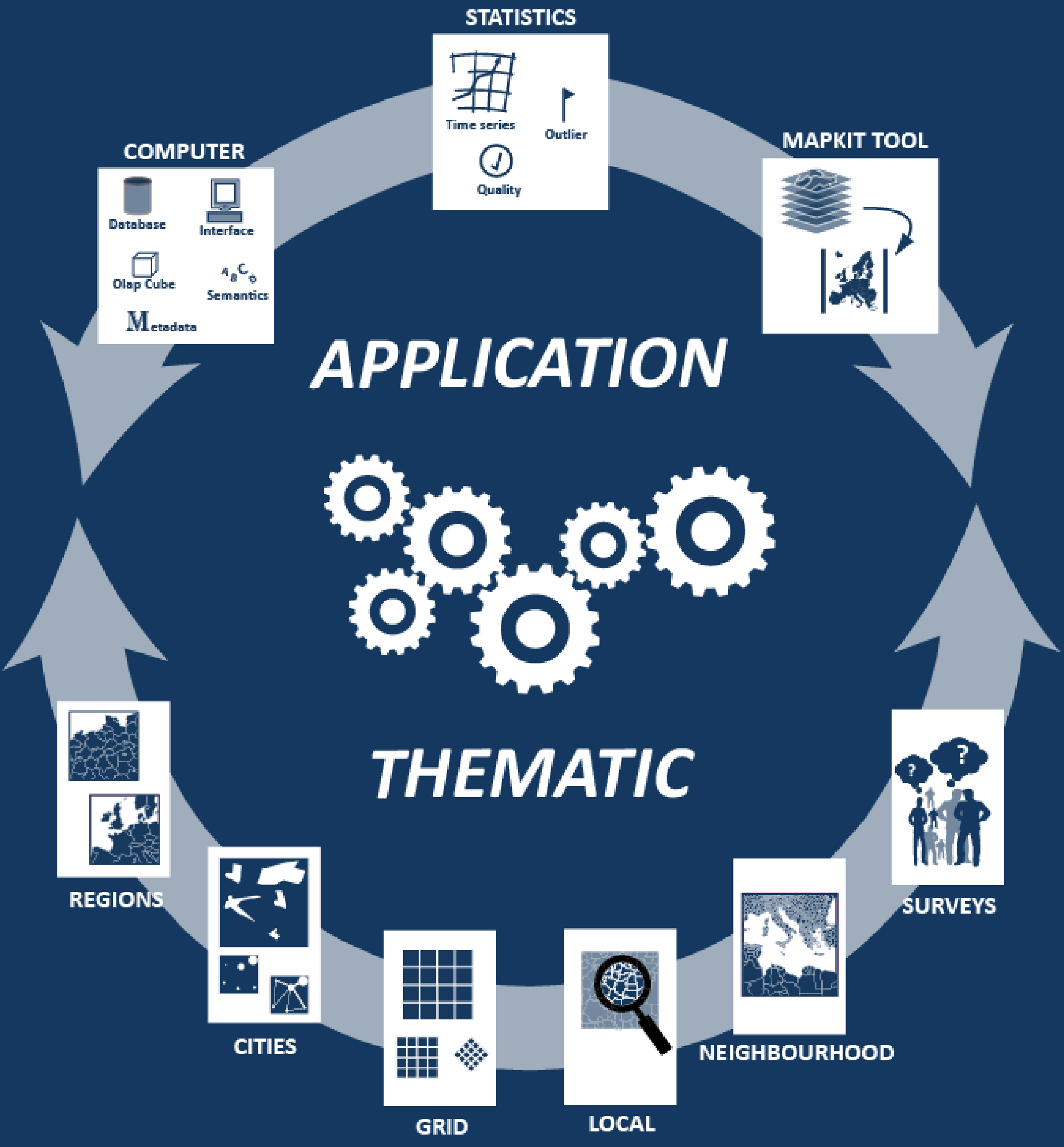
ESPON 2013 DATABASE

QUALITY RATHER THAN QUANTITY...

FINAL REPORT – DECEMBER 2010



This final report represents the 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: 2010

List of authors

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

LIG (FR)
Jérôme Gensel*
Bogdan Moisuc
Marlène Villanova-Oliver
Anton Telechev
Christine Plumejaud

UAB (ES)
Roger Milego
Maria-José Ramos

IGEAT (BE)
Moritz Lennert
Didier Peeters

TIGRIS (RO)
Octavian Groza
Alexandru Rusu

UMR Géographie-cités (FR)
Anne Bretagnolle
Hélène Mathian
Marianne Guerois
Liliane Lizzi
Guilhain Averlant
François Delisle
Timothée Giraud

Université du Luxembourg (LU)
Geoffrey Caruso
Nuno Madeira

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

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

** Expert

TABLE OF CONTENT

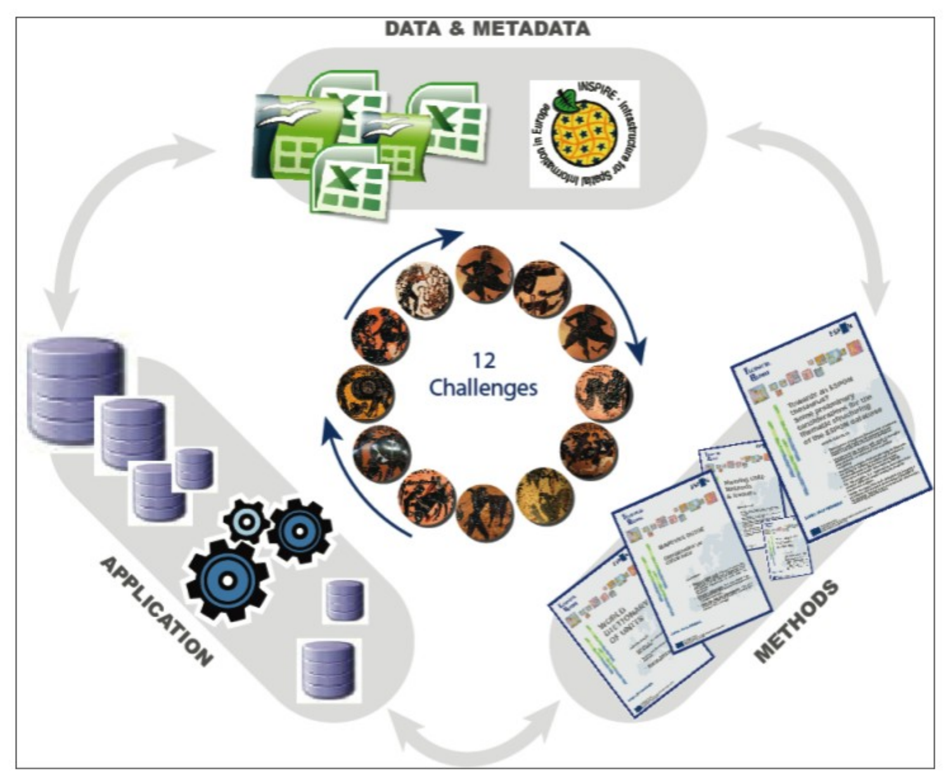
FOREWORDS

The document we deliver here is called the FINAL REPORT.
He that outlives this FINAL REPORT, and comes safe home,
Will stand a tip-toe when the PROJECT is named,
And rouse him at the name of ESPON 2013 DATABASE.
He that shall live this FINAL REPORT, and see old age,
Will yearly on the vigil feast his neighbours,
And say "I WAS IN ESPON 2013 DATABASE PROJECT"
Then will he strip his sleeve and show his scars.
And say "These wounds I had on ESPON DATABASE."
Old men forget: yet all shall be forgot,
But he'll remember with advantages
What feats he did in ESPON 2013 DATABASE: then shall our names.
Familiar in his mouth as household words
RIATE, LIG-STEAMER, UNIVERSITIES OF BARCELONA AND LUXEMBOURG
GEOGRAPHIE-CITES, TIGRIS, NTUA, NCG, UMEA, UNEP, IGEAT
Be in their flowing cups freshly remember'd.
This REPORT shall the ESPON CU teach his NEW PROJECTS;
And ESPON DATABASE 2013 PROJECT shall ne'er go by,
From this day to the ending of the world,
But we in it shall be remember'd;
We few, we happy few, we band of brothers;
For he to-day that sheds his blood with me
Shall be my brother; be he ne'er so vile,
This day shall gentle his condition:
And researchers in European Union now a-bed
Shall think themselves accursed they were not here,
And hold their manhoods cheap whiles any speaks
That fought with us upon ESPON DATABASE FINAL REPORT.

With Special thanks to William S. for inspiration.
Original version available at http://pagesperso-orange.fr/rhetorique.com/azincourt.htm

INTRODUCTION

A division of work in 12 challenges has been the core of the project since the beginning. These challenges provided a simple and efficient division of work between partners and experts, each of them being responsible for one challenge, possibly in association with others. But challenges had also to be integrated in a more synthetic way in the second part of the project, which can be illustrated on the figure below by the three work areas defined as Methods, Application, Data and Metadata.



1. 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 themselves** 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.

2. Methods, presented in the form of standalone booklets called Technical Reports, are the necessary complement of data and metadata. They 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 chunk of 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 elaborated by RIATE that has been made available on the ESPON website.

3. 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 is now available on the ESPON website 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 contributed 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.

The Final report of the ESPON 2013 Database project is therefore not limited to the present document but involves all the above mentioned material (technical reports, applications, data). What we try to present here is a short guide for accessing to this whole set of resources. We have divided this report in two parts:

- **Part 1 Application** presents the software oriented elements produced by the project and also some conceptual elements that drive the software implementation.
- **Part 2 Thematic** presents the different technical reports elaborated in order to improve the scope of the ESPON database in terms of space, time, scale, geographical objects, fields of policy action.

1.APPLICATION

1.1 - INTRODUCTION
1.2 - THE UPLOAD PHASE
1.3 – THE CHECKING PHASE
1.4 – THE STORING PHASE
1.5 – THE DOWNLOAD PHASE
1.6 – CODING SCHEME
1.7 – THEMATIC STRUCTURE
1.8 - OLAP CUBE
1.9 – CARTOGRAPHY IN ESPON

INTRODUCTION – PART 1

The first part of this report presents the software oriented elements produced within the ESPON 2013 Database Project. This concerns not only software elements (e.g. the different components of the ESPON DB Application) but also conceptual elements (e.g.architecture, schemas) that drive the software implementation.

The first section of this part gives a brief overview of the ESPON DB Application and dataflow. The follwoing sections describe, in their respective order, the different phase of the ESPON DB dataflow. Section 1.2 describes the upload phase (i.e. the ESPON DB metadata profile and editor). Section 1.3. follows the different stages of the data checking process. Section 1.4 offers some insights about the storage phase, what are the databases and ontologies that lay behind the ESPON Database Application. Section 1.5 shows the query and download phase, performed by the end users via the Web Download interface.

The next two sections shed more light on the coding scheme (1.6) and the thematic classification (1.7) which are of crucial importance for structuring the ESPON 2013 Database and making available the information for hand-users.

Then, the section 1.8 shows the methodology used for building the ESPON OLAP Cube which allows to combine information described on grid (Corine Land Cover) and socio-economic data in the NUTS nomenclature.

Finally the section 1.9 presents the different map-kits available for ESPON Projects, from local case studies to the World. On top of that, some basic rules of cartography are described in order to ensure harmonisation of maps in the ESPON Program.

1.1. THE ESPON DB APPLICATION AND DATAFLOW

The ESPON 2013 Database Application is a complex information system dedicated to the management of statistical data about the European territory, spanning over a long period of time. The overall architecture relies on two databases: one is used for storing ontology data, and the other, called the ESPON Database, is meant to be queryied by end-users. The latter only is made accessible to users through Web interfaces (see figure on the right, above) that each correspond to the four main functionalities offered by the ESPON 2013 Database Application: registration, administration, upload of both data and metada, query and retrieval of such data and metadata.

The ESPON DB Application data flow describes the path followed by both data and metadata from the moment they are entered in the ESPON DB Application, until they are output as answers to queries expressed by end-users. Four phases are identified along this data flow:

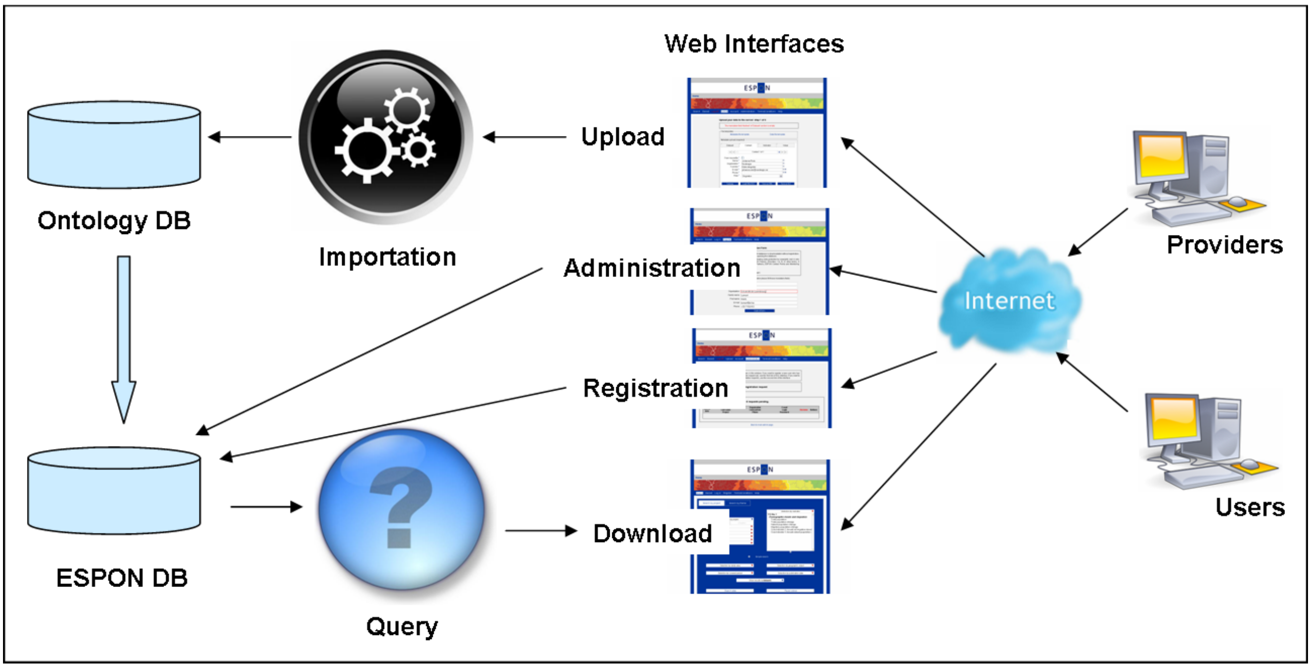
1. The upload phase is handled by the upload Web interface through which users (here, data providers) are guided in the preparation and the transfer of both their data and metadata files to the ESPON Database server. During this phase, users are helped in providing well formated and Inspire compliant metadata through the ESPON Metadata Editor. This phase is described in more detail in section 1.2.

2. The checking phase follows; it aims at validating both data and metadata files provided by users before they are stored in the ESPON Database. The checking process alternates between automatic and manual steps performed either by the application itself or by the expert members of the ESPON DB 2013 Project. If some of the errors detected cannot be corrected or need some additional information and precisions, then both data and metadata files are sent back to providers in order to be fixed. When the checking phase succeeds, then the validated data and metadata files are ready to be stored in the ESPON Database. This phase is described in more detail in section 1.3.

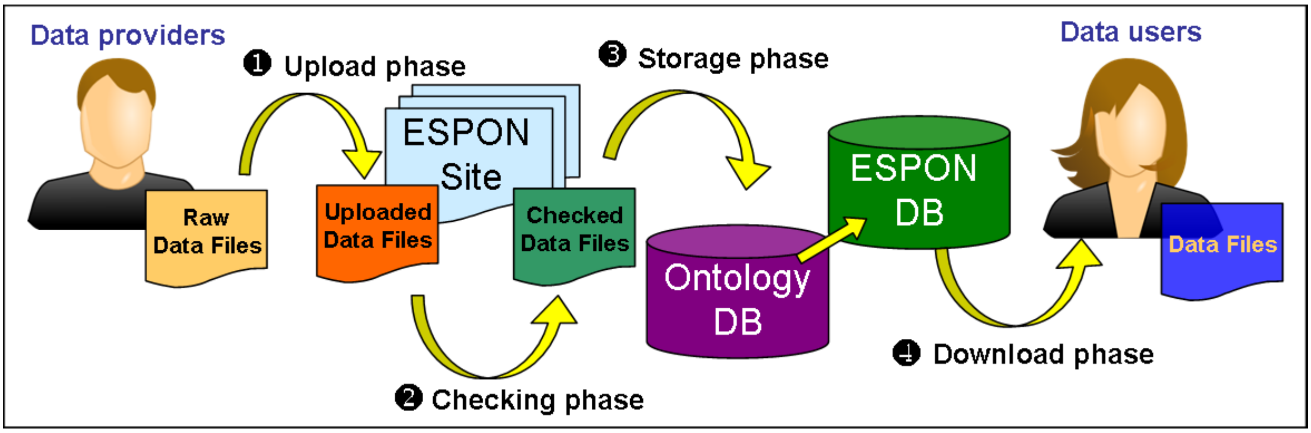
3. The storage phase deals with the management and the maintenance of both data and metadata in the ESPON Database. Flexible database schemas have been designed and built for handling long term storage of statistical and spatial data, considering that both data and metadata may evolve while stored in the ESPON Database, as a result of harmonization and gap filling processes. This phase is described in more detail in section 1.4.

4. During the download phase, end-users of the ESPON DB Application are invited to explore, search and retrieve both data and metadata through a Web interface. Free data and metadata can be accessed and downloaded by any end-user, while data and metadata subject to copyright restrictions are made available for authorized and registered users only. This phase is described in more detail in section 1.5.

The ESPON DB Application Architecture And Data Flow



The ESPON DB Application relies on a Web-based architecture, including two databases (ontology DB and ESPON DB) for long term storage of statistical and spatial data. Data providers and end-users interact with the EPSON DB (register, upload files, query data and download files) via Web based interfaces.



The ESPON DB Application data flow allows receiving data from ESPON Projects (acting like data providers) and returning these data to other ESPON Projects (acting as data consumers). The intermediate phases allow checking and improving data quality and are performed without no interaction with the users.

1.2. THE UPLOAD PHASE

Data and metadata files entered by data providers (mainly ESPON Projects) have to be compliant with the ESPON DB data and metadata formats so that they can be uploaded on the ESPON DB Application server.

The ESPON DB metadata profile has been created because an indepth analysis of the state of the art has revealed that, so far, there is no standard metadata profile aimed at describing statistical territorial data. Indeed, existing spatial data standards (ISO 19115, the INSPIRE directive) offer very detailed description profiles for spatial data, but thematic and statistical descriptions of data are insufficient. The ESPON DB metadata profile covers 3 main purposes:

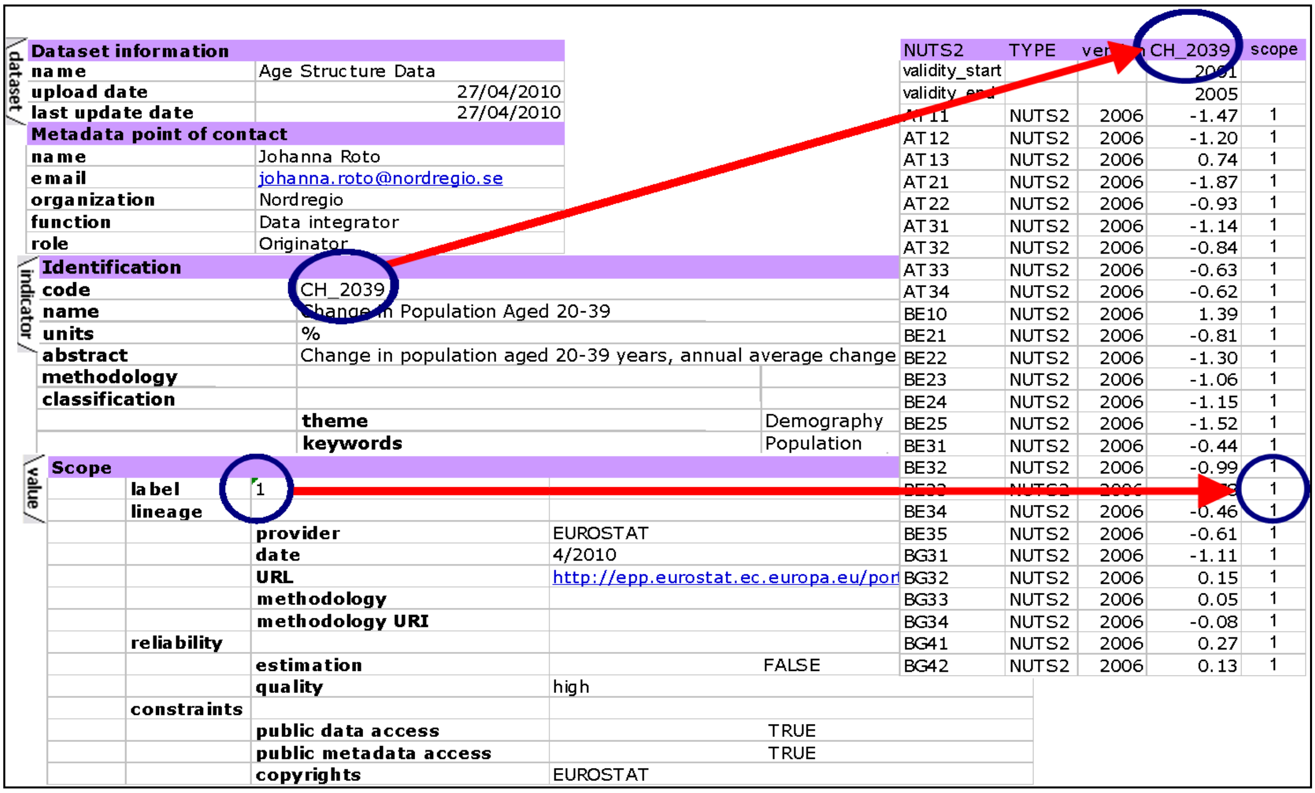
- preserving the compatibility with the existing standards (by INSPIRE, ISO) by integrating the same main elements in the profile.
- minimizing the quantity of work data providers have to do when filling metadata by, for instance, inferring automatically metadata from the associated data when possible (e.g. temporal or spatial coverage).
- providing sufficient information about the content of data (indicators) and about their origin, by including indicator level and value level descriptors in the profile.

The Web metadata editor is an interactive application, which assists data providers in the creation of data descriptions compliant with the ESPON DB metadata profile. The editor can be used to create a new metadata file, or to edit and modify an existing one. It handles, opens, and saves files in both XML and XLS formats. It guides a data provider in filling the three categories of descriptors covered by the metadata profile:

1. Information about the dataset as a whole: contact information, dataset title and abstract, etc.
2. Information about each indicator in the dataset: name, description, indicator methodology, thematic classification, etc.
3. Information about each value in the dataset: the primary source of each individual value, the estimation or correction methods applied to it, the copyright constraints associated with it, etc.

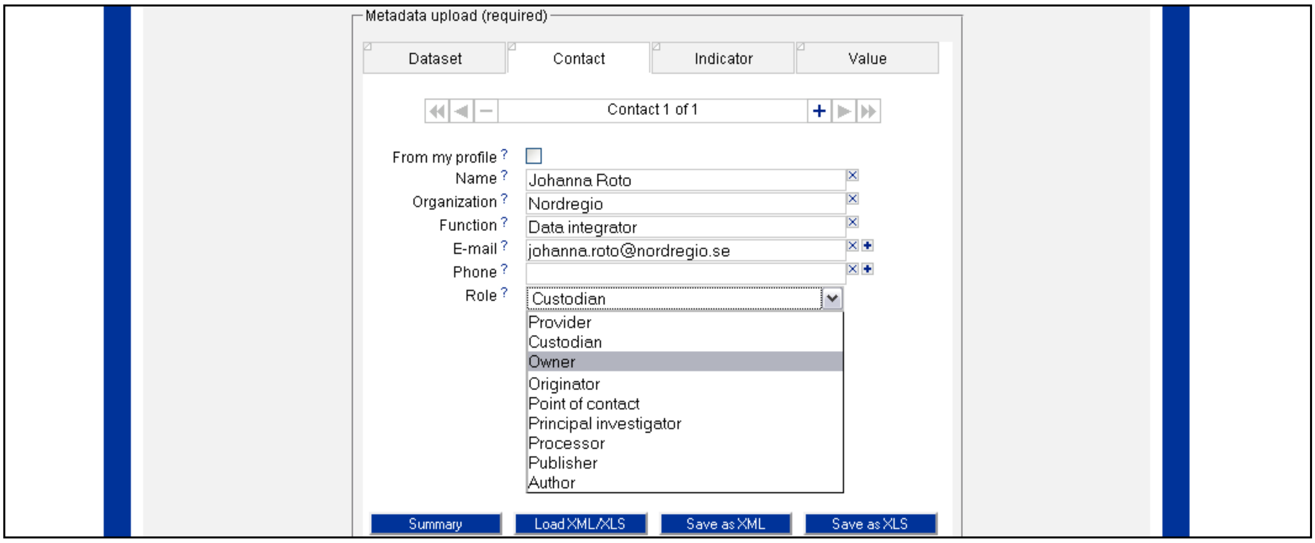
The editor checks and underlines syntactical errors found in metadata and provides dropdown lists that ease the time consuming but valuable task of filling data description (e.g. for personal information, already described indicators, etc.).

The Metadata Profile And Editor



Dataset information			
name	Age Structure Data		
upload date	27/04/2010		
last update date	27/04/2010		
Metadata point of contact			
name	Johanna Roto		
email	johanna.roto@nordregio.se		
organization	Nordregio		
function	Data integrator		
role	Originator		
Identification			
code	CH_2039		
name	Change in Population Aged 20-39		
units	%		
abstract	Change in population aged 20-39 years, annual average change		
methodology			
classification			
	theme	Demography	
	keywords	Population	
Scope			
label	1		
lineage			
	provider	EUROSTAT	
	date	4/2010	
	URL	http://epp.eurostat.ec.europa.eu/port	
	methodology		
	methodology URI		
reliability			
	estimation	FALSE	
	quality	high	
constraints			
	public data access	TRUE	
	public metadata access	TRUE	
	copyrights	EUROSTAT	

NUTS2	TYPE	version	CH_2039	scope
validity_start			2001	
validity_end			2005	
AT11	NUTS2	2006	-1.47	1
AT12	NUTS2	2006	-1.20	1
AT13	NUTS2	2006	0.74	1
AT21	NUTS2	2006	-1.87	1
AT22	NUTS2	2006	-0.93	1
AT31	NUTS2	2006	-1.14	1
AT32	NUTS2	2006	-0.84	1
AT33	NUTS2	2006	-0.63	1
AT34	NUTS2	2006	-0.62	1
BE10	NUTS2	2006	1.39	1
BE21	NUTS2	2006	-0.81	1
BE22	NUTS2	2006	-1.30	1
BE23	NUTS2	2006	-1.06	1
BE24	NUTS2	2006	-1.15	1
BE25	NUTS2	2006	-1.52	1
BE31	NUTS2	2006	-0.44	1
BE32	NUTS2	2006	-0.99	1
BE33	NUTS2	2006	[illegible]	1
BE34	NUTS2	2006	-0.46	1
BE35	NUTS2	2006	-0.61	1
BG31	NUTS2	2006	-1.11	1
BG32	NUTS2	2006	0.15	1
BG33	NUTS2	2006	0.05	1
BG34	NUTS2	2006	-0.08	1
BG41	NUTS2	2006	0.27	1
BG42	NUTS2	2006	0.13	1



Metadata upload (required)

Contact 1 of 1

From my profile ?	□
Name ?	Johanna Roto
Organization ?	Nordregio
Function ?	Data integrator
E-mail ?	johanna.roto@nordregio.se
Phone ?	
Role ?	Custodian

THE ESPON DB metadata profile (upper figure) contains information about the dataset as a whole, about each individual indicator and about each individual value. Metadata and data files are strongly linked, all indicators and scopes described in the metadata file must be present in the data file, and *viceversa*. Metadata can either be provided in the shape of formatted Excel files, or created through theWeb Metadata Editor (lower figure), which adds the benefits of automaticly filling data and checking syntactical errors.

1.3 THE CHECKING PHASE

In order to insure data input in the ESPON DB are error-free, the data and metadata files are first subject to a thorough process of checking. The checking process is fourfold:

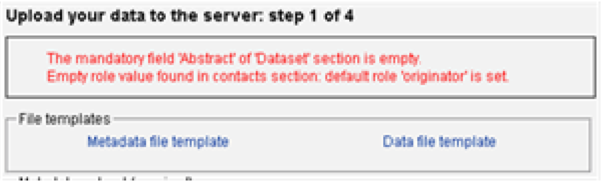
1. The syntactic checking is an automated process that aims at finding and correcting syntactical errors in both data and metadata. It is launched when providers upload their data and metadata files through the metadata editor. There are four categories of errors to be corrected: empty mandatory fields, format errors (e.g. when indicator values are text instead of numbers), typing errors (e.g. when typing the names of metadata descriptors) and data/metadata correspondence errors (e.g. when indicators described in the metadata are not present in the data or *viceversa*). During this phase, the application interacts with the user, then it is possible to solve all syntactical errors before uploading files to the ESPON DB server.

2. The thematic checking is a manual process performed by thematic experts (i.e. lead partner RIATE), which consists in assessing the thematic relevance and completeness of the dataset related to the studied topic. In this phase, the thematic expert assesses whether the indicators and values present in the dataset are well described, whether the completeness of the dataset is satisfactory over the covered area, whether the data resolution is sufficient for describing the phenomenon (e.g. if data are available at a fine territorial division or if a lower NUTS level should be sought). Obviously, there can be no automatic correction for the thematic shortcomings, so if a dataset is considered as unsatisfactory, the data provider is required to make the necessary adjustments.

3. The outlier checking is an automated checking phase aimed at detecting possible errors in individual indicator value. A set of statistical, spatial and temporal analysis methods are applied to find the outliers, values that are potentially incorrect. Outliers may result either from data manipulation errors, or from exceptional but correct values. The difference between the two cases is estabilished by a human thematic expert. If some value errors are detected, the data provider may be required to make the necessary adjustments.

4. The final checking is performed when data and metadata are included in the database by the acquisition tools. If the acquisition is successful, that means that all the integrity constraints of the database are satisfied. This phase consists in checking the consistency of the dataset with itself, but also against the rest of the data already stored in the database. Additional data (especially, spatial and thematic ontologies) help in detecting whether false entities exist in the dataset (e.g. inexistent territorial units), or if duplicated entities appear in the dataset (e.g. the same indicator with different names), or if ambiguous entities are present (e.g. different indicators having the same name, code or abstract).

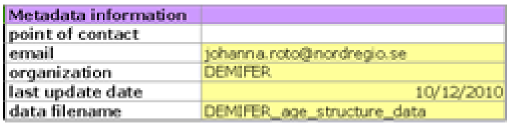
Outputs Of The Data Checking Process



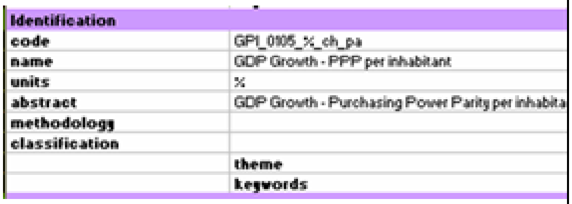
Upload your data to the server: step 1 of 4

The mandatory field 'Abstract' of 'Dataset' section is empty.
Empty role value found in contacts section: default role 'originator' is set.

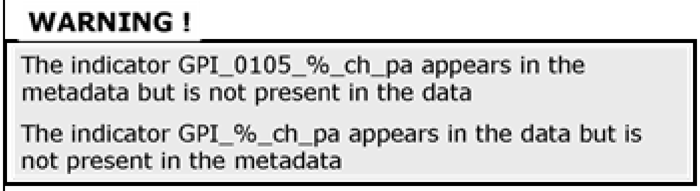
File templates



Metadata information	
point of contact	
email	johanna.roto@nordregio.se
organization	DEMIFER
last update date	10/12/2010
data filename	DEMIFER_age_structure_data



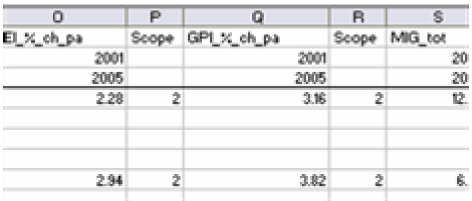
Identification	
code	GPI_0105_%_ch_pa
name	GDP Growth - PPP per inhabitant
units	%
abstract	GDP Growth - Purchasing Power Parity per inhabita
methodology	
classification	
	theme
	keywords



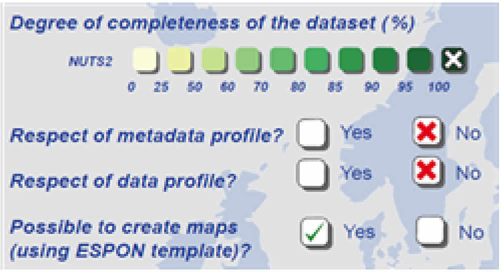
WARNING !

The indicator GPI_0105_%_ch_pa appears in the metadata but is not present in the data

The indicator GPI_%_ch_pa appears in the data but is not present in the metadata



O	P	Q	R	S
EI_%_ch_pa	Scope	GPI_%_ch_pa	Scope	MIG_tot
2001		2001		20
2005		2005		20
2.28	2	3.16	2	12.
2.94	2	3.82	2	6.



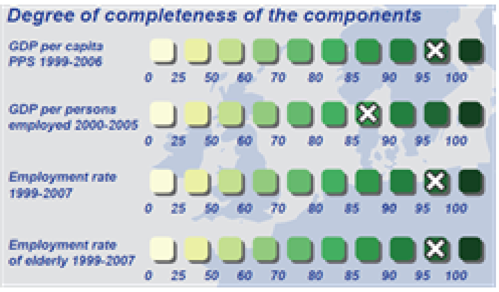
Degree of completeness of the dataset (%)

	0	25	50	60	70	80	85	90	95	100
NUTS2	☐	☐	☐	☐	☐	☐	☐	☐	☐	☒

Respect of metadata profile? ☐ Yes ☒ No

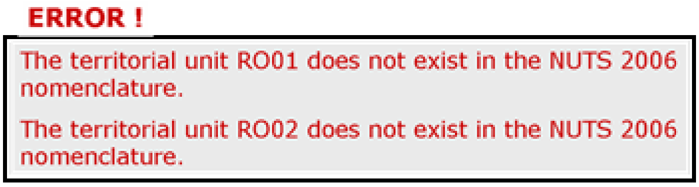
Respect of data profile? ☐ Yes ☒ No

Possible to create maps (using ESPON template)? ☑ Yes ☐ No



Degree of completeness of the components

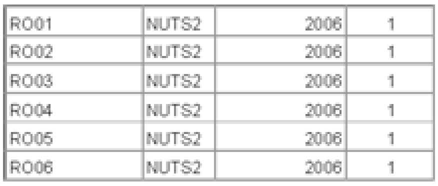
	0	25	50	60	70	80	85	90	95	100
GDP per capita PPS 1999-2006	☐	☐	☐	☐	☐	☐	☐	☐	☒	☐
GDP per persons employed 2000-2005	☐	☐	☐	☐	☐	☐	☒	☐	☐	☐
Employment rate 1999-2007	☐	☐	☐	☐	☐	☐	☐	☐	☒	☐
Employment rate of elderly 1999-2007	☐	☐	☐	☐	☐	☐	☐	☐	☒	☐



ERROR !

The territorial unit RO01 does not exist in the NUTS 2006 nomenclature.

The territorial unit RO02 does not exist in the NUTS 2006 nomenclature.



RO01	NUTS2	2006	1
RO02	NUTS2	2006	1
RO03	NUTS2	2006	1
RO04	NUTS2	2006	1
RO05	NUTS2	2006	1
RO06	NUTS2	2006	1

An illustration of different types of errors and outcomes of the checking process. On the first row, two missing metadata fields reported by the metadata editor. On the second row, a mismatch of indicator code between the data and the metadata file, reported by the upload interface. On the third row, fragments of data quality and completeness assessments, reported by thematic experts. On the fourth row, detection of territorial units assigned to the wrong NUTS version, reported by the acquisition tools upon importation in the megabase.

1.4 THE STORING PHASE

The ESPON DB Application uses two databases for the long term storage of statistical data. The separation is done in order to obtain an application optimized for two different (and conflicting) purposes:

- The ontology database is based on a conceptual schema optimized for data harmonization. This conceptual schema imposes more separation between entities, and separation implies more effort at query time (thus, query processing performance is decreased).
- The ESPON DB is based on a snapshot schema optimized for query performance in the Web interface. The data are structured in such a way that fast query answer is privileged (see a short explanation in the figure to the right, below).

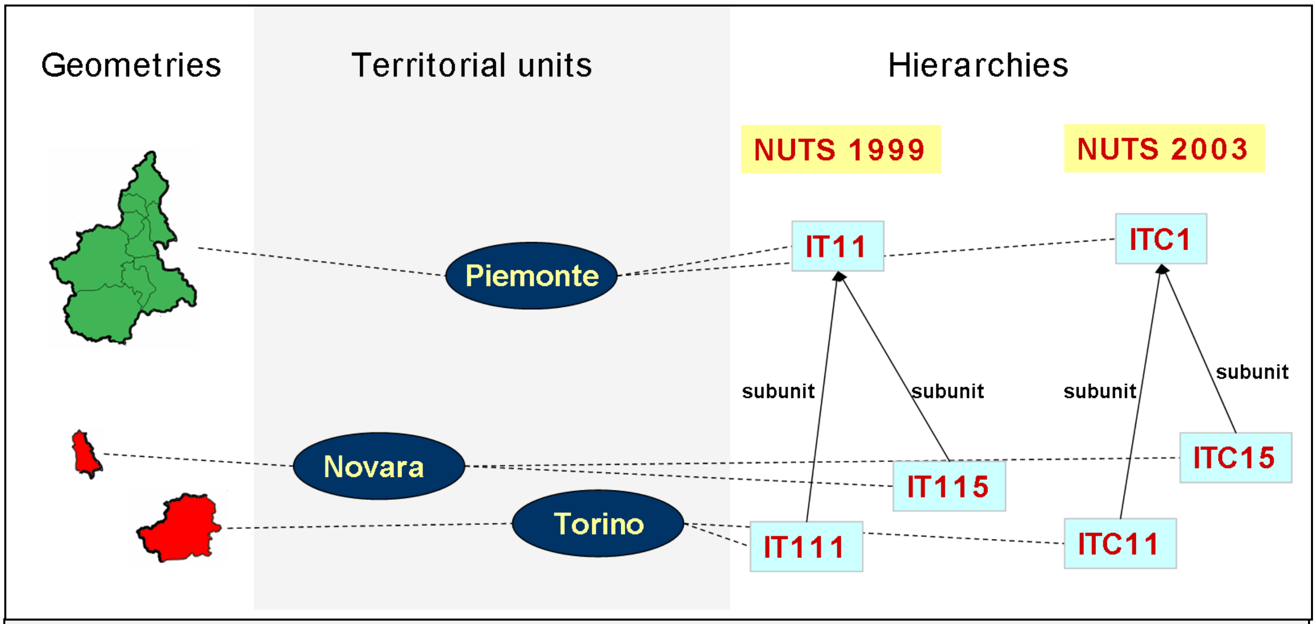
The ESPON DB Application also integrates a standalone Java application that allows inserting the content of paired data and metadata files into the megabase.

In order to enforce data consistency, this ontological database contains two ontologies, a spatial ontology (dictionary of territorial units and changes, see the figure on the right, above for a small example) and a thematic ontology (a dictionary of indicators).

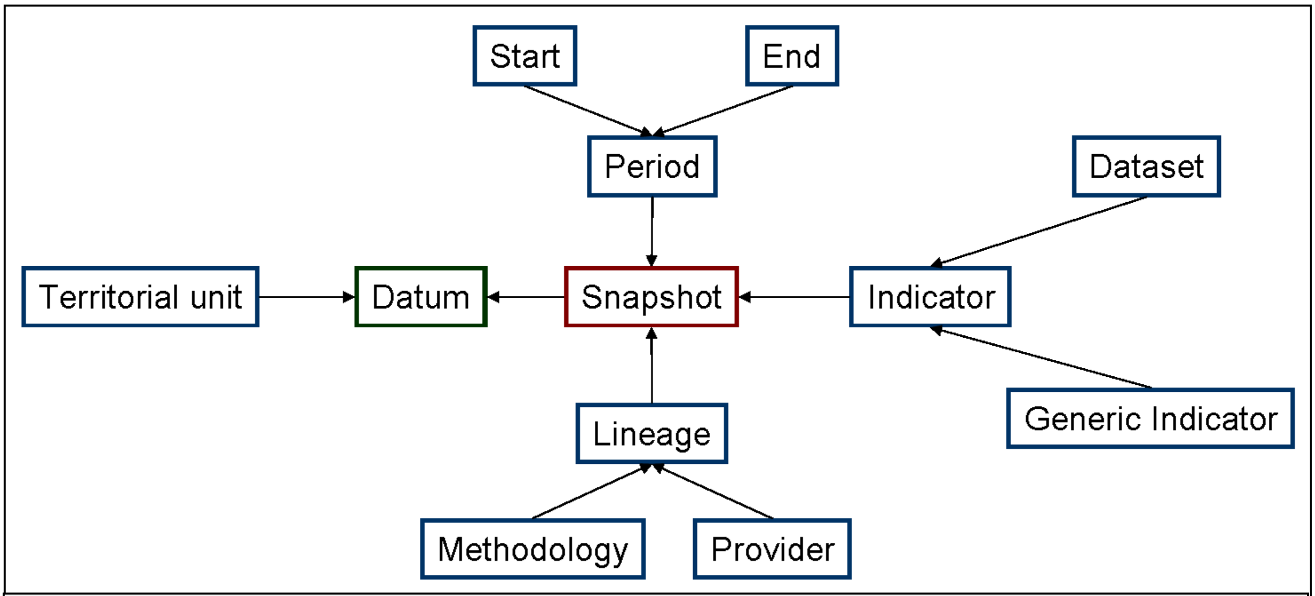
Relying on such ontologies makes it possible to detect fake entities (e.g. a territorial unit code that doesn't exist in a given NUTS revision), duplicated entities (e.g. two codes for the same indicator) and ambiguous entities (e.g. the same code for two different indicators). The existing spatial ontology covers NUTS data and follows the evolution of the different NUTS versions from NUTS 1995 to NUTS 2006. In order to insure database consistency, this ontology is extended to higher levels (world/neighbourhood) but also, as much as possible, towards lower levels (local). The thematic ontology (see Indicator coding and classification section for more clarifications) aims at giving a comprehensive dictionary of indicators stored into the ESPON DB.

Data and metadata that have been made consistent and harmonized in the megabase are transferred towards the ESPON DB. The ESPON DB is a PostgreSQL database implementing a schema targeted at offering high, scalable performance for online exploration and querying of big data quantities (see the fgigure to the right, below, for a brief presentation of the schema). It is designed for storing thematic or environmental data associated with discrete spatial divisions (e.g. NUTS and similar, LAU, etc.).

The schema of the ESPON DB allows storing and retrieving all the content described by the metadata profile. Additionnally, it integrates a user management facility, required for differentiating access to free and copyrighted data.



The spatial ontology makes a clear separation between territorial units and territorial division hierachies. One territorial unit can be part of many hierachies and it may have a different code in each hierarchy. Within each hierarchy, it can have different "subunit" relations with other units. Every attribute (name, geometry, indicators) can evolve in time. This allows a very clear view of territorial division changes.



The ESPON DB schema is optimised for fast querying and for reduced database size. On this simplified representation, we can see how three of the four dimensions of an indicator value (*datum* table) have been merged. Introducing the "snapshot" table allows more than halving the size of the datum table (which is the main table of the database, holding millions of records). It also allows fastening queries, by introducing an additional indexing level.

1.5 THE DOWNLOAD PHASE

The ESPON DB Web download interface is an on-line application designed to offer fast browsing and searching capabilities over the ESPON DB. The Web download interface implements several inovative elements that garanties scalable performance to accomodate the fast growing size of the ESPON DB :

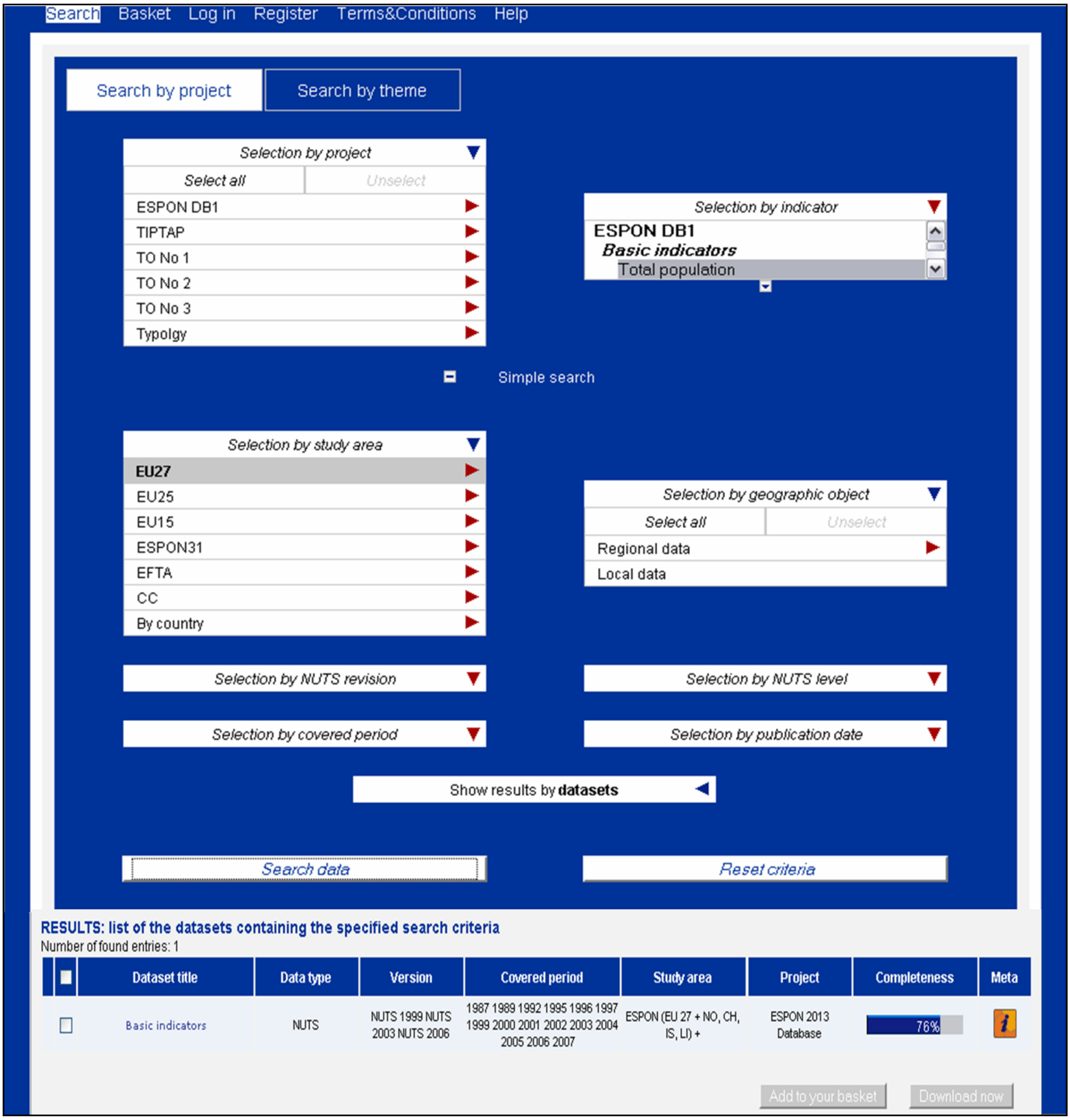
- The use of a server-side application cache system allows the application to avoid querying the database for all browsing tasks excepting the advanced search. This insures fast data searching, whatever the database size.
- The use of an XML exchange format for the answer to queries allows decreasing the size of the data transfers between server and client.
- The use of AJAX techniques (Asynchronous JavaScript and XML) allows further decreasing the size of the traffic between the client (Web browser) and the server (ESPON Web site), by transferring only the parts of query that have changed (in XML) and redisplaying them accordingly on the client (using JavaScript). This allows for load balancing between client and server, as the task of building the presentation from the XML file is performed on the client.
- The dropdown lists used in the interface have been developped as new components in order to match the ESPON look&feel requirements.

The Web download interface (see figure to the right) allows users to search and explore data in two ways: either by project (data provider and dataset) or by theme. In each type of search, an advanced mode is also available, allowing users to add more research and filter criteria: study area (country groups or countries), covered time period, object type (nomenclature versions and levels), and publication date. The search results can be listed as datasets or as individual indicators.

The table of results that is generated as a first answer can be further filtered in order to match yet better the user's needs, by removing unwanted indicators, territorial units, years or versions. Selected search results can be progressively added to a basket as in most of e-commercial Web applications. The basket can be downloaded at the end of the session, under the form of a zip file containing all the datasets selected by the user.

The table of results lets the users see the completeness of the dataset as a whole and also by nomenclature level, under the shape of a percentage bar (see figure). The interface also gives the possibility to the users to consult all the metadata related to the dataset. The three levels of metadata can be viewed: dataset, indicator and value levels. The completeness can be displayed by nomenclature level on a map.

The Web Query And Download Interface



RESULTS: list of the datasets containing the specified search criteria

Number of found entries: 1

☐	Dataset title	Data type	Version	Covered period	Study area	Project	Completeness	Meta
☐	Basic indicators	NUTS	NUTS 1999 NUTS 2003 NUTS 2006	1987 1989 1992 1995 1996 1997 1999 2000 2001 2002 2003 2004 2005 2006 2007	ESPON (EU 27 + NO, CH, IS, LI) +	ESPON 2013 Database	76%	i

The Web Query and Download interface allows users to formulate two types of basic search: ‘by project’ and ‘by theme’). For each basic search, advanced search criteria can be added. This search interface is dynamic: search criteria lists are expanded only if they are used. On the example, two additional search criteria have been added, (study area : “EU 27” and geographic object type “NUTS and similar”). The Web Query and Download interface has been optimized so that building complex queries takes as little space as possible in the Web browser.

1.6 CODING SCHEME

KEY FINDINGS

- The harmonisation of coding schemes is of crucial importance for the ESPON 2013 DB. With this regard, TPGs involved in applied research projects are increasing the level of ambiguity when put into practice their own scheme to code indices, indicators and other measures
- To a certain extent, coding schemes are not used to express the content of data but rather an attempt to homogenise codes. However, some information needs to be provided and, most importantly, it needs to be arranged in a consistent way to avoid conflicts with the web-based user interface
- Despite the diversity of approaches to code data, standards used by ESPON projects were taken into account in the analysis that allowed the creation of the coding scheme

DESCRIPTION

The coding scheme has been elaborated in the context of the ESPON 2013 DB project to provide TPGs with a unique code. Against this background, research teams are encouraged to apply a scheme that comprises three fields. The information to be added in each field corresponds to the subject, restrictions and/or derivations, and level of measurement. Other elements that might be used to classify data should not be considered as they already appear in the metadata file (e.g. time, space).

The procedure is not constrained to a limit of characters, but it is important to respect the above-mentioned structure. As a consequence, the first field should integrate information about the subject. The second part refers to widely used abbreviations that impose restrictions and/or use derivations. Ultimately, the third field specifies the level of measurement so that users can understand the statistical operations that have been carried out on the data. In ascending order of precision, the different levels of measurement are nominal, ordinal, interval, and ratio.

For each field, a non-exhaustive list of acronyms and abbreviations is provided to encourage harmonisation. In some cases, adaptations will be necessary, especially to obtain more degree of freedom when facing rather complex, but similar, data. The coding scheme has been implemented and tested for datasets delivered by the first round of ESPON projects under Priority 1, 2, and 3.

Additional improvements will be needed to further increase the quality of this proposal. At this point, it is not possible to anticipate many of the indices and indicators that will be delivered. That will require the involvement of the ESPON research community through a continuous, dynamic process.

Illustrative examples of harmonised coding schemes

(a)

Subject(s)								Derivations / Restrictions								Level of measurement					
m	i	g	.	p	o	p	_	c	h	.	t				_	r	t	c			

(b)

Subject(s)								Derivations / Restrictions								Level of measurement					
a	c	c	.	a	i	r	_	a	b	s					_	r	t	e			

(c)

Subject(s)								Derivations / Restrictions								Level of measurement					
e	d	u	.	s	c	d	_	t							_	r	t	c			

(d)

Subject(s)								Derivations / Restrictions								Level of measurement					
p	o	p					_	2	0	-	3	9	.	t	_	r	t	c			

(e)

Subject(s)								Derivations / Restrictions								Level of measurement					
C	O	2	.	r	o	d	_	v	o	l					_	r	t	e			

(f)

Subject(s)								Derivations / Restrictions								Level of measurement					
t	y	p					_	r	u	r	a	l			_	n	o	c			

The following examples provide a better understanding of the rationale behind the coding scheme, where (a) reflects ‘Migratory population change’, (b) ‘Potential accessibility by air [absolute level]’, (c) ‘Persons with secondary education degree’, (d) ‘Population aged 20-29 years’, (e) ‘CO2 emissions by road traffic’, and (f) ‘Typology of rural regions’. Each field of the coding scheme should be separated by the underscore symbol. In addition, it suggests a number of cells to be filled in by TPGs.

1.7 THEMATIC STRUCTURE

KEY FINDINGS

- Database structures adopted by international organisations with in-house data constitute an important source of information. Therefore, we apply a visual grouping technique to illustrate, by means of correlation matrices, homogeneous clusters of words that identify those themes
- The rationale for sub-themes derives from text mining methods. We assume that the ESPON 2006 Programme introduced new vocabulary. This assumption is investigated by extracting keywords from a large corpus of textual data. In order to improve the interpretation of the results, we employ visualisation tools of data co-occurrence to understand similarities
- The results obtained suggest that the ESPON 2013 DB should be structured in 7+1 themes and 29 sub-themes

DESCRIPTION

A two-step approach has been developed to structure the ESPON 2013 DB by themes and sub-themes. We argue that database structures adopted international organisations should support the definition of themes. This assumption lies on the fact that, very often, database structures define common topics to allocate data. For this purpose, we employ correlation matrices to analyse similarities and consequently interpret the results through visual grouping techniques. The proposal suggests seven themes. In addition, we add a theme to cover cross-thematic and non-thematic data.

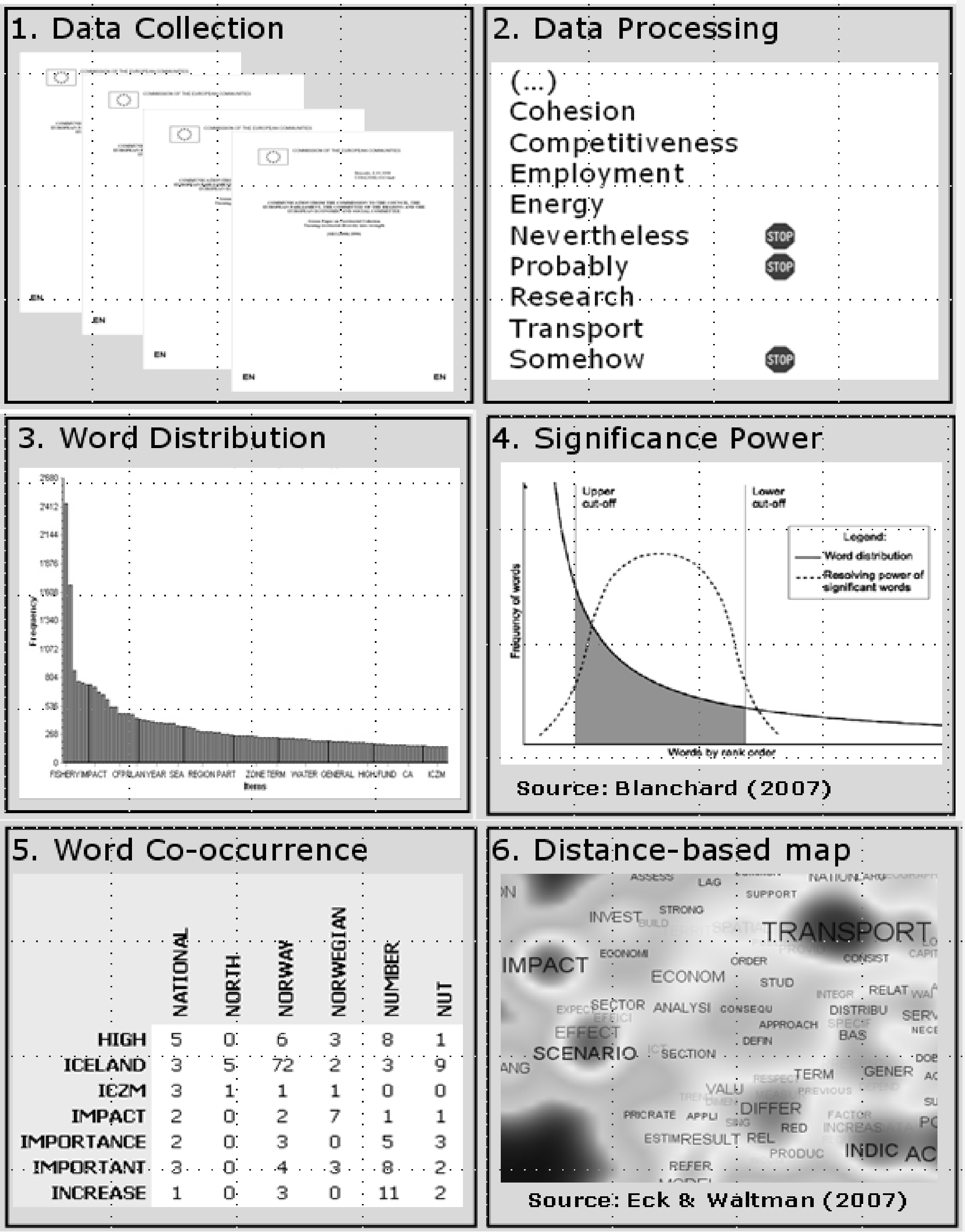
The demand from the ESPON 2013 DB end users will be characterised by immediate, easy and practical access to data. A properly structure is therefore the key to meet this request. The next step comprised the definition of sub-themes. In order to achieve this goal, we explore the potentialities offered by text mining methods. This approach is used to find patterns across textual data that, inductively, create thematic overviews of text collections.

According to Dühr (2010), ESPON introduced new vocabulary of shared spatial concepts in Europe. We investigate this assumption by extracting keyword co-occurrence from texts with ESPON evidence and results.

In order to achieve concrete groups of keyword co-occurrence, textual data needs to be carefully prepared. Similarly, one of the crucial needs in text mining is the ability to visualise the relation of words. Hence, we apply a visualisation tool to construct and view maps of keywords based on co-occurrence and therefore better explore the results obtained from the information extraction phase. The results obtained constitute the basis for decision-making on sub-themes that eventually will facilitate the allocation of variables delivered by TPGs.

Related technical report "Text mining and visualisation tools as means to support the thematic structuring of the ESPON 2013 DB (produced by the University of Luxembourg)

Short description of data preparation and visualization



	NATIONAL	NORTH	NORWAY	NORWEGIAN	NUMBER	NUT
HIGH	5	0	6	3	8	1
ICELAND	3	5	72	2	3	9
IEZM	3	1	1	1	0	0
IMPACT	2	0	2	7	1	1
IMPORTANCE	2	0	3	0	5	3
IMPORTANT	3	0	4	3	8	2
INCREASE	1	0	3	0	11	2

Source: Blanchard (2007)

Source: Eck & Waltman (2007)

The methods used to identify sub-themes on text collections with ESPON evidence considered the above-mentioned steps. These steps have been performed for each of the seven themes that came out from our analysis on database structures.

1.8 THE OLAP CUBE

KEY FINDINGS

- OLAP stands for **On-Line Analytical Processing**. It consists on a **multidimensional** data model, allowing complex analytical and ad-hoc **queries** with a rapid execution time. OLAP technology has been proven to be a useful way to integrate NUTS-based data together with continuous data, such as land cover, over different time frames.
- The **ESPON OLAP Cube** consists on some socio-economic variables which can are integrated and combined within a set of **dimensions**:
  - **Spatial** dimensions (e.g. NUTS regions).
  - **Thematic** dimensions (e.g. land cover).
  - **Temporal** dimensions (e.g. 2003, 2006…).

METHODOLOGICAL ISSUES

An OLAP Cube can be queried online and offline. So far, the online connection has not been implemented. In order to test the cube, we provide a single file .CUB which works offline. The .CUB file can be connected to and queried from Microsoft Excel with a few easy steps. A user manual has been provided, attached to the Technical Report.

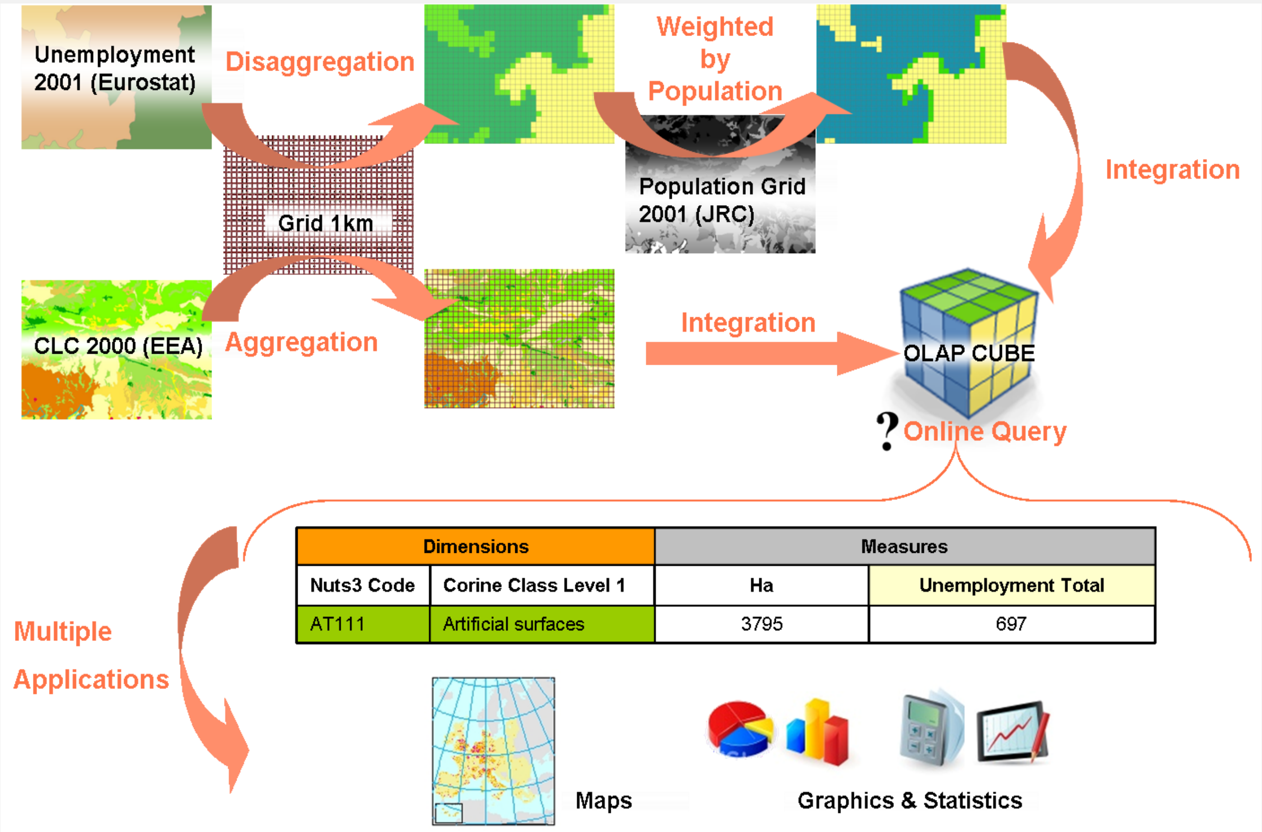
THEMATIC ISSUES

The current version 3.0 of the ESPON OLAP Cube include the following variables and dimensions:

- Socioeconomic variables : GDP 2003, GDP 2006, Active population 2003, Active population 2006, Unemployment 2003, Unemployment 2006.
- Land cover: Corine Land Cover 1990, 2000, 2006.
- Land cover changes: Land Cover Flows 1990-2000, 1990-2006, 2000-2006.
- Measures : Population density 2001 ; Area (ha)
- Geographical dimensions : Elevation Breakdown ; Biogeographic Regions ; Large Urban Zones and City Names ; Massifs ; Nuts 2006 ; Nut 2003 ; River Basin Districts UE.

Related technical report "Disaggregation of socioeconomic data into a regular grid: Results of the methodology testing phase" (produced by the University Autonoma de Barcelona)

Methodological schema of disaggregation of socioeconomic data and combination with other data types



Dimensions		Measures	
Nuts3 Code	Corine Class Level 1	Ha	Unemployment Total
AT111	Artificial surfaces	3795	697

This schema summarises the whole methodology implemented by the ESPON DB project in order to combine socioeconomic data, usually reported by administrative units, together with other types of data, mainly continuous, such as land cover. The OLAP Cube is the core of the methodology as it is used as the integration repository for all the datasets, and allows a prompt querying of the data to get interesting analytical results.

The diagram shows, on the one hand the process of aggregation/disaggregation of data by means of the 1 km Reference Grid, and the weighting by population density whenever it is possible, to add some value to the disaggregation of the source data.

Finally, all the variables reported by 1 km grid cell are integrated into the OLAP Cube, which facilitates their combination and querying as it has been explained, making the creation of maps and graphs straight-forward.

1.9 CARTOGRAPHY IN ESPON

KEY FINDINGS

- The ESPON mapkit is a set of mapkits according to the geographical levels.
- It ensures harmonization of all the maps produced in ESPON projects.
- It is compliant with ~~de~~ ESPON Database application.
- It is available at different format (ArcGis, QGIS, Philcarto).
- A mapping guide is provided to explain main rules for mapping in ESPON.

DESCRIPTION

As a general rule, maps are used to visualize geospatial data and enhancing statistical data to understand phenomena. In ESPON Program, there is a need to produce a lot of maps. This part presents the mapkit developed by the ESPON DB project and follows **3 main objectives:**

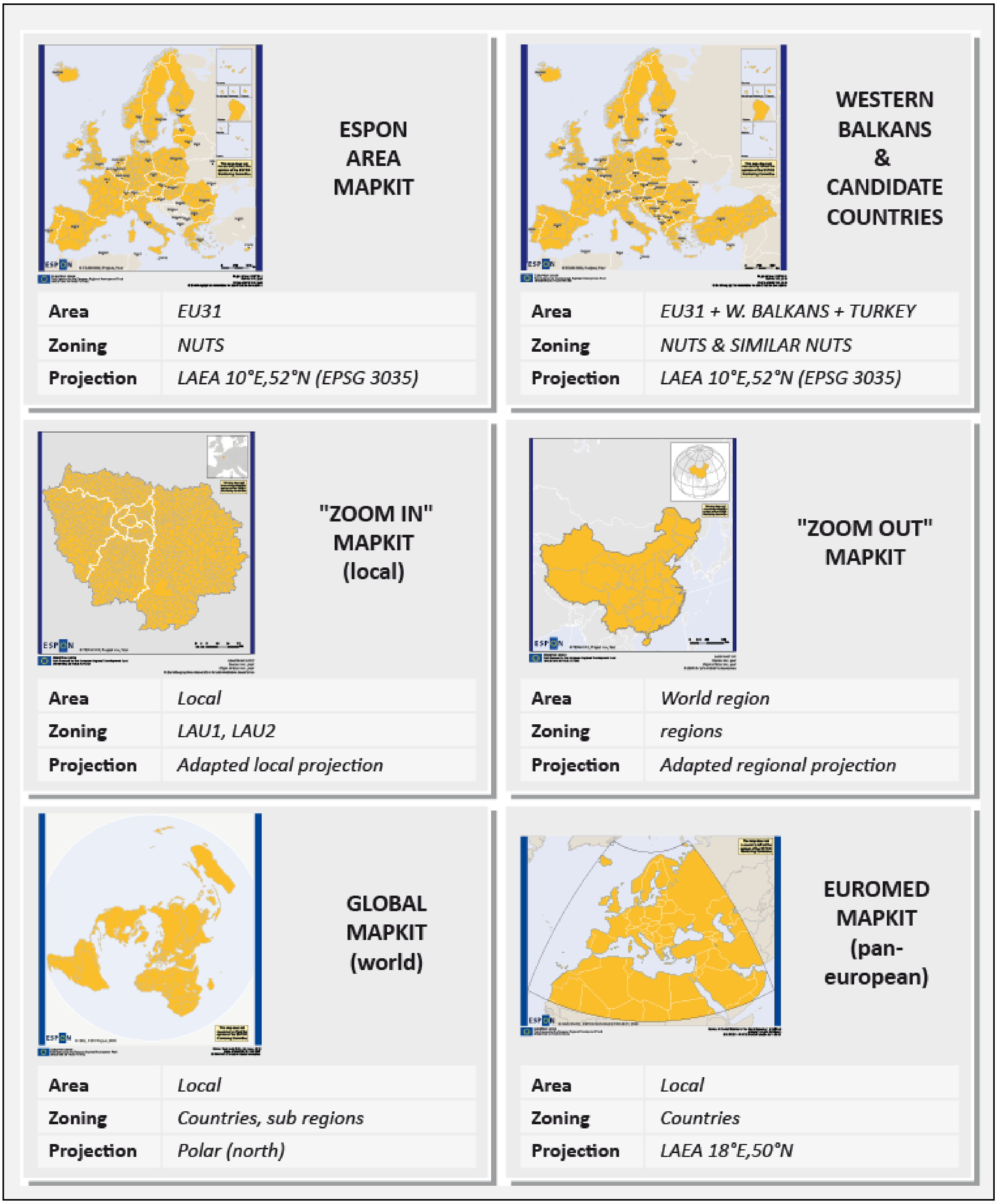
i) Ensuring harmonization of maps. Maps are produced by researchers, engineers or students involved in each ESPON projects. Consequently, we need to ensure graphical harmonization of all maps produced by different authors, with different software. The mapkit tool (consisting of specific mapkits ollection) contains geometries, cartographical templates and graphic elements (logos, disclaimers). When possible, these different elements are available in Arcgis format (mxd + shapfiles), Quantum GIS (a user friendly Open Source Geographic Information System licensed under the GNU General Public License), and Philcarto which is a free software for thematic cartography.

ii) Ensuring compatibility with the ESPON 2013 database application. The ESPON DB application provides indicators at local, regional and global level. It also provides data on different geographical objects (*e.g. dots and grids*). The mapkit ensures the mapping of data on these different kinds of objects. It is compliant with the ESPON Database application and permits to visualize, on a map, the data extracted from the application whatever the kind of data.

iii) Enhancing information (How to make good maps). Many possibilities exist to show data on map. Choosing relevant representation is not an obvious task and has to be considered seriously. Indeed, choosing the wrong way of mapping can completely misrepresent the data. For this reason, a mapping guide was realized to help people to follow "good rules" of cartography. Moreover, it is important to keep in mind that choice in cartography is always dependent on the type of data (and targeting the right audience) and that there is never an optimal solution. Map is always a compromise.

Related technical report "Mapping guide, cartography in ESPON 2013" (Produced by RIATE)

The set of map kits



ESPON AREA MAPKIT

Area	*EU31*
Zoning	*NUTS*
Projection	*LAEA 10°E,52°N (EPSG 3035)*

WESTERN BALKANS & CANDIDATE COUNTRIES

Area	*EU31 + W. BALKANS + TURKEY*
Zoning	*NUTS & SIMILAR NUTS*
Projection	*LAEA 10°E,52°N (EPSG 3035)*

"ZOOM IN" MAPKIT (local)

Area	*Local*
Zoning	*LAU1, LAU2*
Projection	*Adapted local projection*

"ZOOM OUT" MAPKIT

Area	*World region*
Zoning	*regions*
Projection	*Adapted regional projection*

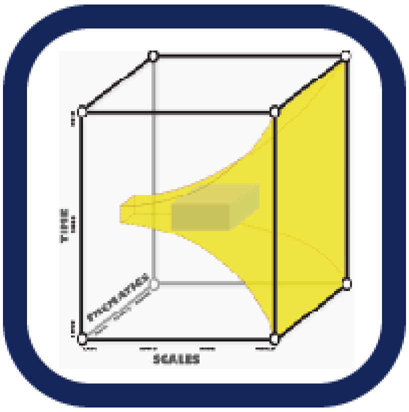
GLOBAL MAPKIT (world)

Area	*Local*
Zoning	*Countries, sub regions*
Projection	*Polar (north)*

EUROMED MAPKIT (pan-european)

Area	*Local*
Zoning	*Countries*
Projection	*LAEA 18°E,50°N*

This picture is an overview of the ESPON mapkit. Actually, it is composed by a set of 6 specific mapkits adapted to different geographical levels, from local to global.



2.THEMATIC ISSUES

2.1 – TIME SERIES HARMONISATION

2.2 – NAMING URBAN MORPHOLOGICAL ZONES

2.3 – LUZ SPECIFICATIONS

2.4 – FUNCTIONAL URBAN AREAS DATABASE

2.5 – SOCIAL/ENVIRONMENTAL DATA

2.6 – INDIVIDUAL DATA AND SURVEYS

2.7 – LOCAL DATA

2.8 – ENLARGEMENT TO NEIGHBHORHOOD

2.9 – WORLD/REGIONAL DATA

2.10 - SPATIAL ANALYSIS FOR QUALITY CONTROL

INTRODUCTION – PART 2

The second part of this report presents the different solutions elaborated by ESPON Database project in order to enlarge the possibilities of territorial data exploration, mapping and prospective. Each of the 10 sections is related to one or several technical reports that are referenced and can be downloaded on the ESPON website. For each topic, we firstly summarize the key findings, the methodological issues and thematic issues on the left page. Then we illustrate the interest of this technical report by a significant and demonstrative example of application on the right page.

The first topic is related to time series harmonization (2.1) and describes how to combine NUTS data with different geometries and, more generally, how to remove holes in time series and enlarge them back to past or toward future.

The second topic is related to urban data, which are becoming more and more a key issue for territorial cohesion and competitiveness. We describe firstly how to cope with two different definitions of cities produced by EEA or Eurostat. A first report demonstrates how to make better use of Urban Morphological Zones, in particular through a naming process (2.2). A second report focus on the specification of Larger Urban Zones and tries to clarify the national definition of cities used by each countries in the different periods of elaboration of urban audit (2.3). Finally, an attempt is made to delineate Functional Urban Areas with an harmonized criterion of polarization (2.4).

The third topic is related to the thematic enlargement of ESPON database in order to overcome the classical focus on economic dimensions. The combination of social and environmental data can be strongly developed through the elaboration of aggregation and disaggregation methods, making possible to transfer information from NUTS to Grid or from Grid to NUTS (2.5). The same idea is applied for individual data based on surveys that can be used for the elaboration of more innovative information on social dimension of territorial cohesion, but with a great attention paid to the problems of sampling errors when data are estimated at regional or city levels (2.6).

The fourth topic is related to the enlargement of the possibilities of ESPON database to upper or lower geographical scales, in order to make more easy the objective of the "five level approach" described in the first ESPON Synthesis Report (p. 17). We analyze firstly the possibility to improve the collection of local data at LAU1 and LAU2 levels (2.7). Then, we propose solutions for the coverage of regional data related to candidate countries and the rest of western Balkans (2.8). Finally, we describe the procedure of data collection for supporting the elaboration of a world database (2.9).

2.1 TIME-SERIE HARMONISATION

KEY FINDINGS

- Review of literature on the various possible solutions for the harmonization of times series. Benchmarking of these solutions and proposal of a general solution that is a development and improvement of the "ESTI" model previously proposed in ESPON 2006 Data Navigator Project.
- Compilation and inclusion in the ESPON 2013 database of data using old NUTS version (1995, 1999, 2003, 2006). In particular, data elaborated in ESPON 2006 program and historical data from Eurostat.
- Elaboration of a systematic dictionary of change of NUTS units based on the concept of "lineage". The concept of lineage is more general than a simple review of modification (as provided by Eurostat) and offers the possibility to follow a regional unit through time, even when names, geometry, codes, etc. are changing.

METHODOLOGICAL ISSUES

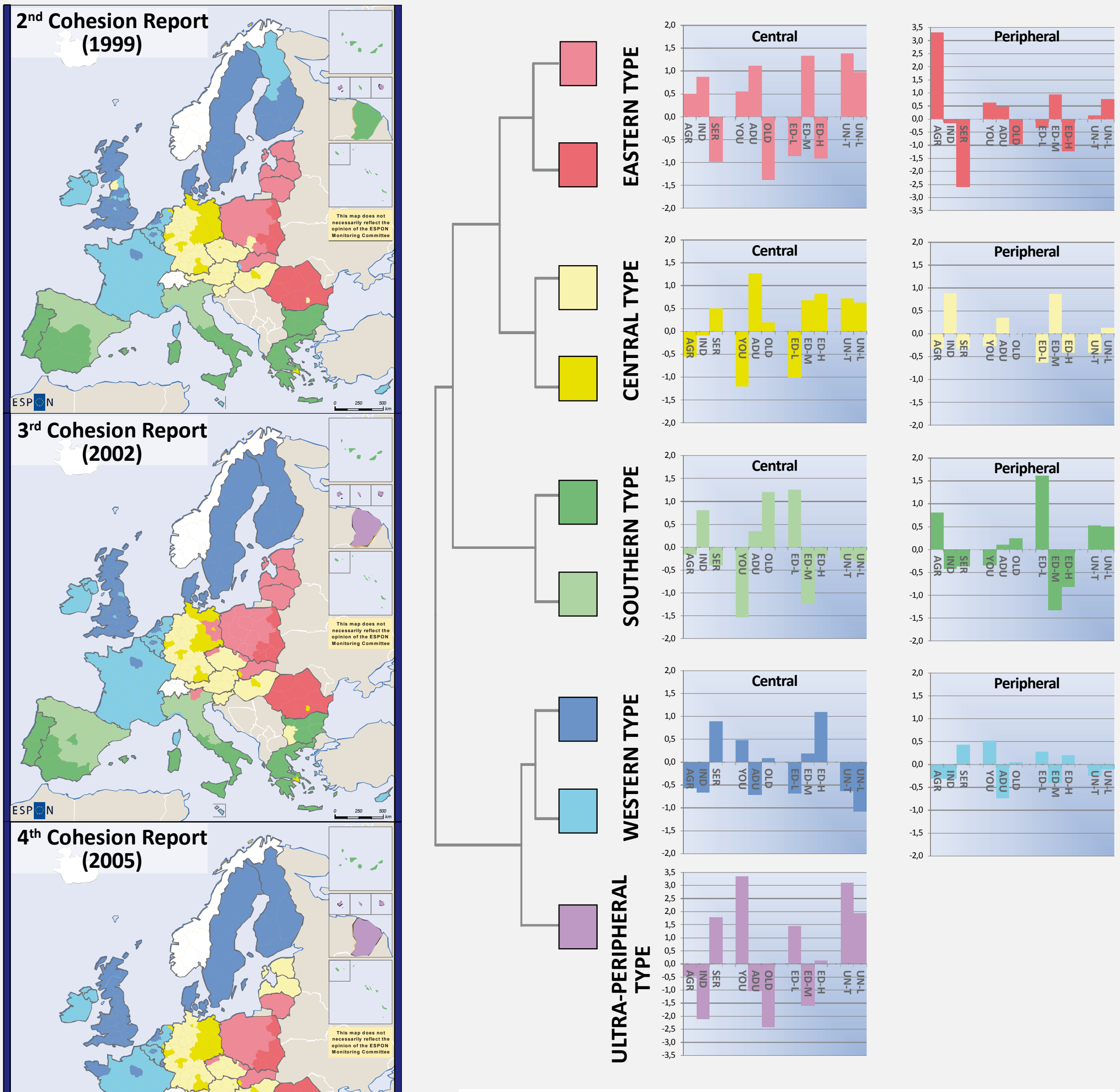
Time series approach can be organized in two main steps. Firstly, the search and the exploration of historical databases (New Chronos from EUROSTAT, cohesion reports from DG-REGIO…) are performed. This step aims at providing a survey of continuous time-data series that could be built from these databases. 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 survey 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 will provide the basis for a future computer implementation of automatic procedures for estimating the values of missing data.

THEMATIC ISSUES

Some of the methods proposed for the estimation of missing values in time series have been directly used in the ESPON Database in order to remove holes or to propose "provisional estimation" of indicators of interest like the Unemployment rate in March 2010 at NUTS2 level (ESPON First Synthesis Report, p. 21). Moreover, we have demonstrated that it is possible to analyze the evolution of spatial patterns of regional unequalities with changing territorial units, through the example of a cross analysis of statistical annexes of 2nd, 3rd and 4th Cohesion Reports (see on next page).

Related technical report "Towards an approach of time series data issues: from empirical methods to application" (produced by RIATE)

Typology of EU Regions according to Cohesion Reports (out of GDP…)



2nd Cohesion Report (1999)
3rd Cohesion Report (2002)
4th Cohesion Report (2005)
ESPON
This map does not necessarily reflect the opinion of the ESPON Monitoring Committee
EASTERN TYPE
CENTRAL TYPE
SOUTHERN TYPE
WESTERN TYPE
ULTRA-PERIPHERAL TYPE
Central
Peripheral
AGR
IND
SER
YOU
ADU
OLD
ED-L
ED-M
ED-H
UN-T
UN-L

Indicators contented in the classification, from the left to the right of the profile:

AGR: Employment in agriculture sector
IND: Employment in industry sector
SER: Employment in service sector

YOU: Share of youngs (0-14)
ADU: Share of actives (15-64)
OLD: Share of old (65+)

ED-L: Share of active with low education level
ED-M: Share of active with medium eduction level
ED-H: Share of active with high education level

UN-T: Unemployment rate
UN-L: Long term unemployment rate

A typology of EU regions, based on structural phenomena used in the cohesion report (demography, economy, education, labour force) but excluding GDP per capita, reveals the existence of 4 basic patterns of strength and weakness in northwestern, southern, central and eastern parts of EU. Interestingly, the exclusion of GDP per capita provides a division of EU territory which does not follow the classical division between old and new member states. Despite some minor changes, this structural pattern is very stable through time and could be therefore used as a basis for “taylor-made” policies of regional development.

2.2 NAMING URBAN MORPHOLOGICAL ZONES

KEY FINDINGS

- A European data base operational for urban studies, containing 4437 cities over 10 000 inhabitants, defined from CLC2000 with harmonized criteria (EEA, last version of Urban Morphological Zone shapes), population (JRC, last version of Population Density Grid), names (ESPON Data Base, see Technical Report) and metadata.

- An automatised process for naming UMZ which allows quick updating with new versions of sources or methods (EEA and JRC) or new dates (2006, 2010…). The methodology is fully explained and documented in the Technical Report.

- A validation of the method at national scale, starting from an expertise which selects the relevant national data base for city names (LAU 2 in majority, but also LAU1 or national urban areas, see Figure 1). Final results have been systematically matched to other sources (Eurostat, Geopolis, Google Earth).

- A powerful data base for exploring the common features of European cities in 2000 as regards to their settlement characteristics (population, surface, density) but also the major regional variations in density patterns and spatial coverages (see for example the general North-South density gradient in Figure 2). Interoperability with other geo-referenced data bases (urban transport infrastructures, urban mobility, socio-economic LAU data…) opens a wide range of environmental and social studies.

METHODOLOGICAL ISSUES

Elaborating a methodology for naming physical entities that are automatically built from satellite images raises a series of redoubtable and very classical issues in data base modelling. The inputs have to take into account a huge set of data, the diversity of sources, technologies and national approaches, and evolving contents. The research that has been developed is constantly based on two main principles: international harmonizationof processes and automation of each step of the process, i.e.:

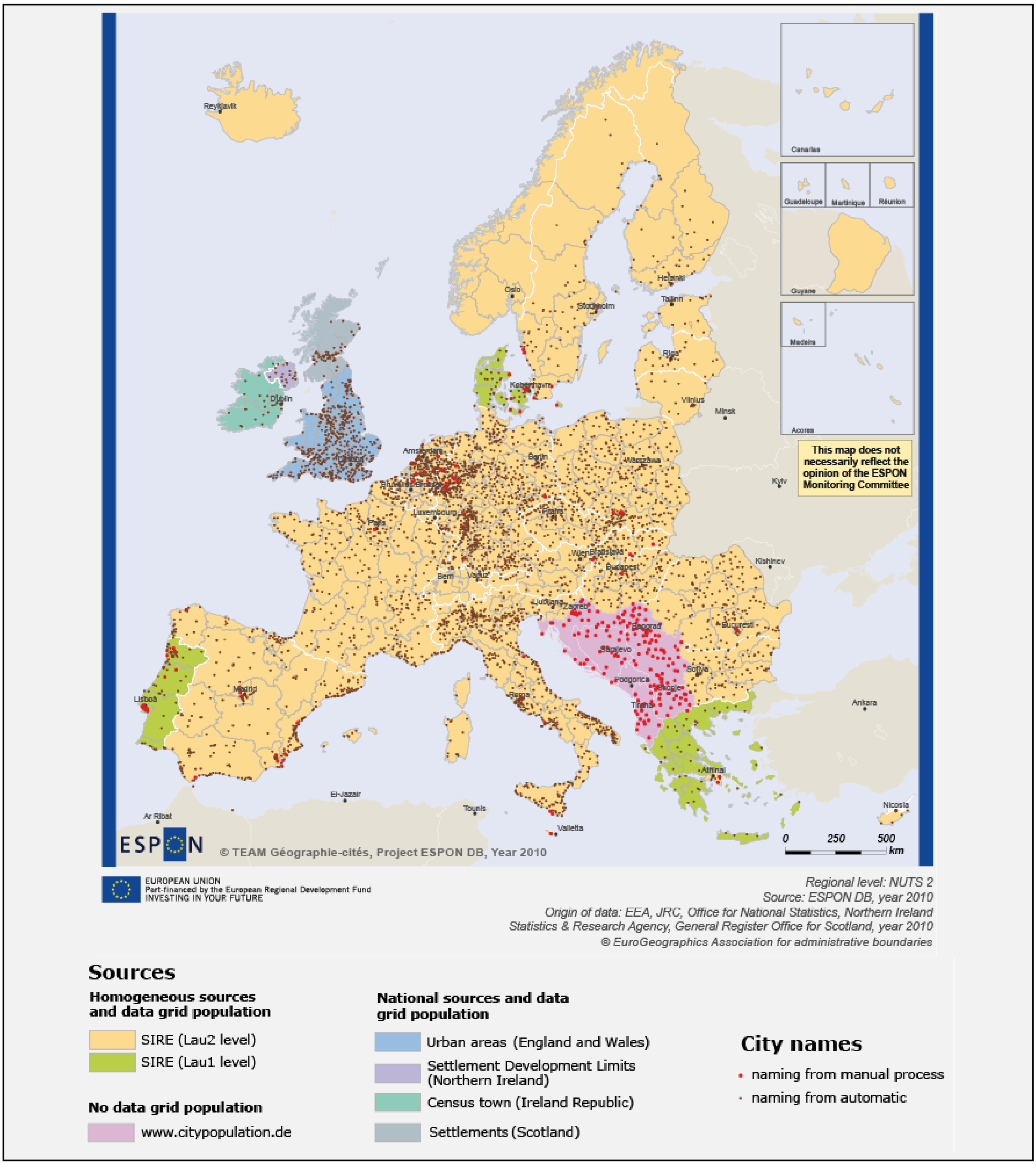
- Automation of the computation of populations intersecting the different sources (UMZ, Population dentity grid and national data base selected for giving the name)

- Automation of the attribution of names given to each UMZ, according to the way it overlaps the national data base elementary units: clearly concentrated inside one unit (then receiving one name), or expanding clearly on 2 or more units (then receiving 2 or more names, like in industrial or littoral conurbations)

- Automation of the final checking, by comparing names to Eurostat compilation of national city names, Geopolis data base (François Moriconi-Ebrard 1993) and Google Earth. A typology of the main inadequations (about 90 cases on 4437) has been realized and solutions proposed, that will make easier the future checks.

Related technical report " Naming UMZ" (Produced by Geographie-cités)

From morphological delimitations (UMZ) to a European set of cities : naming process



© TEAM Géographie-cités, Project ESPON DB, Year 2010

Regional level: NUTS 2
Source: ESPON DB, year 2010
Origin of data: EEA, JRC, Office for National Statistics, Northern Ireland Statistics & Research Agency, General Register Office for Scotland, year 2010
© EuroGeographics Association for administrative boundaries

Figure 2.2.1 illustrates one step of the methodology used for naming UMZ, consisting to choose the relevant sources for naming cities (LAU2, LAU1 or national data sets), according to an urban expertise and to the avaibility of the Population density grid.

2.2. NAMING URBAN MORPHOLOGICAL ZONES

THEMATIC ISSUES

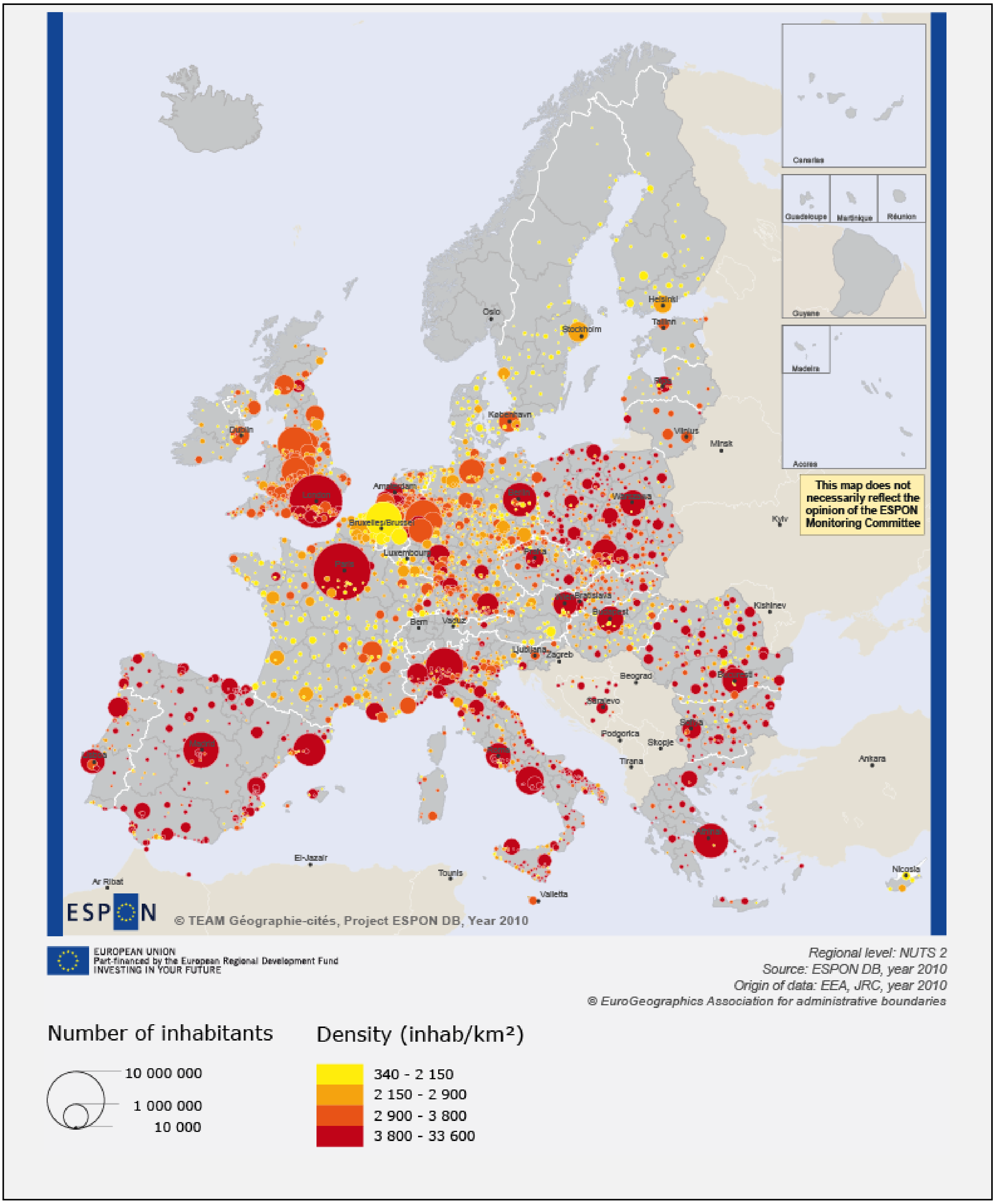
The UMZ data base is now fully operational for a deep exploration of the common features and diversity of European urban settlements. Three types of analyses have been presented in the Technical Report:

- *City size distribution*: the classical rank-size distribution, plotted for the 4437 cities over 10 000 inhabitants, confirms the very high regularity of the hierchical structure at the European level (the determination coefficient R^2 equals to 0.99). The absolute value of the slope, used as an indicator of city size inequality level, is 0.96, very close to other values computed by European researchers with former databases. National studies and computation of primacy index should fruitfully complete this overview of urban hierarchy in Europe.

- *Density patterns*: a multiscalar analysis of density levels in Europe gives striking results, with a major North-South gradient (for exemple, average urban density is lower than 2000 inh./km^2 in Sweden, Denmark, Finland, whereas it reaches 4000 inh./km^2 in Italy and more in Spain or Greece), but also a strong and regular relationship with city size levels (density above 5700 inh./km^2 in cities larger than 2 millions inh., then decreasing regularly until 3000 inh./km^2 for cities between 10 000 and 25 000 inh./km^2). National specificities appear also very strongly (see the French/Spain frontier or the high densities of some Eastern countries like Poland). These studies are of high interest for the future, for example in urban planning issues (transportation or environmental topics).

- *International UMZ*: an index of internalization (% of population living in one or more countries different than the main one) has been computed for UMZ crossing two or more countries. It allows to qualifye in a comparable way to what extent the city is embedded in a multi-national context and completes in a fruitful way the population indicator of these UMZ: for exemple, the most populated international UMZ is Brussel/Anvers/Gand, but it extends in a very small part in Netherlands (international index is only 1%). At the opposite, some UMZ located at the Poland/Germany, Slovakia/Hungary or Austria/Germany frontiers are not very populated but their international index is over 30%.

European City sizes and Densities (UMZ/ CLC 2000)



© TEAM Géographie-cités, Project ESPON DB, Year 2010

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

Regional level: NUTS 2
Source: ESPON DB, year 2010
Origin of data: EEA, JRC, year 2010
© EuroGeographics Association for administrative boundaries

Figure 2.2.2 enlightens one of the thematic exploration that can be lead now with the UMZ data base: the map of the density indicator reveals a major North-South gradient, but also national and size effets.

2.3 LUZ SPECIFICATIONS

KEY FINDINGS

- Larger Urban Zones collected in each country by Urban Audit 2004 represent another way of trying to build a European set of cities. As opposed to UMZ, it is not a top-down approach (starting from identical definition criteria and trying to enrich it by taking into account national diversity) but a bottom-up approach.
- Using LUZ perimeters and indicators requires first a good knowledge of national specifications, in order to be able to identify bias resulting from the national heterogeneityof LUZ definitions.
- Based on an expertise from national reports sent to Urban Audit by the different countries and completed by other sources of information, the Technical Report presents a clear synthesis of the LUZ specifications, through 4 synthetic typologies and maps but also an Annex containing 30 country-sheets which describe in a common vocabulary and “syntax” the general rules used by each country to define its LUZ.

METHODOLOGICAL ISSUES

The national reports are extremely diverse and the synthesis work is very complex: some are not written in English, most of them are very allusive and need to be completed with other sources of informations (collected with the help of Urban Audit). The compilation has led to re-write in a common way the different rules used by each country to define its LUZ, i.e the building blocks, the links between these building blocks when LUZ is an aggregation of elementaty units, the evolution since UA 2001, some particular cases (for example Capital cities) and the degree of concordance with Gisco (that provides shape files).

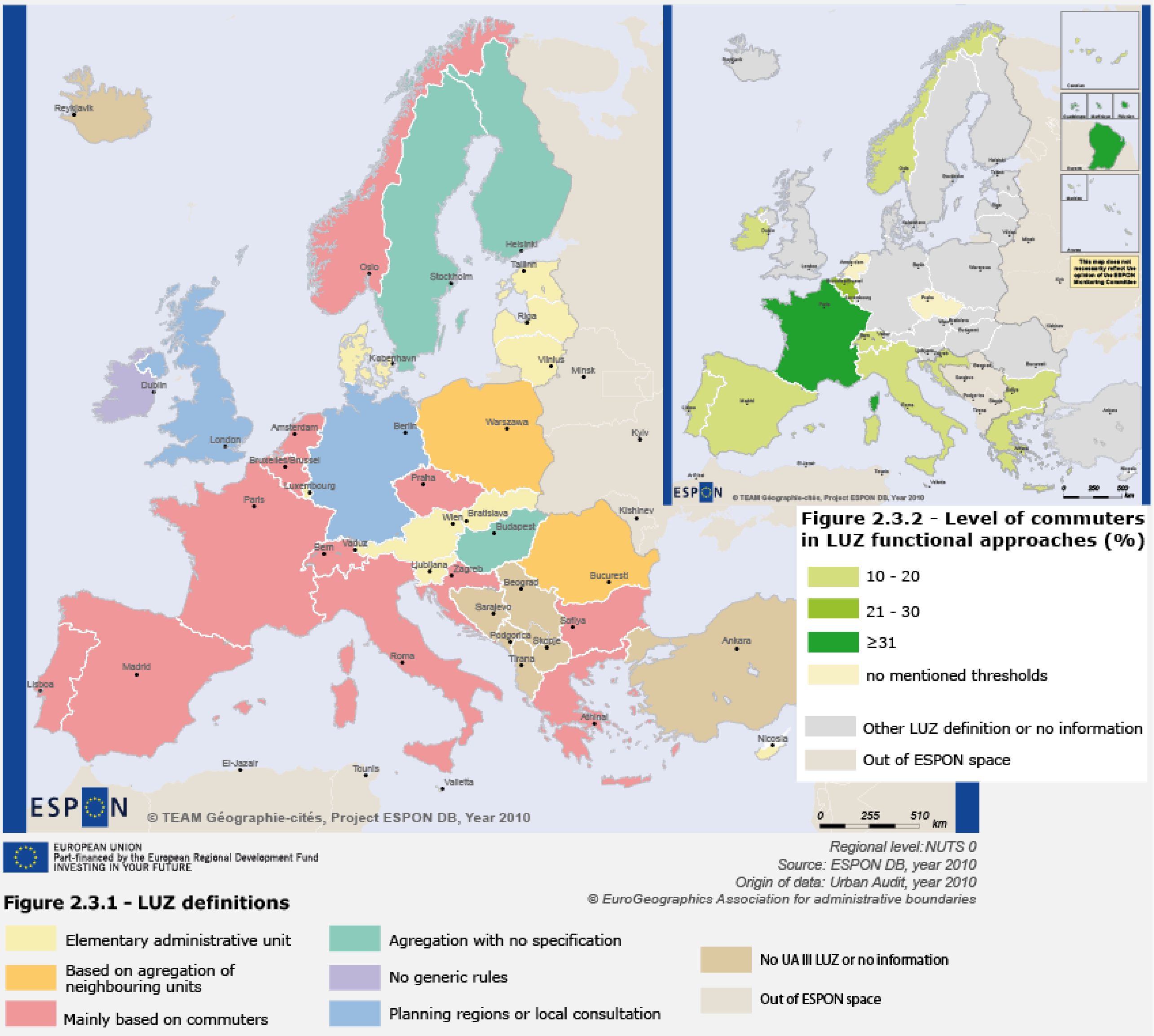
THEMATIC ISSUES

Without surprise, the results enlighten a very large heterogeneity in the national approaches used to define LUZ (Figure 2.3.1) and engage researchers to be very cautious when interpreting some statistical results. But it also enlightens a very interesting evolution between UA 2001 and UA 2004, towards more functional definitions, mainly based on commuters, even if the criteria are very different (see for example commuting thresholds, Figure 2.3.2). It confirms again that harmonization in definitions must not be only guided by the research of a unique rule and criteria for the whole Europe but must be based firstly on a good knowledge of the regional differences in settlement contexts and the political and historical ways each country defines cities. National differerences are not an obstacle to harmonization when the metadata are fully specified.

Related technical report : LUZ Specifications (Produced by Geographie-cités)

Figure 2.3.1: Typology of LUZ delineations (Urban Audit 2004)

Figure 2.3.2: Variety of commuting thresholds in LUZ functional delineations (Urban Audit 2004)



Regional level: NUTS 0
Source: ESPON DB, year 2010
Origin of data: Urban Audit, year 2010
© EuroGeographics Association for administrative boundaries

Figure 2.3.1 gives the result of the general typology of LUZ definitions in Europe, based on a compilation of national reports, which enlightens the diversity of national approaches in LUZ definitions. Figure 2.3.2 shows the variability in the choice of commuter thresholds for countries that explicitely mention this criteria in the rules of LUZ delineations. The map does not enlighten any gradient or regional structure, and statistical analysis did not reveal any correlation between commuting threshold and the average size of administrative units.

2.4 THE FUNCTIONAL URBAN AREAS DATABASE

KEY FINDINGS

- As a joint venture between 3 challenges of the Espon DB project (Urban data, Local data and Time series) and starting from the results of the previous Espon program we provide an update of the database of the Functional Urban Areas (FUAs) and Morphological Urban Areas MUAs), as well as their inter-relations. Not only is it enhanced, it is also fundamentally enriched by the quality of the data provided, as the Functional Urban Areas (FUAs) are now delineated for most of the European countries of the Espon space at the LAU2 level.
- The FUAs are defined as the labour basins of the MUAs which are defined as densely populated areas.
- The main quality and advantage of the FUAs are their simple and universal definition throughout Europe, making them comparable in all the countries where they were delineated.
- And we have also produced a list of indicators for these FUAs.

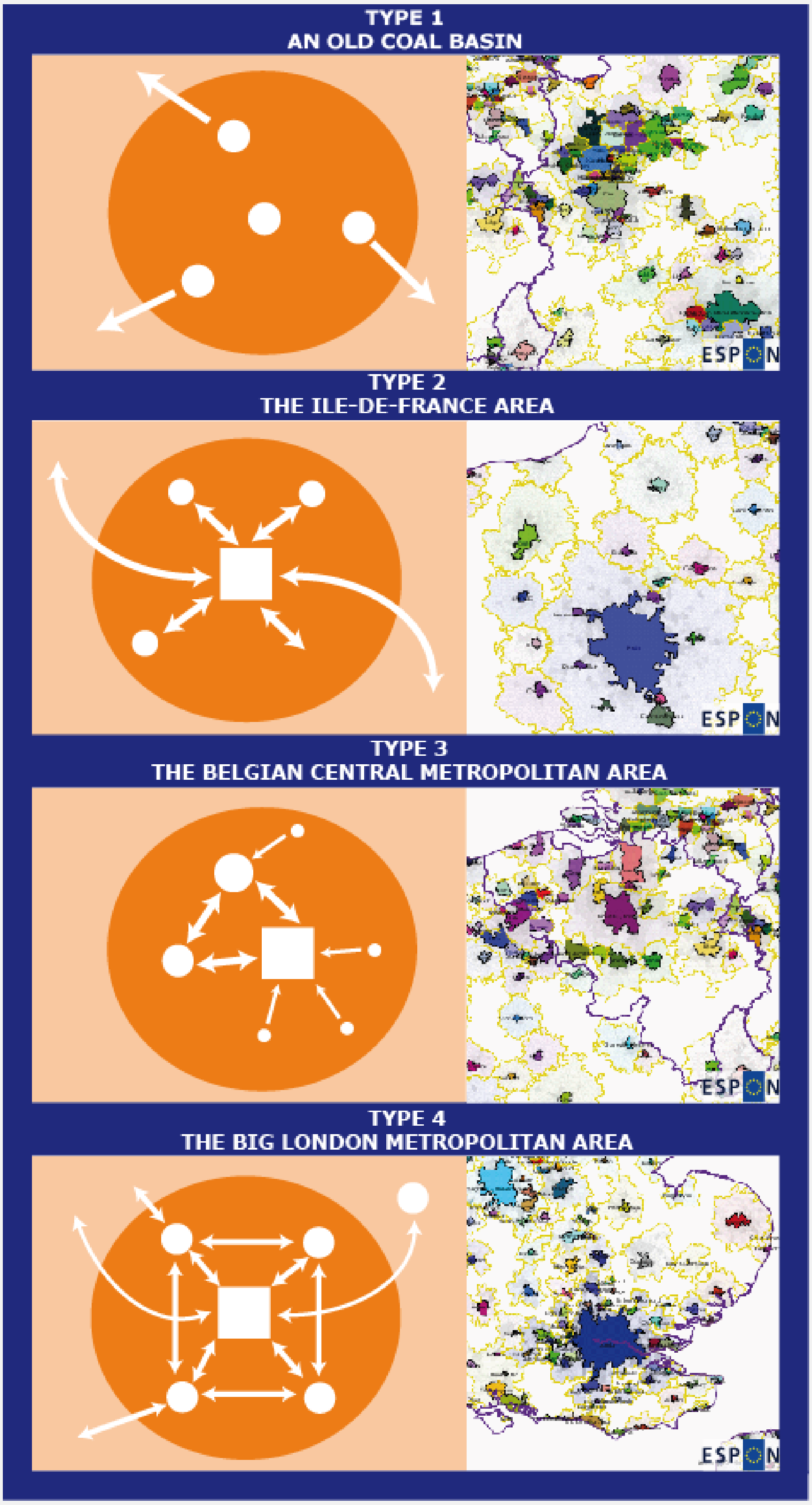
METHODOLOGICAL ISSUES

The MUAs are built by agglomerating the LAU2s having a population density higher than 650 inhabitants per km2. There can be one single LAU2 or hundreds of them. The MUAs kept in the list have a total population of at least 20 000, or actually was it made so by the Espon 1.4.3 project on Urban Functions in 2006.
We have transposed the LAU2s references from the old nuts-5 system from 1997 into the LAU2s of 2008, but without actualizing the population numbers that are still from 2001. The FUAs are the labour basins of the MUAs. They are obtained by agglomerating the contiguous LAU2s having 10 % of their «economically active population» working in the nearby MUA, so to say in the employment place of the neighborhood.
The LAU2s with a population higher than 20 000 but with a population density lower than 650 were also selected, as some - especially in the northern countries or in the south of Spain - and the LAU2s having a built area touching the built area of an already defined MUA where also added even with a lower population density, and finally the LAU2s hosting things such as airports, seaports, industrial plants «near» a MUA where also added to that MUA, so not to leave aside important employment places of cities.

THEMATIC ISSUES

The MUAs are thus independently defined from any cultural, political or administrative definition of the cities throughout the european countries, each having their own urban system. The MUAs have a simple definition which is coherent at the European scale and as densely populated areas they are considered as probable employment places, which will be verified (or not) in the next step, the building of the FUAs.

Different types of FUA structures and related geometries


TYPE 1
AN OLD COAL BASIN
TYPE 2
THE ILE-DE-FRANCE AREA
TYPE 3
THE BELGIAN CENTRAL METROPOLITAN AREA
TYPE 4
THE BIG LONDON METROPOLITAN AREA

The following diagrams summarize for instance four different situations in a high-density area, implying quite different realities as regards functions, economy, management of mobility and territorial planning, but which could be confused if the analysis did not sufficiently explicit the definitions used. Even if these four patterns are purely theoretical, they are respectively globally based on the situation of an old coal basin for the first one (type1), the Ile-de-France Region for the second (type 2, with new cities functionally not much independent from Paris), the Belgian central metropolitan area (type 3) and the big London metropolitan area (type 4), where secondary centers of the external fringe of the FUA have more decisional autonomy and are moreover doubled by a belt of important or specialized cities (cf. Cambridge, Oxford) inside the FUA.

2.5 SOCIAL/ENVIRONMENTAL DATA

KEY FINDINGS

- 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.

METHODOLOGICAL ISSUES

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 three integration methods. The "proportional and weighted" aggregation method is the one that gives better results, plus some added value to the downscaling. Thus, it is the recommended one:

Proportional and weighted calculation: the cell takes an area proportionally calculated value, and 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.

THEMATIC ISSUES

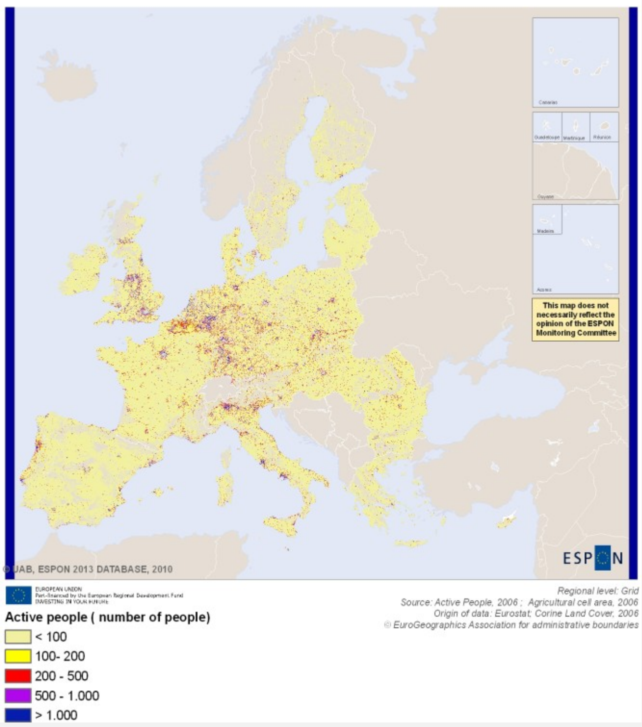
The methodology that has been defined under the Challenge 5 of the ESPON 2013 DB project provides useful tools to combine data provided by administrative units, such as NUTS 3 divisions, together with continous data, namely gridded variables. In the end, it allows the user to go back and forth from one type of data to the other one and viceversa, depending on the purpose of the analysis to be made.

The following maps are just an example to illustrate how gridded data can be used to report by NUTS 3, and how data originally reported by NUTS 3 can be reported by grid, giving an added value to the source data.

Related technical report "Disaggregation of socioeconomic data into a regular grid: Results of the methodology testing phase" (produced by the University Autonoma de Barcelona)

Disaggregation and aggregation of data

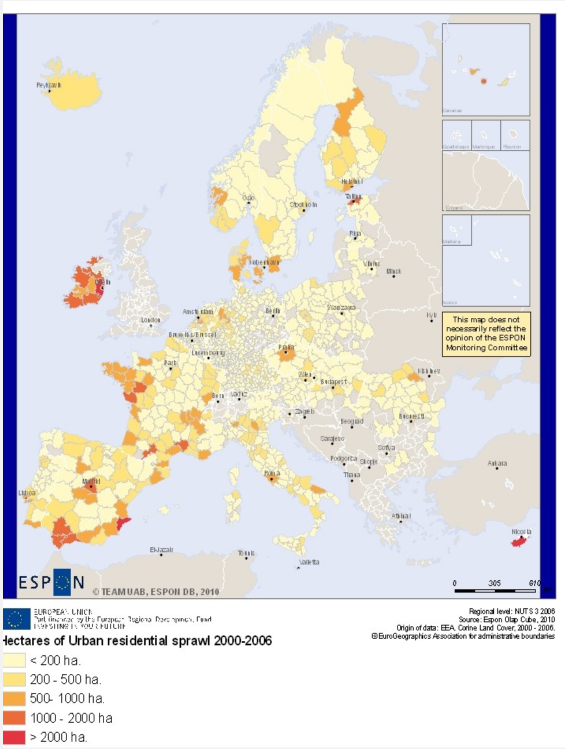
Figure 2.5.1 - Active people 2006 in agricultural grid cells (CLC 2006)



This map does not necessarily reflect the opinion of the ESPON Monitoring Committee

Regional level: Grid
Source: Active People, 2006 ; Agricultural cell area, 2006
Origin of data: Eurostat; Corine Land Cover, 2006
© EuroGeographics Association for administrative boundaries

Figure 2.5.2 - Urban residential sprawl 2000-2006



This map does not necessarily reflect the opinion of the ESPON Monitoring Committee

Regional level: NUTS 3 2006
Source: Espon Olap Cube, 2010
Origin of data: EEA. Corine Land Cover, 2000 - 2006.
© EuroGeographics Association for administrative boundaries

Figure 2.5.1 shows the distribution of active people by 1km grid cells with more than 50 ha. of agricultural landcover. This is an example of **disaggregation** of data by administrative units and the combination of such data with Corine Land Cover 2006 (level 1 agricultural class).

Figure 2.5.2 shows the urban residential sprawl process between 2000 and 2006. This map has been produced by means of the **aggregation** of Corine Land Cover Changes (as Land Cover Flows) by nuts3 regions using the ESPON Olap Cube v3.0.

2.6 INDIVIDUAL DATA ANDS SURVEYS

KEY FINDINGS

- Examination of downscaled population data shows that – despite obvious limitations – it is a quite reasonable tool for certain purposes. For instance, the JRC 1 ha population grid (partly based on CORINE land cover data) is associated with significant uncertainties at the very local level. Nevertheless, it generally provides fairly good estimations of e.g. the population of Urban Morphological Zones (UMZ).
- Eurostat surveys, such as the Labour Force Survey (LFS), are normally not presented with a high degree of spatial resolution, partly due to issues related to sample size and sampling error. In fact, published tables in general present data at NUTS levels 1 and 2 (and occasionally NUTS 3).
- By departing from the JRC population grid, and simultaneously utilizing information from many different survey tables, data sets with a higher geographical resolution can be produced. This can be achieved since the tables together contain more information about the joint distribution over space and attributes than do the single tables side by side.

METHODOLOGICAL ISSUES

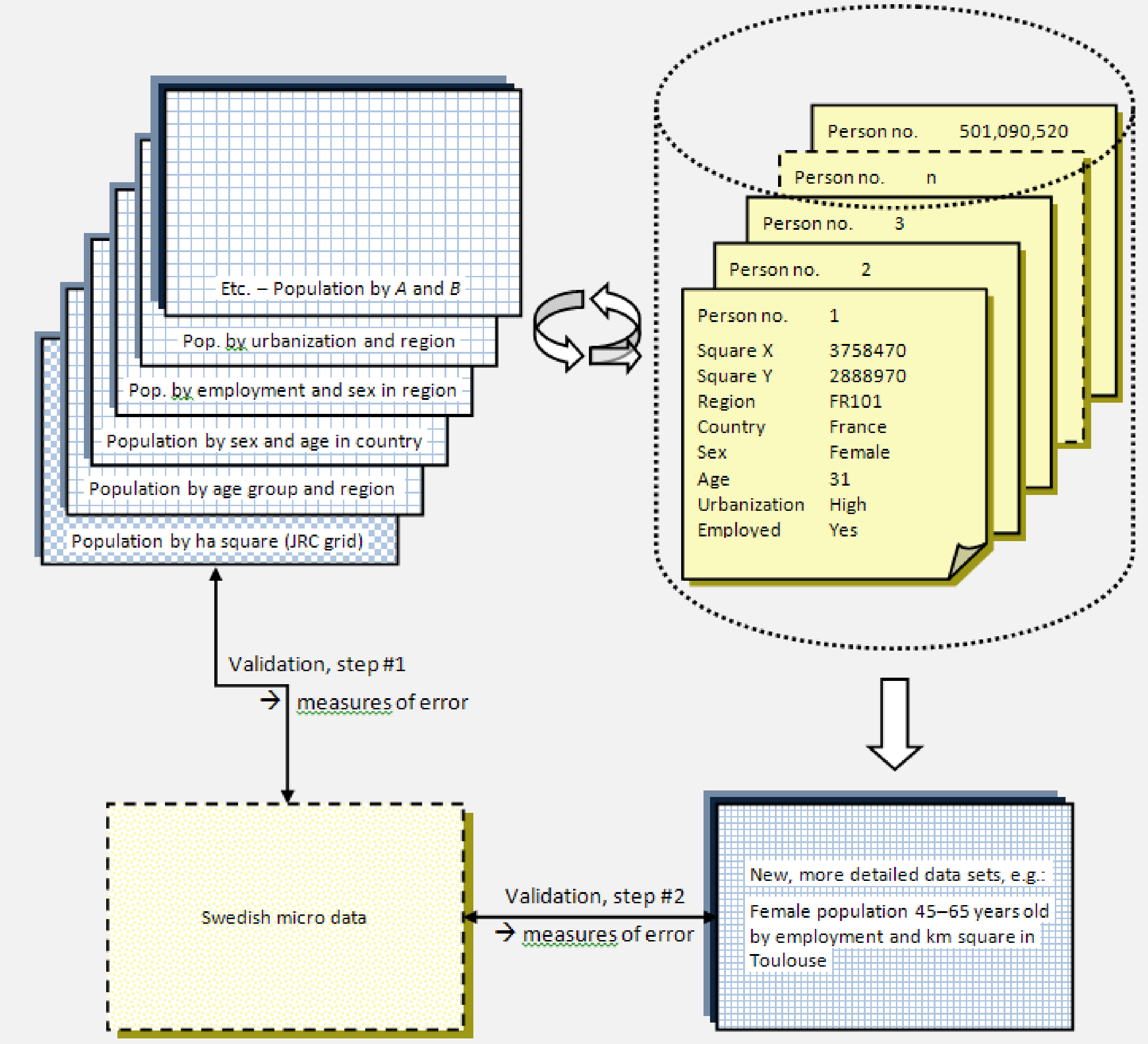
The first step is the creation of a synthetic individual database of the European population. At this stage, the attributes of the artificial individuals are assigned randomly, except being conditioned by totals for nations and regions. After that, in the main step of the procedure, the artificial individuals successively exchange attribute values with each other. Each exchange is performed as long as it improves the total fit between the observed table cells and the corresponding cells in tables based on aggregating the current artificial individuals. The end result is synthetic individuals that are jointly consistent with all information in the supplied tables.

THEMATIC ISSUES

The majority of the population is located in urban areas covering a tiny fraction of space, whereas a small share of the (aging) population is distributed over substantial parts of the countryside; exhibiting a shortage of qualified labour for emerging jobs. Information for analysis and counteraction related to such problems are effectively hidden by current aggregation levels in available data.

Related technical report : Using downscaled population in local data generation – a country level examination (produced by the Umea University)

New data sets from multiple tables via a synthetic population


Etc. – Population by *A* and *B*
Pop. by urbanization and region
Pop. by employment and sex in region
Population by sex and age in country
Population by age group and region
Population by ha square (JRC grid)
Person no. 501,090,520
Person no. n
Person no. 3
Person no. 2
Person no. 1
Square X 3758470
Square Y 2888970
Region FR101
Country France
Sex Female
Age 31
Urbanization High
Employed Yes
Validation, step #1
→ measures of error
Swedish micro data
Validation, step #2
→ measures of error
New, more detailed data sets, e.g.:
Female population 45–65 years old by employment and km square in Toulouse

The figure shows, schematically, how downscaled population is used together with different survey tables to produce a synthetic individual database – as well as new data sets. Despite shortcomings related to disaggregation, survey sampling errors and inconsistencies arising from the method of iterative attribute exchange, the end result represents an efficient way to increase the amount of local information using presently available data resources.

2.7 LOCAL DATA

KEY FINDINGS

- The use of the classical spatial patterns (points, surfaces, networks), mobilized at the local scale and managed with specific methodologies, allows us to provide basic indicators at local level (LAU2), for selected case studies.
- The construction of indicators at local scale using information that is based on some specific geographical objects (grid information or networks) could function as an option to the traditional data sources.
- The exploration of the main sources of data at this level (NSI) is an opportunity to design methods that are able to properly match indicators and geometry. The collection of several LAU2 codes is strongly needed, in this case.

METHODOLOGICAL ISSUES

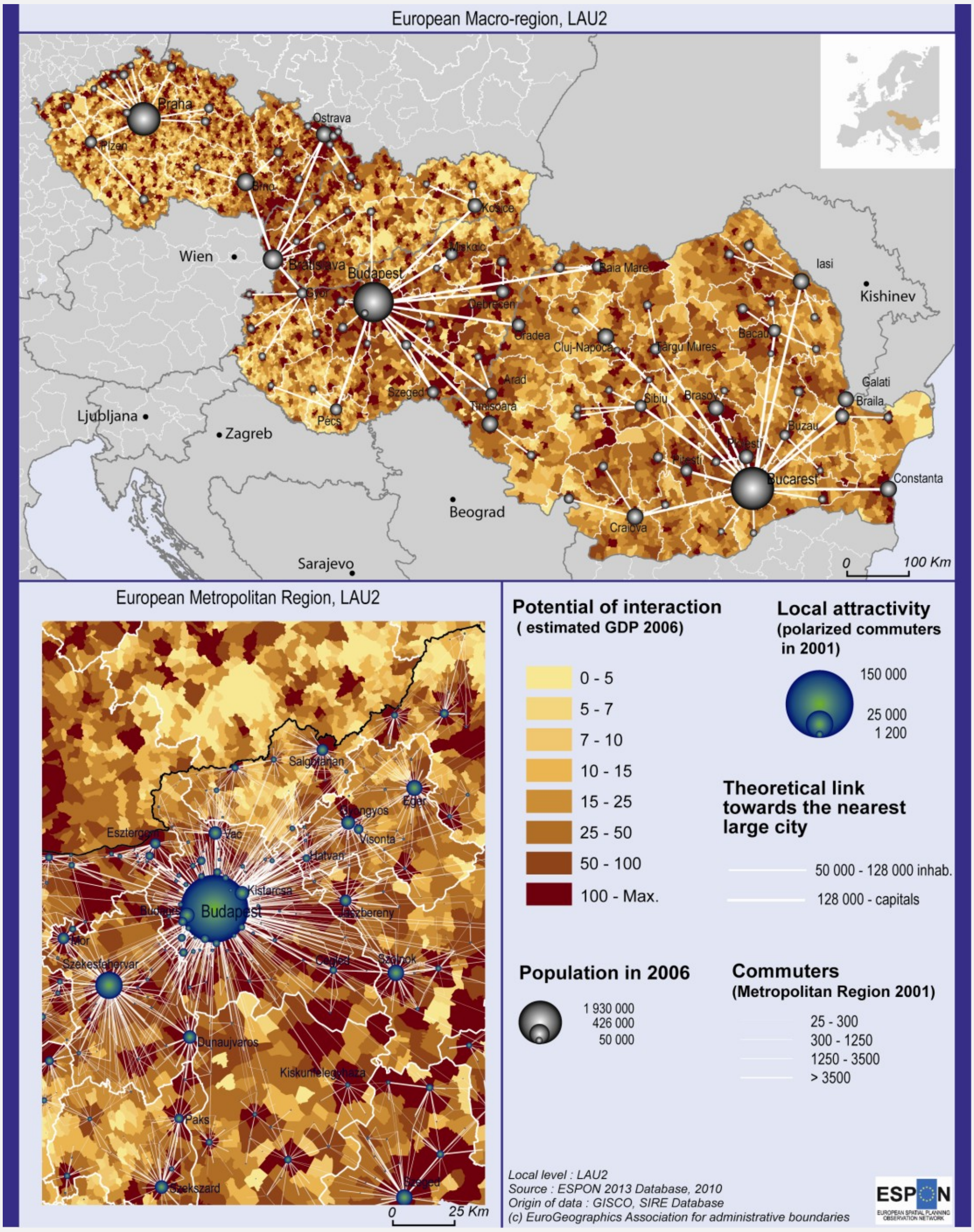
Using the LAU2 as scale of reference for data integration is a double challenge. As a first task, the variety and the quantity of geometries demand not only data mining and GIS techniques of manipulation, but also a good knowledge of the "terrain", sometimes in the classical sens of the expression. Secondly, the data production is time consuming, even when the implementation of simple models is needed. However, solutions and methods to overcome these difficulties can be figured (see Technical Report) and good practices in the local data manipulation are now available. Experimenting the methods is partially depending on the number of LAU2, but in some cases (potential model) is just a matter of time.

THEMATIC ISSUES

The intersection of the grid information with the LAU2 geometry can function as a method of data collection and indicators development. As the GDP is generally available only at NUTS level, reagregating data at LAU scale represent an alternative way to analyze the territorial economic performance. The grid data was provided by ESPON DB project (Challenge 5) and for the moment only values for 2006 were agregated in the LAU geometry. To eventually smooth the high contrast between the spatial units, an option was made for the potential of interaction, with constraints regarding the moving-window and the distance decay. One of the model's parameter was the road network density at LAU2 and the values for Bulgaria are missing, explaining its absence on the map.

Related Technical Report "Local Data" (produced by TIGRIS)

Economic performance and attractivity at local scale



Local level : LAU2
Source : ESPON 2013 Database, 2010
Origin of data : GISCO, SIRE Database
(c) EuroGeographics Association for administrative boundaries

Analyzed at the local scale, the economic performance is an archipel shaped by capitals, active frontiers, transportation corridors (Gyor-Budapest) and old industrial regions (Czech Silesia). In this territorial frame, the rural spaces are not so passive as one might expect, if we except the remoted areas or some specific cases (the Romanian eastern border). When zooming at a metropolitan scale, the same archipel model is this time articulated by a core-periphery gradient, complicated by the presence of privileged axis and secondary poles (Vac, Esztergom, Hatvan), the last ones linking Budapest to regional cities like Szekesfehervar, Szolnok or Szeged.

2.8 ENLARGEMENT TO NEIGHBOURHOOD

KEY FINDINGS

- Availability and quality of the data on the Candidate Countries (CC) and Potential Candidate Countries (PCC): the Western Balkans countries and Turkey, at NUTS2 level is in general terms satisfactory
- Data availability and quality on CC and PCC at NUTS3 level, which is the most challenging for the needs of ESPON is almost fully satisfactory for Croatia, FYROM and Turkey which have adopted the NUTS classification while it is satisfactory for a wide number of issues for the rest CC.

METHODOLOGICAL ISSUES

In order to ensure a sound comparability of data of the CC and PCC which have not adopted the NUTS classification, we have classified the existing administrative units of these countries at different territorial levels in "similar NUTS" territorial units. We have used for this purpose the criterion of population potential of the EU NUTS classification as well as the overall structure of government in these countries with focus on the power of the respective regional and local authorities and the main features of territorial development in each administrative level per country.

The implementation of this method ensured that the "similar NUTS" divisions correspond almost fully with the respective divisions for the EU countries and could be further used in the definition of "similar NUTS" divisions in the Eastern Neighbouring countries (ENC) and the Southern Mediterranean Neighbouring countries (MNC).

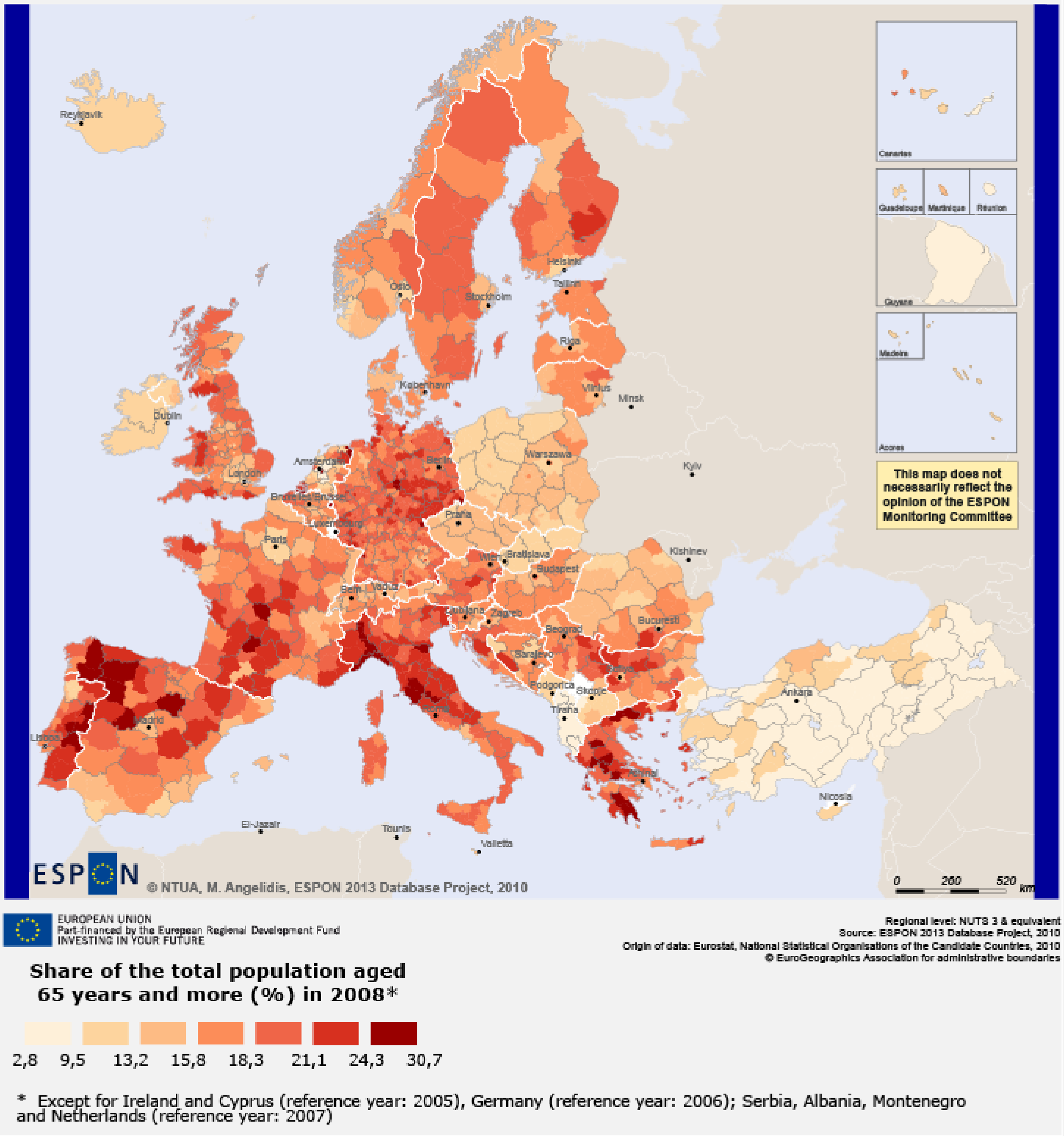
THEMATIC ISSUES

For the CC and PCC which have not adopted the NUTS classification, specific datasets and metadata at NUTS 0, 1, 2, 3 levels have been elaborated and included in the ESPON 2013 database. Main sources of data are the Official Statistical Institutes of the respective countries.

The further development of both formal and informal collaboration of ESPON with Eurostat, DG Regio Official and the Statistical Institutes of the respective countries could ensure a regular bilateral flow of territorial data for these countries.

Related technical report: "Analysis of the availability and the quality of data on Western Balkans and Turkey" (produced by the National Technical University of Athens).

Population aged above 65 years in the ESPON Area, Western Balkans and Turkey



© NTUA, M. Angelidis, ESPON 2013 Database Project, 2010

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

Regional level: NUTS 3 & equivalent
Source: ESPON 2013 Database Project, 2010
Origin of data: Eurostat, National Statistical Organisations of the Candidate Countries, 2010
© EuroGeographics Association for administrative boundaries

Share of the total population aged 65 years and more (%) in 2008*

* Except for Ireland and Cyprus (reference year: 2005), Germany (reference year: 2006); Serbia, Albania, Montenegro and Netherlands (reference year: 2007)

The Map shows the population ageing in the Western Balkans and Turkey as well as in the ESPON space, at NUTS3 and "similar NUTS3" levels in 2008, using the population aged 65 years and over rate % as an example of the **close of the CC / PCC "gap".** This step is very important because it allows study the territorial particularities of these countries which should be taken into account in the future Cohesion and Neighbourhood Policies of the EU.

2.9 WORLD/REGIONAL DATA

KEY FINDINGS

- Elaboration of a provisional ESPON 2013 World Database with "indicators of reference" (Population, GDP, CO2 emissions, …) describing "units of reference"(states or territories) on a long period of time (1960-2010)

- Comparison of the official list of countries and related codification (ISO3) from main international "thematic" providers (UNEP, CHELEM, World Bank, …). Elaboration of tables of correspondences between this database and the data previously collected by ESPON 2006 Europe in the World Project

- Linking of World data with Eurostat Regional data through a methodological tool (named "Gap Tracker") for explaining the differences between global databases and Eurostat data.

- Preparation of maps and graphics at global/regional scales in order to feed the first ESPON 2013 Synthesis Report

METHODOLOGICAL ISSUES

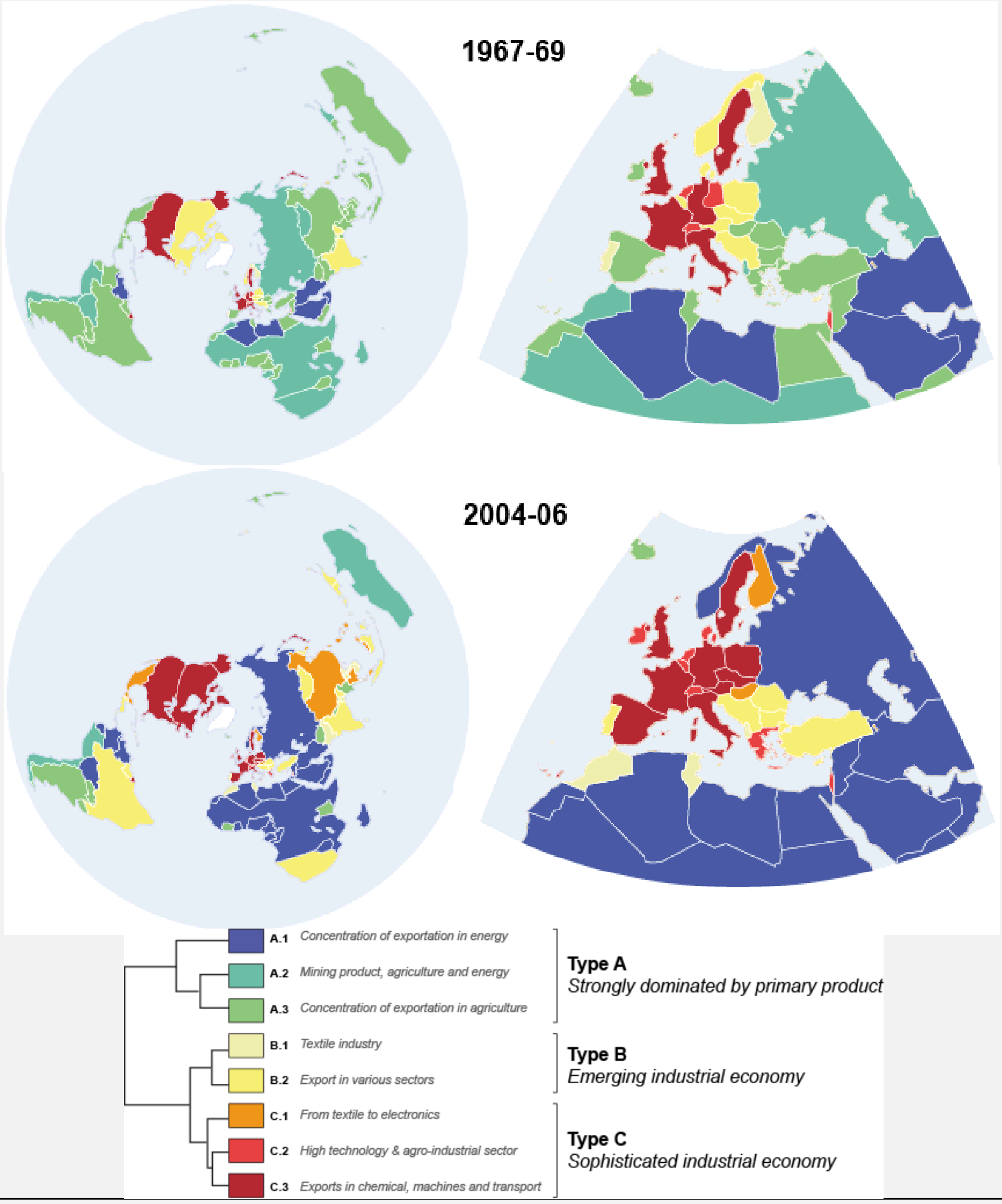
The ambition of the ESPON 2013 Program to support a "Five level approach" implies the elaboration of databases covering the "regional" scale (EU31+ southern and eastern neighborhoods) and the "global" scale (World) for different thematics (environment, demography, …) and different types of geographical objects (cities, countries, flows, …) at different periods of time (1960-2010). The main problem is not the collection of world data (more easy than NUTS data) but rather their internal harmonization (e.g. "France" can refer to various geographical objects, including or not the remote territories) and also the differences between World/European/National data providers (e.g. Population of UK in 2000 is very different according to UN, Eurostat or National Statistical Office).

THEMATIC ISSUES

The ESPON 2013 database project was not directly in charge of thematic exploration at world/regional scale but some experiments have been led in order (1) to evaluate how to map and store results of new projects like TIGER and (2) to support ESPON CU (Synthesis Report or SR). As an example, we shall mention an update of the map of discontinuities of GDP/inh in Europe in 2008 (SR, p. 57), a map of a global cities network in 2008 (SR, p. 32), a comparison of HDI and global footprint for world countries in 2006-2007 (SR, pp. 84-86), a typology of countries for trade export in a 40 year period (see Figure , right), and last but not least a complete illustration of the ESPON "Five level approach" for population growth 2001-2006 (SR, pp. 15-16).

Related technical report "World database – Towards a World Dictionary of unuits "
(produced by the University of Geneva and RIATE)

Typology of country profiles for trade export 1968 – 2005


1967-69
2004-06
A.1 Concentration of exportation in energy
A.2 Mining product, agriculture and energy
A.3 Concentration of exportation in agriculture
Type A
Strongly dominated by primary product
B.1 Textile industry
B.2 Export in various sectors
Type B
Emerging industrial economy
C.1 From textile to electronics
C.2 High technology & agro-industrial sector
C.3 Exports in chemical, machines and transport
Type C
Sophisticated industrial economy

At the beginning of the 21th century, we observe a very intense polarisation between the extended group of countries that export mainly energy or mineral ressources, and the group of industrialized countries which is more and more enlarged to new developping countries. The industrial cores of the 1960's are now clearly in competition with new industrial economy. In Asia as in America, the process of industrialization has clearly spread toward neighboring countries. The situation is fully different in EU, which is surrounded by countries with a lower level of industrialisation that export mainly energy or primary products, in both eastern (Russia, Central Asia) and southern (Africa) directions.

2.10 SPATIAL ANALYSIS FOR QUALITY CONTROL

KEY FINDINGS: DETECTING EXCEPTIONAL VALUES

- Exceptional values can arise from **(a)** data input or manipulation errors and **(b)** data values which are truly outlying (outliers).
- The accurate identification of an exceptional value is important because input errors should be treated differently to outliers.
- Input errors can usually be identified mathematically or sometimes, statistically. Outliers can only be identified statistically.
- A 'weight of evidence' approach to the statistical identification of outliers is proposed and presented in an accompanying technical report.

CONTEXT

It is paramount that the data to be stored in the ESPON 2013 Database should be as reliable as possible. It would be unwise to assume that data supplied to the Database is completely free from error. The activities under spatial analysis for quality control have led to the design and implementation of a battery of filters and tests against which potential input data can be tested. Exceptional values can arise for a number of reasons, so rather than employ a single test, several tests are applied. For outliers, this allows a judgement to be made as to the reliability of a sample observation based on the **weight of evidence** from an application of nine tests.

Exceptional values may arise during the coding, transmission, manipulation or editing of data. They may not be noticed until late in the day or not at all, particularly if the data that is a candidate for input to the Database has been the outcome from a series of sequential manipulations. Exceptional values may also arise in the measurement of data; perhaps a sensor on some measuring equipment has been incorrectly calibrated or is faulty; a respondent can tick the wrong box in completing a survey form. Exceptional values can also be true and valid observations.

DESCRIPTION

Logical input errors made during data input or manipulation in the ESPON 2013 Database can arise for a number of reasons. For example: incorrect NUTS codes might be entered or assigned in a table lookup; incorrect data values could be input; data could be repeated exactly but assigned to different variables; data could be displaced within or between columns; data could be swapped within or between columns. In general, the process leading to the identification of an input error will follow some logical, mathematical approach that can be conveniently coded within the database architecture.

Related technical report : Spatial Analysis for quality control 1 & 2. (Produced by the National Center for Geocomputation)

For example, if a land use class could only take a positive integer value from 1 to 9 say, then an incorrect value of say, -2, 4.5 or 10 would be easily identified. An input error may also be identified statistically. For example, if the number 27 is inadvertently entered as 72 for a region's unemployment rate, the value 72 may lie in the extreme tail of the distribution of values for the variable and as such is statistically-outlying. A difficulty here would be to distinguish between an input error of 72 and a true value of 72.

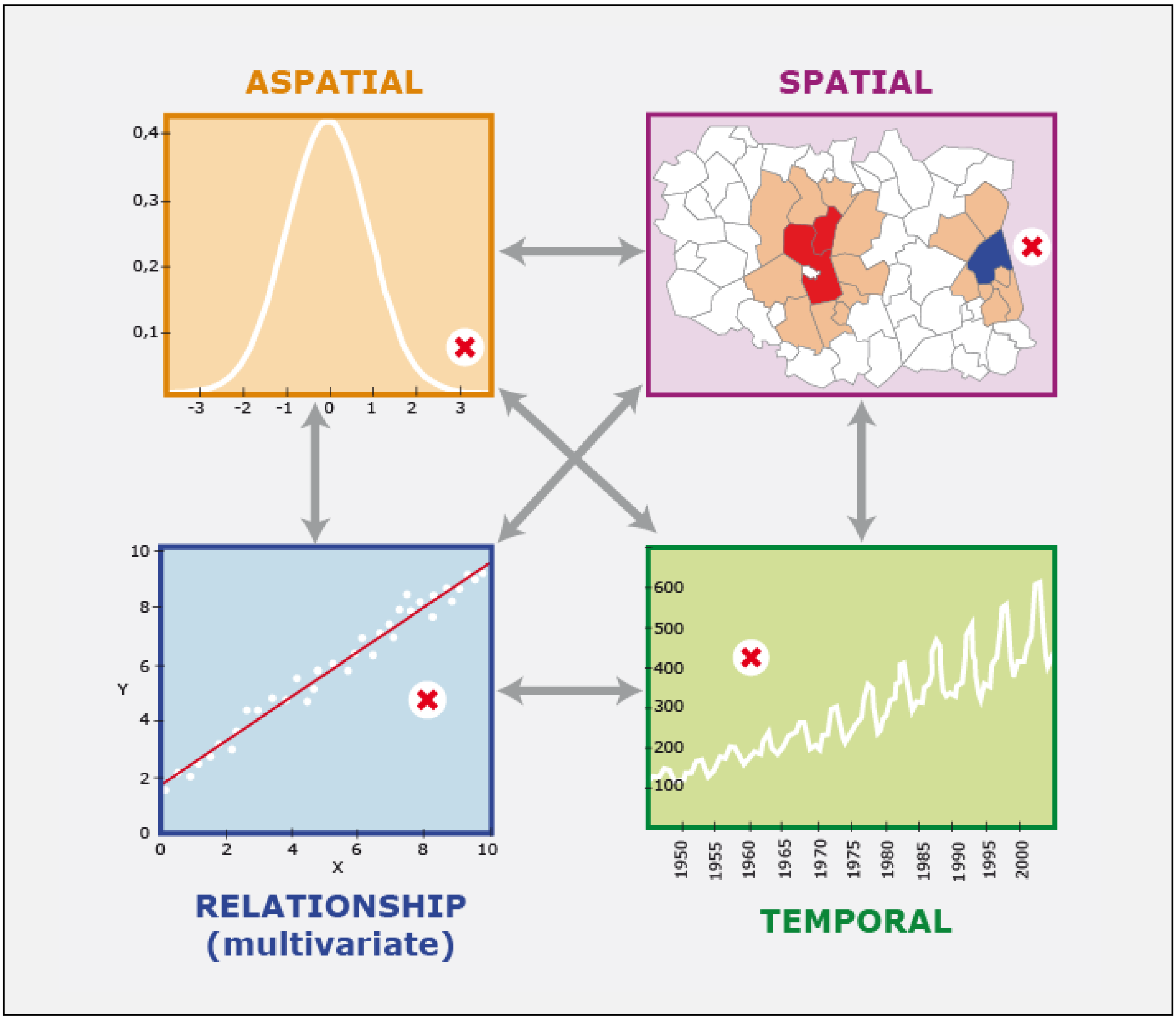
Statistical outliers can similarly arise for a number of reasons in the ESPON 2013 Database. In the accompanying technical report, we propose a novel 'weight of evidence' approach to the detection of outlying values, which has been developed since the Second Interim Report. As demonstration of this approach, it is applied to a real ESPON data set. The approach applies nine representative and complementary statistical outlier detection tests, where observations are flagged as outlying according the outcome of each test. This then builds up a 'weight of evidence' for the likelihood of a given observation being outlying (i.e. the evidence is strongest for an observation that is flagged as outlying for all nine tests).

The nine different tests deal with the identification of outliers with different characteristics which renders the value unusual. These characteristics are:

- **Aspatially outlying:** an unusually low or high value (i.e. an outlier in or near the tail of a statistical distribution).
- **Spatially outlying:** an unusually low or high value in a region when compared with values in neighbouring regions.
- **Temporally outlying:** an unusually low or high value with respect to a time series of observations.
- **Relationally outlying:** a pair (or group) of observations that relate to each other in an unusual manner. That is, the relationships are very different from what has been observed elsewhere in the database. For example, a particular region may have a high unemployment rate coinciding with a high GDP value (i.e. indicating a positive correlation) whereas we would normally expect high unemployment rates to coincide with low GDP values (i.e. a negative correlation). The statistical methods used to detect these kind of unusual relationships are multivariate, whereas in the previous three cases, univariate statistical detection methods are used.

These four major forms of outliers are schematically depicted in the figure below. Note that an observation may be outlying in more than one way; for example, it may be both aspatially and temporally outlying.

Combinated methods for detecting outliers



Individual and combination of variables in the database are evaluated against the nine detection tests. An observation that is found to be unusual on only one or two tests is considered less likely to b e outlying than an observation which is found to be unusual on all nine tests.

The decision of what action to take with regard to an identified logical input error or a statistical outlier (e.g. remove, replace, leave alone) should ultimately reside with an expert on the given data set (or its provider). Mechanisms can be put in place within the database architecture to do this in an efficient and effective manner.

For input errors, a simple removal and (if possible) a replacement will generally suffice. For outliers, considerably more attention is required. Outliers and their detection should not be naively viewed as a data cleaning or screening exercise, but can also be used to uncover interesting or unusual relationships in the database that has not been considered before.

CONCLUSION

In this conclusion, we summarize briefly the most important progress made during the project (section 1) and the most important difficulties encountered (section 2) before to address some proposals to our followers (section 3)

1. Progress made on ESPON Database during the projects

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 is now available on the ESPON 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).

2. Difficulties that have been encountered

The first set of difficulties that we have faced within 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 map kit 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 the difficulties (and the cost) of communication and networking within the ESPON Program in general and with external organization like EUROSTAT or DG REGIO. We had not anticipated the importance of this topic which revealed to be crucial but implies a lot of time of communication, explanation or discussion with a lot of actors. As an example, the ESPON CU asked to the ESPON Database Project to perform a manual data check of data delivered by Priority 1 or Priority 2 projects, which was a very time consuming task, to be done with strong time pressure (as it was a condition of payment for this projects). As a second example, it was also difficult to have regular contact with EUROSTAT or EEA because such a meeting should be jointly organized with ESPON Coordination Unit and not directly by ESPON Database Project.

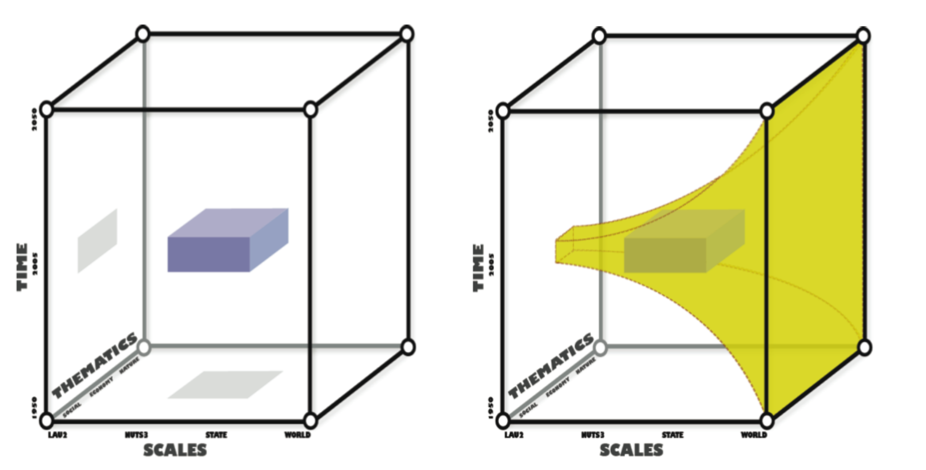
The third set of difficulties has been related to the excess of financial control. We know that the rules of the ESPON program are what they are until 2013. But we also know that the European Commission has insisted in 2008, after the crisis, on the necessity to make the rules of control easier and avoid unnecessary administrative burdens. Our feeling as coordinators is that many times there was a danger of blocking the achievement of the ESPON DB 2013 Project. More and more work time, normally devoted to the productive part of the project, was in practice transferred to the management of administrative burdens related to "every-six-month-reports".

The fourth and final set of difficulties has been related to the lack of Knowledge and Support System. We were very frustrated to observe that our project did not benefit from a Knowledge Support System as the other major project of Priority 1. It was not necessarily because of the size of our project (more than all other Priority 1 project) but rather because of its strategic importance for ESPON that a KSS should have been established, with the best specialists of the domain (*e.g. G. Andrienko, M. Goodchild, …)* as a form of scientific recognition of the quality of the ESPON Database in the fields of computer science, cartography, geomatics.

3. Recommandation for the future

Based on the experience gained during the period 2008-2010, we suggest some recommendations to the followers of our project.

A) Maintain the ambition of enlarging database dimensions



Initial situation of ESPON 2006

Objectives for ESPON 2013

This objective remains fully accurate and a lot of work has still to be done in order to support innovative applied research on territorial cohesion.

B) Reinforce the networking dimension inside and outside ESPON

As explained in previous section, the ESPON Database project is necessarily connected to all other ESPON projects of Priority 1, 2 or 3 during all their lifecycle. In the initial stage, ESPON database provide new projects with data, map kits, technical reports … In the final stage, ESPON database receive data elaborated by this projects and check it before integration in the database. All of this implies a strong cooperation and more contacts than the opportunities offered by the ESPON seminars every 6 months.

Outside ESPON, it is of crucial importance to have more regular communications with data providers at European, national or global levels. It is also important to be in contact with the scientific community working on advanced innovations in GIS, Cartography, Data Modeling, …

C) Develop joint methods for automatic outlier detection and missing value estimation

One of the most important discoveries made during the ESPON 2013 Database project is the fact that estimation of missing values and detection of outliers are not two separated problems but a single one. To be sure, when you declare that a statistical value X is an "outlier", it is necessarily because you have an implicit model of estimation of X that indicates you that the model estimation X* is dramatically different from the observed value X. Therefore, all methods of outlier detection are also, by definition, methods of estimation for missing values. The four methods elaborated by the NCG team for outlier detection (see. 2.10) are in practice equivalent to the four dimension of the ESTI model proposed by LIG, RIATE and Géographie-cités for estimation of missing value and used in the challenge of time series (see. 2.1). At present time, we have mostly used simple method based on only one of the four possible dimensions (statistics, space, time, thematic) but better estimation of missing values or better detection of outlier can be expected by multidimensional methods.

D) Quality rather than Quantity

Repeating one more time the motto of our project, we would like to underline that many "big" databases has disappeared like dinosaurs because they did not make sufficient effort on the quality side and more precisely on the question of metadata. It is certainly true that the pressure made by our ESPON 2013 database project on the other ESPON projects for a very strict check of data and codification of metadata delivered (see. 1.2 and 1.3) has possibly limited the number of data they decided to deliver with their final report… But we fully assume this Malthusian perspective on data collection because it is the only sustainable strategy in the long run. Especially in the case of an applied research project where data can be used for political decision and should be fully subject to control of sources, estimation made, etc.

E) An open ESPON database

Better than storing a lot of Gigabytes, ESPON Database should offer a limited number of original and very high quality data, that would offer him a specific place in the European and world network of data producers. This specialisation should be balanced by the opening of countinuous data flows and exchanges with Eurostat, EEA, Eurogeographics, … but also OECD, UNEP, UNPP and more generally all National Statistical Institute of the ESPON 31 area and the neighboring countries.