Western Balkans and Turkey NUTS geometries have been included in the "European" Map-Kit. For Croatia, FYROM and Turkey, the NUTS system already exists. But, concerning the rest Western Balkans countries (Albania, Kosovo, Montenegro, Bosnia-Herzegovina, Serbia), "SIMILAR NUTS (SNUTS)" has been created – see previously. See in more extent in the "Mapping Guide" Technical Report.

More than the ESPON area, the map-kit includes the candidate countries and the Western Balkans.

Our work should follow the steps of the entire Database project. Therefore, during the stages of the project in the year 2009**,** we were more specifically interested in a first set of "basic" data, delivered to RIATE, which are gradually integrated in the Database. This set includes more specifically, the following (from NUTS0 to NUTS3 levels):

- Total Population,

- GDP in Euros and GDP in PPS,

- Active Population and Unemployment,

- Total Population by sex and age (for the year 2005)

We provided, in addition, data for:

- Total area,
- Land area and
- Population density.

These data were compiled from all available sources (Eurostat, National Statistical Offices and other sources). The respective metadata are included in the Database.

2.10 Challenge 12: individual data and surveys

Coordinator: RIATE & U.UMEA

Objectives: Examine how to integrate individual data based on census or surveys in the ESPON database.

Background

The Department of Social and Economic Geography, Umeå University has been contracted as experts in the ESPON DB project, and has the main responsibility for Challenge 12, "Individual data & surveys". An important background is that the department has access to and thorough experience in using individual, longitudinal population data. In particular, the department manages the database ASTRID, which not only covers the entire population of Sweden for a substantial time period, but also has a high degree of geographic resolution. The initial work plan for the Umeå University team, presented in more detail in the previous interim report, concerned four themes or activities: 1) availability and usefulness of survey data, 2) the modifiable areal unit problem (MAUP), 3) comparisons between Swedish and European-wide data, and 4) integration of survey data in the ESPON database.

State of the work

During 2009, research has focused on the first three themes/activities. In particular, efforts have been focused on the third activity, i.e. comparisons between Swedish and European-wide data. In cooperation with other ESPON DB teams and challenges, available Swedish individual register data has been used for purposes of exploratory studies, comparisons and methodological development.

In the ESPON Database context, there is a need to present or utilize population data at the local level. Work carried out in Challenge 5 encompasses the development of a methodology to disaggregate socioeconomic data into a regular grid. (This is presented in more detail in other parts of this report, as well as in a separate technical report: "Disaggregation of socioeconomic data into a regular grid: Results of the methodology testing phase".) An important tool in the proposed workflow, which aims to present data in a km^2 grid, is the Joint Research Centre (JRC) dataset "Population density disaggregated with CORINE land cover 2000". This grid allocates commune population data to grid squares, mainly using CORINE land cover data. In one of the three proposed integration methods for socio-economic data, "proportional and weighted calculation", the

JRC dataset is used to distribute socioeconomic indicators to the grid. Challenge 6, concerned with urban data, utilizes the same population grid to assign population to Urban Morphological Zones (UMZs).

Tests of the reliability of the population grid in different settings, however, are scare. Previously, a comparison with Austrian reference data at the km^2 level has been performed. This showed an overall reduction by 50 percent in the disagreement with reference data, when compared to a non-weighted distribution of the population (Gallego: "Downscaling population density in the European Union with a land cover map and a point survey"). Using Swedish population register data, the population estimations of the grid has been further examined, focusing on 1) its local predications in different settings, and 2) the overall and internal estimations for UMZs of different sizes. These comparisons are presented in more detail in a technical report ("Using downscaled population for local data generation: A country-level evaluation"), and briefly summarized below.

In the first test of the population grid, register and grid populations were aggregated to km^2 squares, and absolute residuals for squares, individual municipalities and "municipality groups" (9 categories) calculated. In the two latter cases, residuals were summarized, but also presented per area unit and in relation to population size. The second test of the population grid concentrated on Swedish urban areas (UMZs). An important difference is that, in this context, a comparatively a comparatively large proportion of squares are actually inhabited. In this UMZ register population was compared with UMZ grid population, focusing on the overall population for each UMZ.

The first test showed that way local discrepancies are related to different municipalities depends on whether absolute residuals are just summarized or related to area or population. When absolute residual sums are related to municipality population size—arguably the most relevant and interesting way to approach the issue—we find an overall median error (gini-style) of about 50 percent. However, there are substantial differences between different types of municipalities. Rural municipalities are associated with the largest errors (median = 65 percent). Primarily, this is because the grid assigns population to many uninhabited squares that actually are uninhabited. Suburban municipalities exhibit the largest variation (range = 39 percentage points). In other words, the population grid works very well for certain suburban municipalities, but quite poorly in other suburban settings.

Concerning the second test, limited to overall predictive ability for each UMZ, we find that the population grid generally provides fairly good estimations of UMZ population sizes. There is a tendency towards overestimation when it comes to certain small UMZs, while otherwise there is a clear pattern of increased underestimation with increasing population. A plausible reason for the underestimation is that the grid overestimates areas with many buildings but small resident population, e.g. second home areas. The overestimation trend is harder to explain, but may have to do with varying degrees of over- and underestimation of the population of concerned city centers and suburban areas. All in all, while there are obvious limitations to the population grid, it is a quite reasonable tool for disaggregating socioeconomic data and—in particular—assigning population to UMZs.

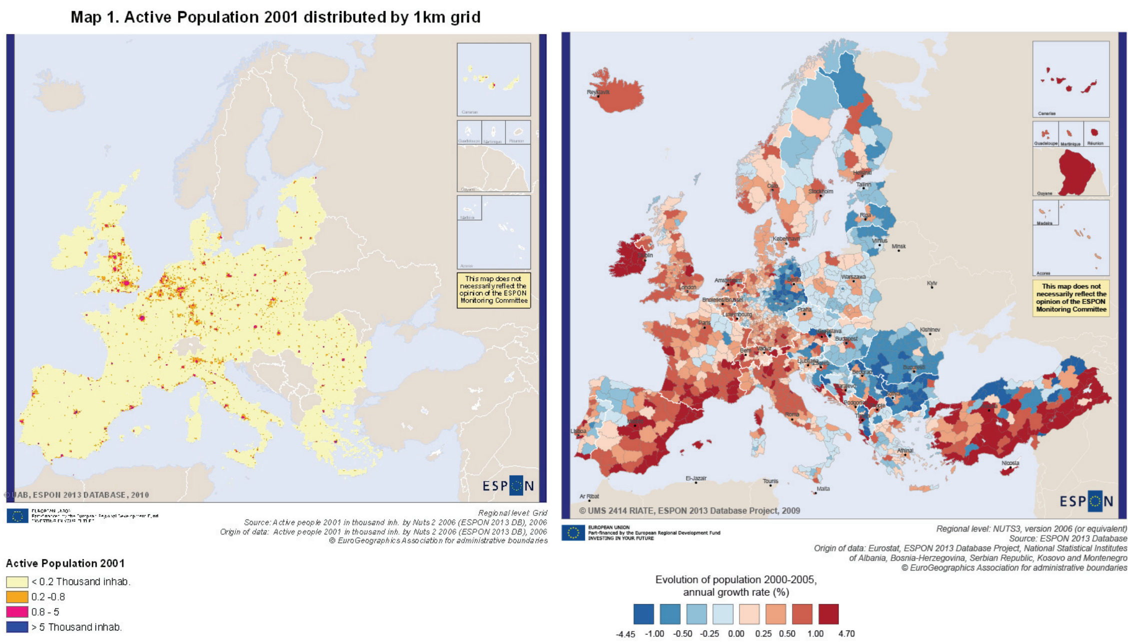
Other work within activity three include cooperation with Challenge 6 in further studying UMZs. Taking departure in the Swedish delimitation of cities and their population, a comparison between the Swedish delimitation of urban localities and UMZs is underway. This examination takes the national context into account (e.g. comparisons of rank-size graphs, number of cities per size class, etc.), but also addresses the local level by comparing the two databases in a GIS.

Within the first activity, concerned with availability and usefulness of survey data, the existing Eurostat surveys have constituted the main focus. The surveys that have been considered for further empirical work include ECHP (European Community Household Panel), LFS (Labour Force Survey), CIS (Community Innovation Survey) and SILC (Statistics on Income and Living Conditions). Taking into account their different focus, geographical and temporal scope, population and sample size, the Labour Force Survey has been identified as the one of most interest to address in the ESPON context. This work within activity four—which relates to the possibilities of integrating survey data in the ESPON database—has recently begun in earnest and expected results and deliveries will be elaborated on in the next section.

2.11 Cross-Challenge activities

Mapkit tool and mapping guide

Regular updates of the ESPON mapkit tool has been designed by UMS RIATE at the request of the ESPON Coordination Unit, and a technical report called "Mapping Guide" has been put on the ESPON website in order to make more easy the creation of maps. In both case, this document was made in order to take into account recent modifications in the ESPON design and corporate identity. Another contribution has been made by UMS RIATE in order to take into account the outputs of Challenge 5 (grid data) and Challenge 11 (Neighborhood). It appears indeed that the actual ESPON map template does not fit with grid data (as the reference meridian is not parallel with the European reference grid) and it does not insure a complete cartographic coverage of candidate countries (in particular Turkey). The new template presented in Malmö solves these two problems (Figure 11) but it has to be approved by the ESPON Monitoring Committee before further used by ESPON Projects.



Map 1. Active Population 2001 distributed by 1km grid

This map does not necessarily reflect the opinion of the ESPON Monitoring Committee
© UAB, ESPON 2013 DATABASE, 2010
ESPON
EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE
Active Population 2001
< 0.2 Thousand inhab.
0.2 -0.8
0.8 - 5
> 5 Thousand inhab.

Regional level: Grid
Source: Active people 2001 in thousand inh. by Nuts 2 2006 (ESPON 2013 DB), 2006
Origin of data: Active people 2001 in thousand inh. by Nuts 2 2006 (ESPON 2013 DB), 2006
© EuroGeographics Association for administrative boundaries

This map does not necessarily reflect the opinion of the ESPON Monitoring Committee
Lisboa
Madrid
Roma
Athinai
Nicosia
El-Jazair
Tounis
Malta
Ar Ribat
© UMS 2414 RIATE, ESPON 2013 Database Project, 2009
ESPON
EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE
Evolution of population 2000-2005, annual growth rate (%)
-4.45
-1.00
-0.50
-0.25
0.00
0.25
0.50
1.00
4.70

Regional level: NUTS3, version 2006 (or equivalent)
Source: ESPON 2013 Database
Origin of data: Eurostat, ESPON 2013 Database Project, National Statistical Institutes of Albania, Bosnia-Herzegovina, Serbian Republic, Kosovo and Montenegro
© EuroGeographics Association for administrative boundaries

Figure 11: A proposal of new ESPON template more adapted to grid data and neighborhood data (to be validated by ESPON MC)

External Networking with other statistical institutes

As stated in Challenge 7, the role of UL within the consortium involves as well the coordination of the networking activities foreseen under the seventh challenge. This challenge is aimed at foresting data exchange and expertise with the main data providers and potential external users of the ESPON 2013 DB. To this end, UL has the responsibility to draw the broad picture of networking activities with external organizations (i.e. Eurostat, DG Regio, EEA) and record all the information and data flows.

The networking activities are a crucial element within this project. For that reason, previous meetings have been organized to present the latest developments and establish new avenues of cooperation for information and data exchange. More recently, UL represented the ESPON 2013 DB project at Eurostat, during the working party on "Regional and Urban Statistics", 8-9 October 2009, in Luxembourg. This occasion constituted as well an opportunity to disseminate among EU representatives of NSIs some of the main scientific deliveries achieved by the project. Besides, it also paved the way to enhance the ESPON-Eurostat Action Plan on important topics, such as data availabilities, data integration, database structures and metadata expertise.

In addition to this, other events have contributed to consolidate the importance of networking activities. The ESPON workshops and seminars have demonstrated to be appropriate occasions to present the originalities and progresses of the ESPON 2013 DB, particularly the workshop on "Use of ESPON Institutions and Commission Services" on 6 May 2009, in Brussels, the workshop on "Approaching New Functional Areas", 5 November 2009, in Luxembourg, and the open seminar on "Territorial Development Opportunities in the Global Economic Recession", on 3-4 June 2009, in Prague. Ultimately, these deliveries represent a concrete application of the work plan defined and co-written by UL for the FIR concerning external networking with institutions and data suppliers.

Internal networking with other ESPON Projects

As it can be seen on (Figure 12) the dataflow inside the ESPON program implies regular contact between the ESPON DB project and the other ESPON projects from priority 1 and priority 2. It is a very important task that involved a lot of human resource, in particular from the lead partner RIATE. Basically, networking with ESPON DB appears at two crucial steps in the lifecycle of other ESPON projects.

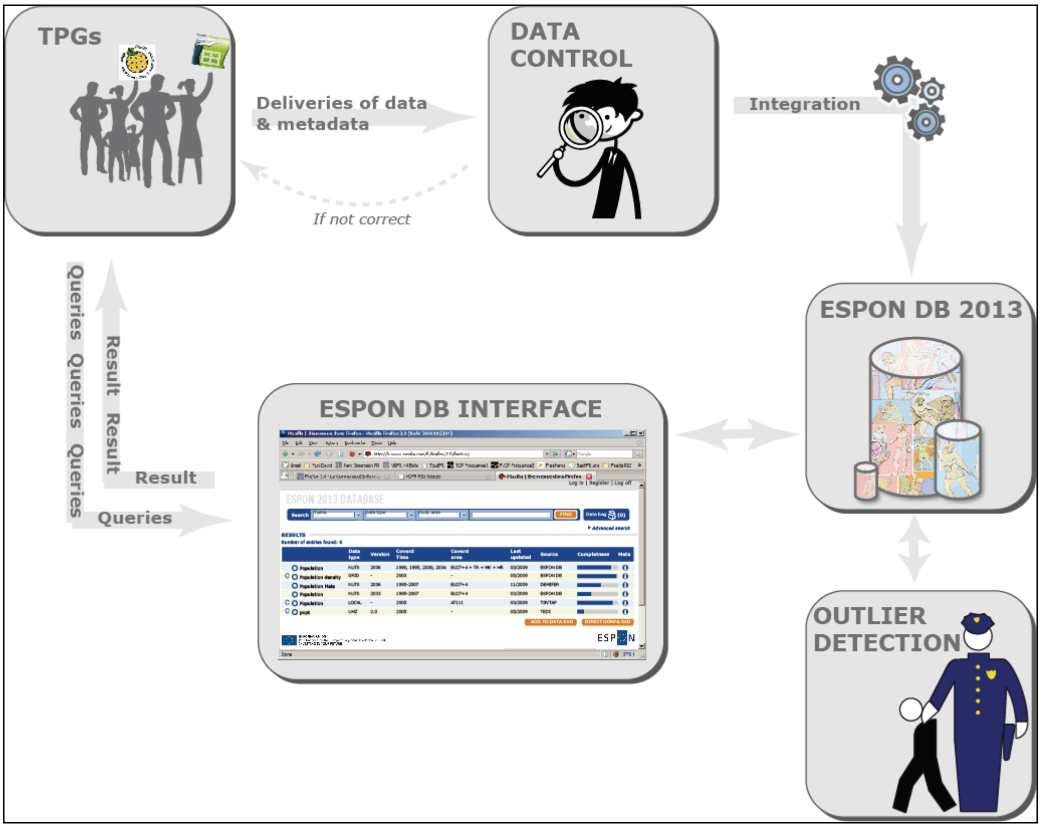
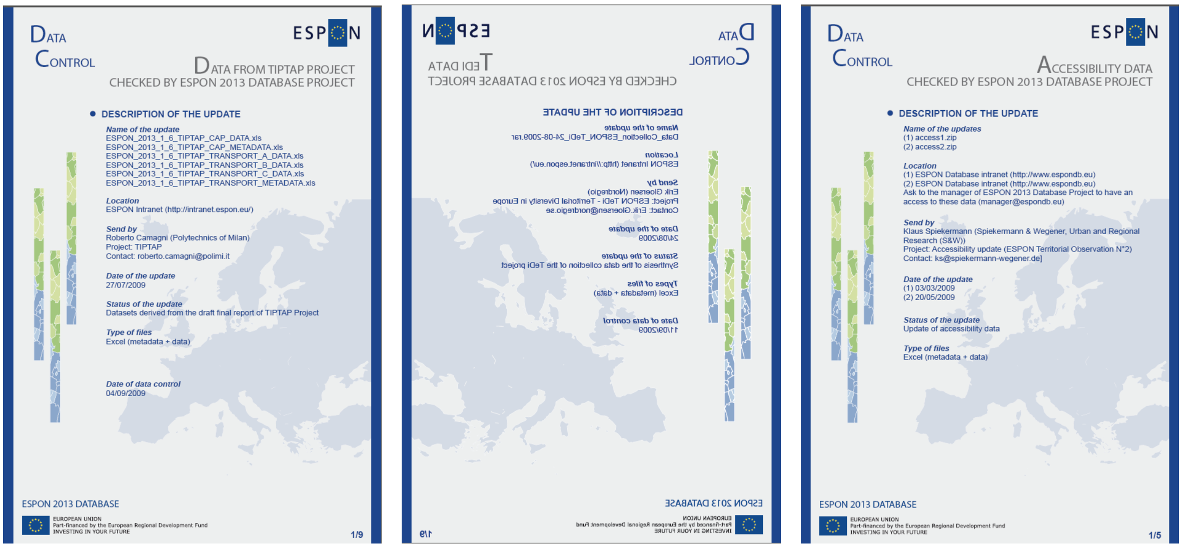


Figure 12: A general view of data flow inside ESPON

At the beginning of an ESPON project, the priority for Transnational Project Groups is to obtain basic data (GDP, Population, Activity…) and basic geometries, before to start the collection of more specific data on selected topics. When projects are dealing with more specific scales (World data, local data), more specific themes (environment, demography, …) or more specific geometries (network, cities, grid), the support from ESPON DB is rather based on the sending of draft technical reports elaborated in the different challenges. Finally, many project addressed very specific questions on the existence of particular type of data at EU level, with specific request on non EU countries (Switzerland, Norway, Iceland, Liechtenstein) and neighboring countries (Balkans). In the beginning of the ESPON DB project, we have allocated a contact person to each project, selected in the different project partners according to their thematic specialization. But it appears that in many cases, project partners addressed directly their request to the coordinator UMS RIATE which was subject to a strong pressure.

At the end of an ESPON project, the priority for the TPG's is to deliver their data and metadata in a form that fulfill the norms elaborated by ESPON DB project. In practice, the data are firstly sent to ESPON Coordination Unit that transfer it to ESPON DB project for validation, with very short delays in general. One more time, this tasks is generally concentrated on the lead partner RIATE who has developed an efficient methodology (Data Check) but remains submitted to a strong pressure because data check is a condition for the achievement of other ESPON projects. In most cases, direct contacts are necessary with TPG's before to obtain data and metadata in line with the objectives. It is clear that this step is actually consuming a huge amount of human resource, despite the fact that only 5 Priority 1 projects are currently running.



DATA CONTROL

ESPON

DATA FROM TIPTAP PROJECT
CHECKED BY ESPON 2013 DATABASE PROJECT

- DESCRIPTION OF THE UPDATE

Name of the update
ESPON_2013_1_6_TIPTAP_CAP_DATA.xls
ESPON_2013_1_6_TIPTAP_CAP_METADATA.xls
ESPON_2013_1_6_TIPTAP_TRANSPORT_A_DATA.xls
ESPON_2013_1_6_TIPTAP_TRANSPORT_B_DATA.xls
ESPON_2013_1_6_TIPTAP_TRANSPORT_C_DATA.xls
ESPON_2013_1_6_TIPTAP_TRANSPORT_METADATA.xls

Location
ESPON Intranet (http://intranet.espon.eu/)

Send by
Roberto Camagni (Polytechnics of Milan)
Project: TIPTAP
Contact: roberto.camagni@polimi.it

Date of the update
27/07/2009

Status of the update
Datasets derived from the draft final report of TIPTAP Project

Type of files
Excel (metadata + data)

Date of data control
04/09/2009

ESPON 2013 DATABASE

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

1/9

DATA CONTROL

ESPON

TEDI DATA
CHECKED BY ESPON 2013 DATABASE PROJECT

DESCRIPTION OF THE UPDATE

Name of the update
Data_Collection_ESPON_TeDi_24-08-2009.rar

Location
ESPON Intranet (http://intranet.espon.eu/)

Send by
Erik Gloersen (Nordregio)
Project: ESPON TeDi - Territorial Diversity in Europe
Contact: Erik.Gloersen@nordregio.se

Date of the update
24/08/2009

Status of the update
Synthesis of the data collection of the TeDi project

Types of files
Excel (metadata + data)

Date of data control
11/09/2009

ESPON 2013 DATABASE

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

1/9

DATA CONTROL

ESPON

ACCESSIBILITY DATA
CHECKED BY ESPON 2013 DATABASE PROJECT

- DESCRIPTION OF THE UPDATE

Name of the updates
(1) access1.zip
(2) access2.zip

Location
(1) ESPON Database intranet (http://www.espondb.eu)
(2) ESPON Database intranet (http://www.espondb.eu)
Ask to the manager of ESPON 2013 Database Project to have an access to these data (manager@espondb.eu)

Send by
Klaus Spiekermann (Spiekermann & Wegener, Urban and Regional Research (S&W))
Project: Accessibility update (ESPON Territorial Observation N°2)
Contact: ks@spiekermann-wegener.de]

Date of the update
(1) 03/03/2009
(2) 20/05/2009

Status of the update
Update of accessibility data

Type of files
Excel (metadata + data)

ESPON 2013 DATABASE

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

1/5

Figure 13: Data check of projects from priority 1, 2 and 3

It is important to notice that Data Checks (Figure 13) are not limited to projects of priority 1 (*e.g. TIPTAP*) but is also requested for all projects from priority 2 (*e.g. TEDI*) or updates of previous ESPON data realized under priority 3 (*e.g. Accessibility Data*). We are afraid that this task is consuming too much time actually and will imply some reallocation between work packages.

Support to ESPON Coordination Unit

Some of the above mentioned tasks can be considered as part of this fuzzy item called "Support to ESPON coordination units". For example, we have attained regularly the workshop organized either by ESPON or by Eurostat and contributed through presentation of ESPON DB results.

Out of this, we also contributed actively to the different ESPON seminars of Bordeaux, Prag and Malmö, either by presentations of our results (in particular in Malmö where we organized a special session and edited a booklet of main discoveries made) or by realization of posters. In the case of Bordeaux's meeting, the poster was collected outside ESPON and could not directly be considered as an extra work. But in the case of Prag, we were invited to prepare posters on selected topics decided by ESPON CU. More than one month of work was implied in this task, with the frustrating result that all of our initial proposals of poster was finally refused and replaced by simple maps extract from the recent report. It was clearly a vast of effort and resource as it would have been more simple to inform directly of the maps to put into poster, instead of letting us prepare a more conceptual and creative solutions.

3 Expected activities until the final report

3.1 Time series issues: from conceptualization to operational results

Time series data challenge activities until the final report will focus on the following main tasks:

The results of the exploration of NUTS territorial changes will be improved concerning the New Member States. These countries did not have NUTS before 2003. In addition, the available sources of data don't allow the reconstruction of earlier NUTS versions. This issue requires a historical analysis of the national administrative boundaries of these countries.

Based on the results of the study of nuts changes, we will implement time series data. This has already been underway since the end of 2009 on the Cohesion Reports data. It consists in collecting data and identifying the latest NUTS versions. The main aim of this work is to produce comparative maps.

The current version of the Technical Report will be improved and updated. Namely, we would like to use the 1 km grid data for modeling the NUTS change. It will make possible to answer to the following question: what is the share of the population concerned by a NUTS change? Another issue consists to continue to explore Eurostat historical database in order to build time-series in the old NUTS versions. This, plus the elements already developed will allow us to develop the framework of estimation of missing values. Some first tests will be done from the results of the exploration of territorial changes and archived databases.

The framework of time series management process automation will be developed with coordination with computer science team. The aim of this task is to produce harmonized time series data.

3.2 Finalized "World Dictionary of Units"

The aim of the challenge focusing on World Dictionary of Units is to provide to ESPON a final version of the World Dictionary of Units, with a clear documentation mentioning the choices made in the Final Technical Report associated to the delivery.

Associated to the this dictionary, we aim to deliver an ESPON World Database 2.0, including statistical data and an updated version of the geometries to ensure the compliance with the codes of the territorial units.

Another issue consists to finalize the "Gap Tracker Tool" by proposing a systemic way for comparing data from World databases and Eurostat regional data. This will be based on the knowledge accumulated during the first period of the project.

All these updates will contribute to update the Technical Report on that topic.

3.3 Focusing on the SIRE database exploration

During the first steps of the project, the local data challenges has essentially focused its research in a bottom-up approach, based on a systemic data exploration and collection by National Statistical Institutes. This demarche is very useful to provide guide-lines and methodological tools for ESPON Projects and avoid many mistakes in the data collection phase. However, the heterogeneity of the information and the great number of units make this strategy very long to implement; it is indeed difficult to establish a concrete agenda for the delivery of complete datasets for this strategy.

Consequently, we aim to move the work by a top-down approach, focusing on existing databases at LAU2 level. The first exploration in that way will be developed by checking what could be done thanks to the SIRE database, delivered by Eurostat in 1998. The main interest of this database is namely to provide basic data (derived from National censuses) for more than 100 000 municipalities in the European Union from the 1980's to the 1990's. However, no geometries are associated to the SIRE Database. In that direction, the work of ESPON Database Project could focus in different ways (non exhaustive):

- Make an inventory of available data in the SIRE database. What is the degree of completeness of this data? What are the indicators contented in the database, which could be updated? Missing values could be estimated?
- Evaluate the concordance of the SIRE database with the geometries provided by Eurogeographics at LAU2 level. It will imply probably to develop coding conversion tables to make the link.
- Link the bottom-up and the top-down approach by working on the concordance of SIRE database with current data available on National Statistical Institutes. One can imagine a case-study applied to Czech Republic or Slovakia.
- Think about existing possibilities for updating the SIRE database in the ESPON Database architecture.

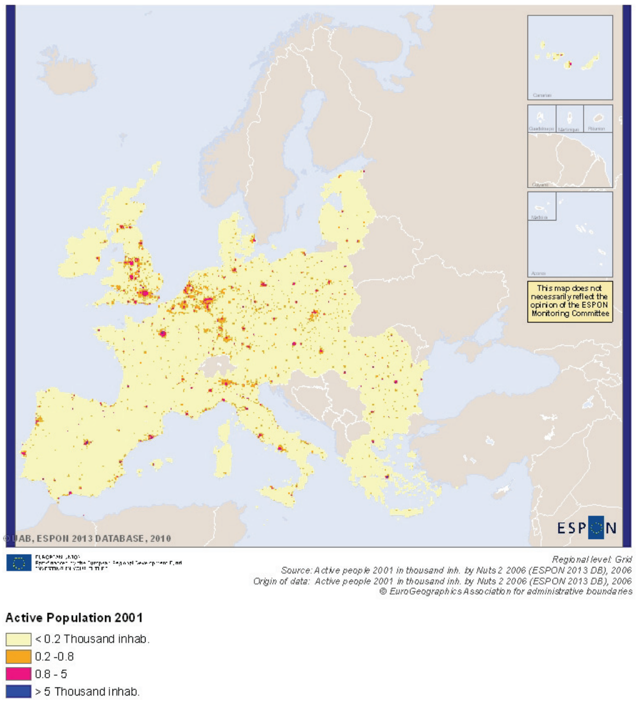
3.4 Improvement of the Integration of socio economic and environmental information methodologies

Since the first interim report (March 2009), the UAB team have been leading the challenge 5 that tries to answer to the need of "Combination of heterogeneous sources-balancing Eurostat data" defined in the ESPON 2013 DB Program. In this sense, three different integration methods have been defined and tested with different ESPON socio-economic variables.

The activities defined for the last period of the Espon 2013 DB project (January – December 2010) will be focus on the treatment of the most updated ESPON socio-economic variables and its integration with environmental data by building and distributing an ESPON OLAP Cube.

New tests based on the most updated information and ESPON projects results.

Map 1. Active Population 2001 distributed by 1km grid


This map does not necessarily reflect the opinion of the ESPON Monitoring Committee
© UAB, ESPON 2013 DATABASE, 2010
ESPON
Active Population 2001
< 0.2 Thousand inhab.
0.2 -0.8
0.8 - 5
> 5 Thousand inhab.

Regional level: Grid
Source: Active people 2001 in thousand inh. by Nuts 2 2006 (ESPON 2013 DB), 2006
Origin of data: Active people 2001 in thousand inh. by Nuts 2 2006 (ESPON 2013 DB), 2006
© EuroGeographics Association for administrative boundaries

New tests have been started using the most updated territorial information (Nuts3 and Nuts2 2006) and new information coming from ESPON 2013 projects.

The new tests will be based on the third aggregation method "Proportional and weighted information" that gives better results, and some added value to the downscaling.

The next map shows one of these new tests developed (Map1. Active people 2001 distributed by 1km Grid).Source: Active people 2001 in thousand inhabitants distributed by Nuts 2 2006 (ESPON 2013 DB). Weighted by JRC's population density grid6)

Integration of socio-economic and environmental information

A first ESPON OLAP cube will be built and distributed containing a delimited small number of Espon socio-economic variables in order to test the possibilities of analysis behind this technology.

[6] http://www.eea.europa.eu/data-and-maps/data/population-density-disaggregated-with-corine-land-cover-2000-1

The OLAP technology[7] use a multidimensional data model, allowing complex analytical and ad-hoc queries with a rapid execution time.

In the case of ESPON, the OLAP cube will consist on the ESPON socio-economic variables as numerical attributes or measures that will be aggregated using a set of dimensions.

The user will be able to ask to the ESPON-OLAP cube questions taking into account socioeconomic variables or indicators and environmental data. For example, if we integrate the GDP and CLC in an OLAP cube, we could analyse which land cover flows occur by different GDP ranges, and, in the end, get the results on a NUTS3, NUTS2 or country basis.

Improvement of the methodologies proposed

The addition of new methodological approaches is not excluded.

Some aspects that have to be deeply analyzed are:

Treatment of administrative units with no data values

Differences between geographical extents, for example between Nuts 3 2006 layer and Corine Land Cover

Improvement of the disaggregation performance in terms of time and manageability of the final layer

Integration with the ESPON 2013 Database

During the next months a discussion around the integration of data grid in the ESPON 2013 Database will be opened between all the project partners involved in order to find the best solution to make the ESPON OLAP Cube available to users.

3.5 Validation of cities databases integration

For December 2010, the work will be progress following three directions.

Technical report on urban ESPON DB metadata (LUZ, MUA, UMZ)

Technical Report will be delivered integrating the results of our semantic expertise on LUZ, MUA and UMZ metadata.

A first part of this technical report will be available for Madrid Meeting (June 2010) and will be organized in two parts:

- A general table with four sections (Production context, Data Contents, Availability and Specification, see Table 1 in FIR for more details)

[7] Some OLAP information resources:

1- http://en.wikipedia.org/wiki/Online_analytical_processing

2- http://www.cs.sfu.ca/CC/459/han/papers/chaudhuri97.pdf

3- http://es.wikipedia.org/wiki/Cubo_OLAP

- Tables and maps showing, for each data base the geographical coverage, the population and name of largest cities (> 100 000 inhabitants)

The second part of this technical report will be available for the next ESPON general meeting (December 2010) and will summarize the results of databases integration. For cities larger than 100 000 inhabitants (about 500 cities), we will overlay the MUA, LUZ and UMZ and will analyze the results. First, we will try to see if there are systematic deviations between the two morphological objects (MUA and UMZ). The purpose will be then to evaluate the differences between UMZ and LUZ delineations, in a geometric perspective at first: how UMZ are embedded inside LUZ delineations, what are the average differences in terms of surfaces? Are there regularities observed according to countries? According to size classes of cities (either in terms of spatial extension or population)? We will try then to qualify the discordances between UMZ and LUZ in terms of volume of population concerned, exploring the same sources of potential regularities.

Finally, for the most specific units (according to the gap between delineations), we will test a methodology for qualifying the population, using SIRE indicators. For instance, we will evaluate the type of activity of people living in these "non included" parts.

Integrating databases into the ESPON DB

The different tasks devoted to storage will be completed in the next months. First, we will finish to fill the metadata corresponding to MUA and UMZ data, and will provide the geometry and the correspondence between UMZ and the LAU2 level. Furthermore we will begin the integration of the most relevant indicators of the UA databases, as well as their geometrical delineation if they are available.

Further work on UMZ (naming and validation)

As specified in the Technical Report "Naming UMZ", a further work is necessary to discuss and probably improve the choices that have been made. For United Kingdom, Ireland and Portugal, LAU1 have been used but we need to check and correct some results, using LAU2 in some cases. We have begun a deeper work, using national sources and checking the names given to cities in national census (as a result of a first investigation, in Portugal the map of cidades available at http://sig.ine.pt, and for Great-Britain the census output area boundaries that will be required at http://www.statistics.gov.uk/census2001. For Ireland, there are less than 20 UMZ larger than 10000 inhabitants and the results will be easily checked without trying to find automatic processes).

We have also planned to go deeper in the validation processes. Working with challenge 12 (Umeå University), we will use the Swedish grid of population (2000), in order to complete the results obtained with the V4.1 population density grid (Gallego 2007). If we have some more time, we will try to enlarge our case studies by adding new countries, with national morphological urban areas, such as Great Britain or Ireland.

3.6 ESPON DB application

Finalizing ESPON Thesaurus

The current deliveries points to considerable future work, of both empirical and conceptual nature. At the empirical level, it is evident that we need to refine our understanding of what is being measured to better allocate each ESPON indicator into a specific theme and sub-theme. With that regard, the quality of metadata is of extreme importance to pursue that goal. Perhaps more fundamentally, there are some open questions at the conceptual level. Primarily, future deliveries should validate the usability of the results presented in previous occasions. Secondly, it should better understand what kind of knowledge in being labeled as such to ease data allocation and therefore optimize naming conventions and coding schemes. That is, extract commonly used words from qualitative and unstructured data to improve ESPON 2013 DB thematic structure and eventually offer some consistency on how to name indicators. Consequently, some of the difficulties that emerged on our latest experiments should be further investigated by means of text mining techniques to uncover complex patterns of semantic relationships and contribute to the definition of an ESPON thesaurus.

As explained in our earlier contributions, text is a collection of unstructured data. In this sense, extracting valuable patterns depends on how data is structured. Due to the fact that textual data is unstructured and often formatted inconveniently it is necessary to follow certain procedures to ensure some consistency to the overall process. The first step is obviously to collect data. In our case this represents any relevant document, study or policy note that addresses ESPON evidence and results, mainly in terms of comparable information on territorial dynamics and potentials for development.

We have initially identified the final reports of the ESPON 2006 Program. This desk research expanded then to documents published by other sources (e.g. European Commission, European Parliament, Committee of the Regions) covering a wide range of themes that touch upon knowledge developed by ESPON. In total, we have gathered approximately 200 documents addressing topics that range from spatial planning to environmental hazards or polycentric development to accessibility in Europe. Altogether these documents constitute a large textual database that needs to be structured as efficiently as possible before addressing any preliminary analysis.

In order to achieve valuable results we have to bear in mind that textual data is a complex conjunction of words and phrases that frequently need to be considered as a whole. Besides, a quite huge amount of dependency is present and should not be ignored. In the same way, it is important to overcome word and semantic ambiguities that may adversely influence our experiments. Altogether, these aspects increase the ambiguity of having solid and robust results. The methodological approach is therefore of vital importance to comply with specific problems that might emerge from our analyses.

For this purpose, a particular emphasis should be given to information visualization techniques. Due to the complexity of mining textual data such tools play a central role in increasing the performance of text analysis process and hopefully reduce the risk of having biased views. Besides, it can also increase the commitment as the process unfolds. Taking into account the conjunction of these different aspects it seems logical that the ESPON 2013 DB needs an appropriate method to allocate data by themes and sub-themes. We argue that text mining techniques, together with visualization tools, have the capacity to analyze the content of pre-defined corpora in effective, communicative way.

Within this context, we propose to evaluate a non-exhaustive group of text mining tools in terms of strengths and potential limitations to analyze unstructured data. The body of literature on this topic has explored different approaches that range from extraction to clustering. Such developments that go beyond frequency ranking could positively be applied to mine unstructured textual data and retrieve information that is difficult to discover by other means.

However, this methodological approach reveals some constraints. The fact that the majority of those tools have been designed for very specific purposes, covering different aspects of text mining and visualization capabilities, constitutes an unfavoured stimulus. However, based on what we could learn from such assessments, including scope of data types (e.g. journal articles, technical reports, policy notes) and visualisation options, it would definitively convey an important encouragement to improve results and, ultimately, identify keywords and clusters of keywords to improve the quality of the ESPON 2013 DB search engine for querying data.

3.7 Consolidating the database

From the implementation point of view, the efforts of the ESPON 2013 Database project will be concentrated on three main directions until the Final Report: the evolution of the Web interface for the database, the continuous updating of the database itself with indicators provided by other ESPON projects and the development of a specific administration interface allowing the maintenance of the databases.

The evolution of the Web interface

Although the delivery for the SIR includes a fully functional version of the ESPON Web Application, it is expected that there will be many modifications to bring to the application, either in order to fix bugs or to improve the usability of the application and to make it as compatible as possible with the requirements of the CU and of the final users. Expected improvements will probably define different data search profiles (and, respectively, different search interfaces) for different use scenarios. These improvements should be based on users feedback and suggestions collected during the ESPON seminars and our discussions with the CU.

The continuous updating of the ESPON database

As the ESPON 2013 program makes progress, more and more data are available for integration in the ESPON database. This requires a considerable effort for data checking and correcting at different levels: a major manual checking and correction is done by lead partner RIATE, and then the files undergo a second automatic checking and manual correction procedure at LIG. Once the files are completely semantically and syntactically corrected, they can be put into the ESPON database. Unfortunately, in practice there are many errors of all types (syntactic and semantic), so the verification and correction process is quite long and tedious. This is mainly due to two factors: first the misunderstanding of the data/metadata profile by the ESPON projects, and second the manipulation errors (mistyping, wrong copy/paste, etc.) which are easy to make in Excel files. We think that when the ESPON projects start using the Web metadata editor, many of these errors will be avoided, at least for the XML format of the metadata files. Until then, the effort for data checking, correction and insertion in the database is significant and it goes beyond what had been initially anticipated.

Implementation of a custom tailored interface for data manipulation within the ESPON database

Although in the technical documents describing the ESPON 2013 Database Application we refer mostly to the ESPON database, the actual application is much more complex. It includes a back office with another database (which we call the megabase) which serves as a warehouse and a factory for the ESPON database, which is only a showcase of finished products (that is, harmonized data) optimized for fast querying. It is in the megabase that the ontologies described above are stored and used for insuring the data consistency and for enriching the metadata. Besides the megabase, the back office part of the ESPON database includes some software for data acquisition, which allows filling the megabase by importing data and metadata files and detecting syntactical errors in the files when importation fails. It also includes some software for data harmonization, completion and exportation towards the ESPON database.

From the beginning, we intended to use existing open source or free data management software for the back office of the ESPON database. For the first version of the ESPON database, in which there was only one database with a simpler schema, we used the open source spatial data management tool GeoKettle (a spatially-enabled version of Pentaho Data Integration, formerly known as Kettle) for data acquisition in the ESPON database.

Our conclusions from using this tool were that this software is too heterogeneous, insufficiently documented and unreliable to be used by others than database experts.

Due to the fact that the megabase has a very complex structure (much more than the first version of ESPON database), it became clear that such a tool would not be sufficient to meet the needs of the ESPON 2013 Database project and for the future maintenance of the ESPON database in general.

As a consequence, we started working on a custom tailored data management application for the ESPON database. In the same spirit as before, we include, as much as possible, existing open source or free software libraries (like the POI library of the Apache project for Excel file manipulation, or the STEAMER-LIG team’s internal library TTSQL for text file manipulation), but it also relies on code and visual interfaces developed specifically for the ESPON database.

Our aim is then to provide a functional database administration interface for the ESPON DB. The interface will be built as a Java desktop application that will allow a database administrator to import easily data files into the megabase. It will also allow some basic data harmonization features (like merging data from different datasets) and to export these data to the ESPON database, making them available to the ESPON community and the public.

3.8 New methods for outlier detections

Our second interim report provides an introduction to the detection of logical input errors and statistical outliers (i.e. exceptional values) for data sets of the ESPON 2013 Database. Some important aspatial and spatial techniques have been introduced and demonstrated within the R statistical computing environment.

The field of robust statistics and outlier detection is both extremely large and diverse, and as such can not be comprehensively reviewed within the terms of reference of this report. However, outlier detection techniques applicable (or designed for) *spatial* data sets are currently not as developed as those for *aspatial* applications.

Robust methods may be useful for the ESPON DB 2013 project. Classical statistical methods may produce results which are distorted by the presence of outliers, as their underlying assumptions do not consider that outliers require a different treatment from the rest of the data. Robust methods, on the other hand, are intended to be resistant to the effects of outliers, and often identify them as part of the analytical process. This enables the development of more relevant modeling techniques where exceptional data are thought to occur, and that we can identify exceptional observations with some confidence, and can start to determine why they might have unusually high or low values.

We cannot ignore the spatial component. Much policy formulation, evaluation and adoption that takes place in the EU is spatially specific; therefore the evidence which is required in the processes must take this into account. The spatial component provides the context for determining which observations have locally exceptional values.

In this respect, our current research is focused on this specific area of model development. Here robust versions of geographically weighted summary statistics (GWSS), geographically weighted regression (GWR) and geographically weighted principal component analysis (GWPCA) are to the fore, as they allow the detection of outliers in both univariate and multivariate spatial data sets. We have already tested some of these methods on 'real' datasets and will continue to do so.

Our expected deliveries for the final report of this phase of the ESPON project will be firmly based on the analytical techniques described and applied in our second interim report. However we will now hone these procedures using a concrete, real-life data set rather than the fabricated data set used here. This new data set will no doubt present some new analytical and practical challenges that have not been considered. This should enhance the detection methodology, which may need to include the addition of further techniques.

To this end, we will also introduce a selection of the robust geographically weighted techniques that we are currently working on (see the next section). An improved version of Hawkins' spatial outlier test is also under development, as is a robust version of the local Moran's I statistic (with respect to outlier identification).

3.9 Improve the quality and the quantity of data in neighbouring countries

The continuous updating of the ESPON database

The establishment of contacts and regular dataflow with the Western Balkans countries and Turkey's National Statistical Organisations (NSO), Eurostat and DG Regio are strongly required to ensure a regular dataflow between them and ESPON.

The assessment of the availability and the quality of the data for the WB countries and Turkey (Candidate Countries / CC and Potential Candidate Countries / PCC) has almost advanced considerably on the basis of the data available in the official websites of the CC / PCC NSO, other sources etc (a corresponding Technical report is included in the SIR).

Further improvement of this work could be possible with the establishment of contacts and regular dataflow with the CC / PCC NSOs, Eurostat and DG Regio.

A first possible strategy on this issue is to create a cooperation scheme among the ESPON Database project, the ESPON program (MC, MA and CU), Eurostat, DG Regio and the CC / PCC NSOs.

In case it appears that the implementation of this strategy is difficult, we could alternatively implement a second strategy of establishment of systematic non – official contacts between the ESPON Database project –and the ESPON programme- with the CC / PCC NSOs.

We will use the cooperation with the CC / PCC NSOs:

- To further check the conformity of the administrative divisions of the CC / PCC (except Croatia, FYROM and Turkey) with the EU NUTS rules.
- To clarify the concepts used in some datasets provided by the NSOs.
- To ask some additional datasets and updates.

We expect to deliver on that topic a short Technical Report on the cooperation with the CC / PCC NSOs, Eurostat and DG Regio at the end of 2010.

Evaluation of intermediate values

We will evaluate the values of several indicators (population etc) for the interval between the years for which we have official data.

The deliverable will be tables (Excel files) of estimated values and respective metadata.

Time of submission: End of the 1st semester 2010.

Additional quality checks

We have already compared the values of several indicators compiled from different sources. We will make additional checks/ comparisons, for example between the NSOs data and the respective ONU data.

Deliverable:

Short Technical Report describing the results of this task, respective Tables.

Time of submission: End of the 1st semester 2010.

Enlargement of the scope of the issues to be studied / Urban data

According to the enlargement of the content of the Database for the ESPON countries and the requests from other ESPON 2013 projects, we will collect statistics on similar issues for the CC/ PCC.

We will focus in particular on the urban data, making it possible to enlarge the urban database elaborated by ESPON 2006 project and further developed by ESPON 2013.

Deliverable:

Short Technical Report describing the results of this task, respective Tables to be included in the Database.

Time of submission: October 2010.

Final Technical Report on the Territorial data for the Western Balkans and Turkey

This Report will integrate the assessment of the availability and quality of data on Western Balkans and Turkey as well as definitions used for some of the datasets.

Deliverable:

Technical report

Time of submission: End 2010 (before submission of the Final Report).

3.10 Analyse the relation between regional dimension and existing surveys

A central point of departure for activity four is the extent to which regional dimensions are captured in existing surveys. First, the extent to which information with spatial meaning or connotations is collected is interesting as such, by indicating for which subject areas and data the regional dimension is considered of special interest. Second, such information can be utilized to facilitate or enhance geographical disaggregation of survey data.

When it comes to the Labour Force Survey, it appears that only very few attributes are possible to access combined with the NUTS 3 level. One pivot table provides population by sex, age group, nationality, degree of urbanization (three categories of number of inhabitants/km^2) and labor status. Beyond that, there is a series of tables labeled "LFS regional series" with several tables presenting economically active population by sex and age; employment by sex and age, education level, economic activity, professional status, full-/part-time and hours worked; and unemployment by somewhat disparate categories. Most of this information, however, is only presented at the NUTS 1 or NUTS 2 level.

So, on the first issue, as revealed by the tables possible to construct from Eurostat's net database regarding labor force properties from national surveys, quite a gloomy picture emerges. Only the most basic information regarding labor participation and unemployment has been deemed interesting enough to publish with a high degree of spatial detail. Two explanations are possible: 1) Eurostat's judgment of the insecurity and sampling error in the survey data prohibits presentation of other labor force properties at a detailed regional level. 2) Considering the space and effort available for constructing tables, the choice reflects current (lack of) interest in spatial outcomes within countries compared to between countries.

Concerning the second issue, we work on two different paths. The first and simpler one investigates whether or not information in the available tables can be used to produce some kind of spatial statistics at the regional and local level with any degree of reliability. As usual, the test case is Sweden and entails a comparison with observables for the same spatial subdivisions. For the national area and county levels, the test mainly reveals the extent of the sampling error in the Labour Force Survey. For the km^2 level, available data is allocated with the help of the population estimates of the JRC population grid. One idea to test is whether the information regarding "degree of urbanization" imputed on km^2 per county improves estimates beyond a pure random assignment. The reason for performing this simple analysis is to get an understanding of what is possible to achieve only based on data that is available to anyone.

The more ambitious path requires access to the underlying individual observations in the labor force surveys (for our test only Sweden), including address for every such person. Then, it will be up to us to apply statistical procedures for merging survey and aggregate register data into spatial estimates for several indicators in the surveys, such as unemployment, hours worked, etc. Would survey data like civil status, education level, profession, etc. improve the estimation of the spatial outcome on labor force participation and performance?

3.11 Cross-Challenge activities

Generally speaking, the cross-challenge activities for the remaining period will follow the same pattern than what has been described in section 2.13 (update of mapkit tool, data check of reports, internal and external networking). The only important points to mention are related to the so-called activity of "support to

ESPON Coordination Unit" for which two specific request has been addressed in February 2010 by our project officer.

Organisation of an ESPON Workshop

We have been invited by ESPON Coordination Unit to take the responsibility of the organisation of an ESPON Workshop in Luxembourg on a subject of interest related to database. In principle, we have agreed to do it and selected the question of "Time series and estimation of missing values" that appeared to be the most relevant. We are currently discussing with ESPON Coordination Unit on the choice of an agenda for this workhop and a selection of researchers and stakeholders to be invited.

Support for the elaboration of the ESPON Synthesis Report

We have also been invited to contribute to the preparation of the ESPON Synthesis Report that ESPON Program as a whole should deliver for its mid-term report in the autumn 2010. After discussion with the direction of the ESPON Coordination Unit, we have agreed in principle to realise the illustrations (maps, figures, diagram) of this report but fully excluded to contribute to the written part of the work which is delegated to external experts of the CU. We have also mentioned as a necessary condition of a participation to have some freedom in the elaboration of maps and figures, even if our proposals has to be finally controlled and approved by ESPON MC and ESPON CU (otherwise, we had also proposed to execute simply what was decided with "zero initiative", but it was not the wish of ESPON coordination Unit). Last but not least, we consider as very important that the map initially prepared by other ESPON projects would be sent for approbation to their initial authors after we have introduced transformation in order to harmonize and fill eventual statistical gaps. We will indeed be obliged to redraw some maps in order, for example, to adjust the choice of palette in all the synthesis report, but we have to check with the authors that it does not alter their initial creation.

It is important to add that we have accepted this two tasks "in principle", but we have to check if it is really feasible, according to the amount of human resources actually spent on the work package "Support to ESPON Coordination Unit". Actually, a lot of human resources is vasted for administrative tasks, due to the complexity of ESPON reporting, and we do not exclude to transfer part of the other resources of the project to these administrative tasks (*see. 1.3*). in this case, we cannot guarantee the realisation of the above mentioned actions.

4 Perspectives: needed improvements

4.1 General options

In this section, we address firstly some general question concerning the general structure of the follower of ESPON DB project for the period 2011-2013

4.1.1 OPTION 1 : One large ESPON DB II project or several medium-sized ?

It is not obvious to decide what is the best solution between the two options of (1) reconducting a large scale project for ESPON Database or (2) splitting it in several medium-sized projects.

The option of re-conducting large-scale projects has serious advantage from scientific point of view but is, according to our experience (*see. 1.2*), not sustainable from administrative point of view. It is clear that a really integrated database, able to take into account all dimensions of territorial cohesion, implies a network or project partners specialized in different fields, both thematic (environment, social, economic, transport) and technical (cartography, computer science, ergonomic, spatial analysis, statistics). And this network of project partners has to be completed by experts specialized in specific topics like exploitation of surveys (urban audit, labor force) or specific territories (Balkans, remote territories). It is the solution that was chosen in ESPON DB I project, but we consider it difficult to recommend, as it appears to be too difficult to manage from the administrative point of view.

The option of splitting tasks in different medium-scale project has of course some clear shortcomings from the scientific point of view (no general view, lack of integration of results, multiple partners for other projects). But it would make incredibly easier the management, both for ESPON coordination unit and for projects coordinator, as they would be limited to 2 to 4 partners each instead of 6 to 8. With this option, we could imagine having (for example), one project specialized in the storage of data (control, check, diffusion), one project specialized in the evaluation of time series and estimation of missing values, one project specialized in the integration of various geographical objects (grid to nuts and nuts to grid, cities and networks, etc). To be efficient, this project should organize regular meetings and exchange experiences on a regular basis. But they would be autonomous from administrative point of view and therefore more able to face the administrative burden.

4.1.2 OPTION 2 : Building an open database network

Because of the problems with the agenda of the beginning of ESPON (see. 1.2) it was not possible to establish such a network between ESPON DB projects and other projects of priority 1, 2 and 3. It is only during the ESPON Seminar of Bordeaux, Prague and Malmö that direct contacts could be established in order to discuss general questions like metadata structure, data checking issue and, exchange of results and experiences. With the enlargement of the ESPON Program to 30 new funded projects in February 2010, it appears crucial to build such a database network based on some practical tools.

Building a web forum for discussion on data and tools: The experts of the ESPON DB project can propose many answers to the questions of the other projects, but there are many situations were the best answer for the question of a project X can be given by another project Y. There are also common questions (about cities, grid data …) that could serve a general topic where specialists from different projects can offer vivid ideas and inputs. We propose therefore that the next responsible of the ESPON DB take the initiative to build such a forum.

Opening the collection of technical reports to other ESPON projects. The ESPON DB project has opened a collection of technical report that appears as a very important element of promotion and diffusion of ESPON results. This collection should not be closed in 2011 with the end of our projects, but followed by the future responsible of ESPON DB and, ideally, enlarged to the other ESPON projects, in particular of priority 1 and priority 3. The technical reports are complementary to the final reports that focused on political aspects and empirical results and have increased their credibility. We suggest therefore that projects which have designed interesting methods for data collection or elaboration could contribute to the collection of ESPON technical reports.

Establishing more direct contacts with statistical institutes at national and European levels. Currently, the contacts between ESPON DB projects and official institutes are strictly controlled by the ESPON Coordination Unit and submitted to an annual action plan in the case of Eurostat. Contrary to the other ESPON project of priority 1 or 2 (who developed direct informal contacts), the ESPON DB project is the only one that cannot launch contacts easily. This is obviously a counter-productive strategy and we strongly recommend for the future to apply a more flexible framework and give more degrees of initiative to ESPON DB project, as it is the case for the other ESPON projects.

Make ESPON an integrated actor of the European statistical system. The data that are stored in the ESPON database are not all copyright free and cannot be all disseminated outside. Some restrictions are introduced, which are necessary in order to obtain certain data for internal use by ESPON projects. But many data produced by ESPON can be disseminated and we suggest that all ESPON metadata should be freely visible to the non-ESPON world, even if the data themselves are subject to restrictions. Sharing metadata is necessary according to INSPIRE directive. But it is clearly a crucial way for a better awareness of ESPON work toward external organisations.

4.1.3 OPTION 3 : Associate MC and ECP to the challenge of local data

One major difficulty that we had anticipated in our answer to the ESPON DB tender was the problem of collection of local data at LAU1 and LAU2 levels. Despite our efforts in the current ESPON DB project, we knew that no complete collections could be achieved in the short term of 3 years, and that further work should be done by our successors. We have also noticed that the results produced by the projects from priority 2 (ex. Metroborder, EuroIslands) are difficult to help (provide data) and to valorize (store data) because we do not have a complete and coherent coverage of data at local level. In order to cope with these difficulties, we have several alternate proposals:

***Launch a specific priority 3 project on data collection at local level**:* this proposal is consistent with the option 1, if it is decided to split the next phase of the ESPON DB in different small and medium size projects. Our experiences indicate that collecting data at LAU level is a full task that implies a network of specialized partners all over the ESPON area. It cannot be done as sub-part of a general project on databases. It probably implies a specific database design that could store the results of Priority 2 projects more easily than the current ESPON database, which is more adapted to medium geographical scales (NUTS2, NUTS3, Cities, …).

Use the Monitoring Committee and ESPON Contact Point Networks as support for national collection of local data: many questions have to be addressed for data collection at local level that are generally simple for a national specialist but complicated for a foreign researcher. What are the criteria for delimitation of cities? What is the translation of this variable name? Are there any historical databases at local level?, etc. In our first interim report, we had suggested the use of the ECP network to help our project in this task of data collection at local level. But it was answered by the ESPON CU that it was rather a function of the MC members. After experimentation, it appears that MC members can be useful for institutional contacts (but they have limited time to help !) but that focal points are certainly more efficient for empirical questions on data existence or data definition. Therefore, we suggest again that it could be of great interest to launch priority 4 projects on this question of local data, where all interested ECP could participate. These priority 4 projects would be coupled with the above mentioned priority 3 projects on local data.

4.2 Specific recommendations

In this section, we provide more specific expert advice transmitted by project partners of ESPON DB on tasks that could not be completed in this phase of the project but which could be of great interest for our in the next phase of the ESPON DB Project.

4.2.1 Toward an automation of time series reconstruction

Based on the interest of time series data, the innovative approach proposed in the ESPON DB project needs to be improved and developed along two main ways:

The automation of time series data harmonization: this means the integration of the dictionary of changes in a data model. A conceptual approach is required to accomplish this task. Besides the dictionary of changes, the model also has to integrate estimation methods able to complete missing values. The main objective of this automation is to enable the user, via the query interface, to retrieve datasets with complete temporal series by choosing/selecting a version of nuts and one or more dates. This tool will allow the user to have a prospective or a retrospective vision of a given phenomena.

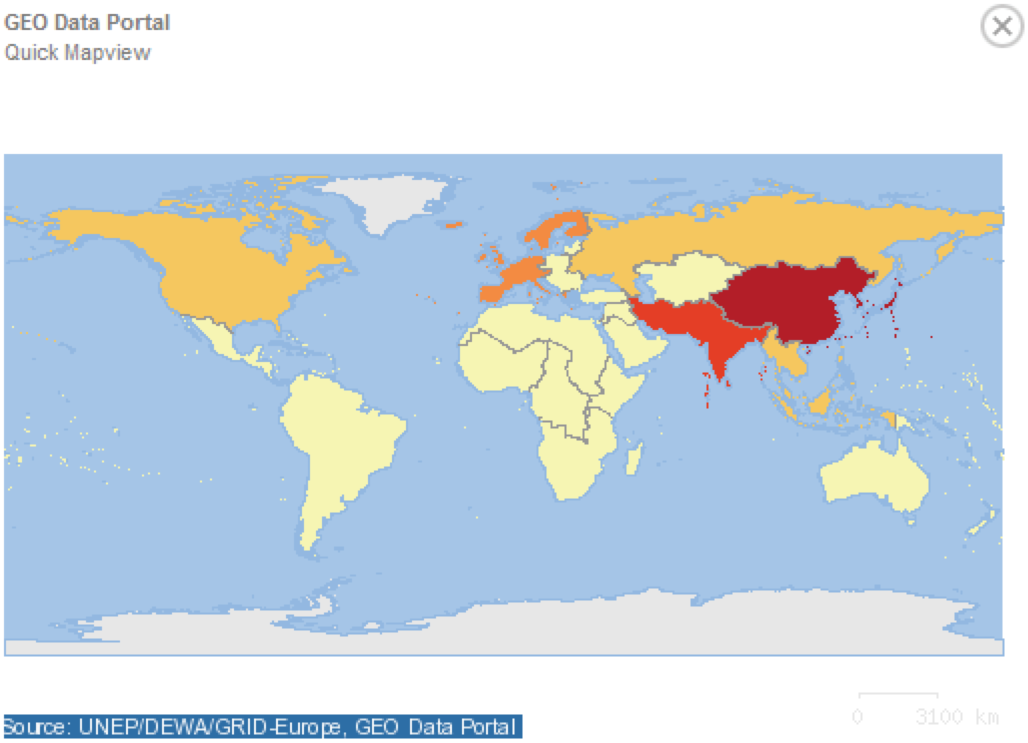
The issue of time series data could be developed as a transversal question in the ESPON DB project 2011-2013. The methodology implemented on the European NUTS can be extended to other spatial scales and other geographic objects. The results on nuts 1, 2 and 3 levels should be improved by the integration of LAU 1 and LAU 2 levels. World temporal databases might be also very useful. Concerning the extension to other types of geographic objects, we can propose for example cities and land use. This enlargement will require the adaptation of the data model developed on a hierarchical spatial structure.

4.2.2 Integrating European and World databases

Fundamentally, the perspectives development of world and regional databases articulation could be oriented in two main directions: the integration of a visualization tools of data (thematic maps) and the expending of ESPON DB thesaurus for world databases.

Based on Geo Data Portal multimedia experience in visualization applications, it should be useful to launch a work on visualization tool in the ESPON interface. It implies to develop closed links between computer scientists and geographers. The Figure 14, extracted from Geo Data Portal web site application, represents global distribution of population aged from 15 to 64 years. The visualization of features gives advance information before access to the data set. Some other institutions like, OCEDE and EUROSTAT, have also developed visualization tools. Graphic design of this application must be improved by adding graphics to the map and legend.

Secondly, thematic structure of World Databases is not exactly the same than the ones which are progressing in the ESPON Database (adapted to regional data issues). To facilitate comparisons of indicators, datasets and terms of reference, an expending of ESPON DB thesaurus is required.



Source: UNEP/DEWA/GRID-Europe, GEO Data Portal

Figure 14 : Map viewer from Geo Data Portal

4.2.3 Local data as a key challenge for territorial cohesion.

One simple query on the U.S. Census database will provide the user with a large amount of local oriented information. Ethical dilemmas such as knowing the revenue of the only doctor in an anonymous county in the Midwest are solved by replacing the basic indicators with others better chosen (the median of the revenues, standard deviation of the yearly income or with some statistical tricks masking the idiographic temptation). One also could test the gravitational model on inter-counties (and not inter-states) flows of population, based on datasets extracted from the last Census (which is quite a Holy Grail of the quantitative geography). The utility of this feature, both from an academic and financial point of view, could be considered as a lucrative externality of the geographic component of the R&D system. It is certain that this access to basic strategic information should have an impact on the general economic competitiveness. When strategies for increasing ESPON countries global competitiveness are drawn, it is of utmost importance to take into consideration the local dimension as well, in order to not to lose the cohesive one.

When the local dimension means dealing with more than 100 000 LAU 2 units in the ESPON space, the TIGRIS team has just built the database bricks and the containers for this specific information. It should continue with the wall. These bricks, meaning already available different deliverables, were somehow projected by us, shaped by RIATE and tested for quality by the NUI (National University of Ireland). The Challenge 4, managed by TIGRIS, includes now a strong networking work component and focus on future tasks such as:

- an updated (and updatable) database of local indicators translated into a lingua franca, allowing different LAU 2 scale studies.
- a few amount of indicators is better than nothing, even if chronologically and semantically unharmonized. That's why continuing with a country by country sedimentation of indicators at LAU 2 scale should be considered as an option until the 2011 EU Census uniform data publication (maybe in 2013).
- thirdly, information about NUTS 3/2 geometry modifications are already explored. Transposing an appropriate frame of analysis at lower scales (LAU) could be quite an academic challenge and a policy-drive issue, when the mention process of modifications decelerates the economic convergence process in fragile regions of the ESPON space.
- in the pyramid of the ESPON geographic information, the LAU base could prove essential for validation and verification tasks by summarizing for wider reference spaces, such as the NUTS.
- the economic territorial convergence in the ESPON space is largely depending on the local base (insufficiently known). It also depends on what we call now the proximity measure. Having access to this proximity by a local to global approach could also be a policy-driven philosophy. Testing this way and the necessary steps might be an essential task in a new Challenge 4.

As one could read and imagine, the tasks proposed above are intimately linked with (C) challenges already explored within the actual project and the purpose is to capitalize on the inner networking effort already made. It's not just a simple persuasion exercise; we rather aim to show that the diffusion of acquired knowledge shall serve to build new better bricks.

Considering the progress of the work in Challenge 4, the TIGRIS team is able to deliver a set of LAU2 databases derived by means of sources of information and user-priority. Eventually, our deliverables are the results of the compilation of wider data sources, such as GISCO, MUNIS, SIRE, EUROSTAT and the NSI. Chronologically marching, the integration of the SIRE database involved the fragmentation of the proper database in new fields, organized by spatial and semantic attributes. This new database is theoretically oriented towards users concerned with changes between the early period of data collection in the EU and some mid-term chronological marks (2001 or 2006). It is relational database, generally involving 5 variables or indicators by theme. These themes are organized in specific thematic: population in 1981 or 1991, active population, dwellings or demographic indicators.

In the spirit of the former ESPON DATABASE project, we have considered that it is useful to integrate a cartographic platform at local spatial scales. Taking into account the fact that the next European Census will be realized in 2011, a backward chronological mark such as 2001 seems quite appropriate for further comparison. A GISCO based geometrical database, country by country was consequently built. Some of the basemaps and files are under reserve (e.g. for France) because of the large amount of computer memory needed for processing. For almost all the countries, the themes are composed of the LAU2 polygons and centroids, together with the network datasets for rail and road infrastructure, according to the not-upgraded spatial scale of 1 : 250 000. Some countries, such as Cyprus or Iceland miss the rail network dataset. Other countries (Bulgaria) present only the LAU 2 geometry. Integrating the network datasets eventually allow us to build local indicators based on "real" distances – the interaction potential of population or the accessibility of the LAU 2 units in national or trans-national spatial contexts.

Derived from the prior deliverable, a LAU 2 database concerned only with factual information about the LAU spatial units is also available. Linked with the map-kit at micro-scale for each country in the ESPON space, this database should be considered as a frame for integrating the results of the Priority 2 Projects. At the same time, this database could be one of the necessary steps taken in order to enlarge geographic knowledge at local scales of analysis. Already containing trans-scalar information, such as NUTS belonging, different labeling of LAU 2 units or basic variables (population and density), this derived database can be transformed in a territorial investigation tool.

For superficial LAU analysis, using our next TIGRIS deliverable should be quite convenient. Based on EUROSTAT list of local administrative units, we can provide a database filled with lists of administrative situations for each (known) local spatial unit in ESPON space. Not linked with geographic geometries, but being somehow *official,* this database is likely to provide basic information about LAU scale in Bulgaria, compared with Portugal (e.g.).

Deeping into the realities of LAU 2 building databases, the final deliverable focused on the collection of indicators for some ESPON countries – Czech Republic, Slovakia, Romania, Bulgaria, Norway and others. Specific themes are available as an example of what Challenge 4 should provide. These data could be integrated in the ESPON Database, provided that metadata quality is judged good enough and provided that scaling issues (due to large datasets) are overcome.

Partly superposing on the objectives described in FIR and partly derived from the evolution of the work on this Challenge (local and/or regional data), the mentioned list of deliverables is much more a compromise between *possible* and *not probable* issues, rather than a linear approach on the tasks. Maybe the *uncertainty* of the result and sometimes the *undefined* below the method is the touch of reality needed in the local spatial approach.

4.2.4 Developing the use of grid data in ESPON research for a better integration of social and environmental dimensions

Geostatistical Analyst

A Geostatistical Analyst will be made in order to make an in-depth comparison with other integration strategies:

- The ESPON 2006 integration strategy based on transferring all the information that it is not delivered on the basis of administrative units (NUTS 2 or NUTS 3) toward administrative units.

- Disaggregation strategies proposed by new projects developed in the European context. In this scenario, a direct contact with the European Grid Club project will be necessary to define an appropriate channel of exchanging information.

Improvement of the methodologies proposed

- Following the results from point the Geostatistical Analyst.

- Taking into account the comments, necessities and problems found by the other Espon projects.

Build new ESPON OLAP cubes

New Espon OLAP Cubes will be built depending on the needs of the others ESPON projects.

The integration of new variables inside the cube will be analyzed. Socio-economic and environmental variables will be integrated, like, for example, ESPON typologies, the height of each cell, or the average temperature.

Integration with the Espon 2013 Database

Close collaboration with all project partners in order to build an appropriate framework to make available the ESPON OLAP cubes.

Smoothing

Close collaboration with other project partners in order to generate a complete grid database, including smoothing calculations with the amount in the neighborhood.

ESPON Projects help desk

Advise and support others ESPON projects at combining ESPON socio-economic data with environmental data due to our experience at the UAB, as European Topic Centre on Land Use and Spatial Information, supporting the EEA in monitoring the land use/land cover change in Europe and analyzing the environmental consequences.

4.2.5 Toward an integrated ESPON urban database

Starting from our different experiences and researches on urban databases in the ESPON DB, we are more than ever convinced than any harmonized database on European cities has to be seen as a snapshot model, a simple representation of an urban system at a given time moment. These different urban databases may enrich each other, most of all are complementary. For example, some databases are, from an "expertise" point of view, very coherent (the LUZ, which are based on national delineations, i.e. national analyses) but some others are more coherent from a "statistical" point of view (the UMZ, which are created with the same criteria for all the countries, so that they present a strong statistic homogeneity). During the two last years, we have confronted these different approaches, using bottom-up models (LUZ, starting from a good knowledge of national settlement contexts and a strong semantic basement) but also top-down models (MUA and UMZ). For the same objet "city", the attempts can be completely different if one takes the point of view of an urban planner, a scientist researcher, a manager or a simple resident, and this diversity of approaches is much more important when one takes into account the historical and cultural backgrounds. The semantic contents are very rich and diverse in the bottom-up approach, whereas they are much poorer in the top-down approach (and need to be enriched).

Accordingly, we suggest for the next ESPON Data base a few striking stakes. First, after this first period consisting in the in-depth analysis of the differences between databases from semantic (metadata) and quantitative (surfaces, populations) points of view, it could be relevant to go deeper to "qualify" the UMZ as residential entities containing not only populations but also activities, amenities, services…, and as cores and poles of larger urban areas.

Thus two challenging focuses could be done on:

1) Describing the UMZ with new indicators in order to confront them to the observations done in some bottom-up logics (for example LUZ). That would lead to work on methodologies for transferring socio-economical data from LAU2 level to UMZ objects, using not only the density population grid but also other elements (for instance the work done by Challenge 5).

2) Delineating larger functional areas polarized by the UMZ, if the data on daily commuting are available all over the European countries at the same geographical level (LAU2), for example in the last version of SIRE databases. This work could be confronted with the LUZ delineations that correspond to functional national definitions. At least for the largest cities (more than 100 000 or 200 000 inhabitants), these LUZ delineations could be used for calibrating some homogenous functional definition at European scale, centred on UMZ cores.

4.2.6 Making querying data simpler for various types of end users

The ESPON 2013 DB project is, so to speak, the backbone of the ESPON 2013 Program. Its importance lies on the potential to deliver easily accessible data to the wider audience on cross-cutting topics. This interdisciplinary approach to applied research is driven by policy demand and therefore it represents a major opportunity to underline the importance of defining appropriate themes, sub-themes and potential keywords that could improve search engine rankings.

With this regard, it is likely that end users do not want to spend a significant amount of time on ESPON DB search engine looking for relevant information. Besides, it is common accepted that potential users in most cases only view the first results. As explained in previous occasions, an interesting feature offered by text mining techniques is the possibility to perform cluster analysis of words and expressions. The output of such experiments could be easily integrated in the database as an additional feature to improve search queries and therefore offer end users different possibilities to refine their search processes.

This opens the new areas of research that could bring more value to the ESPON 2013 DB. First, because the enlargement of data will likely represent more textual corpora for analysis (e.g. journal articles, ESPON scientific reports, EC reports). And second, because future deliveries will improve the definition of themes and sub-themes. Such contributions, however, might disclose new associations of words that often are subject to change. Given this, search engines should encourage end users to validate data allocation.

Indeed, ESPON data will expand considerably and for that purpose is important to organise concepts that could be used as keywords to refine search queries. This constitutes a challenging task that is often subject to criticism due to the difficulty in grasping the exact relationships between terms. To this end, thesauri-based search facilities have the potential to assist in organizing terminology. Once the attempt to simply some of the terms is achieved, this would give the legitimacy to propose accurate, correct, and appropriate definitions through glossaries that express empirical, theoretical vocabulary developed by ESPON. Such dynamics constitute as well as opportunity to improve metadata structure and content. The knowledge embedded in the metadata would suggest more technical details describing data and therefore improving the quality of search queries.

Building a thesaurus requires expert knowledge. However, recent studies have shown similar approaches that facilitate the work of compiling and assembling corpora by means of tools that record, define, and establish relationships between terms[8]. Once the ESPON 2013 DB has defined its metadata editor, through a web-based interface, other applied research projects will be able to provide more information.

The idea of having a thesaurus incorporated into the ESPON 2013 DB with controlled vocabulary deriving from text mining techniques and used to indicate semantic relationships between terms would provide something better than a

[8] Tindall, I.; Moore, V.; Bosley, J. et al. (2006). Creating and using the urgent metadata catalogue and thesaurus. *Science of the Total Environment*, 360(1), 223-232.

simple keyword search. Moreover, it would allow end users to enter familiar words that could lead them to preferred terms associated with themes, sub-themes and eventually datasets. At later stage, this could also represent an interesting topic for analysis by focusing on the patterns and behaviors adopted by end users to retrieve data.

4.2.7 Seamless integration of data of different types

The most important improvements for the next phase of the ESPON 2013 Database from the data management point of view are: i) the complete integration of raster data (what is usually called environmental data) with the vector data in the ESPON database, ii) the complete integration of local data in the ESPON database, and ii) the development of automated methods for data harmonization, checking and estimation.

Seamless integration of environmental data within the ESPON database

Storing environmental data into the ESPON Database Application is possible and environmental data files will be available for download from the ESPON DB, at the end of the first phase. However, the complete integration of environmental (or raster) data is a difficult task to solve, which raises three types of issues:

Difficult conceptual issues need to be solved in order to harmonize environmental data with statistical NUTS data, either by storing these types of data in the same common data structure, or by keeping the two types of data separately and developing dynamic conversion methods from one structure to the other.

Non-trivial scaling issues also have to be answered. Realistic and thorough performance testing is required in order to evaluate properly what are the computing power requirements of such conversions, but we can assume that it is virtually impossible to make such transformations on complete datasets in real-time.

Seamless integration of local data within the ESPON database

In the same way as with environmental data, it will be possible to store and download local data files from the ESPON DB Application, at the end of the first phase. However, information stored within these files will not be connected to the rest of the database, so it will not be possible to harmonize this data with the NUTS data. In order to perform the complete integration of local data several issues are to be overcome:

Scaling issues, due to the large number of potential units, are to be overcome. In order to be able to give fast response to queries involving the roughly 150 000 local units database and/or hardware optimizations are to be taken into consideration.

Spatial ontologies are needed in order to harmonize the local data with the NUTS data. This would allow for vertical (top-down and bottom-up) harmonization and estimation methods to be applied to indicator values. Ideally, these ontologies should also describe the temporal evolution of the local units, at least during the period of time covered by the different NUTS versions. From a practical point of view this requires a tremendous amount of work, which cannot be automated.

However, some automated change detection methods could be implemented in order to assist the thematic experts in the elaboration of this ontology.

Query interface issues, due to the large number of available units, would need to be addressed. A special interface design would need to be coupled with the spatial ontology (which should for instance define rural or coastal areas, etc.) in order to allow the user to describe and find easily the desired subset of units. This is an important issue not only from a user (application ergonomics) point of view, but also from a server performance point of view (users downloading large datasets would be very demanding on the database server's side).

Development of automated data harmonization, estimation and checking methods

Extending the ESPON Database data management software with the integration of automated harmonization methods is an interesting development direction. This would allow capitalizing the thematic experts' knowledge and dynamically linking different types of information (e.g. world data with NUTS and local data) or various temporal versions of information (e.g. NUTS 2000 data with NUTS 2006 data). These methods should be used in order to automatically complete (estimate) missing indicator values. However, an important prerequisite for the elaboration of effective harmonization and estimation methods is the availability of harmonized local data and grid data, because estimation can be accurate only if it is combined with high resolution spatial and thematic data.

Development of geographic Web services for publication of the ESPON Database content

Another interesting direction for improving the visibility and accessibility of the ESPON Database is providing access to the database not only through a Web query interface, but also via standard OGC compliant Web services, like CS (web cataloging services), WFS/WMS (web feature and/or mapping services) would allow the content of the ESPON database to be directly integrated into third party applications, including spatial data portals.

4.2.8 Automation of quality control and outlier detection

The large and diverse field of robust statistical methods and outlier detection can not be comprehensively reviewed within the terms of reference of the second interim or for that matter, the final report of the first phase of the ESPON DB project. It is clear therefore that further research in this area during the second phase of the ESPON Database project would be appropriate to improve and refine the detection methodologies described in this report.

Here it is also envisaged that our relatively advanced geographically weighted robust techniques should not be fully presented in the final report of the first phase of the ESPON project, but instead left for the second phase, when the development of these robust spatial methods has properly matured. Work in this second phase should also include the packaging of the R code for these robust spatial methods, so that techniques are fully portable, transferable and openly documented.

Among the improvements we wish to see are

the development of outlier handling strategies – whilst we can detect outliers we have to decide how they are handled – removed, edited, imputed.

What methods are used for altering exceptional data – could rollback techniques be implemented to allow a researcher to return to the 'unedited' data, albeit with the proviso of caveat emptor?

The development and implementation of the methods we have considered in the reports, and the appropriate context for their use – at what point in the database creation are they employed?

The development of appropriate metadata for describing the methods that have been used to determine whether a particular value is exceptional and what operations have been carried out on them.

The evaluation of Bayesian methods for dealing with missing or incorrect data – for example, Monte Carlo Markov Chain (MCMC) methods can be used which treat missing data as parameters to be estimated. With such methods we can relax the assumptions of normality for much of the data. It would be helpful to evaluate in terms of accuracy and cost (computational feasibility) whether MCMC represents a significant improvement over ad hoc methods or other techniques such as use of the expection-maximisation (EM) algorithm.

4.2.9 Enlarging the data collection for the European neighbourhood

The challenge dealing with data collection in the European neighborhood should focus its work in two main directions: (1) Develop the data collection in the other geographical dimensions and levels (LAU1 and 2, cities, flows). (2) Enlarge the data collection in the Eastern and Southern Neighborhood.

Exploration of other geographical levels

The data collection in other geographical dimensions is very useful both for studying territorial issues and visualizing the respective results in the following very important spatial analysis and planning issues.

In particular, defining a strategy for the enlargement of the work on W. Balkans and Turkey to cover LAU1, LAU2 or "similar to LAU1" and "similar to LAU2" data should be very useful The most appropriate strategy seems to focus in a first time on the assessment of the availability and quality of data. As this is very difficult, not for technical reasons but mostly concerning data comparability issues, it would be eventually preferable that the terms of reference discern the LAU1 level (formal obligation for the project) from the LAU2 level (exploratory work).

Another interesting issue concerns the analysis of flows and networks in the European Neighborhood. Within a context of political integration of Western Balkans and Turkey in the European Union, demographic and economic flows between these countries and European memberships should grow up in a near future. It is really important to develop the material to analyze such dynamics. The work should focus in different ways: Analysis of transport's networks (rail, road, air); energy; provision of services; economic, social and political networks; migration flows. This short listing is not exhaustive. Anyway it will be important to ensure the compatibility of this data with European Statistics provided by Eurostat and the different DG's.

In term of methodology, provide such information in GIS format is also very important to map this information. However it is not an obvious task (provide a "SLAU2" shape) which will be probably time-consuming.

Data collection in Southern and Mediterranean neighborhood

Defining strategy for the enlargement of the scope of the ESPON Database to cover regional data for the Eastern Neighboring countries (ENC) and the Southern Mediterranean Neighboring countries (MNC) should also be considered as a key-element for the next ESPON DB project. In order to ensure the continuity between the ESPON Program, the geographical target of such work could be the Euromed Area (38 States in the Southern and Eastern neighborhood + Western Balkans)

Succeed in this challenge implies to take care of different elements:

- Background: Review of existing databases at that level of analysis
- Identify data sources at regional level (National Statistical Institutes, International organizations…)
- Identify levels "Similar to NUTS 1-2-3" for each country
- Overview of common indicators available in each neighboring country
- Data harmonization with international data provider (Geo Data Portal, United Nations databases…)
- Creation of map layers to show the information

It is quite clear that this improvement will benefit of conclusions provided by the existing challenges 3 and 11 (World database – harmonization of values between heterogeneous databases – and data collection in Western Balkans and Turkey)

Moreover, this enlargement of the data collection should go faster if regular contacts would be elaborated between ESPON and the National Statistical Institutes of Neighboring Countries. These contacts missed to ESPON Database Project (phase 1) in the Western Balkans and Turkey. It should be interesting to develop a proactive demarche in that sense. It is also quite clear that the contacts between ESPON and international organizations (OCDE, United Nations) should continue and be improved to create a real data flow on the topic of region data in the European Neighborhood.

4.2.10 Integrating synthetic samples of individual data for in depth analysis

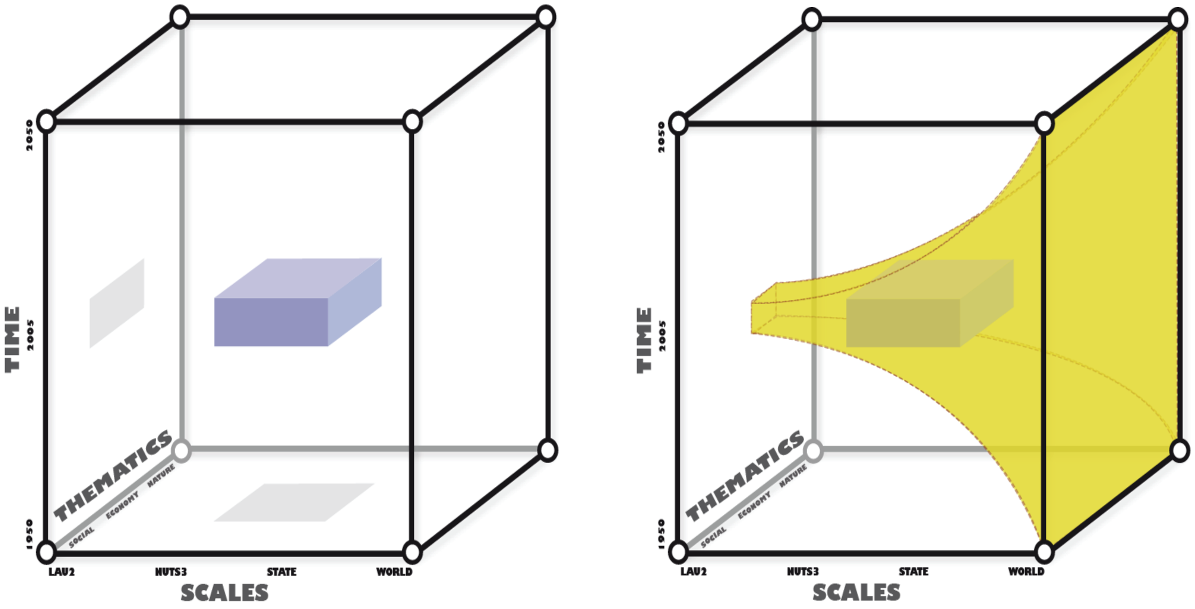
One vision for further analysis of the spatial development of Europe would be to create a localized artificial micro dataset mirroring several properties of each individual in the European population. On the wish list are attributes like settlement/km^2 square, sex, age, civil status, education level, labor market participation, profession, income, etc. Such a continuously updated database would preferably represent 100 percent of the population, but a coverage of 10 percent or even 1 percent would also be useful. The idea would be that the database (in effect, constituting a synthetic sample) captures and uses all (table) information from censuses, register data and surveys available to Eurostat, also drawing on additional available spatial information. Examples of the latter kind of data could be the partly CORINE-based JRC population grid, but perhaps also new interpretations based on, for instance, services like Google Maps. Such a database would, in a sense, replace all the different fragmentary sources it is based on, since it contains all information in the combined sources. A resource of this kind would enable ready access to data for a lot of new in-depth analyses of the European spatial situation and development.

Such an endeavor will meet two challenges, one practical and one methodological. The practical quest is to gain continuous access to all relevant sources and prepare them for the synthesis. The methodological quest involves developing a robust technique for using and merging a lot of disparate sources, so as to successively improve an initially more or less random distribution of the population over the selected dimensions—including space—in a reasonably consistent manner. In this context, a balance has to be struck between on the one hand preserving the precise information in each table, and on the other hand maintaining the overall consistency of the created population. The project can draw on several experiences of developing such techniques for certain tasks, including a number of efforts carried out in order to enable spatial micro simulation in Sweden, France and the UK.

5 Conclusion : Toward Final Report

With this Second Interim Report, we arrived at the final step of the project where the question is no more to explore new directions but to finalize and secure the results obtained in order to transmit them to our followers (as it is generally admitted that a new ESPON DB II project will be launched for the period 2011-2013). Now, it is interesting to come back to our initial ambition of enlarging geographical scales, time period and themes (Figure 15) in order to evaluate what has been really achieved and what still remains to do.

Figure 15 : Enlarging geographical scales, time period and themes



Initial situation of ESPON 2006

Objectives for ESPON 2013

New directions for ESPON database have clearly been explored and the collection of technical reports will simplify the task of our successors in the ESPON program. The stabilization of the metadata model appears of particular importance as it will simplify the data management in ESPON.

The increasing number of ESPON projects will be very challenging for ESPON database in the years 2010 and 2011, at the precise moment when our project ESPON DB I will be achieved and replaced by a new project ESPON DB II. On one hand, this increasing number of projects will be very helpful to test the operational dimension of the new directions introduced by ESPON DB I (e.g. test of World dimension with new project on Globalization). On the other hand, this great number of projects will complicate data management, in particular the question of data supply (for projects at the beginning period) and data check (for projects at the final period).

The improvement of the interface of data query/ data upload will be a crucial step in 2010-2011, simply because the increasing number of ESPON projects will oblige the future responsible of ESPON DB II to use automatic procedures more often than we did in ESPON DB I. And also because new geographical objects will certainly be more requested, beside of NUTS2 and NUTS3 units that are the easiest to manage at the moment.

The ESPON DB I will therefore try to consolidate the results obtained so far, but no doubt that much work remains to be done by our successors in 2011, if we want to reach the objective of a fully integrated and multipurpose ESPON database (Figure 16).

Figure 16 : A vision of future ESPON DB