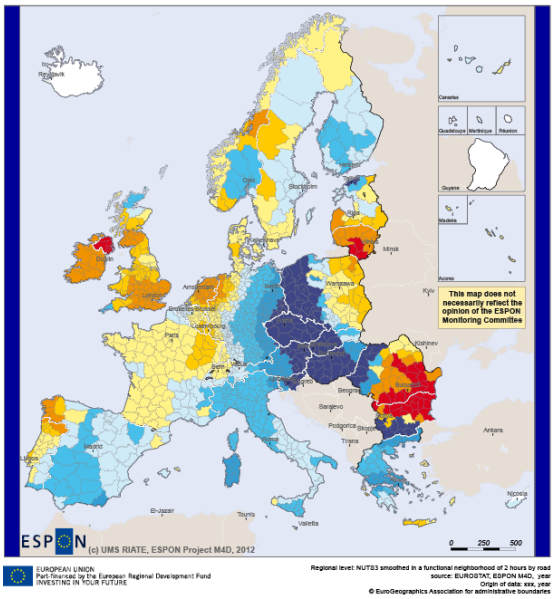
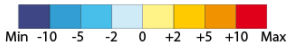
ESPON M4D
Multi-dimensional Data Design and Development

FIRST INTERIM REPORT
REVISED VERSION

21st December 2012

Local convergence of GDP/capita in 2 hours functional neighbourhood between 1999-2000 and 2007-2008





Difference between levels of sigma convergence of GDP/inh between 1999-2000 and 2007-2008

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

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

The partnership behind the ESPON Programme consists of the EU Commission and the Member States of the EU25, plus Norway, Switzerland, Iceland and Liechtenstein. 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 on-going ESPON projects.

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

List of contributors to the FIR

UMS RIATE (FR)

Claude Grasland

Ronan Ysebaert

Isabelle Salmon

Nicolas Lambert

Timothée Giraud

NCG (IE)

Martin Charlton

Alberto Caimo

LIG-IMAG (FR)

Jérôme Gensel

Benoit Le Rubrus

Anton Telechev

Marlène Villanova-Oliver

UAB (ES)

Roger Milego

Maria-José Ramos

UMR Géographie-cités (FR)

Anne Bretagnolle

Hélène Mathian

Marianne Guérois

Antonin Pavard

TIGRIS (RO)

Octavian Groza

Alexandru Rusu

Daniel Tudora

Aurelian Nicolae Roman

TABLE OF CONTENT

Annexes

Annex 1: Linkage between Core Indicators and the Regional Database

Annex 2: Guidance on how to deal with the two NUTS divisions in practice

Annex 3: Linkage between Core Indicators and the Regional Database

Annex 4: Guidance on how to deal with the two NUTS divisions in practice

Annex 5: Clarification on the work envisaged on cities and urban data

Annex 6: Clarification on the Statistical section

List of figures

Introduction

After the Aalborg meeting and with the delivery of the first interim report, the members of the ESPON M4D projects are very tired but also proud of the work done during the last year.

Firstly tired, because the pressure exerted on the project team since the Cracow meeting has been extremely high, for both accidental and structural reasons. Concerning accidental reason, two important factors has contributed to the difficulties encountered by our project in 2011. The first one was the departure of the leader of NCG team that created delays in the delivery of tasks by this partner and reallocation of work toward other teams (RIATE and LIG) which where at the same time under pressure for the realisation of their own contributions. This shock was difficult to absorb and explain why the results presented in Cracow meeting were not as important as we had planned initially. The second accidental factor was the delay in the allocation of ESPON project on Territorial Monitoring which was planned to start at the same time than ESPON M4D but will finally not be allocated before the autumn 2012. Many of the questions addressed by the ESPON MC to the ESPON M4D project (and further relayed by ESPON CU) could have been more easily solved through a partnership with the Monitoring project, especially when it comes to the production of an interface of access to data for policy makers. We have done our best and spend a lot of time and human resource to fulfil these requests of easier access to ESPON database for policy makers. But we are still not convince that it was the responsibility for M4D project to achieve alone such a task which will be much better fulfilled when the Monitoring project will deliver a specific tool implemented as a kind of "plug-in" on the ESPON database. More generally (and we are now speaking from structural and not accidental facts), we consider that the ESPON data portal should be considered as an integrated tool linking efficiently the contribution of all priority 3 projects together. The example of OECD has been regularly presented as a model by ESPON MC but it is precisely a case where two different applications are combined: (1) a statistical data explorer which is classical and finally not very ergonomic for policy makers as well as for researchers and (2) a data explorer which has been developed separately by a team of specialists and is plugged on the database. If we transpose the example to ESPON, the real question in our opinion is not to add a "policy button"[1] to the statistical data explorer but to accelerate the development of the web mapping tool project (why was it launched so late?), to connect the HyperAtlas to the database, to create a specific area for territorial monitoring derived from INTERCO selection and based on the work of the future monitoring project.

Secondly proud, because despite the above mentioned difficulties we have been able not only to absorb the supplementary work induced by the accidents of 2011 but also to develop the main objectives of the ESPON M4D project and to present many innovative results at the Aalborg meeting in three different sessions and the following week at Eurostat in a two days workshop. In both case, we have register very positive reactions from other ESPON TPG's as well as from EUROSTAT's contact. We will not develop here all the reasons of satisfaction but simply point what are according to us the major achievements of M4D that can contribute to the external promotion of the ESPON program as a whole:

[1] The addition of a select button "*I am rather a policy maker/ I am rather a scientist/ I' am both*" on the front page of the new interface of the ESPON database has been decided by the ESPON CU against the opinion of ESPON M4D team which had obey to the commitment but consider still that it is not a good idea. Many interfaces try to adapt to user's profile, but without such a kind of categorical declaration. It is for example not the case on OECD which is presented as a reference model to be followed …

- *The elaboration of a new metadata model that is an important complement to INSPIRE in the field of urban and regional social and economic data.*
- *The creation of a database that is not limited to regional data but can also store new nomenclatures (i.e. new types of geographical objects like cities of different types, grids, networks, flows, …)*
- *The elaboration of an efficient process of data integration and selection which insure a maximum level of quality without creating too much burden for delivery of data by projects.*
- *The development of multiple channel of information on data with newsletter, portal, FAQ, technical reports, …*
- *The introduction of multiple way of request of data by projects, thematics, key-words and policies.*
- *The development of longer time series for selected indicators and the possibility to transfer this selected data to various geometries through the pivot of an OLAP cube.*
- *The synthesis of the various options for the delimitation of urban objects (morphological or functional) and the further elaboration of correspondence tables.*
- *The elaboration of new indicators at LAU2 level covering all ESPON territories, derived from CLC or transportation network.*
- *The proposal of new indicators based on functional neighbourhood that are more in line with the perception of local actors and more accurate for the evaluation of local effects of EU policies (convergence).*

This list is not limitative but provides a general idea of the importance of the work done by the ESPON project M4D since the inception report.

For the next period, our hope is to be less subject to accidental pressure in order to have more time for the support of other ESPON Projects and ESPON CU.

Working Package A – Application

First results

This section reports on the computer sciences issues that have been coordinated by the LIG STeamer partner for the first year of the M4D project. Referring to Figure 4 of the Inception Report (page 14), "*Division of the project in activities and work packages*", these computer and application issues match the activities of the Working Package A.1 (Database, Interface), A.4 (Updates), and C.1 (Platform). After a first section covering a general overview of the ins and outs regarding these issues, and presenting main encountered difficulties, two sections give an account, respectively, of:

- The accomplished tasks, detailing the progress for each working package.
- The remaining tasks.

Overview

One major key point of our objectives was to respond to the remarks and expectations of the ESPON Monitoring Committee (MC) and Coordination Unit (CU) concerning the user-friendliness and easy usage of the ESPON Database Interface.

As presented in the Inception Report version 1 in July 2011, and revised version 2 in February 2012, efficiently responding to this major issue required a consequent study and review of the existing tools delivered at the end of the ESPON Database Phase 1. Consequently, this first year of the project was especially devoted to focus on the underlying input data workflow processes, and to re-design the schema of the database. These two preliminary tasks were necessary and indispensable in order to support, on the one hand, a much better quality of data (content of the database), on the other hand, the attractiveness, easy and efficient retrieval of the data thanks to the output interface.

Thus, the targeted goal of a user-friendlier interface is about to be achieved, especially thanks to major improvements in the input data process phase. Also referenced as the "Platform" functionality in the Inception Report, the M4D Partners have designed, defined and set up the input data workflow, including a paramount specification for the expected inputs. This workflow and management of the input data is precisely described in the "**How To Deliver My Data**" document, attached as an appendix to this First Interim Report (FIR). Targeted to ESPON projects, this document clearly explains what the M4D Partners expect from TPGs, depending on the type of their projects: Priority 1 (Key Indicators) projects, Priority 2 (Case Studies, aka Zoom-In), and all the other type of data called into the document the "Background data".

Besides the data integration, the most prominent work of this first year is certainly the delivery of the **Data and Metadata Specifications**. Indeed, scarcity of data documentation within the ESPON Program Phase 1 has been seen as an important impediment to the building and use of the ESPON database. Difficulties stemmed from uncertainties about legal constraints, sources, units, etc. Moreover, in the Phase 2, the ESPON database aims at being enriched and expanded in the time, spatial and thematic

dimensions. Information about the data made available is thus even more crucial. Building a rich database would have been useless without a strong effort to describe precisely the data that have been gathered and integrated into the database. Creating and organizing metadata is therefore an additional, important, and transversal challenge for the ESPON community.

To be useful for ESPON projects and other end-users, data must always be accompanied by metadata, including information about their quality and sources. It is also particularly important that the metadata are created in a manner that is consistent with international (ISO) and European (INSPIRE) standards, so as to ensure the use of the database in the longer-run, and to make it compatible with other national and international database initiatives.

As a leading edge improvement in this beginning of the Phase 2, "*The Data and Metadata Specifications"* document availability, attached as a FIR deliverable, represents a key evolution regarding the socio-economic indicators description.

Additionally, consecutively to this reference document, Application Programming Interfaces (API) have been designed and implemented to check the syntactic consistency of the delivered input data received from ESPON Projects. This dataset syntactic checker has been integrated into the Web Interface Application, now entitled ESPON Database Portal. Restricted to registered users, this "Platform" functionality has been merged within this new Web Application. ESPON Projects can now upload and on-line check the correctness of their data.

The **ESPON Nomenclature Support** specification is another (new) indispensable reference document novelty. This document, also a FIR deliverable, precisely describes how to define a new nomenclature based on spatial units. Indeed, as mentioned previously, during this Phase 2, we are aiming to extend the support of different geographical objects in the database. To the existing library of NUTS units of all public official versions will be added the libraries covering the nomenclatures like WUTS (World Unified Territorial System), UMZ (Urban Morphological Zones), LUZ (Larger Urban Zones), FUA (Functional Urban Areas), LAU (Local Administrative Units) and, probably, some others. Several datasets, already created during the first phase of ESPON Database project, use statistical units of these nomenclatures, but do not provide information about these units. Before the integration of such datasets, the database needs to know some details about the statistical units used, their hierarchy and relationship with other nomenclatures. This information, collected into the ESPON database, allows afterwards integrating any number of new datasets using the dedicated statistical units library. To build the library of statistical units, the automatic analyser created for this purpose reads the data on a nomenclature of statistical units, checks for its validity and coherence and then puts the analysed information into the ESPON database. In summary, targeted at ESPON Projects, this « Nomenclature Support » document describes the structure of the package and the layout to use in order to prepare a new nomenclature (or a new version of an existing nomenclature) for the integration into the ESPON Database.

Besides these preliminary requirements for the input process, taking into account the high level of priority regarding the availability of data by both "all-users" and ESPON projects, we have designed a **complete new interface**. Actually, the design of the output interface passed through two steps. One first proposition was presented during the ESPON Internal Seminar in Krakow, in November 2011. Following several studies, discussions among M4D Partners, brainstorming, and meetings with the CU, we have written the **Web Interface Specifications** document (FIR deliverable). This document aims at proposing a description of the main concepts and functionalities of the new

interface. Based on these specifications, validated by the CU at the end of March 2012, we could start implementing this new interface. A first version is delivered at the end of June 2012.

Regarding the encountered difficulties during this first year of the project, the main issue was to manage the delivery of an improved output Web interface in a short implementation period (less than three months), while its requested user-friendliness and easy use widely depended on the hidden part of the iceberg, e.g. the data input management.

Not only the new interface may now respond to different profiles of users, but it has also been redesigned to support the integration of input data. Merging the "Platform" functionalities into the Web Portal allows now ESPON CU, TPGs, and M4D involved partners, to monitor the input data integration process. In particular, the "Upload" section of the ESPON Database Portal (requires login), has been designed so to allow an easy monitoring of the projects integration process ("Tracking" page), despites, as described in the "Web Interface Specifications" and "How To Deliver My Data?" documents, the complexity of this workflow.

Lately, besides the much easier query search possibilities (depending on the user profile, by Theme, by Policy, by Project or by Keyword), the application has been designed for an easier access to the available resources (ESPON Map Kits or OLAP Cube Data Tool page, for example). The "Help" menu now proposes an access to useful documents:

- User Manual,
- Data and Metadata Specification
- Nomenclature Support (aka Dictionary of Spatial Units)
- "How To Deliver My Data?" document

As a conclusion to this overview section, the first year of the M4D project regarding the computer sciences issues has mainly consisted in laying down the foundations in terms of design and specifications for an innovative spatial data infrastructure: the ESPON Database Portal interface has begun to be implemented to reach the targeted objectives of user-friendliness and attractiveness. Based on these solid foundations and better-defined objectives, its development can now be completed to achieve the expected full set of expected functionalities.

WP A1 Computer

WPA1 Computer – Database

Regarding the Database, the main work accomplished during this first year of the project has consisted in improving the design and the quality of the database model for the following key points:

- The INSPIRE directive compliance.
- The support of different geographical objects in the database.
- The availability of several classifications of indicators (policy classification for example)
- The ESPON Priority 1 Projects Integration workflow.
- The ESPON Priority 2 Projects (Case Studies) Integration workflow.
- The query possibilities from the Web interface.

In order to reach these objectives and to insure a high quality of content of the database, this first year was devoted to the study, the design and the specifications of needs.
Main results of these studies are:

- Reference documents for ESPON Projects delivering data:
  - The **Data and Metadata Specifications**. This document aims at responding to the metadata challenge, described below.
  - The **ESPON Nomenclature Support**: this document precisely describes how to define a new nomenclature based on spatial units.

- An integrated **Web Portal** for the data integration (work in progress). Projects and practitioners are invited to login the Web Application, to upload their data and to follow the steps of its integration ("Tracking" functionality). Please consult the "How To Deliver My Data" document, attached as an appendix to this FIR, for further information about the data integration workflow process.

In order to achieve the objectives cited above, a deep refactoring has been made on the entire model of the ESPON database, resulting in producing a new version that may be characterized as more flexible, more robust and optimized than the database of the previous phase of the project.

WPA1 Computer – Interface

Taking into account the remarks and recommendations of the ESPON Monitoring Committee (ESPON MC) and of the ESPON Coordination Unit (ESPON CU), a complete review of the design of the Web interface has been achieved. In summary, the concept of the new interface aims at improving the attractiveness and general user-friendliness of pages and functionalities.

In order to reach this objective, the deliverable entitled "*Web Interface Specifications*" has been written by the M4D Partners in order to precisely describe the concept, design and layout of the Web Application pages. This document also aims at triggering an iterative dialog with the ESPON CU that validated the first version in March 2012.

The M4D delivery in June 2012 includes a first version of the Web Application. Its implementation is based on these specifications.

Regarding the integration of data and support to the ESPON Projects, the Web Application has also been designed so to merge Users and Projects needs in an integrated Portal.

Once logged in, the projects are proposed targeted functionalities (previously called "Platform" Working Package WP C 1) such as:

- dataset upload and syntactic check according to the Data and Metadata Specifications.
- tracking of the dataset integration process.

A complete description of the integration workflow is available in the attached support document entitled "How To Deliver My Data?". This document targets ESPON Priority 1 Projects (Key Indicators), ESPON Case Study Projects ("Zoom-in" indicators of Priority 2 Projects) and ESPON Background Projects.

WPA1 Computer – Semantics

The development of a new web interface required a hard work on the metadata templates and the metadata themselves. At first, it was absolutely necessary to build a table with all the indicators produced and delivered by the ESPON projects, even by the oldest ones, in order to have a general view of these indicators. The indicators from the past projects as well as the newer one had to be given new metadata, as specified with the templates developed by the Laboratoire d'Informatique de Grenoble (LIG) in cooperation with the RIATE in Paris. The metadata of the datasets were transformed to be compatible with the request parameters on the interface. It involved also building a keyword system. We used mainly keywords from the GEMET thesaurus. We also adopted a thematic classification of the indicators which mixed EUROSTAT's thematic classification and the ESPON's one. Concerning the political classifications, the CU decided to integrate these parameters in the metadata and to emphasize on the relevance or not of the the ESPON projects indicators for specific policy reports such as "Europe 2020".

WPA1 Computer - OLAP Cube integration

The ESPON OLAP Cubes and all the documentation, tools and current and former versions have been integrated into the ESPON Database Portal by means of a specific ESPON OLAP Cube page. From this webpage users can found:

- Access to the current and former versions.
- Access to documentation plus User Manual.
- Link to the ESPON OLAP Cube Viewer.

Below are displayed two screenshots of the future aspect of ESPON OLAP Cube tool webpage (implemented in the next steps of the M4D project).

Figure 1 – Screenshots of the ESPON OLAP Cube Web Page which will be developed for the end of 2012

A new ESPON OLAP Cube has been created, version 5.0, by means of the new data sources and procedures specified under point 2.2.5.3 of the current document. This is the list of the different data dimensions and measures available in the new update:

Measures

- Area in hectares
- Active People 2000 thousand inhabitants
- Active People 2001 thousand inhabitants
- Active People 2002 thousand inhabitants
- Active People 2003 thousand inhabitants
- Active People 2004 thousand inhabitants
- Active People 2005 thousand inhabitants
- Active People 2006 thousand inhabitants
- GDP 2000 Million Euros
- GDP 2001 Million Euros
- GDP 2002 Million Euros
- GDP 2003 Million Euros
- GDP 2004 Million Euros
- GDP 2005 Million Euros
- GDP 2006 Million Euros
- Population 1990 thousand inhabitants
- Population 1991 thousand inhabitants
- Population 1992 thousand inhabitants
- Population 1993 thousand inhabitants
- Population 1994 thousand inhabitants
- Population 1995 thousand inhabitants
- Population 1996 thousand inhabitants
- Population 1997 thousand inhabitants
- Population 1998 thousand inhabitants
- Population 1999 thousand inhabitants

- Population 2000 thousand inhabitants
- Population 2001 thousand inhabitants
- Population 2002 thousand inhabitants
- Population 2003 thousand inhabitants
- Population 2004 thousand inhabitants
- Population 2005 thousand inhabitants
- Population 2006 thousand inhabitants
- Population 2007 thousand inhabitants
- Population 2008 thousand inhabitants
- Population 2009 thousand inhabitants
- Population 2010 thousand inhabitants
- Unemployed People 2000 thousand inhabitants
- Unemployed People 2001 thousand inhabitants
- Unemployed People 2002 thousand inhabitants
- Unemployed People 2006 thousand inhabitants

Spatial dimensions or LARUs (Land Analytical and Reporting Units)

- Biogeographic Regions 08
- City Names
- CLC00 Hierarchial
- CLC06 Hierarchial
- CLC90 Hierarchial
- Dominant Land Cover 2006
- Elevation Breakdown
- Large Urban Zones
- Land Cover Flows 1990-2000
- Land Cover Flows 1990-2006
- Land Cover Flows 2000-2006
- Massifs
- Nuts 1999 code
- Nuts 1999 name
- Nuts 2003 code
- Nuts 2003 name
- Nuts 2006 code
- Nuts 2006 name
- Nuts 2010 code
- Nuts 2010 name
- Regional Sea Basins

WPA1 Computer - User-friendly OLAP visualization tool

The ESPON OLAP cubes are a great tool to explore and analyse European indicators at different aggregation levels. A user friendly, web-based visualization tool has been set up in order to extend the possibilities of the Cubes. Based on open source technologies, it offers universal access (through a web-browser) and simplifies the creation of maps or extraction of datasets from OLAP queries.

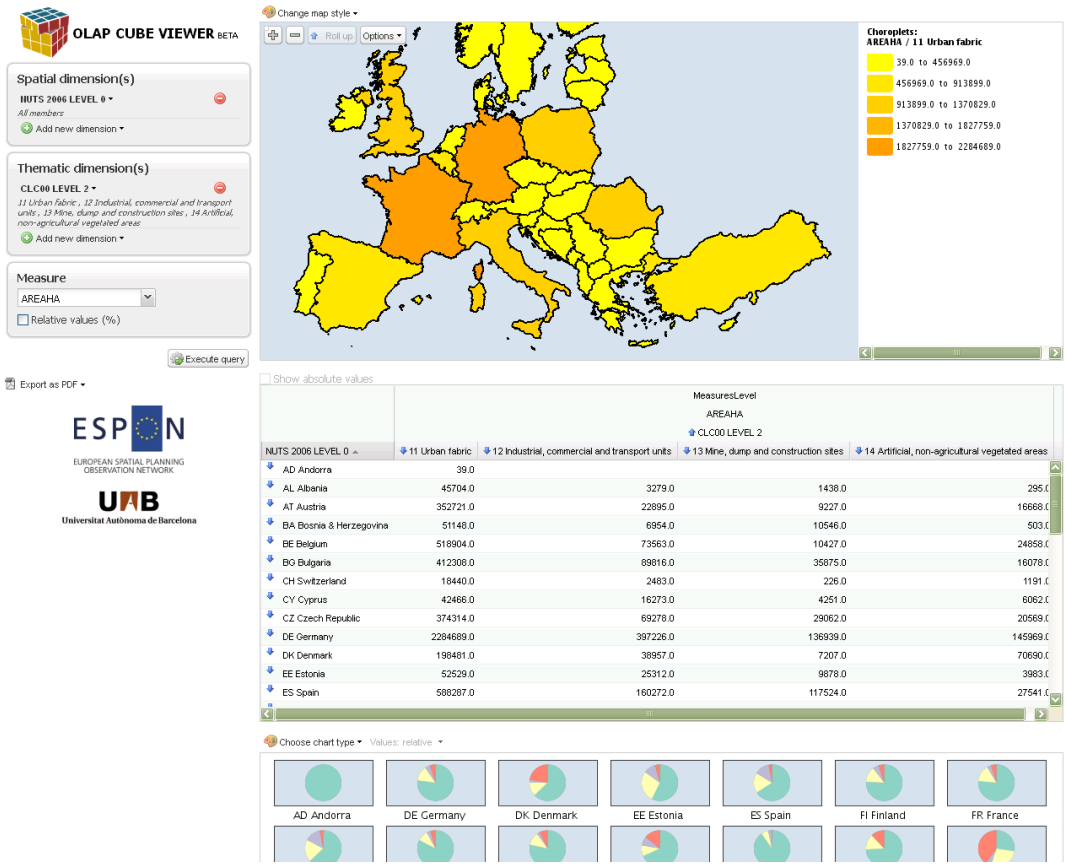
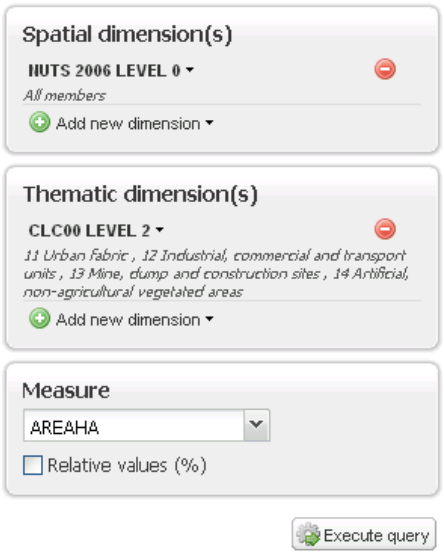


Figure 2 - Screenshot of the OLAP Cube Viewer

This software exposes the content of an OLAP cube as an interactive web application, showing users the potential of OLAP technologies for data exploration and analysis. The application **combines map, table and chart visualisation** for the same data, improving the interpretation of results. The user can build custom queries in an easy and interactive way, and can also modify the query by interacting with the maps and tables.

Note that the tool is still on a beta stage, which means that it is available for testing and evaluation purposes, but it is still being developed and improved. At this time, the application is using the ESPON OLAP Cube version 4.0 (December 2011).

The application can be accessed through the following address:
http://158.109.174.100/webbi/



USER INTERFACE

The tool has been designed to be **user-friendly and intuitive**, providing a simple user interface to build the queries, which can be further refined by the **dynamic interaction with the map and the table**.

The **query builder** is laid out on the left side of the screen. It is used to select the dimensions and measures that will be included on the query. Dimensions have been categorized on two groups: **Spatial Dimensions** (also know as Reporting Units, such as NUTS, Sea Basins or Biogeographic Regions) and **Thematic Dimensions** (such as Corine Land Cover

stocks [1990, 2000 and 2006] and Land Cover Flows [1990-2000, 2000-2006 and 1990-2006]).

The **measures** are the statistical data (for example Hectares or Gross Domestic Product) that will be analyzed for the selected reporting unit and land cover classes.

The results of the query are simultaneously shown as map, table and chart. The user can modify the query by **interacting with the map and the table**, drilling down to more detailed levels (e.g. changing from country level to regional level) or up to more generic levels (e.g. from regional level to country level). Note that map, table, chart and query builder will always be **synchronized**, so a change on any of them will also change the rest of them.

The map also provides standard navigation tools such as zoom and pan, and several **symbolization** legends can be applied, including overlay symbols (proportional symbols, bar chars and pie charts) and choropleths.

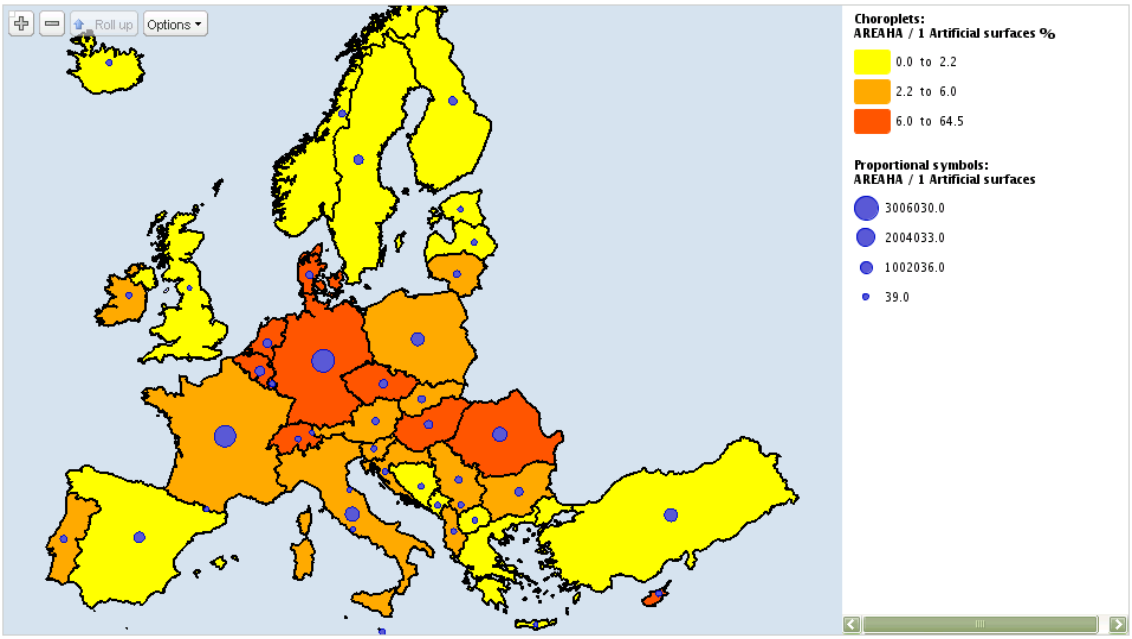


Figure 3 - Example of different map symbolization

TECHNOLOGY

The cube viewer has been built using a number of free software technologies, making possible to re-use the application for different thematic areas. This approach also encourages improvements and customizations from any interested party. In this way, the tool has filled an important gap in the software field, as no comparable open source tool is currently available in the market.

The main technologies used on the project are:

- PostgreSQL + PostGIS, as spatially-enabled relational database management system.
- GeoMondrian, an Spatial-OLAP server.
- OpenLayers javascript library in order to provide map visualization and navigation.
- ExtJS and GeoExt in order to build the user interface.
- GeoTools and JTS for geographic data processing and delivering.

- Olap4j, which provides a uniform programming access (API) to different OLAP servers
- JFreeChart, used to generate charts.

Note that all those applications and libraries are open source technologies.

NEXT STEPS

A number of improvement paths are envisaged, and they will be steadily implemented during the project:

- Several approaches will be tested to **improve the performance** of the tool for queries involving a big number of dimension members.
- The database will be updated to the **last version of the ESPON OLAP cube**, and regularly updated whenever a new Cube update is delivered.
- The **chart component** will be improved, as current visualization is suboptimal for big queries.
- The **symbolization options** will be adapted and improved according to the ESPON Mapping Guide (within the limits provided by the architecture and purpose of the application)
- The **user interface** will be modified to better fit big tables and charts (on the result area) or bigger number of dimensions (on the query builder).
- The **print template** will be customized

The first fully functional and updated (not beta) version of the tool is foreseen by December 2012.

WP A2 Statistics

WPA2 Statistics - Introduction

Work Package A2 deals with the problems which are attendant on the presence or otherwise of anomalies in the data for the Database. In the Inception report the three subsections of this package are Time Series, Outlier Handling, and Data Quality. As work on these has progressed it is clear that some adjustment to the original somewhat vague intentions is required, and that these intentions need to be made somewhat clearer.

In their wide ranging survey Chandola et al (2007)[2] assert that "Anomalies are patterns in data that do not conform to a well-defined notion of normal behaviour". They also point out that such notions are often domain dependent, and this is reflected in the terminological variation which is encountered; this includes: anomaly, outlier, discordant observation, aberration, surprise, peculiarity, and contaminant. There are a wide variety of anomaly detection approaches which have been developed over the years, but many of these are specific to particular domains, such as the detection of credit card fraud. They suggest that there are several challenges in anomaly detection:

- defining a 'normal' region is difficult
- normal behaviour tends to evolve
- the notion of an anomaly is different in different application domains
- the availability of training ('labelled') data is a major issue

[2] Chandola V, Banerjee A and Kumar V, 2007, *Anomaly Detection: a survey*, Technical Report TR07-17, Minneapolis: Department of Computer Science, University of Minnesota

- the noise inherent in the data may make anomalies difficult to detect

Hawkins (1980)[3] opens his monograph with the pragmatic definition of an outlier to be "an observation which deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism". Another view of outliers is that they are one or more observations that greatly influence the value of a statistical estimator. For the ESPON Database it is important that any outliers which arise as a result of some data error are identified and dealt with. There will be other outlying observations which are genuine – however, they will have been identified as such and agreed with the data supplying partner[4].

It is clear that a single general purpose anomaly detector into which an ESPON project dataset can be placed and which prints a report highlighting the NUTS regions in which there appear to be anomalous data is perhaps a chimera. The range of ESPON projects and the diversity of the data which is offered as potential input to the database mean that we have to think of a range of generic techniques which can be selected as appropriate and applied to a particular dataset. It is also desirable that the anomaly detection that is run on a dataset could be re-run at a future date, and which would yield the same results: in other words, the process should be reproducible[5].

ESPON datasets and data quality checking

The data sent from the ESPON projects has a well-defined layout. Each spreadsheet contains four labelled worksheets, containing (a) general information about the dataset (b) metadata about the indicators (c) metadata on the sources and (d) the data itself. The data tables are organised such that rows represent observations (usually NUTS regions) and the columns represent indicators. The columns are logically organised into pairs such that the first contains the data values and the second contains the source key.

Our approach has been to provide a *library* of detection methods which can be assembled in a series of templates. The templates provide a basic outline for the different sorts of data which we expect to encounter in the data quality activities. Currently there are two templates, one for time-series data, and one for cross-sectional data. An *instance* of a template is the result of its adaption for a particular dataset. Eventually there will be at least as many instances as there are ESPON datasets supplied for data quality assessment.

The output of data for the Database by an ESPON project is in the form of one or more Excel spreadsheets. We have adopted the principle that we do not work directly on the spreadsheet: rather, we copy the data and work on that copy.

The following sections outline (a) the issues and approaches involved in handling time series (b) the issue and approaches involved in handling cross-sectional data and (c) the modus operandi for dealing with the data quality operation in the M4D workflow.

WPA2 Statistics - Time-series

A characteristic of time-series is that there are a series of measurements of some characteristic of a NUTS region taken at regular intervals: Y_1, Y_2… Y_T. The characteristic might be for example: annual average unemployment, estimated population on January 1st, per capita domestic product. Such data are sometimes referred to as longitudinal data in contrast to cross sectional data in which measurements of a characteristic are taken across all NUTS regions at a single time period.

[3] Hawkins D, 1980, *Identification of Outliers*, London: Chapman and Hall

[4] The GDP/capita of Liechtenstein or Luxembourg are typical example of such values that are exceptional – compared to the neighboring regions - without being false.

[5] For example, when time series with breaks are rebuilt on the basis of census results.

There are two challenges for time series in ESPON. The first is the identification of anomalous values in a time series and the second is the imputation of missing data in a time series.

There is a rich body of statistical theory which deals with modelling the behaviour of a times and permits future forecasts to be made. This again poses a challenge, since the ESPON series are generally short in comparison in those typically encountered in modelling. If we have 20 or so observations, then our estimates of the parameters of the model will be less sure than those from a series with 200 or 2000 observations. However, we have been successful in fitting a class of models, known as ARIMA (AutoRegressive, Integrated, Moving Average), to some example series from RIATE, and have devised a method of identifying potentially anomalous values based on these model fits.

With the ESPON data there is the possibility that we will encounter series of a merely few observations. It is tempting to think of the outputs from the decennial census as forming a series, but these are more akin to panel data. With very short series, the ARIMA route is not the most appropriate[6]. There are alternative approaches, some of which were developed during Phase I of the Database project. These treat the series as in the space domain rather than the time domain. There are also techniques for dealing with space-time data – our urgent priority is the assessment of these as potential anomaly detectors.

The second challenge arises when we encounter missing values in a series, at say NUTS3 level, but we have complete series at NUTS2. This is akin to having a matrix - the rows are the observations and the columns are the time periods - but we also have the column totals. The problem is then one of trying to determine the most likely values for the missing cells in the matrix with the constraint supplied by the column totals. We will liaise with the RIATE team in the development of these methods, and this is an urgent priority (see the Technical Report on the Core Database Strategy and the second annex on the ESTI model).

Deliveries June 2012:

Technical report; Data quality methods for time series data of absolute indicators
Check of data related to the Core Database Strategy

WPA2 Statistics - Outlier handling

A priority this year has been the development of a coherent set of tools for the identification of anomalous values: outliers. The reason of this priority is the fact that outlier detection will be introduced as a compulsory step in the process of data delivery by ESPON TPG, at less for the "10 best indicators".

We took as our starting point the testing approach developed by Charlton and Harris[7] in Phase I of the Database Project. These were very closely tied to the identification of outliers in a dataset of GDP estimates which had been deliberately contaminated by the RIATE team. The approach was somewhat inflexible; we have removed some of the tests and added others in order to create a library of detectors which can be applied to a wide variety of data.

Typically data arrives in an Excel spreadsheet, organised into a rectangular matrix. The rows represent spatial units for a particular NUTS level and date, and the columns represent the indicators. There is metadata to describe the characteristics of each indicator. Two challenges have been (a) the development of a wider range of tests for the different data types which we encounter and (b) the implementation of these tests in

[6] Especially in the current period of world crisis where very strong change of trends has been observed since 2008 for May criteria (unemployment, GDP/capita, migratory balance).

[7] Harris P and Charlton M, 2011, *Spatial Analysis for Quality Control Phase 1: the identification of logical input errors and statistical outliers*, ESPON 2013 Database technical report, March 2011

a coherent and convenient form which will permit reasonably prompt assessment of the candidate datasets.

The tests are both mechanical and statistical. The mechanical tests include checks for the existing of missing values (usually flagged as NA[8]) in the data. We observe that some spatial units have missing values for all the indicators in the dataset – these are reported. A second mechanical test is the identification of omitted spatial units: we check against those units which are present in the shapefiles for the NUTS units available at the Eurostat portal and reproduced in the ESPON mapkit.

The statistical tests we use are both exploratory and confirmatory, and they are conditioned on the level of measurement of the data that we find in each indicator. The exploratory approaches involve simple data summaries and visualisations, the latter including both barcharts/boxplots and maps. The confirmatory approaches include aspatial univariate bivariate and multivariate tests, as well as spatial tests. Two challenges arise here. The count data (population, GDP, area,…) are not directly comparable for administrative units which are of different sizes and subject to Modifiable Area Unit Problem (MAUP). It is conventional to account for this geographic variation by normalising the data which some suitable denominator or at-risk population: for example gross domestic product is usually normalised by dividing by the population to yield per capita gross domestic product. This indicator is "normalized" because it is now independent in principle of the size of spatial units and can be submitted to outlier checks.. It is the reason why, where possible we will normalise with a suitable denominator, and if this is not possible area normalisation can be applied (i.e. measuring the spatial densitity of the phenomena).

The spatial tests, which deal with how different the value in an area is in comparison with its neighbours, work well for spatial units which have neighbours. The definition of neighbouring is based on either adjacency or a pre-specified local sample size. Spatial units which are islands, or which are detached (such as La Réunion, Ceuta or Melilla) are necessarily omitted; they will, if present in the dataset, have been subjected to the spatial tests. The model based aspatial tests can be extended into a spatial form. However, we are aware that the behaviour of such models demands acre in their application (Wall, 2004)[9].

We can deal with one attribute at a time or consider several simultaneously. Penny and Joliffe (2001)[10] suggest that relying on a single technique is unwise, and that the analyst should consider the results from several tests. Where possible we follow this prescription. The testing strategy has been implemented in the R language, and is described below.

Deliveries June 2012:

> Technical report: Outlier detection techniques for spatially normalized values (completion)
>
> Technical report: Detection techniques for spatially un-normalized data, interval, ordinal and nominal (completion).

These two reports will be coalesced into a single Technical Report report on outlier detection.

[8] NA : Non available. But we have also introduced in the metadata model some variants like the case of data that are simply "Impossible" to measure (e.g. length of maritime coast for regions of Austria).

[9] Wall, M, 2004, A close look at the spatial structure implied by the CAR and SAR models, *Journal of Statistical Planning and Inference*, 121, 311-324

[10] Penny KI and Joliffe IR, 2001, A comparison of multivariate outlier detection methods for clinical laboratory safety data, *The Statistician*, 50(3), 295-308

WPA2 Statistics - Data quality

This section is concerned with the implementation and operation of the Outlier check process. The position of the outlier check in the data flow process is shown in Annex 1 of M4D Guidance Paper on data delivery (Ysebaert et al, 2012)[11]: it follows the data delivery, and syntactic checks, and the metadata semantic check. A.2.3 operationalizes the battery of tests developed by Harris and Charlton (2011)[12] in a manner which allows for the rapid assessment of each dataset as it arrives at the NCG. Harris and Charlton describe a series of tests and present the code in the *R*[13] language for 6 worked examples. The code for each worked examples is presented as a stand-alone entity (that is, running the code for example 2 does not require the user to have run the code example 1). This leads to a great deal of redundancy and repetition. The codes themselves were developed assuming that the main inputs to the database were to take the form of ratio or stock data, and ignored the existence of other data such as typology data. As a consequence, additional tests have been adopted.

We have therefore designed a new strategy for dealing the each dataset as it arrives in the data flow: there now exists a single unified R *function* which uses *control information* about the dataset to determine which tests are appropriate. Thanks to the new metadata model, the control information is extracted from the Dataset and Indicator worksheets in each supplied dataset, and identifies the indicators present in the dataset, their types (e.g. text, ratio, count, typology); for the typology indicators the list of valid codes is extracted from the metadata. It also identifies the spatial units (e,g, NUTS level[s] and date). The tests to be applied are collected together into a *library*. The library consists of high level functions to implement the various test strategies (exploration, univariate, bivariate, spatial and multivariate), with lower level functions to deal with the individual tests (e.g. boxplot statistics, Moran's I). This redesign and implementation allows for considerable flexibility in the maintenance of the over outlier check process, and also for extension to incorporate additional tests.

The output from the outlier check function is a *report*. The report has two sections. The first consists of the commented output from the outlier check function. The second is a spreadsheet in which there is on row for every spatial unit in the input Dataset, and a series of columns with the results for the individual tests. The individual test results in each are expressed as a series of 1s and 0s (for passed/did not pass the test). The summation of these across each row provides a score on the overall performance of each spatial unit. The smaller the score, the more likely the spatial unit is to contain outliers.

This approach will be applied to the data that arrives from the ESPON Project partners in the data flow. We will also continue to identify further tests for outliers and implement them in a programme of continuous improvement.

An important point of debate for the future is the decision to delete or store the results of outlier checks. With the first option, we consider that outlier check is only a step of the quality control process, which is achieved when the authors of the data decide to validate or correct the flagged information. With the second option, the different flags resulting from outlier checks are considered as information of interest *per se*, to be kept after validation and correction by the authors of data. We can indeed imagine a user wanted to find quickly a list of regions that are "*exceptionally different from their neighbours*" in order to realize case studies on econometric local convergence or cross border cooperation. In this case, the flag "spatial outlier" would be a very useful information.

Deliveries: June 2012:

[11] Ysebaert R, Salmon I, Le Rubrus B and Telechev A, 2012, *How to deliver my data: Guidance Paper for ESPON TPGs*, ESPON M4D Technical Report, June 2012

[12] ibid

[13] R Development Code Team, 2005, *R: a language and environment for statistical computing, reference index version 2.14.1*, Vienna: R Foundation for Statistical Computing, ISBN: 3-9000051-07-0

Web function; Operationalisation of data quality check on Platform and reporting procedures

A number of tasks were reallocated between Partners in 2011, one of which was the development of the Platform. The development of the Platform on the NCG servers was originally an NCG task – this is not now the case. The data quality checks will take place at the NCG on the NCG's own servers and not on the Platform. The data quality check and reporting functions have been operationalised, and will be further developed.

WP A3 Mapkit tool (RIATE)

From June 2011 to June 2012, two new Mapkits have been created (Seas and Neighbourhood) and one Mapkit has been revised (ESPON Area). All these Mapkits are available under the resource part of the ESPON Data Portal. The Mapkit Web page contains a screenshot of the mapkit (as shown below), a description of the content of the Mapkit and a link to the mapping guide, produced within the ESPON Database 1 project. The download of the shapefiles is restricted to registered users.

- **Mapkit Seas:** The ESaTDOR Regional Seas tool kit (figure 4) is the Mapkit developed by the ESaTDOR project.

Figure 4 Screenshot of the Seas Mapkit

- **Regional Neighborhood mapkit (provisional):** The Regional neighborhood mapkit is defined in order to make maps on the paneuropean area as defined by the ITAN project, at regional level. It contains a generalized version of the NUTS regions on the ESPON area and also generalized geometries of the regions on the neighborhood area (Similar to NUTS). This mapkit has been delivered to the ITAN Project for validation in October 2012. It will be adjusted after the ITAN feedbacks. Then, it will be available under the ESPON Database Portal

Figure 5: Screenshot of the regional neighborhood Mapkit

- **ESPON Area narrow mapkit:** NUTS 2010 geometries have been added to this mapkit (territorial changes in United Kingdom, Netherlands, Germany, Finland, Italy and Greece).

Figure 6: Screenshot of the updated version of the ESPON Area narrow Mapkit

WP A4 Updates (LIG)

This section proposes the list of main changes for each distribution update of the ESPON 2013 Database Web Application, deployed on *infeurope* server, from the latest to the oldest revision number of the project. This listing concern the updates that have been carried out since the ESPON Database Phase I delivery of the 31st of March 2011. For each item, a brief comment describes the main issues of the update.

Update 2012-02-28 r903
This update concerns:

- The update of the Metadata Specifications reference document.
- The integration of the FAQ under the Help section.

Update 2012-02-22 r899
Important changes in the code of the application, they fix the problem with the database connection. This update requires a Java Data Base Connectivity (JDBC) Data Source configuration on the server, which has precisely been described in a "Installation Instructions" document that has been delivered to infeurope.

Update 2012-02-17 r897
As agreed with ESPON CU on the 7th of February 2012, this update concerns modifications of available projects from the Web Application. Details are given below:

- the modification of a data file in the "projects" folder (the [PROJECTS] directory on the server that is pointed by the value of the ESPON_PROJECTS_PATH environment variable):
  - the [PROJECTS]/data_UMZ.rar can be deleted;
  - the attached file data_UMZ_Named_10K.zip must be copied pasted to this "[PROJECTS]" folder.
- some modifications in the descriptions of the available files in the "Urban Data" category:
  - the note mentioning that a new archive zip file will be soon available has been removed (above file replacement);
  - the description of the FUA database;
  - the link to the "LUZ specifications (Urban Audit 2004)" has been modified, the previous version used to redirect to an invalid URL on the Eurostat Web Site, it now points to the Urban Audit Web Site.
- the modification of a data file in the "projects" folder (FUA Database item, the data is now public but it does not contain the protected Eurogeographics geometry anymore) and the modification of the visibility for the report and data:
  - the [PROJECTS]/data_FUA.rar can be deleted;
  - the attached file data_FUA_noGeom.zip must be copied pasted to this [PROJECTS] folder.

Update 2011-12-21 r886
List of improvements implemented for December 2011 delivery:

- **General improvements**
  - A new search algorithm has been implemented. Search requests are now processed about 30% faster than in the previous version.
  - Duplicate indicators have been removed in the indicators selection list.
  - The entire layout has become more convenient: it is focused on thematic search, not on the projects or datasets.
  - Pages with indicators overview have been added to the search interface in the "Overview" menu.
  - Search results table is made clearer and without the controls that turned out to be useless in the previous version.
  - The most recent version of `Data and Metadata Specifications` has been integrated into the help menu of the application.

- **Minor improvements**
  - Directly downloadable projects: file sizes are now displayed.
  - Contact e-mail address added into the Terms&Conditions page.
  - The list of years in the period's selection is centred on the current year.
  - Period's selector: minor bug correction.
  - Explicit super-themes have been added to indicators metadata.
  - Study area selection: countries are now sorted by names, not by their codes.
  - Reported misspellings have been corrected.
- **Functionalities removed from the previous version**
  - Publication dates criteria.
  - Geographical objects criteria (because only NUTS are supported in this version).
  - The old metadata editor has been disabled until the implementation of a new version respecting the new Specifications.

Update 2011-09-27 r828
This update aims at fixing a bug in the generated xls files that can be downloaded by the user: in the generated "data" sheet, the codes of the indicators were sometimes misplaced. In the revision 826 (see Overview of changes in r826), this bug triggered the disabling of the *Basic Indicators, June 2009* dataset. The previously noticed errors about the values inconsistencies are now fixed, this dataset is tagged "visible" back in the database, it is available back for any search.
The generated espondb.war file now embeds a standard custom manifest file, including for example the Implementation-Version property, showing the version and build date of the application.

Update 2011-09-21 r826
This update aims at enabling/disabling the visibility of a dataset integrated in the database, including its associated indicators, themes and keywords. This functionality has been added to disable the visibility of the *Basic Indicators, June 2009* dataset, in which several values inconsistencies have been detected. Once fixed, this dataset will be made visible again.
This functionality to disable/enable the visibility of datasets is a first step towards the Dataset Administration tool that has been scheduled to be delivered later.

Update 2011-09-19 r824
- The delivered Web Application war file is now pre-configured with JDBC parameters in order to avoid the "application Configuration" step, while deploying the application, or restarting the server.
- The functionalities of the administration side now include a new menu item to execute embedded JUnit tests. Currently, only two tests are launched (checking the availability of the documents referenced as ESPON projects), but the tests suite will be completed in the next revisions.
- A new data file, ESPON_OLAP_Cube_Update.zip, has been added to the "projects" folder of the Web Application.

Update 2011-06-17
Minor modifications. Aims at proposing a stable version for the ESPON Seminar in Budapest from the 21st to the 23rd (MC Members Meeting demonstration) of June 2011.

Update 2011-05-27
Minor modifications only.

Current situation (30 June 2012)

- Web Application version 2 (online access to the current version under development)

- The search interface contains 30 datasets, 443 indicators and 1 070 213 values (cf part networking for the description of indicators integrated in the database)

Working Package B - Thematic

WP B1 Regions

All the elements displayed in this part are described in detail in the Technical Report on the Core Database Strategy (delivered in December 2011 and available under the ESPON Database Portal). On top of that, the first inputs of the Core Database have been presented during the ESPON Seminar in Aalborg (June 2012) and the Eurostat's meeting held in Luxemburg the 19th June 2012. This presentation is available under the ESPON Website[14] ("Transformation of regional indicators with functional neighborhood").

Development of the Core database strategy

The Core Database Strategy (CDS) is based on the assumption that it is possible with a limited number of selected indicators to create an innovative, evolving and sustainable database specifically targeted to the monitoring of territorial cohesion. This objective is not contradictory with the storage of the rich material produced by all ESPON projects but it is complementary. To be sure, we can distinguish two different kinds of objective that ESPON M4D is obliged to address simultaneously;

1. **Keeping the memory of ESPON 2013 program** suppose that the readers of reports produced by ESPON between 2007 and 2013 should be able to obtain an access to the most important indicators that has been elaborated. Typically, a synthetic typology or a new measure of IT connections derived from an exceptional survey. In both case, the indicators are important because they have been the results of an important intellectual effort supported by public funds and they are proofs of the effort realized. But we have to keep in mind that some of the indicators are sometimes exceptional which means (1) of very high interest but (2) impossible or difficult to reproduce easily in the future.
2. **Making possible space-time-scale comparisons** is a different objective where the added value is not necessarily the originality of the indicator but the possibility to follow the evolution through time, both in the past (estimation of historical values) and in the future (forecast of expected evolutions). A good example of this approach is offered by the different "map updates" produced by ESPON 2013 in order to compare evolution with indicators measured during the ESPON 2006 program. But we suggest exploring also the possibility to compare the same indicator at different scales (local, regional, national, global) and for different types of geometries (grids, cities, administrative units, smoothed maps). The progress of ESPON 2013 as compared to ESPON 2006 is not only related to the enlargement of time series but also to the possibility offered to benchmark urban, regional and local dynamics.

The Core Database Strategy (CDS) is an ambitious attempt combines the two objectives through the identification of a specific part inside the ESPON database on which efforts should be more concetrated. The idea is to deliver a label "Core data" to a very limited number of indicators that are at the same time (1) the most relevant for EU Cohesion Policy but (2) should be regularly updated in the future for the purpose of territorial monitoring but also cartographic applications like Web-mapping tool, HyperAtlas, etc. This core part of the database is the one that should be used in the future for the regular delivery of a monitoring report, following the model that has been initiated by INTERCO project between 2010 and 2012 and will be enhanced by the ESPON monitoring project that will start its activities in autumn 2012. The role of ESPON M4D is to support actively what will be the major achievement of the priority 3.

[14] http://www.espon.eu/main/Menu_Events/Menu_OpenSeminars/openseminar12061314_after.html

From a pragmatic point of view, the CDS is based on five basic rules:

- Keep only count variables (dark green circle on the figure 8a): In concrete terms, it means that a variable like GDP per capita is initially excluded from the core database as it is not a count variable but the ratio of two count variables (total population and total amount of GDP (in $ or pps). As a consequence, the CDS focuses only on a limited number of count variables.
- Store formula of indicators of interest derived from count variables (e.g. a dependency ratio is a ratio between "young" and "old" divided by "adult" (shown in light green on the figure 8a). They are therefore virtual data that are computed on the fly when a user requests it. As an example, thanks to the core data collected and the methodology stored, it has been possible to update and extend to the Candidate Countries the "Map update on demography and migration".
- Enlarge time series of count variables in past and future with estimation of missing values. The core data should be absolutely complete in order to make possible the computation of total aggregates at different levels.
- Develop automatic procedure to exchange count variables between geometries of various types (figure 8b: For example, thanks to the ESPON OLAP Cube, the GDP data at NUTS 3 level will be firstly estimated at grid level (using ancillary variable like JRC 2001 or GEOSTAT 2006 population grids) and then re-aggregated to other geometries like water basin, FUA, etc.
- Propose innovative procedures of multi-representation and multi-level analysis of indicators for territorial monitoring and political decision: As displayed in the blue circle of the figure 8a, whatever the geometry of interest (cities, regions…) is, an isolated indicator can be transformed into a family of variants targeted to answer to specific political questions (comparison to the EU average, territorial discontinuities…).

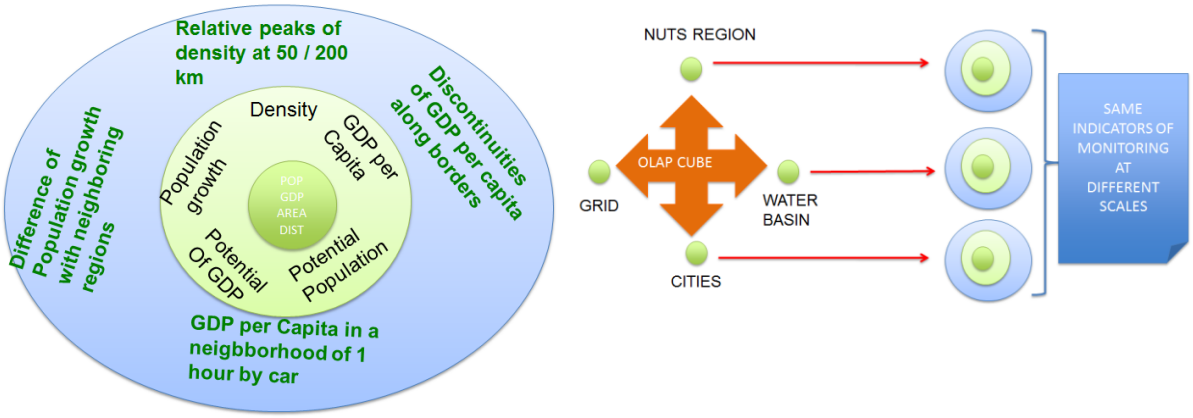


Figure 8a: "Three circles" of indicators strongly interconnected themselves: the core (green), the derived indicators (light green) and innovative indicators of high policy relevance (blue).
Figure 8b: Conversion of core indicators in several geometries thanks to the OLAP Cube technology.

The CDS may become an efficient strategy for the monitoring of regions and cities within the ESPON area. It focuses precisely on the indicators that can be regularly updated in past and future. It may be useful to understand the current dynamics and the forecast evolutions.
The fact that CDS only stores a limited number of count variables allows yearly updates. And then, all the other derived or complex indicators will be automatically recomputed as they are based on a formula.
Yet, any innovation adopted by policy makers (such as maps of deviation, analysis of discontinuities, potential based on time distance…), can be easily transposed to the most recent data.

Last but not least, when modification occurs in the geometry of spatial units (change of NUTS delineations for instance), the revision of indicators is feasible by the use of a limited number of count variables and formula.

Collection of time-series data for a first set of core indicators

Starting from the technical report produced in December 2011, 7 datasets of basic count data have been collected in the NUTS 2006 delineation, for the ESPON Area and the Candidate Countries. All this data have been collected yearly for 10 years at least.
The table below summarizes the data collected and integrated in the search interface of the ESPON Database Portal. For finding it easily, the user has to activate the option "core data" under the "general filter".

Indicator name	Level	Year	Degree of completeness of the table (ESPON Area + CC)
Total population	NUTS 0-1-2-3	1990-2011	100 %
Population by 5-years age-class	NUTS 0-1-2	2000-2009	100 %
Births and deaths	NUTS 0-1-2-3	2000-2010	99,9%
Total land area	NUTS 0-1-2-3	1990-2011	100 %
GDP (euro and pps)	NUTS 0-1-2-3	1999-2008	100 %
Active, employed and unemployed persons	NUTS 0-1-2	1999-2010	100 %

The main data source of these datasets is Eurostat. However and following the Core Database principles, most of the missing values have been estimated thanks to the ESTI framework, developed within the ESPON Data Navigator 2 framework. All the methods used for estimate the missing values have been indicated in the source part of the metadata. As a consequence, the user who wants to use these datasets will be able to quickly define:

- What are the "official data", e.g. coming from Eurostat
- What are the estimated values produced by M4D and with which method.

Once this time series are achieved and checked (by NCG partner), they are transmitted to UAB in order to be integrated to the OLAP cube where the are disaggregated at grid level with the more relevant ancillary variables. They can be therefore recombined into different geometries like FUA, LUZ, water basin, elevation levels … Of course, this procedure of disagregation/reaggregation introduce some noise and the estimations obtained in the new geometries are less accurate or at less subject to a more important margins of errors. But they open a great diversity of new possibilities for the monitoring of territorial diversity and the development of successful strategies in line with Europe 2020.

An illustration of the Core Database Strategy : the functional neighborhood transformations

Looking at the data offered by OECD which is the major competitor of ESPON in the field of regional databases, we have noticed that one of their most important comparative advantages was their freedom to propose new territorial divisions mixing official levels of NUTS nomenclature. Being an independent body of international expertise (not limited to EU) the OECD is not obliged to follow EU regulation and can decide to mix NUTS2 and NUTS3 units and, even more, to create non official units based on addition of NUTS 3

regions that are not part of the same NUTS2 and NUTS1 (example of Hamburg or Bremen in Germany).

Such freedom is not really possible in the framework of ESPON which is an EU funding program, contractually obliged to use in priority the official NUTS nomenclature. Of course, the ESPON program had demonstrated in the past the limits of this official nomenclature (Cf. ESPON 2006 MAUP Project) and has introduced a proposal of NUTS2/3 level which is very near from OECD. But it is fair to admit that this new level has finally not be very frequently used in ESPON program because of its normative obligations and its responsibility to support EU policies, especially since the adoption of the territorial agenda and the development of the Europe 2020 strategy.

The question that we have explored was therefore the following one: is it possible to improve the meaning of regional data without changing the official limits ? Can we propose simple transformation of data that eliminates most of the default revealed by the MAUP study, without moving to a non-official nomenclature? We suspect that the answer is yes and we will briefly summarize here the solution that we have found and presented for the first time at the ESPON meeting in Aalborg.

If we take the classical example of GDP/inhabitant, we know that the allocation of wealth to a precise region is conventional and does not necessary reflect the real economic situation of inhabitant. In a urban agglomeration, the inhabitants are travelling from residential location in periphery to city centre where job is located and what does really matter is not the amount of GDP in the territorial unit but the accessibility to GDP, eventually located to other administrative units. In many cases the inhabitants of the periphery look poorer than they are in reality (Flevoland) and the inhabitants of the centre look richer than they are (Brussels). Aggregation of NUTS units in a non official manner could be a solution, but, as explained before we decide to exclude this possibility because of political constraints of the ESPON Program.

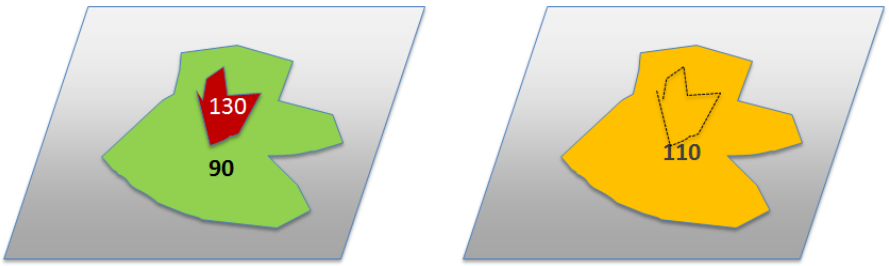


Figure 9: Screenshots displaying the Modifiable Area Problem (MAUP)

Fortunately, the ESPON program has an important comparative advantage on OECD which is the participation of many researchers specialized in the measure of functional accessibility by road, rail or air between regions. It is therefore possible to evaluate the cost of interaction between regions (time) and to derive from a time-distance matrix the situation of indicators not only inside a given region but in its neighbourhood. We just have to decide on the hypothesis that govern the decrease of accessibility of actors to resources located outside their region according to realistic assumption. On the basis of a recent matrix of road distance available in RIATE (source: Carsten Schürman, RRG), two concrete examples have been explored.

a) A functional view of growing regions

When mapping the evolution of GDP and population regarding to the ESPON average for the period 2000-2008 at NUTS3 level (figure 10, left map), the map reveals a mosaic of situation, in particular for little size units (Germany, Benelux, Portugal, England). The

fact to smooth these results in a neighborhood of 2 hours by road distance[15] (figure 10, right map) allows to generalize the observed trends and minimize the MAUP effects described above. The resulting map shows that a large part of the functional neighborhood of Germany, Benelux cross-border area, Southern Italy and Northern United Kingdom have experimented a relative loss of their share of EU31 GDP and population as regard to the rest of the ESPON Area. It is the reverse situation for Norway, Spain and Ireland. Interesting is the case of Eastern Europe which has converged with the rest of the ESPON Area in term of GDP but has lost a relative weight of population in the same time.

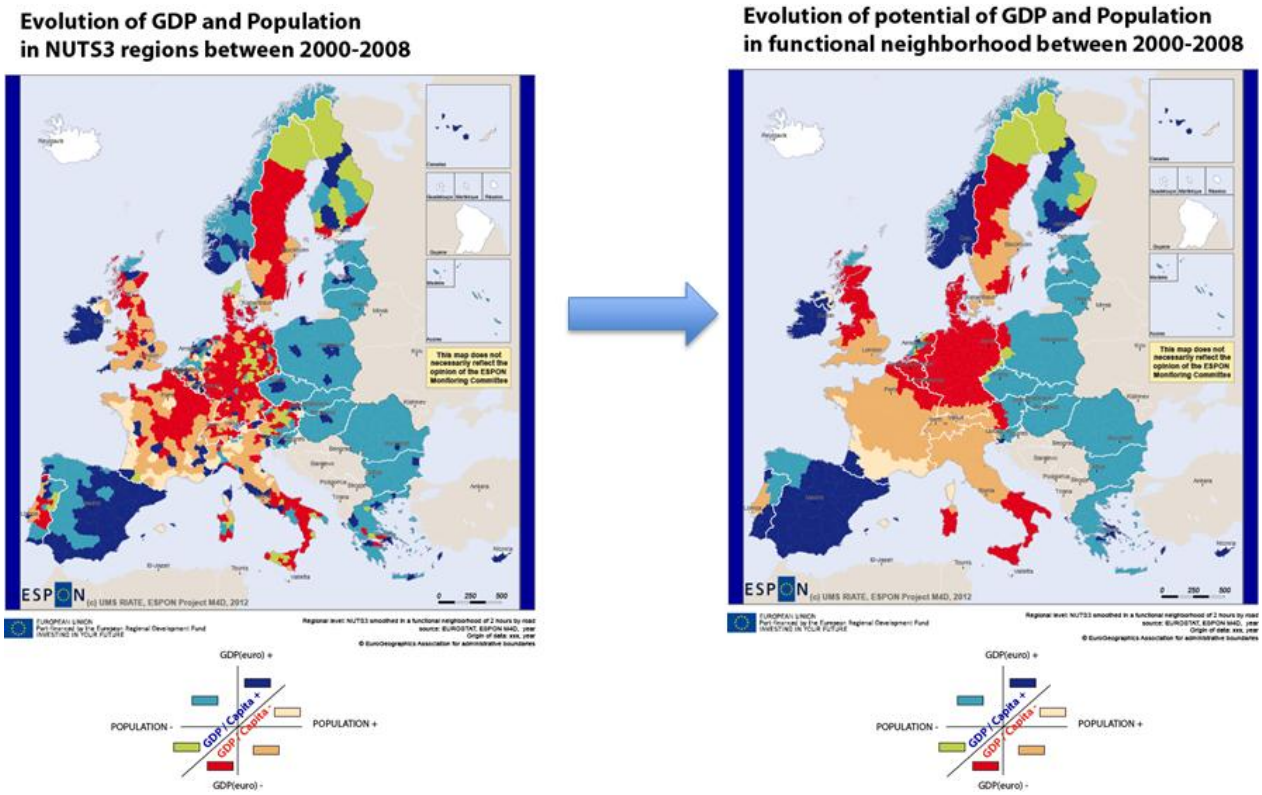


Figure 10 – A functional view of growing regions (evolution of GDP per capita smoothed in a functional neighborhood, e.g. 2 hours).

The key-point is the fact that the functional transformation is not considered as a better solution than the classical map by NUTS3 but a complementary information. In the region of Budapest, the classical map reveals that the situation is pretty good inside the limits of the region (dark blue), but the functional map reveals that it is not so good in the functionnal neighborhood of 2 hours by road (light blue).

b) Local convergence of the European regions

Following the recommendations of the DG Regio, we can move a step forward and try to evaluate not only the regional average but also the regional heterogenity in functional neighbourhood areas. The idea here is to propose a geographical vision of the processus

[15] With a decrease of influence with distance in order to simulate the concrete behaviour of actors.

of economic convergence or divergence at local levels with a local sigma convergence measure[16].

The first step consists by measuring local heterogeneity indexes in 2 hours functional neighborhood at two time periods for GDP per capita (figure 11, left maps). Regions displayed in brown or orange are the one where heterogeneity with their neighborhood is the higher: Whatever the time (1999-2000 or 2007-2008), this index is maximal at the border between old and new Member States, but also between Ile-de-France and London regions with their functional neighborhood.
More interesting is the analysis of the evolution of local heterogeneity between 1998 and 2008 (figure 11, right map): it displays that the local heterogeneity has considerably decreased at the border between Eastern and Western Europe. In the same time, local heterogeneity has considerably increased between Bulgarian and Romanian regions, between Lithuanian and Polish regions and between Northern Ireland and Ireland.

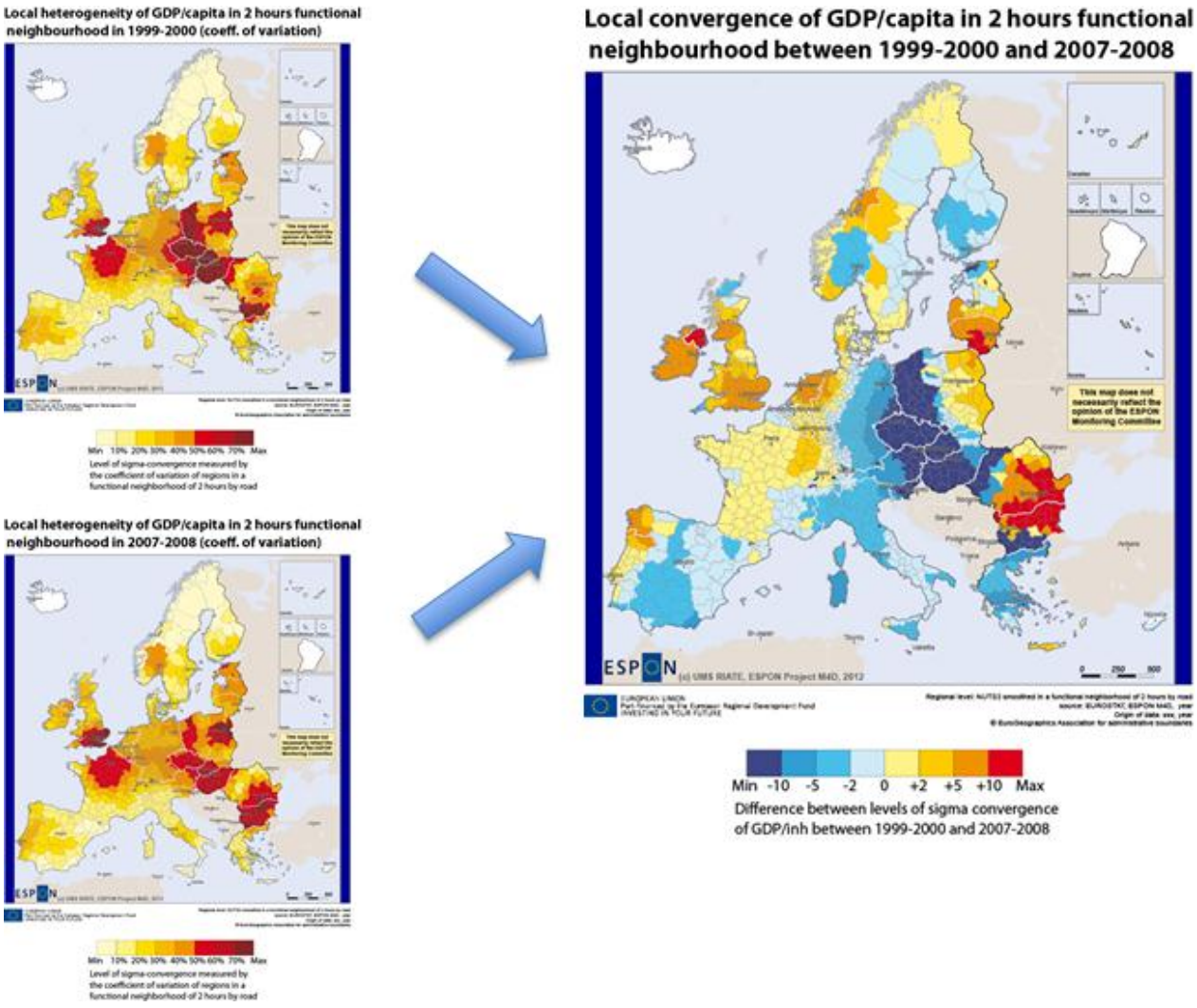


Figure 11 – Measure of the evolution of local converge of GDP per capita in 2 hours functional neighborhood.

[16] Montfort P., 2006, Convergence of EU regions, measure and evolution, Working Paper, DG. Regio, n°01/2008, 20 p.

What is behind this demonstration is to show that with a limited number of count variables, a time-distance matrix and some basic methods of spatial analysis, it is possible to develop new indicators which are at the same time empirically innovative (they reflect better the behaviour of individuals or firms) and policy relevant (they indicate where the EU objective are locally fulfilled or not

Next steps

With respect of these initial results convincing, we aim to continue the work in two directions:

- Improve the technical report on the core database strategy by proposing a review of innovative and policy relevant methods for using ESPON Core Data.
- Enlarging the data collection of basic count data to other indicators: employment and gross value added by economic branches will be the next (and complex) next steps of the data collection.

WP B2 Cities

WPB2 Cities - Integration of urban databases in ESPON DB

A/ Integration of specifications

The aim of this work is to formalize the metadata in order to help the users choosing the most appropriate database regarding their scientific targets. In December 2011, the metadata ontology model for urban agglomerations had been delivered. We have since added the metadata ontology model for functional areas defined by a top-down approach (i.e. excepting Urban Audit 2004 LUZ) (Figure 12a and 12b). Concerning the new harmonized LUZ, we have worked with documentation available on Circa and have now to contact directly the producers in order to fulfil the missing information. For the FUA database, we have to notice that it consists only in a set of urban names, population and surfaces without any geometry. We have considered it in the metadata ontology model as it had been used as a sort of "laboratory" for constructing the FUA_IGEAT a couple of years later.

These two different models will be available on the ESPON DB portal when all the fields have been completed.

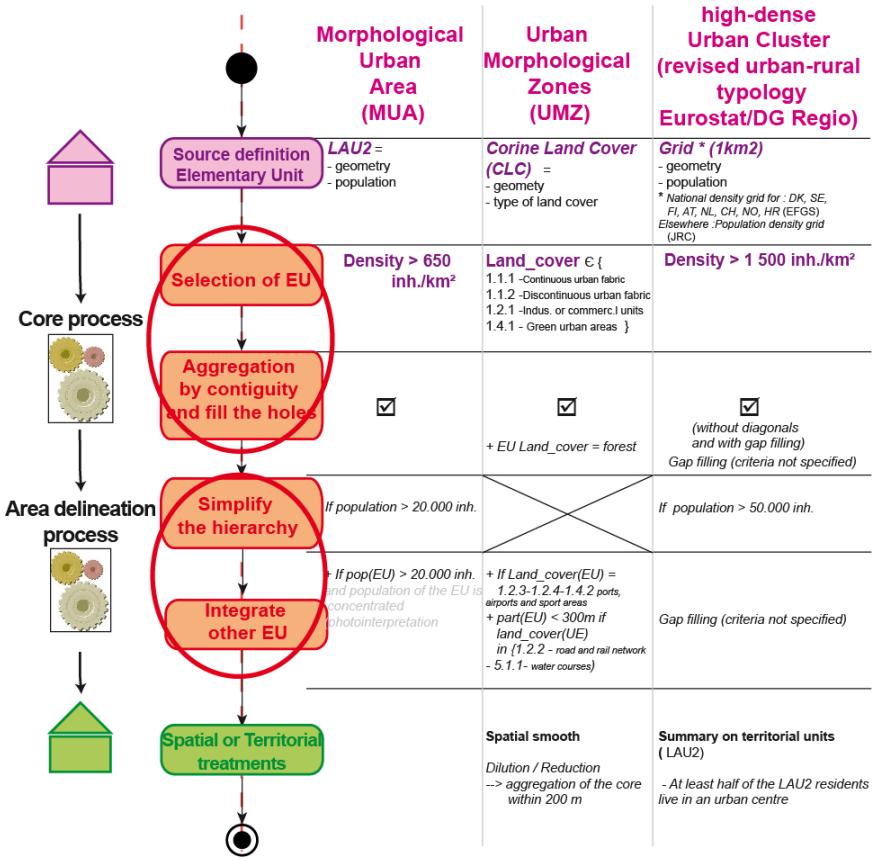


Figure 12a: Integration of specifications for morphological urban areas

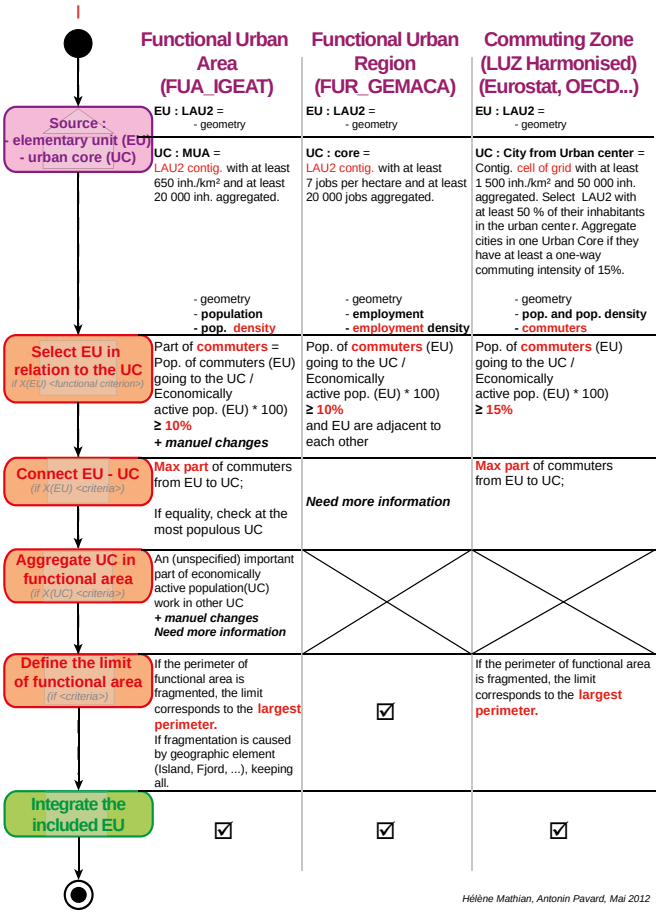


Figure 12b: Integration of specifications for functional urban areas (work in progress)

B/ Integration of data

The aim of this work is to evaluate if it has some sense to compare some indicators measured for different databases or to enrich a database using the data of another one, and to prepare the elaboration of correspondence dictionaries between entities of each database.

The methodology that has been constructed is generic and can be applied to a variety of urban databases. For instance, the indicators of interoperability are currently used to build a dictionary of correspondence between UMZ and LAU2, that will be of high interest for evaluating the relevance of using socio-economic indicators (for example from the SIRE database) inside UMZ delineation or using grid based indicators inside MUAs or other urban objects built as a collection of LAU2. In the same way, these indicators could be applied to functional objects such as new LUZ and their indicators crossed with other functional objects (for example FUAs from IGEAT).

For the particular example of MUAs and UMZ databases, this method allows to create a new indicator that enriches the MUA database and that consists in the percentage of built-up area per MUA. We have selected the MUA larger than 100 000 inhabitants (476 objects) and overlaid them with UMZ.

The first step consisted to establish the link between the two databases and, on this basis, to build statistical indicators that could enrich the matching process and properly describe the different types of overlapping (of population and of surface). That leads to consider not only intersections between the two objects (Figure 13, upper part) but also more complex case of non spatial inscription or multiple inscriptions (Figure 13, lower part).

The next step consisted to apply these indicators to an "a priori" typology of spatial configurations (Figure 14) in order to check their sensibility. The first two cases (one UMZ in one MUA and several UMZ into one MUA) give very similar results and are both characterized by a good interoperability. The concerned urban objects are mainly located in Sweden, Denmark and south of Spain (Figure 15). When the indicators enlighten particular complex cases (such as the case 3 or the case 4 in the legend of the map page 37, that extends mostly in the European Megalopolis, in England, Belgium, Netherlands, Germany and Italy), it means that the enrichment of one database by using the other is particularly difficult and that the results have to be interpreted with caution. The complexity of urban patterns in this region results from the conjunction of high densities of settlements and historical accumulation of wealth, skilled work and urbanisation. The exploration of different urban databases, through indicators of interoperability, is not only a way to enrich databases but also to capture urban features that can be useful in the policy context of ESPON program.

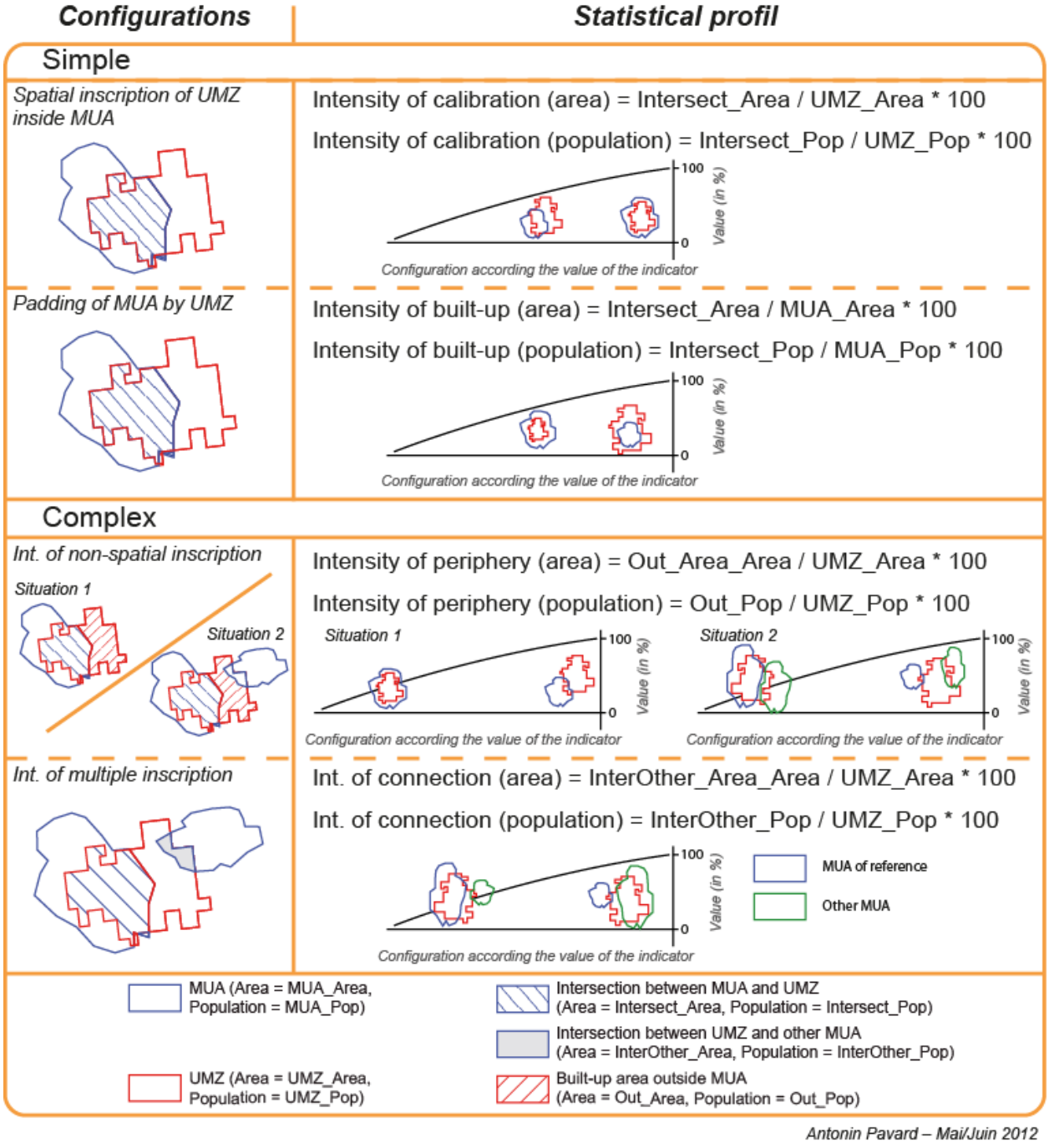


Figure 13: Defining four statistical indicators for evaluation of interoperability

The next step consisted to calculate these indicators for each entity and to summarize them at the level of the classes of an "a priori" typology of spatial configurations (Figure 14) in order to check their heterogeneity, intra and inter-classes. The first two cases (one UMZ in one MUA and several UMZ into one MUA) give very similar results and are both characterized by a good interoperability. The concerned urban objects are mainly located in Sweden, Denmark and south of Spain (Figure 15). The third case (one UMZ partially covering one MUA) is more complex and needs to be further explored as it integrates a diversity of localisation and city sizes. The last case (one UMZ that straddles several MUA) is the most complicated and is related to the polycentric areas and conurbations of the Pentagon (from England to north Italy). In this latter case, there is no interoperability at all between the two databases.

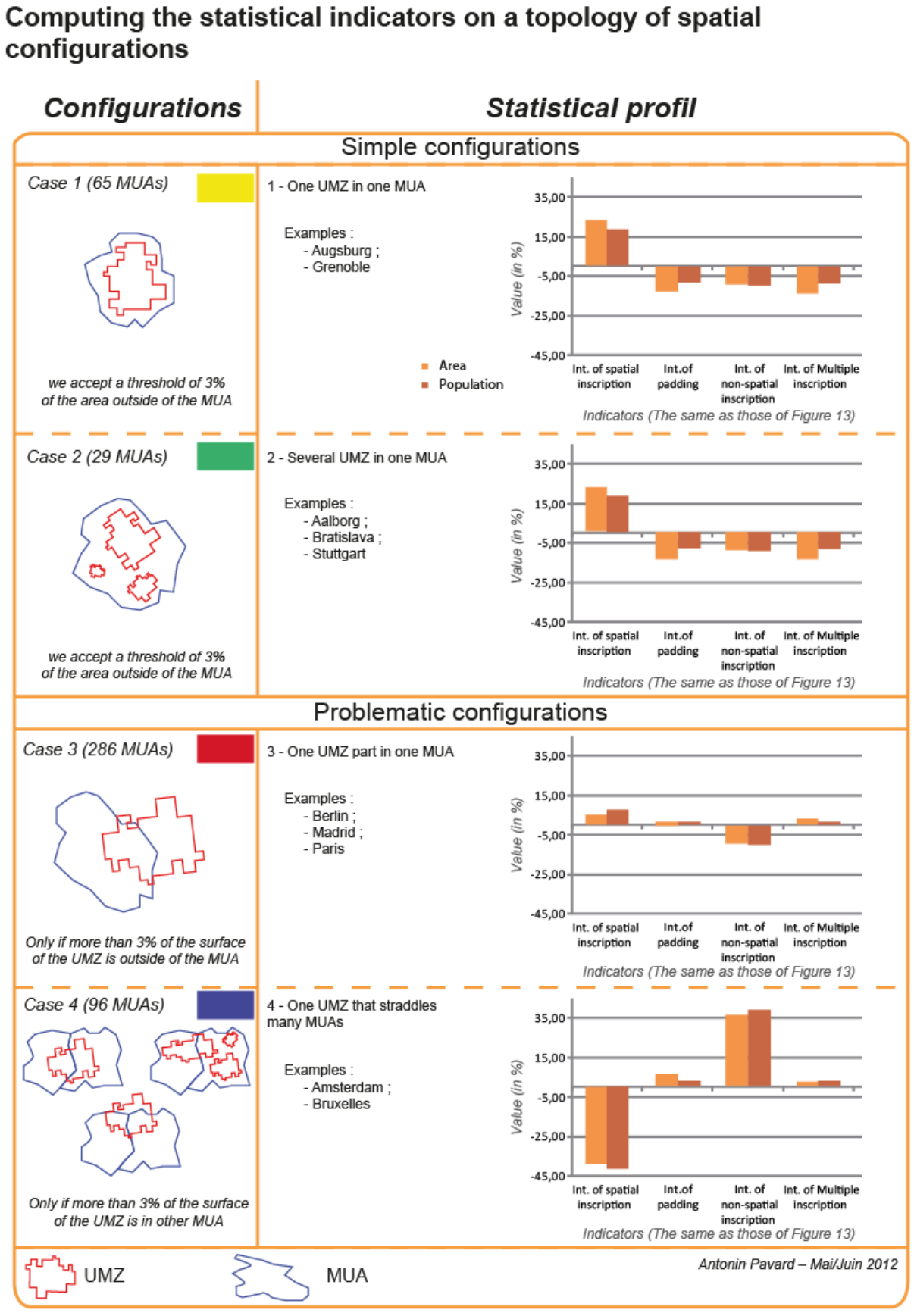
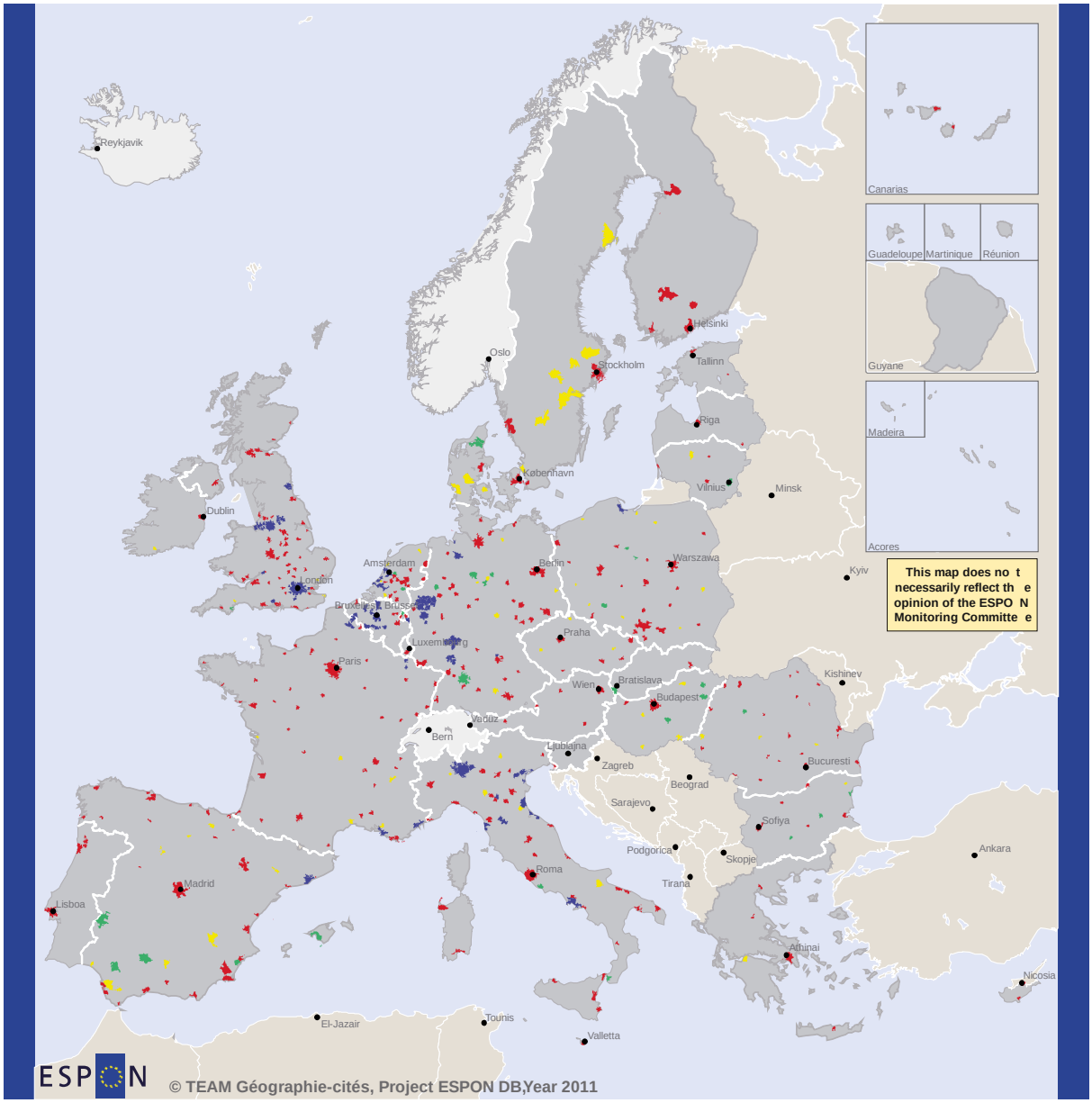


Figure 14: Computing the statistical indicators on a classification of spatial configurations

Typology (a priori) of MUA according bulit-up area



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

Regional level: NUTS 0
Source: ESPON DB, year 2012
Origin of data: IGEAT (MUA V.2011), Joint Research Center (Density Grid V.5), he European Environment Agency (UMZ 2000 V.3
© EuroGeographicsAssociation for administrative boundaries

Configurations of interoperability between MUA and UMZ

- 1 - One UMZ in one MUA
- 2 - Several UMZ in one MUA
- 3 - One UMZ portion in one MUA
- 4 - One UMZ that straddles many MUAs

MUA unselected

The population density grid does not contain any data for Norway, Switzerland and Iceland. Thus we have no UMZ named for these countries. And we can not compare the UMZ with the MUAs of these countries.

Figure 15: Mapping the spatial configurations of interoperability between MUA and UMZ

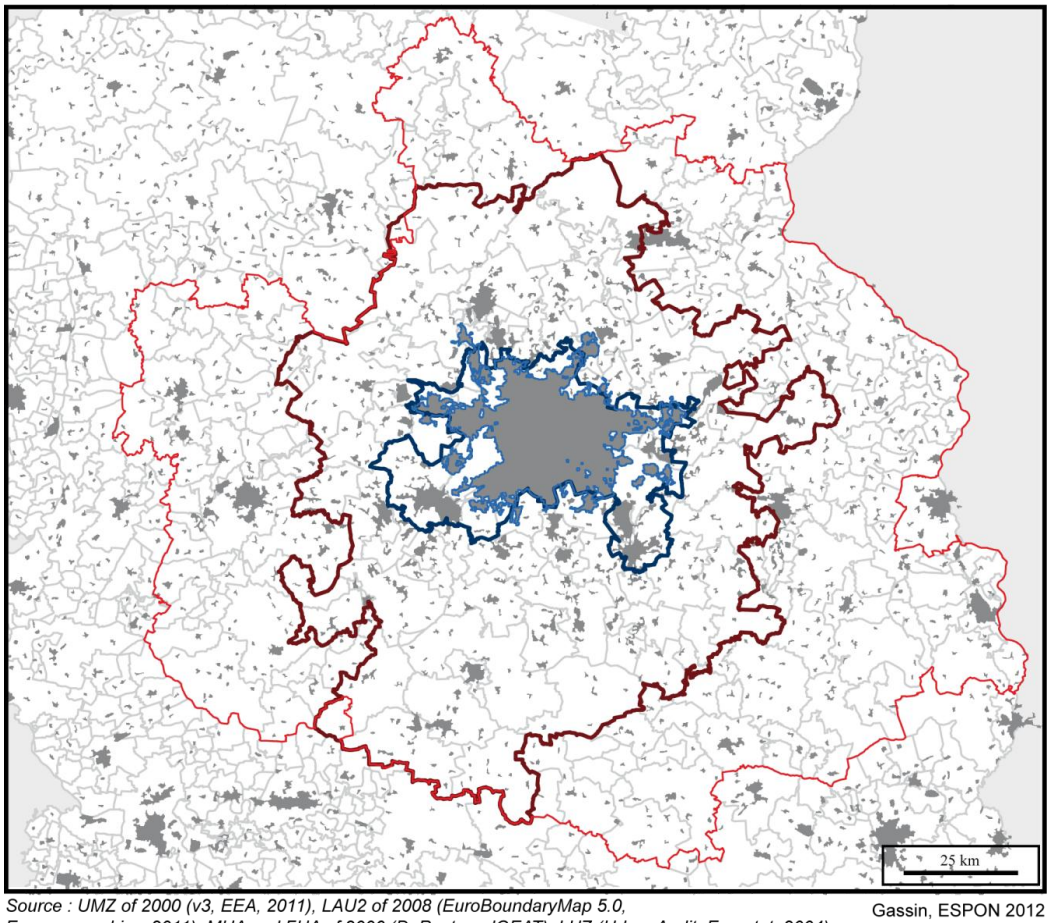
WPB2 Cities - Expertise on FUA construction methods

A/ Data collection for the selected sample zones

Different cities have been selected (see December 2011 Deliverables) among which two are prototypes (Barcelona and Paris) and two others have been explored through a Master work (Prague and Berlin, including a one week field for data collection and interviews with local actors). Other sample zones should be selected in the next few months.

Four types of data have been systematically collected and integrated in a GIS for the four sample zones:

- ✓ Urban delineations (MUA, FUA_IGEAT, UMZ, Urban Audit 2004 LUZ) (see Figure 16 for Berlin)
- ✓ LAU2, LAU1 and SIRE (see Figure 17 for Prague)
- ✓ Transportation networks (see December 2011 Deliverables)



Source : UMZ of 2000 (v3, EEA, 2011), LAU2 of 2008 (EuroBoundaryMap 5.0, Eurogeographics, 2011), MUA and FUA of 2006 (D. Peeters, IGEAT), LUZ (Urban Audit, Eurostat, 2004)

Gassin, ESPON 2012

Figure 16: Data collection of urban delineations, Berlin sample zone

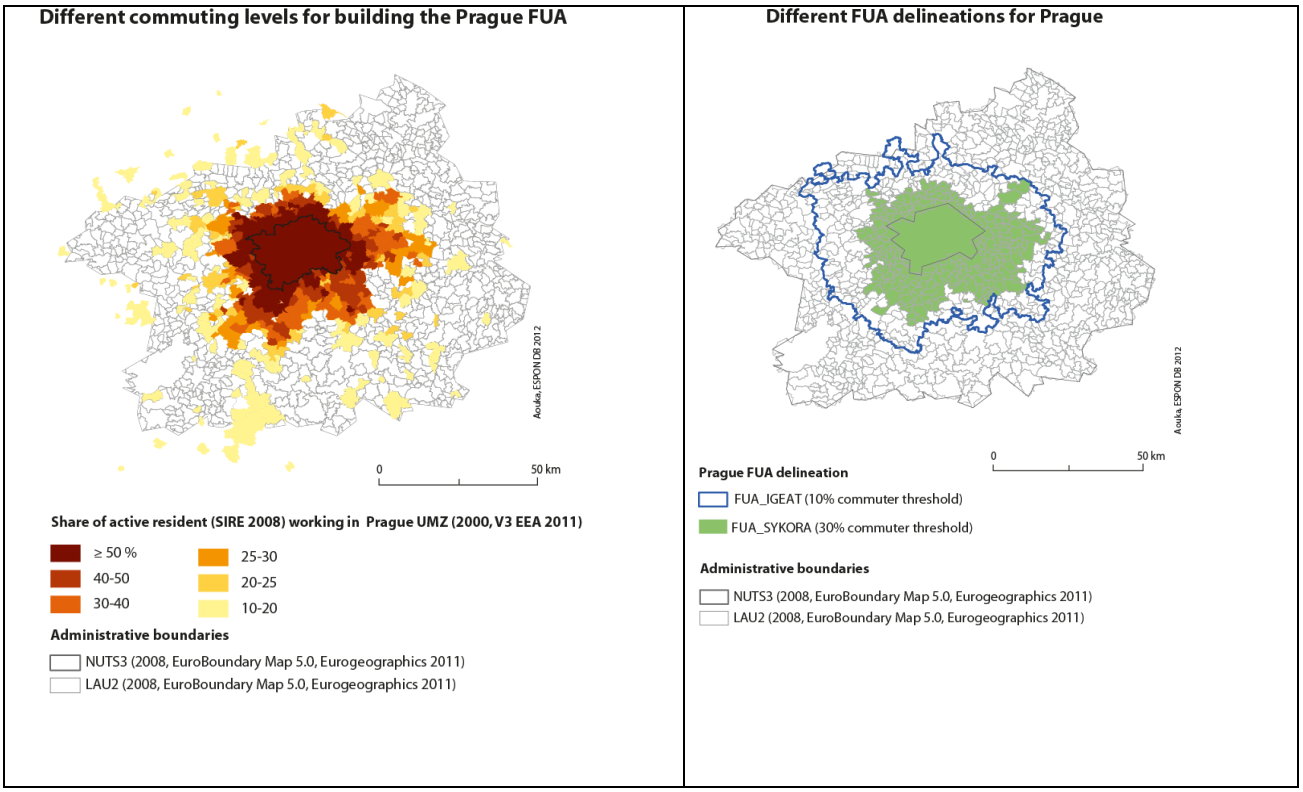


Figure 17: Different commuting levels and FUA delineations for building the Prague FUA

B/ First elaboration of a generic methodology for building the FUAs based on accessibility

Different steps for modelling the one hour isochron have been defined with the Expert on transportation data (MCRIT, Barcelona). The aim is to start from a specific model calibrated for Barcelona (characterized by rich and fine data on multimodal transportation and traffic peak hours) and derive it progressively toward a generic model for European cities, based on the use of Euro Regional Map (ERM) for transportation graph and speeds (Figure 18).

ESPON M4D

Different steps of modelisation of the zone of 1hour accessibility : from a specific model calibrated for Barcelona to a generic model for larger european cities

MCRIT / Géographie-cités

sources / Models	Structural data	Functional data	Results
Objective Model	European data	easy collected data or parameters	
M1 / M2	ERM → ERM*	speed ~ category roads	
M3	ERM*	speed ~ average speeds and peak hours	
M4	ERM* + rail services	speed ~ average speeds and peak hours + commercial speed	
M5	ERM* + rail services	Navstreet + ??????	
Reference Model	Traffic data MCRIT	Hourly data MCRIT	

Figure 18: Modelling steps for the construction of Fuas based on the one hour isochron (work in progress)

WP B3 Grids (UAB)

After launching in December 2011 a survey amongst ESPON projects in order to know first-hand which would be their needs in terms of data integration, a few answers have been received so far. The projects who have answered have not expressed a particular interest in some datasets to be integrated within the ESPON OLAP Cube. However, we will keep the survey open in order to keep collecting ESPON projects needs.
On the other hand, DG-Regio was very much interested in having a broader representation of NUTS breakdowns, plus a time series on population data.
In this phase of the project, and following the general interest of ESPON and the specific demands of DG-Regio, we have proceed with a new update of the ESPON OLAP Cube, launching version 5.0.
The following datasets have been processed in order to convert them into the 1 km^2 grid and build the updated Cube, together with the former data:

- NUTS 1999
- NUTS 2010
- GEOSTAT 2006 POPULATION Data
- Population data between 1990 and 2010, yearly, from the ESPON Database (different sources, compiled by RIATE).

The procedures which have been applied are summarised below:

Update of NUTS breakdowns (1999, 2003, 2006 and 2010)
The first step is to create proximity raster datasets for each year (1999, 2003, 2006 and 2010). The original NUTS data are converted from vector to raster ensuring that output file is aligned with EEA reference grid with LAEA ETRS89 projection.
Afterwards the proximity tool is applied in order to ensure that the coastal grid indexes are taken into account in the selection process. This tool extends zones to the near empty (no data) spaces, covering all the raster dataset with values till the end of the extent.
The proximity raster should be converted then to points. The intention is to have in the end both x and y index values to make possible to know the relation between grid indexes and NUTS Codes.

After some coordinate simplification, the final step is to join the point dataset with the proximity rasters and recreate the index field, exporting to table the final product, including just the grid index rows together with the NUTS code value and COUNT field, that is the number of hectares.

Disaggregation of population from 1990 to 2010
Taking as starting point the proximities created previously, the 1km grid (raster) and the GEOSTAT 2006 population raster[17], all 3 layers are combined in order to have the number of hectares, and population 2006 by NUTS3.
From the output combined raster we get 3 identifiers of each layer. To get the final table it is needed to join the 3 identifiers to the source rasters and recreate the useful fields there, population, grid index and NUTS3 codes.
After having in the same table the 3 fields with their associated COUNT field that will be used later on to count the number of hectares, we export the table to dbf and import it afterwards in Microsoft Access. In this step we add as well the population data from 1990 to 2010 in excel format as a second table within the same Access database.

[17] http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco_Geographical_information_maps/popups/references/population_distribution_demography

We create new fields for each year (1990 to 2010), double format, in the combine table. Then we query both tables to calculate a new field with the percentage that each row (pixel) represents in the total population of the NUTS3 he belongs to. With this calculation we get the proportional percentage that each pixel have in terms of population against the total population of the NUTS3 it is included in. To finally get the disaggregation of the population by each reference year, the same query has to be done, updating the previously created year fields, using the proportional percentage and each total population of the NUTS by year:
(Proportional percentage by 1km pixel * 100) / Total Population of the NUTS 1990
This example is just for year 1990, but this should be done for every year till 2010.

Additional, we have added new time series (new Cube measures) for GDP, Unemployment and Active Population.

WP B4 Local data

Objectives

Working with local data and trying to integrate indicators in the lowest administrative level of geometry is a multiple challenge: methodological, technical and conceptual. From a methodological point of view, the construction of indicators aims to a compromise between coverage (all the ESPON Space) and relevance (zoom-in variables that have added value for scientific and policy analysis). From the technical point of view, we are facing a large amount of data and geometries that implies a special management and that are time consuming, even for simple operations (a ratio, for example). The conceptual aspect is related to the methodological issues and it refers to the strategies one will adopt when working at local scale.

In accordance with the Inception Report specifications, we have adopted a working strategy based on three topics of interest. This strategy allowed us to deliver several indicators and to extract a large amount of new information about the local data integration.
Our first intention (objective) was to **integrate data from the CLC 2006 land use layers in the LAU2 geometry**. This intention is not new, it was already tested as methodology on selected countries from Eastern Europe and it provided some interesting insights regarding the territorial organization at local scale. More recent tests on Spain, Portugal and France also provided good information about the land use, at local scale. This time, we have somehow switched the classical approach. Instead of working with relatively small territories or regions (about 20 000 spatial units) and with all the indicators, we tried to integrate one indicator - arable land from the agricultural areas CLC 2006 for all the LAU2. The operation was a success and the result can be easily mapped. Integrating this kind of data will provide two basic indicators: surface of arable land by LAU2 and share of the arable land in the LAU2 surface. The last one is very useful for two reasons: it allows basic data verifications (if larger than 100 or negative, it is a problem) and it functions as a basis for other indicators construction, such as the location quotients.
The secondary objective was **to explore the possibility to create data about the accessibility to networks in the LAU2 geometry**. The road network density at local scale is an indicator of territorial endowment showing how structural differences between territories are distributed in the ESPON Space. If from the methodological point of view is not very challenging, the solutions to the technical difficulties are more interesting. Without having the intention, these technical solutions lead us to an ad-hoc qualitative classification of the LAU2, based on the intersection between the local spatial units and different types of roads (local roads, regional roads, long distance roads and motorways).

The interstitial spaces that appear on the map could work as a key of lecture for the deficiencies in the transportation networks endowment.
Other issues of interest spontaneously appeared and needed exploration and verification (integration of data from different sources, applying different theoretical models at local scale, refining layers of centroids etc.) The solutions and the walkthrough these problems are explained in the technical report.

Work done

In geography, the term structure is a topic of debate. The source of this debate is triple: the attribute of structure is spatial or territorial (another source of debate because of the confusion we made between them), the structures can be geometrically approached and measured and the word itself has a philosophical and ideological background. Conceptually navigating between these three approaches is challenging. Integrating data from the CLC 2006 in the LAU2 frame means for some geographers to intersect two areal spatial structures, an amorphous one (the arable land) and an administrative one. This is the geometrical (quantitative) approach of the work we done. Concerning the indicators chosen to be integrated, we consider that the agricultural land use is a stake of major importance and that the possibility to zoom-in at local scale is crucial for a better understanding of the territorial organization induced by this economic activity.

Arable land at LAU2 level

The work involved in obtaining this type of indicators should be separated in two parts: calculation of arable surfaces by LAU2 and the construction of the location quotient. The calculation of the surfaces is the output of two separate processes of spatial analysis - an intersection between two layers containing the data (CLC 2006 arable land areas and LAU2 geometry for 2006) and a summarization of the surfaces by LAU2 code. The data is joined as table to a mapping layer and cartographic verification can start. The construction of the location quotient involves some supplementary steps and might be considered a form of multi-scalar analysis. The location quotient was build using two scales of reference, the national and the European level. It functions as a double report between the local context and the superior context. In the classical forms, the value of 1 is the central reference. We have preferred to re-center on 0, subtracting 1 in the formula.

$$LQ = \frac{(\text{Arable land by LAU2} / \text{Surface of the LAU2})}{(\text{Total arable land at national scale} / \text{Total surface at national scale})} - 1$$

The interpretation of the values is simple. The negative values can be translated as relative absence of arable land (local ratio inferior to the national ratio), the positive values reflect trends of concentration and the values close to 0 should be read as proximity to the national or European share. The value of -1 is specific to the LAU2 where the arable land is completely missing.

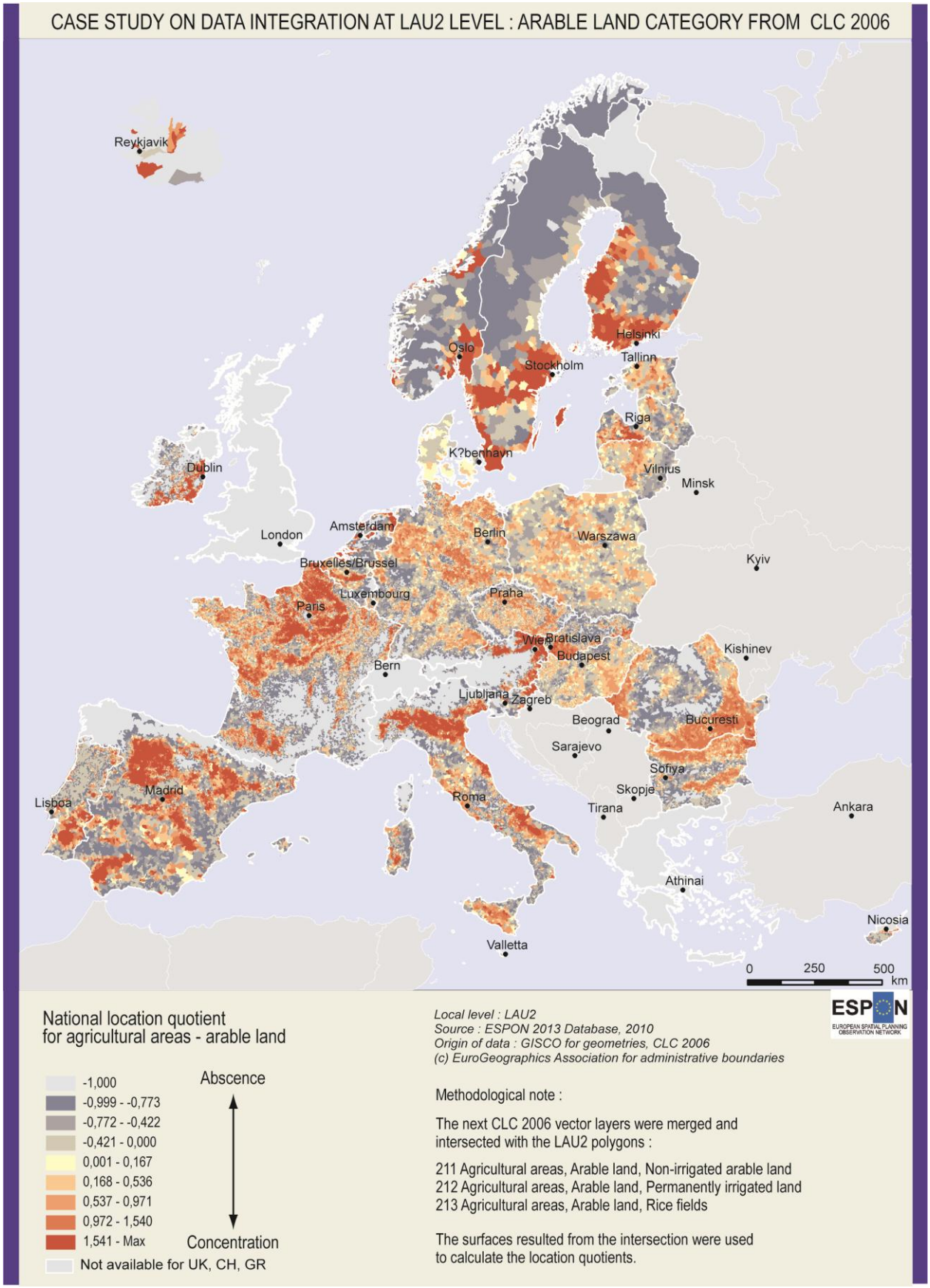


Figure 19 - Location quotient of arable land surfaces in 2006 - national reference

The map shows three major areas of concentration of arable land. Theoretically, the concentration trends should be influenced by natural conditions and ecological constraints. This influence is only partially visible, the social and economic structure of the agricultural practices interfering with the spatial repartition of the indicator. The delineation of high areas of concentration is more visible in the Western countries of the ESPON Space, while the recently integrated states from the East present more moderate values of the indicator. Should this opposition still reflect logics of incomplete transition of the Eastern countries? Despite similar ecological constraints and natural conditions, in Romania, Hungary, Czech Republic and Poland the location quotient of the arable land is smaller.

Road density at LAU2 level

The road's density spatial distribution is explained by several factors - the homogeneity of the populated areas, the national logics of the transportation system and the major opposition between the core of economic activities and the peripheral regions. In areas where the system of populated places is homogeneous, the road density is relatively higher because the spatial repartition of the populated places will encourage multiple interactions between places. As the transportation systems are difficult to implement, the national policies of territorial endowment also explain the road network densities. Countries having sacrificed the road network development in the name of the industrialization, like Romania, present lower density. The opposition between the core areas and the peripheral regions of Europe could be translated in terms of gradient development of networks, explaining why countries like Ireland, partially Spain or Romania show lower densities of the road network.

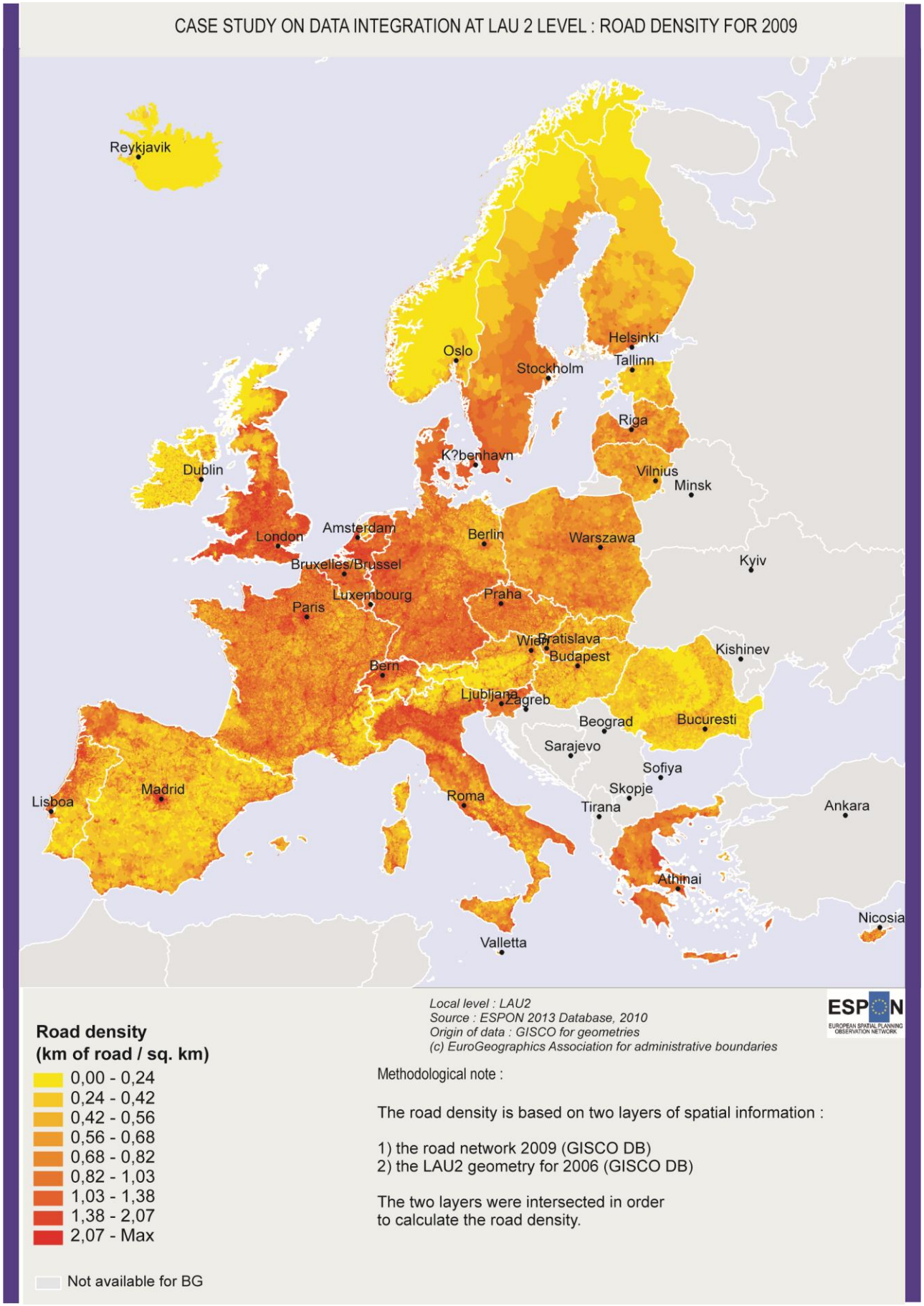


Figure 20 – Estimation of road density in 2009 at LAU2 level

Difficulties

The technical difficulties we met in the CLC 2006 data integration are limited. The main steps and a walkthrough describing the procedures are available in the technical report. Being more a problem of time management than an intellectual challenge, the intersection between the CLC 2006 layers and the LAU2 geometry requires a perfect preparation of the files (proper projections, a good organization of the tables and attention to the field names resulting from the analysis). With more than 100 000 spatial units, even a simple operation (ratio between CLC 2006 surfaces and the LAU2 area) needs patience. Working with large files from CLC 2006 (arable land or forests) complicates the situation. There are two ways to solve these problems: a country by country approach or the aggregated method. The country by country approach has the disadvantage of the multiplication of operations needed (more projections, more intersections, more table joins). Mastering all the steps becomes a problem. We have adopted the second method and we have obtained decent results regarding the time of operation. The mapping process and the cartographic analysis helped us to better understand the spatial logics that organize the repartition of arable land in the studied area. However, the map is also a source of questions. Why there is no representation of arable land in the North-West of Spain? Answering to these questions forced us to integrate other CLC 2006 indicators, using the same methodology. It looks that there is a strong correlation between the lack of arable land and the presence of heterogeneous agricultural surfaces. It could be a matter of photo-interpretation during the creation of the CLC 2006 files or it could reflect the reality. More verification was done using Google Earth, but the results are inconclusive.

More problems were encountered in the moment we integrated data about the road density. The unusual large time of operation forced us to split the road network by the hierarchical code. We have obtained for network files declined as follow: local road, regional road, long distance road and motorways. For each network we have calculated the length of the segments and we have intersected them with the LAU2 geometry. The final summarization enabled us to calculate the road density and to map the results. The four networks also gave us the opportunity to classify the LAU2 by the type of road crossing them. More interesting than the classification is the map presenting the interstitial areas in the ESPON Space, areas that are not connected to superior segments of the road network (long distance roads or motorways).

SIRE Database and related shapefiles

The research of a coherent information at LAU2 linking Eurogeographics geometry and Eurostat LAU2 data is a kind of "*Loch Ness monster*". Since many years, many attempts has been made by DG Regio, research teams and private firms to produce such an information, but without real success or at less without general public diffusion because in most cases a lot of "adjustments" was necessary to obtained an "approximative" coherence. In the ESPON program, many of this databases has been used historically, but they have been difficult to store in the ESPON database because (1) the authors did not wanted necessarily to offer a free use or (2) the metadata was not correct or (3) many empirical problems remains not solved and (4) everybody was waiting for the "final solution" to be at less delivered by Eurostat …

The common workshop between Eurostat and ESPON M4D held in June 2012 has definitively clarified the situation, in the sense that, according to GISCO team, the objective to produce a "perfect" database at LAU2 level is recognize at impossible for the time being. ESPON can take benefit from the work of harmonization done by GISCO and DG Regio, but it is a chimeric objective to expect something perfect in the short term. The recommendation of GISCO team to ESPON M4D project is to be "pragmatic" and do not hesitate to contact directly national statistical offices and/or to use mixture of LAU1

and LAU2 units in order to remove holes. It is important to consider that the regulation rules on LAU levels are more flexible than for the NUTS, Eurostat says, and it is precisely the reason why the previous names of "NUTS4" and "NUTS5" has been replaced by "LAU1" and "LAU2".

Questions and answers derived from the GISCO-EUROSTAT Workshop

> ** Is there an available documentation about the data collected in the EU new member states?*
> *No.*
> ** We are currently working on an extract of the SIRE database formerly realized for the ESPON DB1 project. Is there any update of the SIRE DB since this extraction?*
> *No.*
> ** We saw some discrepancies in the flows data, is there a specific documentation related to this particular part of the SIRE DB?*
> *No.*
> ** What shapefile should we use to map the SIRE DB?*
> *The COMM_CENS_2001 geodatabase associated with a pivot table is designed to be used with the SIRE DB.*

In conclusion, the difficulties results from the fact that the constitution of the SIRE DB was agreed among the EU states on a non-constraint basis. The Member States were not obliged to use harmonized methods to gather and disseminate their data to Eurostat. Beside, the agreement explicitly forbade the dissemination of the DB outside Eurostat. These two facts explain the lack of documentation, the lack of harmonization, and the discrepancies in the SIRE DB. Considering that the next census round for most of the European countries approaches, the SIRE DB contains data from censuses around the year 2001 and the need of an incredible amount of time-consuming harmonization work, there is no push for developing the SIRE DB in Eurostat. Nonetheless, the shapefile named COMM_CENS_2001 was build with the administrative LAU2 limits around the census years in each country. Used with a correspondence table constituted by the GISCO team it is possible to map most of the LAU data from the SIRE DB.

This meeting with Eurostat appears as a milestone and a validation of the hybrid strategy initiated previously by ESPON Database project and followed by ESPON M4D : on the one hand, making the best used of data harmonised by Eurostat and DG regio ; on the other hand, using direct contact with National Statistical Office and trying to create coverage of LAU2 geometries with non census data like CLC or European Population grid.

WP B5 Neighborhood

WPB5 Neighborhood - Regions

Regarding to the agreement defined between the ESPON ITAN project and M4D during a meeting held in Paris the last April 12th (cf annex 1 of the document), some activities of this working package have been exchanged or transferred to the ITAN Project. The aim of this reallocation of activities was to better manage human resource and the timetable of the deliveries for both projects.

In this way, all data collected in Neighboring countries by the M4D Project: Morocco, Algeria, Tunisia, Russia, Belarus, Moldova and Ukraine have been transferred to the ITAN

Project for checking and enlargement of the data collection, thanks to the experts network involved in the ITAN Project.

Consequently, the ITAN Project has also received the Country Reports produced by the M4D Project: It concerns Algeria, Tunisia, Morocco, Belarus, Russia and Ukraine. The ITAN Project is now in charge to improve and complete them.

Now, the M4D project will concentrate its efforts in the following directions:

- **NOMENCLATURES DEFINITION**: The M4D Project takes in charge the harmonized SNUTS definition. The M4D Project will create the SNUTS nomenclatures regarding the information available in the Statistical Yearbook of each Neighbouring countries. It means that the territorial units defined will have a statistical existence in each country. In this way, we can expect several updates of the data collected in the future (census, surveys etc.). All the information gathered regarding to the SNUTS territorial units will be summarized in a dedicated technical report. These nomenclatures will be further checked by the ITAN Project, and used for delivering the different datasets.

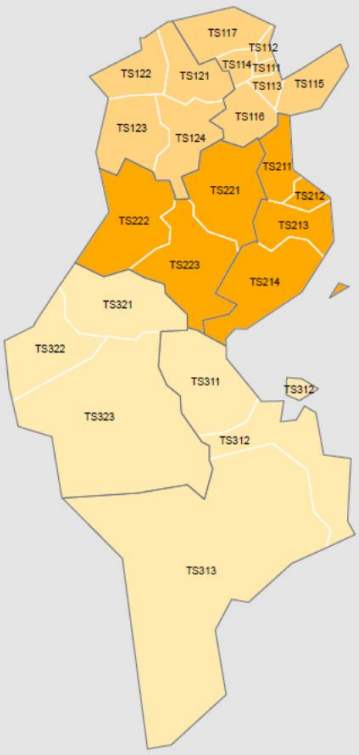
- **GEOMETRIES-MAPKIT:** The ESPON M4D Project will provide geometries at 2 levels of map generalisation (one adapted to the mapping and another one adapted to GIS calculation) for the SNUTS units identified. A dedicated map-kit called "European neighbourhood and ESPON Area" will be produced in that order at the SNUTS2 level at least (SNUTS3 level if possible, depending of the administrative divisions existing in the ENR countries). Regarding to additional map-kits produced by the ITAN Project for case-study data (e.g. Northern neighbourhood, Eastern neighbourhood, South-Eastern neighbourhood, Mediterranean neighbourhood, North calotte case-study, Baltic Sea case-study, Black Sea case-study, Gibraltar case-study and Balkans case-study), the ESPON M4D Project will provide its expertise and – if necessary – its support for ensuring the full compatibility with the ESPON mapping rules.

- **DATA COLLECTION:** It is impossible to deliver in a short time both a complete overview of available information and datasets in ESPON Metadata format for Neighbouring countries. As a consequence, the M4D team is no longer in charge of collecting the core data for the neighbouring countries. This task is now transferred to the ITAN Project. Nevertheless, the ESPON M4D project has engaged to provide the structural characteristics of the SNUTS territorial units. In other terms: total population at two census dates, area and - if possible – urban/rural population according to the countries' administrative definition.

At the moment, this part of the work has been realised for Tunisia, Morocco and Algeria. The resulting table[18], which will be integrated into the ESPON Search Interface during the second semester 2012, contains the following information:

Countries	Identified levels (SNUTS)	Area	Total population (years of data availability[19])
Tunisia	0,1,2,3	YES	1984, 1994, 2004, 2005, 2006, 2007, 2008, 2009, 2010
Morocco	0,1,2,3	Not at level 3	1994, 2003, 2004, 2006, 2007
Algeria	0,1,3	YES	1987, 1998, 2008
Lybia	0,1,2,3	YES	1995, 2006, 2010

[18] M4D_basicsneighb_20120622.xls, enclosed to the First Interim Report
[19] Depends generally of the years of National censuses.



• Territorial Units (T.U.) listing

SNUTS3	NOM	pop 2004
TS111	Tunis	984
TS112	Ariana	422
TS113	Ben Arous (Tunis Sud)	506
TS114	Manubah	336
TS115	Nabeul	694
TS116	Zaghouan	161
TS117	Bizerte	524
TS121	Béja	305
TS122	Jendouba	417
TS123	Le Kef	259
TS124	Siliana	234
TS211	Sousse	544
TS212	Monastir	456
TS213	Mahdia	378
TS214	Sfax	855
TS221	Kairouan	546
TS222	Kassérine	412
TS223	Sidi Bou Zid	396
TS311	Gabès	343
TS312	Médénine	433
TS313	Tataouine	144
TS321	Gafsa	324
TS322	Tozeur	98
TS323	Kebili	143

SNUTS2	NAME	pop 2004
TS11	Nord-est	3627
TS12	Nord-ouest	1214
TS21	Centre-est	2233
TS22	Centre-ouest	1354
TS31	Sud-est	919
TS32	Sud-ouest	564

SNUTS1	NAME	pop 2004
TS1	Nord	4841
TS2	Centre	3587
TS3	Sud	1483

Figure 21 - Territorial units at SNUTS level (Tunisia)

NEXT STEPS: In the same way than for the work already done for the Maghreb countries, the M4D Project has engaged to provide the same information for the ITAN Project study area in December 2012 (figure 22). Anyway, regular contacts will occur until this deadline to make available to the ITAN Project each improvement regarding to these topics.

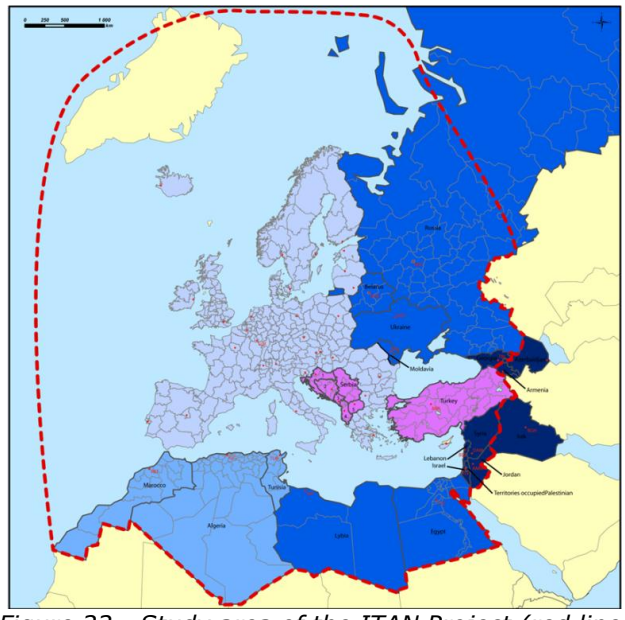


Figure 22 - Study area of the ITAN Project (red line)

WPB5 Neighborhood - Cities and grids

As mentioned in the Inception Report, urban data and grids activities for the neighborhood countries will be launched in the second semester of the year 2012.

Concerning cities, we will first expertise the available urban data bases at world scale, then we will collect and study, as far as possible, the definitions of cities that are currently used in the national census boards of that countries. But the analysis of situation in selected countries of neighborhood (Tunisia, Algeria, Russia, Ukrainia, …) has definitively convinced us that it is not possible, according to the limited ressources available, to try to define precise geometries of urban area in this countries. In many cases, the share of "urban population" is provided in the census for regional units without any details on the detailed geometry of local units that are the building blocks of urban object. And the accession to "urban" can be based mostly on a political decision[20]. In this situation, the best option is to use world databases that provides a relatively good evaluation of the size of most important metropolitan area but are clearly not able to provide information on small and medium cities of these countries.

About grids, a first attempt of crossing the Global Land Cover and the SNUTS geometries will be made, in order to estimate the land cover in the SNUTS nomenclature. The technical progress made with the new OLAP cube on the ESPON territory could be theoretically transposed to the territories of the neighborhood. But we have to keep in mind that the spatial resolution of grids will be less important for many criteria than for ESPON territory. And even when grids looks precise (square of 1 km) they can be in fact the results of very rough estimation made on the basis of informations less accurate than in the case of EU. As in the case of cities, we can conclude that only macroscopic territorial organisation can be really described during the current period of the ESPON program.

[20] In Tunisia, the gouvernorat (equivalent to NUTS3 region) of Monastir is supposed to be occupied by 100% of urban population, which is obviously not the case according to morphological or functional criteria. But it was the place of origin of former president H. Bourguiba.

Working Package C - Networking

WP C1 Platform/ESPON Data Portal

As described in the section entitled "WPA1 Computer", the "Platform" issue has been re-designed to be merged as a subset of functionalities of the Web Application. Please consult the "WIS" (Web Interface Specifications[21]) for a complete review of the content of the platform, now called "the ESPON Database Portal" (figure 23).

Figure 23 – Homepage of the ESPON Database Portal

[21] See the file "webinterfacespecifications.pdf" in the First Interim Report delivery

As described in the WIS, the ESPON Database Portal is structured in 5 main parts:

- **The Welcome page**, where it is possible to find some news related to the M4D Project, the M4D Newsletter and a dedicated interface for researchers or policy-makers (pp. 11-16 of the WIS).
- **The Search interface,** where it is currently possible to retrieve ESPON data following four types of query (by policy, by themes, by project and by keywords) and four types of filters (general, where?, what? and when?). It describes also the solution adopted for zoom-in data, which will be further developed (pp. 17-42 of the WIS).
- **The Resource part,** which contain map kits, data tools, technical documents, territorial data and document of interest (pp. 45-48 of the WIS).
- **The Upload page,** dedicated to ESPON Projects for uploading in a normative way their data in the ESPON Database portal. In this section are also contained all the documentation of interest dedicated to ESPON Projects in order to give support to the data integration the metadata specifications[22], the nomenclature support[23] and guidelines for the data delivery[24] (pp. 49-52 of the WIS).
- **The Help page,** where ESPON users can find the most common questions asked to the ESPON M4D Project (pp. 53-54 of the WIS)

WP C2-3-4 Follow-up of ESPON projects (ESPON priorities 1, 2, 3)

New documents and tools for on-going ESPON Projects

A major criticism addressed to the M4D Project by other TPGs at the ESPON meeting in Krakow (November 2011) was the lack of information regarding to M4D work (What are the latest data available? How shall I deliver my data? What is the last information sent by Eurogeographics regarding to local data? Where it is possible to have a look to the SIRE database? What are the indicators collected by the M4D Project? etc.)

During the last 6 months, between the ESPON meeting in Krakow and Aalborg, the M4D consortium has try to better structure and centralize its work in order to better inform the ESPON Community concerning the ESPON Database requirements (in terms of data delivery) and proposal (in terms of information contained in the ESPON Database).

As a result, some documentation (integrated into the ESPON Data Portal) has been produced to answer to these requirements for on-going ESPON Projects:

a) A newsletter

The M4D Newsletter will be written twice-a-year and distributed at each ESPON Seminar. The aim of the newsletter consists by describing the M4D latest developments (functionalities implemented in the ESPON Data Portal, last data collected…), the ESPON Data news (latest data integrated in the ESPON search interface) and data coming from other institutions, which may be useful for the ESPON Community (Eurostat, Eurogeographics…). The first newsletter delivered in Aalborg has been very well appreciated.

Figure 24: First page of the M4D Newsletter n°1

[22] See the file "MetadataSpecifications.pdf" in the First Interim Report delivery
[23] See the file "NomenclatureSupport.pdf" in the First Interim Report delivery
[24] See the file "HowToDeliverMyData.pdf" in the First Interim Report delivery. These documents are also available under the ESPON Database Portal.

b) Creation of news displayed with RSS feed under the ESPON Data Portal (figure 25)

Information of interest should be delivered also between the ESPON meetings. It is the reason why when a new dataset will be integrated into the ESPON Database, a new map-kit will be available, a new report of interest will be published by other European institutions, or any useful information related to data will be published, a news will be created under the ESPON Data Portal and email will be sent to registered users for notifying the news.

Figure 25– News part of the ESPON Data Portal

c) Tracking tool (figure 26)

Both for the TPG, the ESPON Coordination Unit and the M4D side, a tracking tool has been implemented under the ESPON Database Portal to better follow the data integration process and the result of the syntactic, the semantic and the outlier checks (cf part 6 of the WIS).

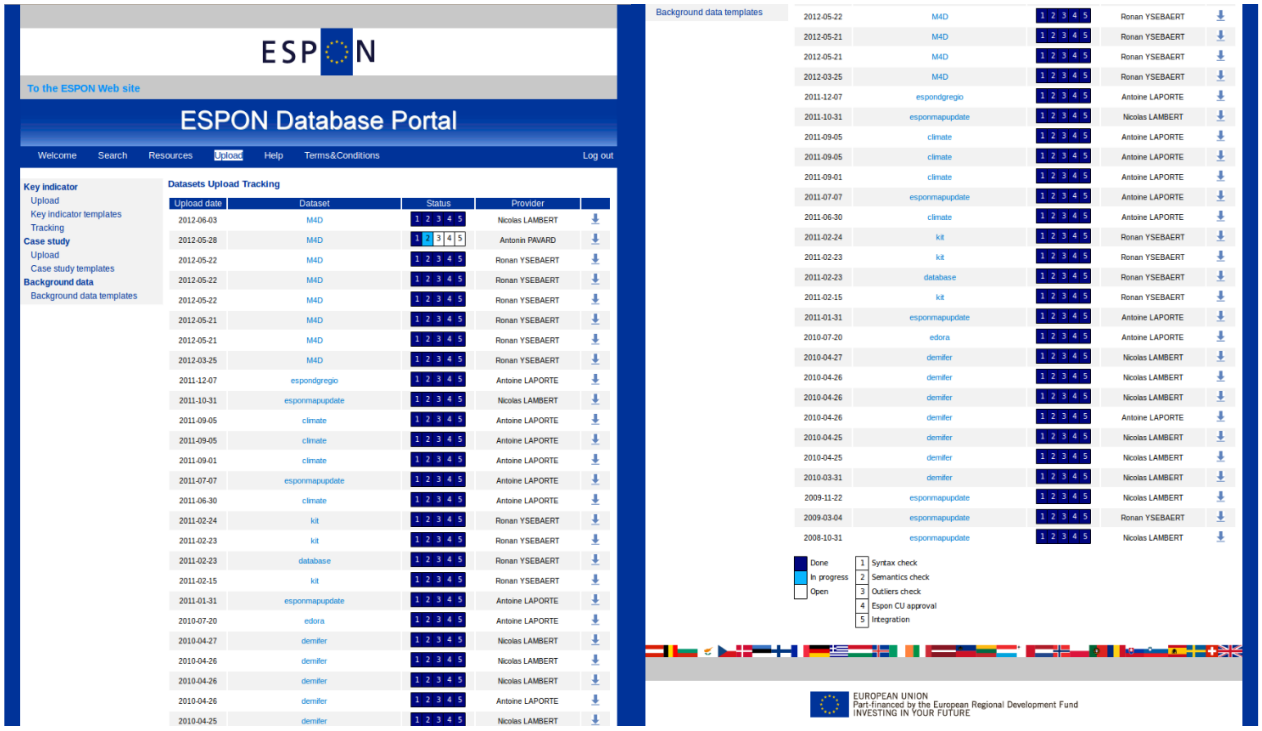


Figure 26 – The tracking tool: a tool for displaying the data integration process

d) How to deliver my data

As mentioned in the M4D Newsletter, there is three ways to deliver data to ESPON M4D: key indicators (previously called "10 best indicators"), zoom-in data (for case-studies) and background data (also called "the entire TPG database").
The "How to Deliver my Data" document, available under the ESPON Data Portal (part upload) in an html or pdf format, gives some guidelines regarding to the data delivery. It explains in details the template to be used by the ESPON Projects depending of the nature of the delivery, the data integration process and gives some advices for filling ESPON data and metadata in a correct way. All the content of this document is fed by concrete situations for a better understanding of the ESPON Database expectations.

Selection of data and recodification of metadata for ESPON Projects which have already finished their activities

Besides this intensive work of communication structuring for on-going ESPON Projects, a complex work of data selection and metadata improvement has been led by UMS RIATE for ESPON projects which have already delivered their data. The aim of this work was to better structure the data integrated in the database. Among other, a specific attention was paid to avoid to integrate duplicated indicators (three total population for the year 2008 for instance), improve the quality of metadata (cf part "semantics" of this report)
Please find below the state of advancement of the data/metadata integration/improvement.

Projects integrated in June 2012 in the search interface of the ESPON Data Portal	Projects which will be integrated in the next steps of the M4D Project (Most of them for December 2012)
• Map Update on Demography and Migration • Map Update on Telecommunication • Map Update on Accessibility • Map Update on Creative Workforce • Map Update on Lisbon Strategy • ESPON-DG REGIO Typologies • DEMIFER Project • KIT Project • Climate Project • EDORA Project • ESPON Database 1 indicators (Corine Land Cover data) • Core indicators from the M4D Project	• ESPON 2006 indicators • ReRisk Project • TIPTAP Project • ARTS Project • ATTREG Project • ESPON DB1 indicators • FOCI project (complex task! – integration of data related to cities) • TIGER project (complex task! – need to develop the nomenclature adapted for world data)

Other datasets derived from ESPON Priority 2 projects will further feed the Zoom-in part of the Search Interface. It implies also a huge work of metadata creation/adaptation for this specific part of the query interface. This part of the work has been done for the TeDi Project and will be extended to the other ESPON Priority 2 projects which have already delivered their data (MetroBorder, SS-LR, CAEE).

WP C5 External Networking

Eurostat

In order to exchange on visualisation tools and methodology related to data management developed both by M4D teams and Eurostat, a meeting was planned.

This meeting was held on the 19 and 20 June 2012.

Many topics have been presented and discussed: Data visualization and online mapping tools; presentations of the UMS RIATE and/or LIG tools (HyperAtlas, AIRE, Subjective Mapper); presentations by Eurostat on their tools and discussions; presentation of the ESPON Database Project: Metadata model and INSPIRE, Query Interface, OLAP Cube and Grid data, Innovative regional indicators, Outlier detection and quality check, Urban data, Local data, Neighbourhood and World data.

Those presentations on each topic were followed by debate and exchanges with Eurostat specialists. Obviously, the presentations are considered as a delivery and will be given to access to ESPON CU for June 2012.

This meeting allowed us to reinforce and deeper provides contacts with Eurostat. As a result we have now a better understanding of how Eurostat is structured and to whom we could ask for specifics topics.

Furthermore, we have then a better expertise on Eurostat data related issues. These expertises will benefit to all ESPON projects. We have also planned to repeat the

experience realized this year and to associate more ESPON TPG' s to the next event, in particular the projects dealing with very specific types of data or geometries.

EEA

No specific meeting has been organised as in the case of Eurostat, because direct contacts are insured by two of the project partners (LIG and UAB) that are involved in EEA activities. Thanks to this team, a lot of information are circulating in both direction between ESPON and EEA, with common benefit of experiences and good practices (OLAP cube technology, HyperAtlas versions…).

OECD

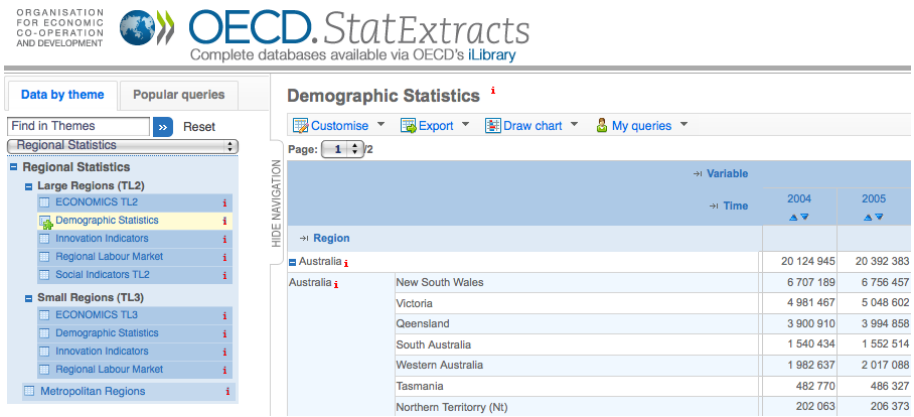
No direct contacts has been launched until know, but this organization has been used as a reference point for many activities developed by ESPON M4D project, in order to adopt their best practices and also offer comparative advantages to ESPON.

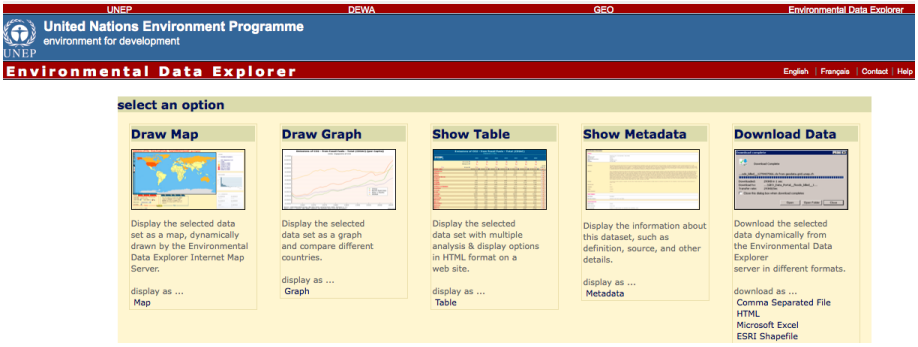
DG REGIO

Many contacts are maintained with DG REGIO, in particular during the ESPON meeting but also concerning the question of the new urban definition elaborated in common by DG REGIO, Eurostat, DG Agri and JRC. It is worth to mention here that the proposal of new indicators based on functional neighbourhood and evaluation of local sigma convergence has been recognised as particularly promising by P. Montfort, representative of DG Regio, during the Aalborg meeting. We suggest therefore to ESPON CU to organize in the future a specific workshop on spatial analysis of economic convergence and spillover effects with DG Regio, OECD and international specialists of spatial analysis and regional science.

UNEP

As OECD, UNEP is firstly used as a reference point in the elaboration of the interface for data access and delivery. In our opinion, OECD has offered very innovative solution for data visualisation (OECD explorer) but this tool is not of interest for ESPON M4D but rather for ESPON Webmapping application. Concerning data access stricto sensu, the application of OECD is not so efficient (very difficult to extract data, limited set of geographical object) and the GEO data portal of UNEP is certainly a better model for ESPON M4D. For this reason, we have involved UNEP team as expert in our project for the development of the database interface but also for the enlargement of data collection on neighborhood.





The OECD StatExtract application which is limited to very few indicators (demography, economy) and very few geographical object does not offer a proper model for the ESPON database which covers much more dimension and different types of geographical object. The GEO data portal of UNEP is much more in line with the current content and ambition of the ESPON program.

Conclusion – From the First Interim Report to the Final Report

Considering, on the one hand, the embedded functionalities in the June 2012 delivery for the Web Application and the state of advancement of the rest of the Working Packages (thematic and networking); and on the other hand, the terms of the contract and expected deliveries mentioned in the Inception (February 2012), please find below an indicative agenda regarding the M4D deliveries until the Second Interim Report (June 2013).

December 2012

- [WPA1] Web Application version 2 with all functionalities.
- [WPA1] Online Metadata Editor for ESPON Priority 1 ("Key indicators") Projects, replacing the "Upload-xls filled template" current alternative.
- [WPA1] SWOT Analysis about OGC (Open Geo Consortium).
- [WPA1] Web Services to access the ESPON Database (one month after the SWOT).
- [WPA1] Update of the ESPON OLAP Visualisation tool.
- [WPA3] Update of the Mapkit on European Neighbour regions covering the ITAN Study area.
- [WPB1] Next steps for the Core Database Strategy (integration of new indicators and description of methods for building innovative indicators).
- [WPB2] Urban databases – Database integration.
- [WPB3] Update of the ESPON OLAP Cube.
- [WPB4] Data, area and share of permanent cultures (location quotient); typology of LAU2 based on the proximity to major rail-road transportation corridors, average distance towards the closest 3 universities.
- [WPB5] Basic data collection for the European neighbourhood at regional level (population, area) and integration in a dedicated part of the Search Interface.
- [WPB5] Draft version of a technical report on European Neighboring cities and – if possible - data collection (depending of the data quality and availability)
- [WPB5] Draft version of the GIS layer (point locations) on European Neighboring cities.
- [WPB5] First draft version of the GlobCover 2009 by SNUTS units and draft proposam for Grid enlargement into the neighborhood.
- [WPC1] M4D Newsletter 2

June 2013

- [TRANSVERSAL] Second Interim Report
- [WPA1] Delivery of the final query and download application
- [WPA1] Maintenance and minor evolutions
- [WPA1] Final version of the ESPON OLAP Visualisation tool and integration of new indicators
- [WPA3] Update and stabilization of the Mapkit on European Neighbour regions in close cooperation with the ESPON ITAN Project
- [WPB1] Technical Report on the empirical feasibility of revised functional divisions at different scales and the Core Database Strategy

- [WPB1] Basic indicators delivery collected with principle of Core Database Strategy for cities and other delineations (water basins, grids…)
- [WPB2] Populating the selected urban databases
- [WPB2] Construction of FUAs for selected cities
- [WPB2] Time series on urban data – data model
- [WPB3] Update of the ESPON Olap Cube
- [WPB4] Distance to the closest FUA; and other indicators provided during the lifetime of the project
- [WPB4] Proposal of alternative functional geometry at local level
- [WPB5] In relationship with the ESPON ITAN Project, revision of the basic data at regional level
- [WPB5] Technical Report – Final version of the data collection on European Neighboring cities
- [WPB5] Final version of the GIS layer (point locations) on European Neighboring cities
- [WPC1] M4D Newsletter 3

In between June 2012 and December 2013

- [WPA1] Regular updates of the Web Application with new implemented features
- [WPC] Continuous integration of ESPON data into the Web Application
- [WPC] Continuous updates of the News part of the ESPON Data Portal
- [WPC] Continuous support to ESPON Projects and ESPON CU
- [WPC] Integration of policy themes, keywords and thematic themes in close cooperation with the ESPON Coordination Unit
- [TRANSVERSAL] Continuous integration of new datasets via the new integration tools

June 2014

- [WPB5]: Final technical report on the ESPON Neighborhood database
- [WPB5]: Final version of the GlobCover 2009 by SNUTS units and by grid
- [WPB6]: Technical report on the evolutions of attractiveness between 2009-2012 and a sample of students located inside and outside EU
- [WPC3]: Implementation of the web-interface dedicated to ESPON Priority 2 projects
- [TRANSVERSAL] Draft Final Report
- [TRANSVERSAL] Continuous integration of new datasets via the new integration tools
- [TRANSVERSAL] Support to ESPON Coordination unit

December 2014

- [TRANSVERSAL] Final Report
- [TRANSVERSAL] Continuous integration of new datasets via the new integration tools
- [TRANSVERSAL] Support to ESPON Coordination unit

Annex 1 – Division of tasks between ITAN and M4D Project regarding to the Neighbourhood working package (WP B5).

ESPON 2013
M4D / ITAN: work division on the European Neighbouring Regions (ENRs) covered by both projects.
Following the discussions/meeting held on April 12th 2012.

The ESPON ITAN and M4D Projects want to work in close cooperation following a clear division of tasks. They have organised a meeting in order to define a precise division of tasks related to:

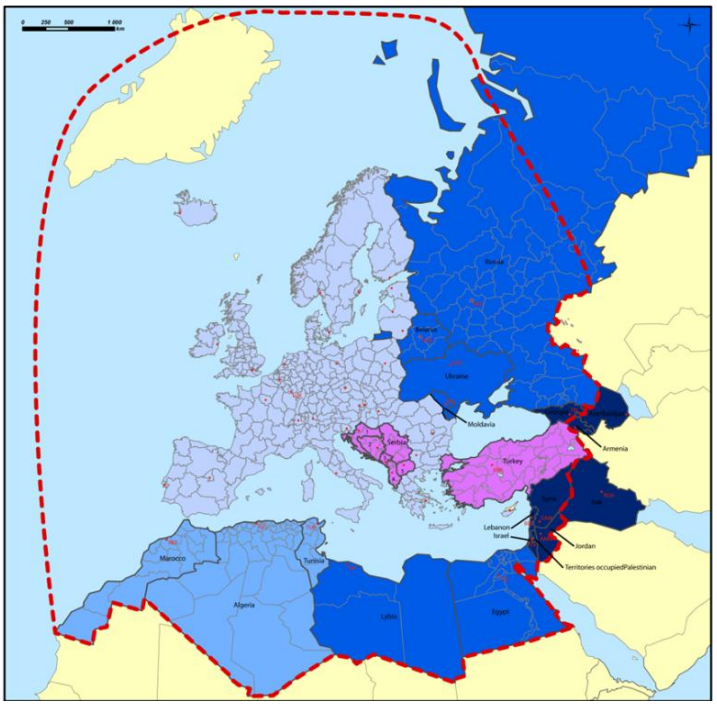
- ***Nomenclature definitions (SNUTS units by countries)***
- ***Geometries and map-kit creation***
- ***Data collection at regional level***

One of the starting points of the discussion was the need for the ITAN Project to have a quick access to nomenclatures and geometries in order to begin the work of data collection in Neighbouring countries. Regarding the agenda defined in the Inception Report of the M4D Project (WP B5, pp 50-55, delivery of data and geometries for Lebanon, Israel, Syria etc. in June 2013 namely), some re-allocation of work was needed by the ITAN Project in order to begin its analysis in quicker.

Taking into account the needs of the agenda and the limited resources available in the ESPON M4D project for the European Neighbourhood (e.g. 55 k€ for the WP B5 and 25 k€ for Regional data + Western Balkans and Turkey), the results of the discussions are the following:

A - Tasks to be done by the M4D Project
The re-allocation of tasks only concerns the regional and Western Balkans parts of the neighbourhood working package (WP B.5.1 and B.5.4). The working packages related to cities (B.5.2) and grids (B.5.3), which will provide an expertise on specific topics are not affected by these discussions. In practical terms, the work of the M4D project on regional data for the European neighbourhood could be summarized as below:

- ***STUDY AREA:*** *The ESPON M4D Project will initiate the collection of the regional information on the European Neighbourhood in a way to fit with the ITAN study area (figure 27), e.g. Greenland, Morocco, Algeria, Tunisia, Libya, Egypt, Moldova, Russia, Belarus, Ukraine, Bosnia-Herzegovina, Serbia, Albania, Kosovo, Georgia, Syria, Lebanon, Israel, Jordan, Occupied Palestinian Territories, Egypt and Macedonia, Croatia and Turkey.*



- Study area of the ITAN Project (red line)

- ***PROVISIONAL NEIGHBOURHOOD DATABASE OF THE ESPON M4D Project:*** *Delivery of all the material already gathered by the M4D Project at the moment: it concerns mainly provisional geometries collected from the GADM database, a selection of basic data collected from National Statistical Institutes websites for the Maghreb area (Algeria, Morocco, Tunisia) and Eastern Europe (Ukraine, Belarus, Russia) and Country Reports on Neighbouring countries. This information may be further checked and/or completed by the ITAN Project following their need or scientific aims. The M4D Project will provide support to the ITAN Project regarding to potential questions related to the content of this provisional database, in particular the metadata and quality check.*

- ***NOMENCLATURES DEFINITION****: The M4D Project takes in charge the harmonized SNUTS definition. The M4D Project will create the SNUTS nomenclatures regarding the information available in the Statistical Yearbook of each Neighbouring countries. It means that the territorial units defined will have a statistical existence in each country. In this way, we can expect several updates of the data collected in the future (census, surveys etc.). All the information gathered regarding to the SNUTS territorial units will be summarized in a dedicated technical report. These nomenclatures will be further checked by the ITAN Project, and used for delivering the different datasets.*

- ***GEOMETRIES-MAPKIT:*** *The ESPON M4D Project will provide geometries at 2 levels of map generalisation (one adapted to the mapping and another one adapted to GIS calculation) for the SNUTS units identified. A dedicated map-kit called "European neighbourhood and ESPON Area" will be produced in that order at the SNUTS2 level at least (SNUTS3 level if possible, depending of the administrative divisions existing in the ENR countries). Regarding to additional map-kits produced by the ITAN Project for case-study data (e.g. Northern neighbourhood, Eastern neighbourhood, South-Eastern neighbourhood, Mediterranean neighbourhood, North calotte case-study, Baltic Sea case-study, Black Sea case-study, Gibraltar case-study and Balkans case-study), the ESPON*

M4D Project will provide its expertise and – if necessary – its support for ensuring the full compatibility with the ESPON mapping rules.

- ***DATA COLLECTION:*** *It is impossible to deliver in a short time both a complete overview of available information and datasets in ESPON Metadata format for Neighbouring countries. As a consequence, the M4D team is no longer in charge of collecting the core data for the neighbouring countries. This task is now transferred to the ITAN Project. Nevertheless, the ESPON M4D project has engaged to provide the structural characteristics of the SNUTS territorial units. In other terms: total population at two census dates, area and - if possible – urban/rural population according to the countries' administrative definition.*

- ***AGENDA:*** *The ESPON M4D Project has engaged to provide all the information mentioned above for December 2012. Anyway, regular contacts will occur until this deadline to make available to the ITAN Project each improvement regarding to these topics.*

B - Tasks related to the ITAN Project

- ***STUDY AREA:*** *The ESPON ITAN Project will study the countries listed in this document Part A / STUDY AREA. The delimitation for the Russian Federation's coverage has to be specified with the ESPON M4D Project team since the whole country will not be studied. The European Neighbouring Region's neighbours will also be included in some of the ITAN Project's analyses, but only on a national scale.*

- ***NOMENCLATURES DEFINITION****: The ESPON ITAN Project is in charge of the nomenclature definition for lower levels' SNUTS starting from the geometries the M4D team will gather (SNUTS 3 if not deliver by M4D, SNUTS 4 and 5 (i.e. equivalent to LAU1 and LAU2) whenever needed – e.g. in the case studies' framework. We will make sure the nomenclature for these levels of analysis is compliant with the ESPON M4D Project's one. The ESPON ITAN Project will hire external experts to get more precise data than what is available online.*

- ***GEOMETRIES-MAPKIT:*** *The map-kit called "European neighbourhood and ESPON Area" will be produced by the ESPON M4D Project and used in the ESPON ITAN Project. However, the ESPON ITAN Project's team will take care of additional geometries whenever needed (e.g. case studies or geographic divisions lower than the SNUTS 2/3 level). Additional map-kits will also be created with the cooperation of both projects' teams to choose the appropriate coverage and projection to display the targeted territories. The ITAN project will need at least 10 specific mapkits.*
  - *5 regional mapkits:*

  The ENR and ESPON
  The Northern neighbourhood
  The Eastern neighbourhood
  The South-Eastern neighbourhood
  The Mediterranean neighbourhood
  - *5 zooms in (for each case study):*

  The Baltic Sea
  The Black Sea
  Gibraltar
  The North calotte
  Three regions in Albania, Macedonia and Greece
  The project partners might also need country map-kits; this will be discussed in due course.

DATA COLLECTION: *the ESPON ITAN Project team is now in charge of the core data collection for the ENR, except population, area and urban/rural population. Infant mortality will be added to the already collected core data for every country.The ITAN team will ensure the coherence of the core database by collecting these data for all neighbouring countries (included in the ITAN project's study area, for Candidate Countries and Croatia, the core data have been partly collected by the M4D team) Desired data (environment, investments, etc.) for the ENR will be collected by the ITAN teams only. Since the collected data so far have been done only through National Statistics Institutes, the ITAN Project team will try to get the core data that was not available online, with the help of external experts. The division of both teams' work is displayed in the chart below.*
The ITAN project will gather very diverse datatype. The ITAN project team will make sure to collect these data following the ESPON DB metadata model so the planned-to-be delivered data will be fully compliant with the ESPON database structure.

- ***AGENDA:*** *the ESPON ITAN Project team will try to gather the majority of core data for all countries' before December 2012 matching the SNUTS levels delivered by the ESPON M4D Project. For other data, the collection phase will not be completed before late 2013.*

		Albania	Algeria	Belarus	Bosnia-Herzeg.	Croatia	Egypt	Georgia	Israel	Jordan	Lebanon	Libya	Macedonia	Moldova	Montenegro	Morocco	Occ.Pal. Terr.	Russia	Serbia	Syria	Tunisia	Turkey	Ukraine
CREATION OF NOMENCLATURES	GEOMETRIES	***M4D team's work***																					
	CORE DATA	**Infant mortality added to the core data list** (cf. M4D country profile)																					
	DESIRED DATA	For all desired data, the **ITAN team will use the same nomenclature than the core data's.**																					
CREATION OF GEOMETRIES	SNUTS0 to SNUTS2/3	***M4D team's work***																					
	Lower scale geometries	The **ITAN team will provide lower scale geometries**, including those **used in the specific case studies**																					
	Geometries of cities	**No geometries will be built but the M4D project team will deliver a report dealing with the urban and the cities delimitation issues.**																					
DATA COLLECTION	Geometry-related DATA	To deliver geometries for all ITAN countries faster, **the M4D team will provide less data than originally planned**. M4D only provides the following data : - The total of population, for two census dates - The land area (for each geometry provided) - Rural/urban population according to the targeted country's administrative definition																					
	CORE DATA	*Work now handled by ITAN**	*Work partly done by M4D*	*Work now handled by ITAN*	*Work now handled by ITAN**	*Work now handled by ITAN**	*Work now handled by ITAN*	*Work now handled by ITAN*	*Work now handled by ITAN*	*Work now handled by ITAN*	*Work now handled by ITAN*	*Work now handled by ITAN*	*Work now handled by ITAN**	*Work now handled by ITAN*	*Work now handled by ITAN**	*Work partly done by M4D*	*Work now handled by ITAN*	*Work now handled by ITAN*	*Work now handled by ITAN**	*Work now handled by ITAN*	*Work partly done by M4D*	*Work partly done by M4D**	*Work now handled by ITAN*
		** The data for Western Balkans and Turkey will be collected accordingly to the first collection performed by the ESPON Database Project Phase 1. For Croatia, Macedonia and Turkey, the data collection will be shared among the two project teams.*																					
	DESIRED DATA	***ITAN team's work***																					
MAPKIT	MAPKIT	**All the mapkits** (for M4D and ITAN projects) **will be built by the M4D team.** **ITAN project mapkits will be created with the collaboration of the ITAN team.**																					

M4D team's work | **ITAN team's work**

– Division of tasks between ESPON M4D and ITAN projects regarding to neighbouring countries

Annex 2 – List of deliveries for June 2012 with associated comments

This document describes the MULTI DIMENSIONAL DATABASE DESIGN AND DEVELOPMENT project list of deliveries for June 2012.

Application

WPA1: Construction of an extended classification covering the three types of criteria defined: statistical, thematic and political challenge: **All the dataset integrated in the search interface follow now this extended classification.**

WPA1: Design and develop JAVA API for metadata
A full description of this delivery is explained within the FIR (i.e. parser for the integration of data and metadata).

WPA1: Update the implementation of the metadata editor (and data check).
The implementation of the metadata editor is expected now for December 2012 as the focus was mainly made on the application. Nevertheless, the Data check process and activity were done.

WPA1: Design and extend the database model for new geographical objects
The "ESPON Nomenclatures Support" specifications document is explaining this expected delivery.

WPA1: Design and create some software for integrating and checking more efficiently data from all ESPON Projects (ESPON DB Admin)
The "Web Interface Specifications" document is explaining this expected delivery.

WPA1: Delivery of a version of the Web application including new functionalities
The "Web Interface Specifications" document is explaining this expected delivery.

WPA1: Update of the query and download application
The web application has been delivered to Info Europe the 30 June 2012 (cf email)

WPA1: Construction of an extended classification covering the three types of criteria defined: statistical, thematic and political challenge.
The p.12 and pp 53-54 of the First Interim Report describes this activity. As a result, it is now possible to query data contained in the database as regard to these topics.

WPA1: Draft version of the ESPON OLAP Visualisation tool
The draft version is available under the following url: http://158.109.174.100/webbi/

WPA2: Web function: Operationalisation of data quality check on platform and reporting procedures
A powerpoint presented in Eurostat (20 June 2012) explains the state of advancement of the work.

WPA2: Technical report: Detecting and Handling Anomalous Data in M4D **TR_Outliers_june2012.pdf Available under the resource part of the ESPON Database Portal**

WPA2: Technical report: Time Series Analysis **Available under the resource part of the ESPON Database Portal TR_Time_Series_june2012.pdf**

WPA3: Stabilized mapkit on Seas (ESaTDOR), final version of the mapkit with NUTS 2010 delineations, Provisional mapkit on ESPON Area and Candidate Countries + European Neighbour Regions (including Bosnia Herzegovina, Albania, Serbia, Morocco, Algeria and Tunisia): **Available under the resource part of the ESPON Database Portal**

Thematic

WPB1: Technical report on preliminary results of the Core Database Strategy.
This technical Report has been delivered in December 2011.

WPB1: Basic indicators collected with principle of Core Database Strategy in NUTS 2006 Division.
All M4D basic indicators are now available under the search interface of the ESPON Database Portal.

WPA1-B2: Based on the delivery in December 2011 by the Géographie-Cités partner, the schema of the ESPON Database will be able to support some nomenclatures of Urban Geographic Objects, as well as a glossary of city names used by the Metadata Editor.
Based on strong relations with the LIG team, an UMZ nomenclature has been created with associated data (using the "ESPON Nomenclature specification"), making possible future queries on urban data and making possible the future integration of other urban nomenclatures (LUZ…). All this information is contained in the zip file "data_UMZ_Named_10K.zip", enclosed to the delivery

WPB3: Update of the ESPON Olap Cube
The version 5.0 of the ESPON Olap Cube has been integrated in the resource part of the ESPON Data Portal.

WPB4: Technical report: strategy for creation of data and geometries at local level
This technical report is joined with the First Interim Report (TR_local_data_june2012.pdf)

WPB4: Data, area and share of artificial and arable surface (location quotient); road density at LAU2 scale
The data used to create maps in the Technical Report will be integrated in the resource part of the ESPON Data Portal in a short term.

WPB5: Country reports for Russia, Ukraine, Belarus, Libya and Egypt
Have been transferred to the ESPON Itan project for check and improvement.

WPB5: Geometries, documentation and data for Tunisia, Morocco, Algeria, Albania, Serbia and Bosnia-Herzegovina
Data and nomenclature are enclosed to the delivery (SNUTS.zip) and will be integrated in the search interface during the end of the year 2012. Geometries are available under the resource part of the ESPON DatabasePortal.

Networking

WPC1: M4D Newsletter 1
Distributed during the Aalborg seminar and available under the ESPON Database Portal

WPC1: Implementation of the M4D Platform
The ESPON Data Portal and its functionalities are the result of the work related to the M4D Platform (cf Web Interface Specifications).

WPC3: Specifications for the data delivery of ESPON Priority 2 projects
The "How to deliver my data" document (available under the upload part of the ESPON Database Portal) answers to questions raised by the integration of zoom-in data.

Dissemination

ECP Belgium (October 2011)

LIG/RIATE: How to use ESPON DATABASE and related tools (HYPERATLAS & OLAP Cube) ? (Ronan Ysebaert and Benoit Le Rubrus)

ESPON Seminar – Krakow (November 2011)

LIG/NCG: ESPON Database (Jerome Gensel and Martin Charlton)
UAB: Territorial Monitoring (Roger Milego)

GISCO Meeting (March 2012)

The ESPON OLAP Cube ppt (Roger Milego)
The M4D Project and the ESPON Database (Claude Grasland and Jerome Gensel)

ESPON Seminar – Aalborg (June 2012)

UMS RIATE: Transformation of regional indicators with functional neighborhood (Claude Grasland)
UMS RIATE-LIG: Presentation of the ESPON Database, the dataflow and the ESPON HyperAtlas - (Jerome Gensel and Claude Grasland)
GEOGRAPHIE-CITES and UAB: Urban delineations and data bases in Europe, ESPON Data Base M4D(Anne Bretagnolle and Roger Milego)

Eurostat meeting (June 2012)

UMS RIATE: Core Database Strategy (Claude Grasland)
GEOGRAPHIE-CITES: Urban Delineation in ESPON and local data (Anne Bretagnolle, Marianne Guérois and Hélène Mathian)
UMS RIATE: HyperAtlas tool (Claude Grasland)
NCG: Outliers' detection (Martin Charlton)
UAB: ESPON OLAP Cube (Maria-José Ramos)
LIG: ESPON Metadata specifications (Jerome Gensel)
UMS RIATE: Neighborhood countries (Ronan Ysebaert)

Annex 3 – Linkage between Core Data and the Search interface.

Framework of Core Data *As it was clearly mentioned in the contract but also the inception report and the first interim report, the M4D project is based on the existence of* ***two pillars*** *in the activity related to data for the ESPON program:*

1) Pillar 1: Storage and diffusion of data produced by other ESPON project is the activity that has focused the most important attention from ESPON CU in the period 2010-2012, in relation with the development of a new interface for data storage and data extraction (briefly said, the activity of Work packages A and C). These data intend to demonstrate the ***added value of ESPON 2013 program as a whole*** *and are characterized by a high level of* ***originality, diversity and ability to answer to a wide range of political and scientific questions****.*

2) Pillar 2: Creation of harmonized data in space, time and geometries is another activity that has been clearly defined in previous report in relation with a "core database strategy". This work which is developed by the thematic group of M4D (Work Package B) try to develop a coherent set of indicators of priori importance that could be available during long period of time (1990-2010), for various scales (World, States,NUTS, LAU …) and last but not least for different types of geographical objects (regional units, cities, grids. These data intend to demonstrate ***the sustainability of ESPON database*** *in the perspective of* ***regular monitoring of territories at different scales****. .*

The **data that are identified as "core" in the database** belong clearly to the second pillar. They are certainly **not the most original** in the ESPON production but they are the **basic building block** of all elaboration of more complex questions (population, activity, land-use). They are also the only variables that can be elaborated on the long term in past and future. And they are also likely to be computed and store at different scales making possible to **elaborate policies of multilevel governance**. Finally, this "core" data can be easily connected with various applications of cartography (web mapping tool) and spatial analysis (Olap cube, Hyperatlas) and they are without any doubt the **common part of all activities developed by the priority 3 of the ESPON program**.

The introduction of a flag "Core data" is therefore something internal for project partners and researchers. The "Core data" has no reasons to be visible for policymakers because it is generally related to count variable (number of inhabitants, number of sq kilometers, amount of Euros) which cannot be directly compare from one region to another. "Core data" (population, gdp, area) are useful to produce "indicators" (Gdp per capita, density of population) but are something clearly different.

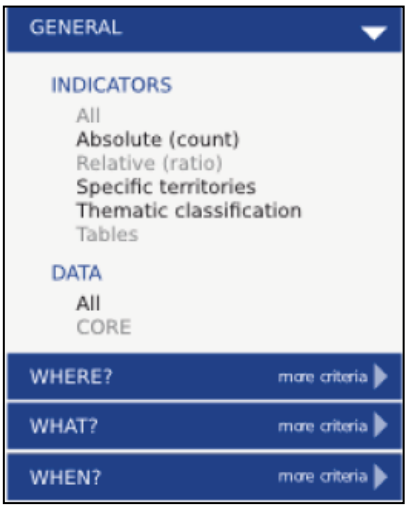
All the M4D Core Data have been integrated into the search interface. But it is obvious that at this stage of the interface development it is not so simple to query these indicators: they are currently mixed with other

ESPON indicators, which are not necessarily available for a large time-period and at each NUTS levels.

To improve the visibility and the user-friendliness of Core data in the Search interface (for researchers only), the M4D Project expect to develop three new and innovative functionalities for the next steps of the project.

1. Filter the query by core data only

In the *general filter* that will be developed in the next steps of the project, it will be to sort indicators by statistical type (count data, ratios) but the user will be able also to obtain only core indicators by clicking on the "Core" item of the data part of the general filter. This functionality is made possible thanks to the metadata field called "Core" in the indicator sheet of the metadata template. When "true" is mentioned, it means that the indicator is a Core indicator.



Indicator Identification		
Code	Name	Abstract
pop_t	Population, total	Total population, both sexes
Policy		
Core	true	
NAT Type	AS	
Theme	Population and living conditions	

Figure 1 – The «core filter » in the query interface which use the core field in the indicator sheet of the ESPON Metadata

2. Download the entire dataset in 1-click

One of the limit of the interface currently available is the difficulty to obtain the complete time-series for a given Core indicator. A user has to select "total population 1990, then 1991, then 1992 etc." to obtain the complete population time-series. To improve the downloading phase and to make it more user-friendly, we propose to download the entire dataset from which the variable has been extracted.
It will appear on the right side of the result page in the search interface (figure 2). Consequently, a user choosing the functionality "*download the entire dataset including this indicator*" will have the possibility in 1-click to obtain the complete time-series.

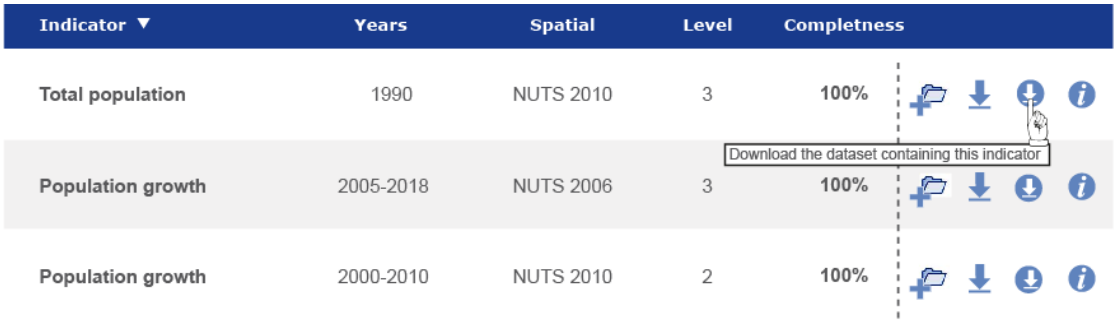


Figure 2 – Proposal of a new downloading option

3. Factorization of indicators by years and level.

At this moment, when querying total population, for example, in the ESPON Search interface, the resulting page returns 80 lines (1x20x4 in fact)! In practical terms, this indicator (1) is available for a time period of 20 years (1x20) and at four NUTS levels (1x20x4). This kind of output makes the navigation in the resulting page very heavy…
To improve this situation, we are studying the feasibility to introduce a kind of factorization of some columns of the resulting page (figure 3). In this way, the total population indicator will appear in a single line.

However, the implementation of this functionality is not conceptually simple. It could have significant impacts on the Where and When filters (for which it is possible to play with these two dimensions). The M4D Project proposes to explore the possibilities which could be concretely implemented for the Second Interim Report.

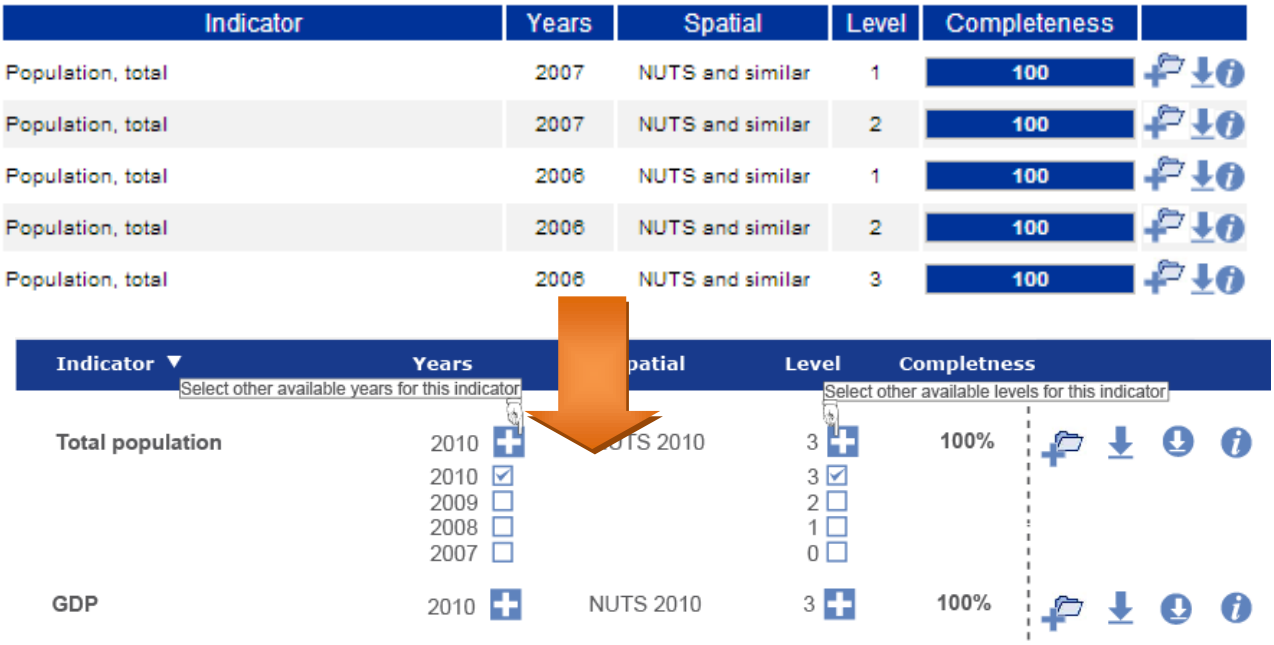


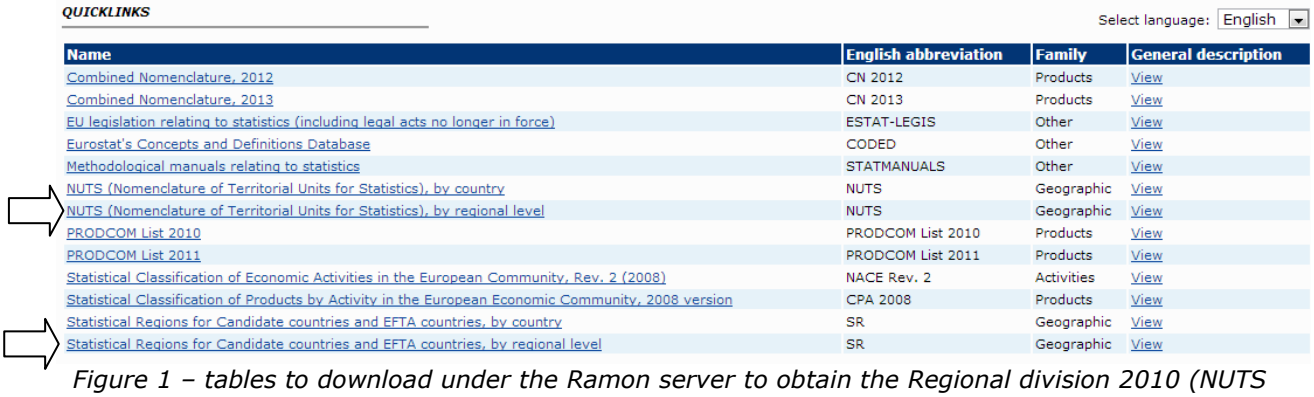
Figure 3 – Factorization of years and level columns for the same indicator

Annex 4 – Guidance on how to deal with the two NUTS nomenclatures in practice (2006-2010)

1. The NUTS 2010 Nomenclature becomes the reference

Since the 1st January 2012, the NUTS classification has been amended. The current version to be used is now the NUTS 2010 classification[25]. It **means that data delivered by ESPON Projects needs now to fit with the NUTS2010 nomenclature**. What does it means for the ESPON Program?

Data must be delivered in the 2010 version, the complete nomenclature can be retrieved under the Ramon Eurostat's metadata server: http://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC (figure 1).



QUICKLINKS

Select language: English

Name	English abbreviation	Family	General description
Combined Nomenclature, 2012	CN 2012	Products	View
Combined Nomenclature, 2013	CN 2013	Products	View
EU legislation relating to statistics (including legal acts no longer in force)	ESTAT-LEGIS	Other	View
Eurostat's Concepts and Definitions Database	CODED	Other	View
Methodological manuals relating to statistics	STATMANUALS	Other	View
NUTS (Nomenclature of Territorial Units for Statistics), by country	NUTS	Geographic	View
NUTS (Nomenclature of Territorial Units for Statistics), by regional level	NUTS	Geographic	View
PRODCOM List 2010	PRODCOM List 2010	Products	View
PRODCOM List 2011	PRODCOM List 2011	Products	View
Statistical Classification of Economic Activities in the European Community, Rev. 2 (2008)	NACE Rev. 2	Activities	View
Statistical Classification of Products by Activity in the European Economic Community, 2008 version	CPA 2008	Products	View
Statistical Regions for Candidate countries and EFTA countries, by country	SR	Geographic	View
Statistical Regions for Candidate countries and EFTA countries, by regional level	SR	Geographic	View

Figure 1 – tables to download under the Ramon server to obtain the Regional division 2010 (NUTS 2010 + EFTA + Candidate Countries 2008)

Next, NUTS nomenclature and EFTA nomenclature has to be combined to obtain the ESPON Regional division in the 2010 version. It is important to remind that key indicators delivered in the ESPON Program must cover at least EU27 + EFTA countries (Switzerland, Liechtenstein, Norway and Iceland) and if possible, Candidate Countries (Montenegro, Macedonia, Turkey and Croatia).

For June 2012, the ESPON M4D Project will make available the complete NUTS nomenclature for ESPON Users, under the Help part of the ESPON Data Portal.

2. NUTS changes between the 2006 and the 2010 version

In practice, the change between the previous NUTS 2006 nomenclature and the NUTS 2010 nomenclature concerns few territorial units: Out of name change and for EU27, it implies 175/1292 territorial units at NUTS3 level (figure 4), 32/266 territorial units at NUTS2 level (figure 5), 7/94 territorial units at NUTS1 level (figure 6) and 1/27 territorial unit at NUTS0 level (figure 7).

[25] More information concerning the history of NUTS change on Eurostat website : http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/history_nuts

Inspired from the ESPON Database Project conceptual framework[26], Eurostat has made available a table of correspondence between the different NUTS version (1999, 2003, 2006, 2010, figure 2).

	Code 20	Code 20	Country	Labels (version 2010 except in *italics = version 2006*) NUTS level 1	NUTS level 2	NUTS level 3	Change	NUTS lev	Sorting order of countri	NUTS 200	NUTS 20
997	ITC4	ITC4			Lombardia			2	11	983	965
1013	ITD1	ITH1			Provincia Autonoma di Bolzano/Bozen		Code change, name change	2	11	996	1030
1015	ITD2	ITH2			Provincia Autonoma di Trento		Code change, name change	2	11	998	1032
1017	ITD3	ITH3			Veneto		Code change	2	11	1000	1034
1025	ITD4	ITH4			Friuli-Venezia Giulia		Code change	2	11	1008	1042
1030	ITD5				*Emilia-Romagna*		Boundary shift	2	11	1013	
1031		ITH5			Emilia-Romagna		New region	2	11		1047
1044	ITE1	ITI1			Toscana		Code change	2	11	1024	1058
1055	ITE2	ITI2			Umbria		Code change	2	11	1035	1069
1058	ITE3				*Marche*		Boundary shift	2	11	1038	
1059		ITI3			Marche		New region	2	11		1072
1067	ITE4	ITI4			Lazio		Code change	2	11	1043	1078
1074	ITF1	ITF1			Abruzzo			2	11	1050	979

Figure 2 – Table of correspondence between NUTS versions proposed by Eurostat

Thanks to this table, it is possible to categorize the territorial unit change between the different versions at each NUTS level (figures 4, 5, 6, 7):

- **Code change**: Implies only a change. The boundaries of these statistical units stay the same between the two NUTS division (displayed in yellow in the maps) *example : EL434 instead of GR434 (case 3).*
- **Merge**: Several statistical units in the NUTS 2006 division are aggregated in a single statistical unit in the 2010 version (in orange); *example: FI181 and FI182 become FI1B1 (case 2).*
- **Split:** a statistical unit in the NUTS 2006 division is divided in several statistical units in the 2010 version (in red); *example: ITF41 and ITF42 become ITF46, ITF47 and ITF48 (case 1).*

3. From the NUTS 2006 to the NUTS 2010

A lot of ESPON Projects have begun to gather data in the NUTS 2006 division. Ideally, they must to be converted in the 2010 version to comply with the European standards. Taking into account the change definition, it is possible to transform NUTS 2006 to NUTS 2010 using different methods. But the existing possibilities depend on the statistical type of indicators created:

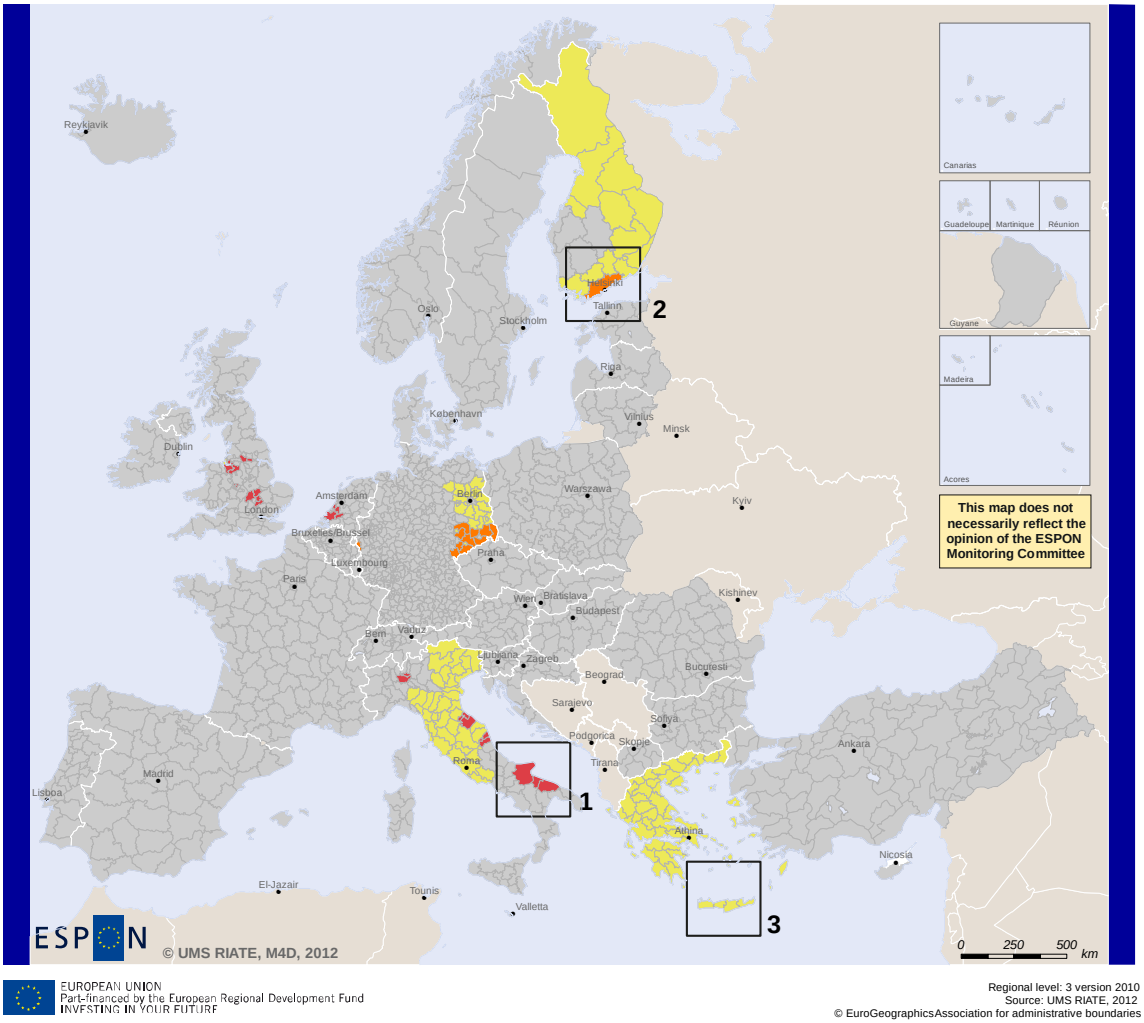
- **Transformation for all type of data**: Change the code by using the Eurostat conversion table. By this way, statistical units displayed in grey and yellow on the maps will be in the NUTS 2010 version.

- **Easy estimation for count data**: The "correspondence folder" in the Eurostat NUTS 2006 – NUTS 2010 table allows to define how aggregate NUTS 2006 in the NUTS 2010 version (figure 3). When using count indicators (total population, GDP in Euros for instance), it is possible to define how add values described in the NUTS 2006 version to obtain NUTS 2010 data. But it is obvious that this methodology is inappropriate for the other data types (typologies, ratios etc.)

[26] Dictionary of NUTS change, available on the ESPON Database Portal : http://db2.espon.eu/db2/resource?idCat=46

Code 2006	Code 2010	Label	Change	Explanation (new = old)
	DEA2D	Städteregion Aachen	New region	DA2D = DEA21 + DEA25
	DED2C	Bautzen	New region	DED2C = DED23 + DED24 + DED2B
	DED2D	Görlitz	New region	DED2D = DED22 + DED26 + DED28
	DED2E	Meißen	New region	DED2E = DED25 + DED27
	DED2F	Sächsische Schweiz-Osterzgebirge	New region	DED2F = DED29 + DED2A
	DED42	Erzgebirgskreis	New region	DED42 = DED14 + DED18 + DED1A + DED1B
	DED43	Mittelsachsen	New region	DED43 = DED16 + DED19 + DED33
	DED44	Vogtlandkreis	New region	DED44 = DED12 + DED17
	DED45	Zwickau	New region	DED45 = DED13 + DED15 + DED1C
	DED52	Leipzig	New region	DED52 = DED34 + DED35
	DED53	Nordsachsen	New region	DED53 = DED32 + DED36

Figure 3 – Explanation on how merge NUTS 2006 values for obtaining NUTS 2010 data



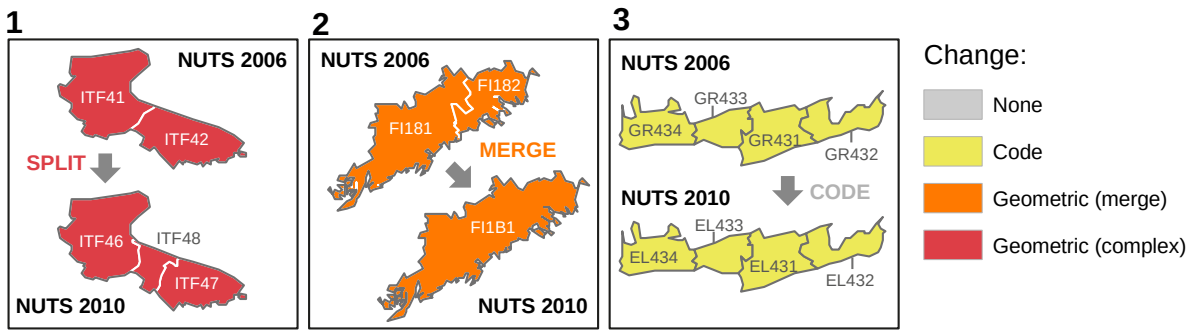


Figure 4 – NUTS3 changes (2006-2010)

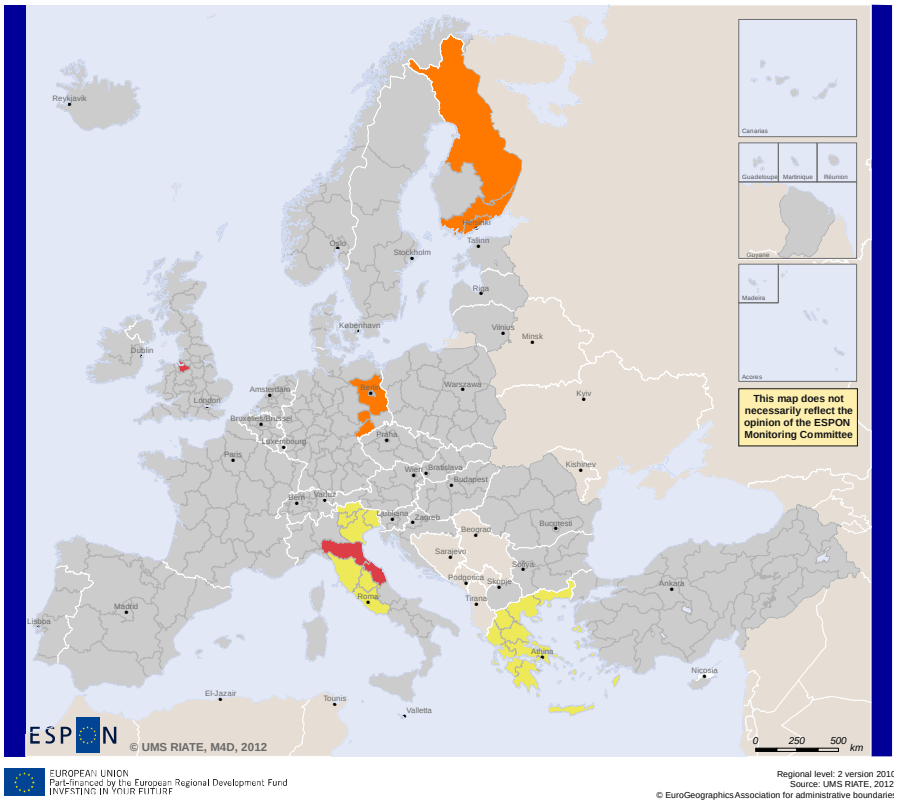


Figure 5 – NUTS2 changes (2006-2010)

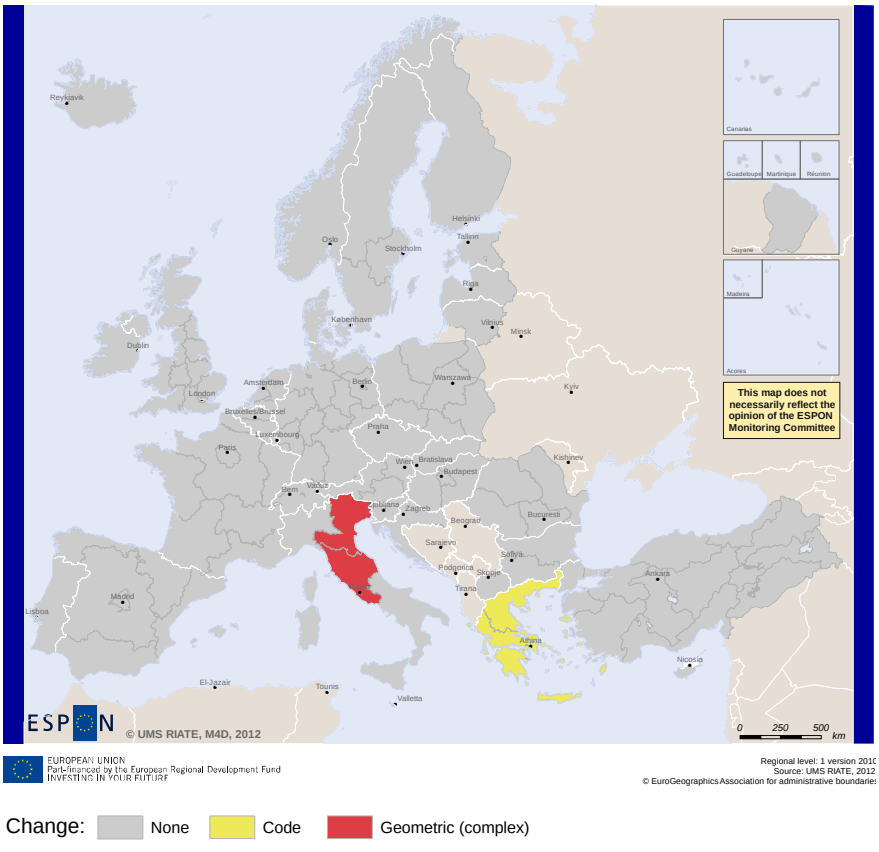


Figure 6 – NUTS1 changes (2006-2010)

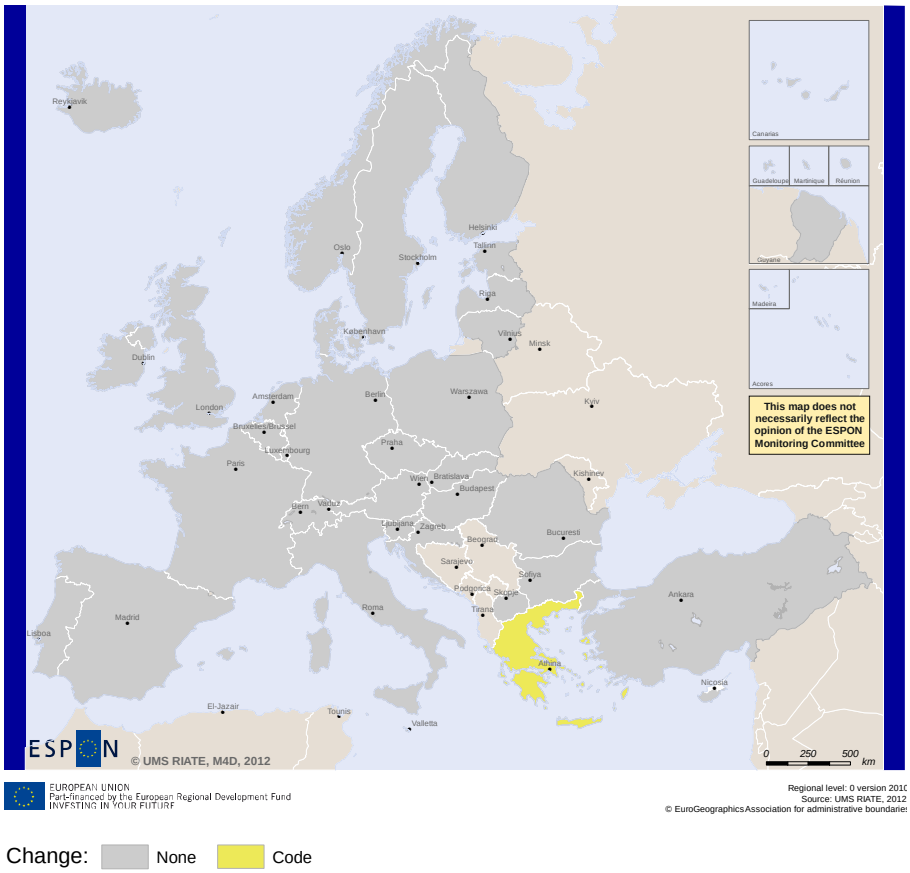


Figure 7 – NUTS0 changes (2006-2010)

- **More complex estimations for territorial units which have experimented a territorial split**: For this kind of figure there is no single solution to propose, but a couple of possibilities depending of the purpose of the analysis. Some of them are described below and refer in a large way to the ESPON Data Navigator 2 framework (ESPON 3.2. project).

➢ ***Possibility 1: Use the ESPON OLAP Cube***

As described in the Technical Report *"disaggregation of socio economic data into a regular grid and combination with other types of data"*[27], the methodology used by the ESPON OLAP Cube (figure 9) makes possible the transformation of count data into different territorial divisions (like NUTS 2010) using a gridding process. Based on this methodology, the last version of the ESPON OLAP Cube[28] proposes the time-serie population 1990-2010 provided by M4D in the NUTS 2006 version in different geographic delinations (NUTS 2010, LUZ…)

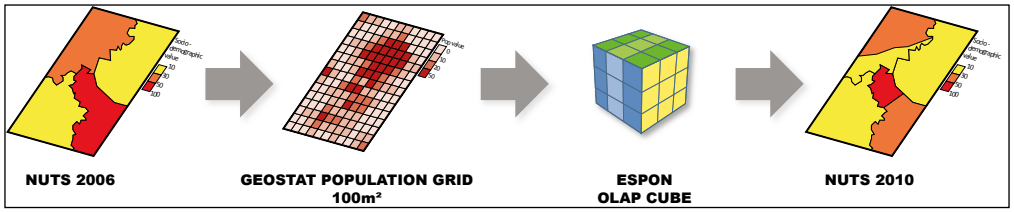
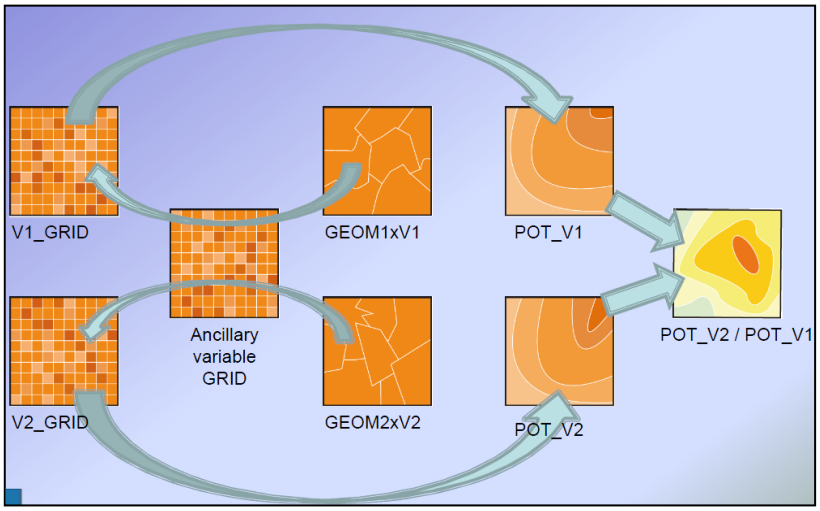


Figure 8 – ESPON OLAP Cube methodology: using gridding for estimating split data

[27] http://db2.espon.eu/db2/resource?idCat=31
[28] http://db2.espon.eu/db2/resource?idCat=21

➢ ***Possibility 2: Use gridding and smoothing***

Another robust possibility consists by using smoothing methods for combining heterogeneous geometries (figure 9a). The first step consists by transforming data coming from two different geometries (like NUTS 2006 and NUTS 2010) in a comparable grid (1 km grid for instance). Then, each grid is smoothed to obtain two potentials which can be made in relation to obtain a smoothed evolution, which is based on heterogeneous geometries. Among other, this methodology was presented during a ESPON workshop on time-serie issues in May 2010[29]. A concrete illustration using this methodology was a map produced a EU-Parliament study[30] where the smoothing procedure was used to ensure compatibility of the non-concordant territorial breakdowns dating from the years 1980 and 1999 (figure 9b).



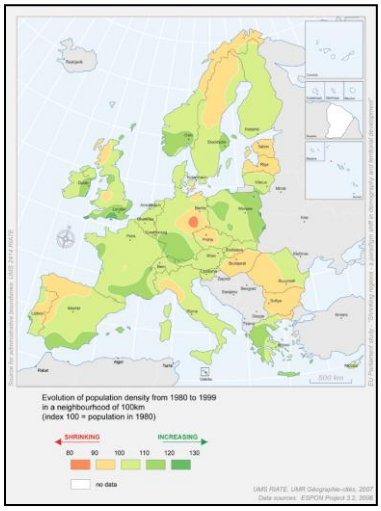


Figure 9a (on the left) – Methodological procedure to obtain smoothing evolutions based on heterogeneous territorial divisions
Figure 9b (on the right) – 100 km smoothed population growth 1980-1999 based on two regional divisions

➢ ***Possibility 3: The smallest common part***

It is also possible to identify the smallest division that two territorial divisions share, which is called the "smallest common part" (figure 10). Applied to the NUTS 2006-2010 example, it implies namely to aggregate some territorial units in Italy, Netherlands and UK at NUTS3 level. The only problem raised by this possibility is the need to create new territorial units, which have generally no official existence.

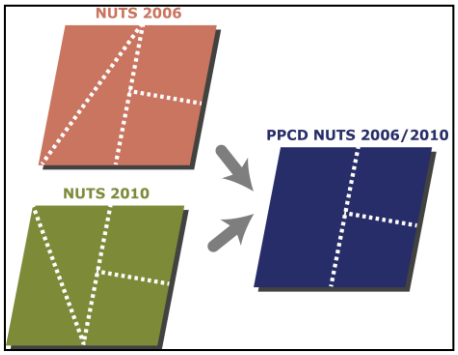


Figure 10 – The smallest common part methodology

[29] Presentations are available at the following url: http://www.espon.eu/main/Menu_Events/Menu_Workshops/workshop1005.html
[30] European Parliament, 2008, Shrinking regions, a paradigm shift in demography and territorial development (http://www.ums-riate.fr/documents/Shrinking_Study_EN.pdf)

- ***Possibility 4: Estimation of missing values thanks to other statistical dimensions***

Based on the development below is largely inspired from the Data Navigator 2 framework produced within the ESPON Project 3.2 (2007). The annex 2 of the Core Database Strategy technical report proposes to formalize in a comprehensive way estimation methods by using the spatial (e), source (s), temporal (t), and thematic (i) dimensions[31] (figure 11). These dimensions can be combined together, and provide indeed various methods for estimating missing values introduced by a change of NUTS nomenclature.

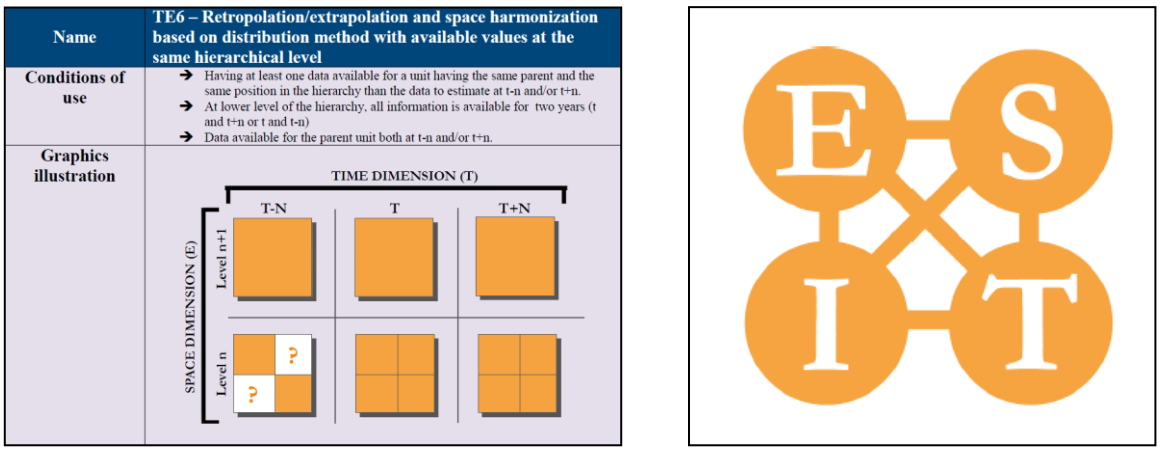


Figure 11 – Output of the Core Database Strategy technical report (annex 2)

4. Next contributions of the M4D Project for this topic

The M4D Project has produced a time-series datasets for some core indicators (cf p 28 of the M4D First Interim Report) in the NUTS 2006 version.

The M4D Project proposes in 2013 to update these core indicators (with the latest data available) and transform them in the NUTS 2010 version.

On top of that, this paper will be improved in the next step of the M4D Project. A technical report will be created on this subject, with concrete examples and solutions. Associated with this technical report, we aim to propose some automatic tools based on the R technology allowing to transfer data from heterogeneous territorial division, which will be complementary to the ESPON OLAP Cube.

[31] http://db2.espon.eu/db2/resource?idCat=32

Annex 5 – Clarification of the work envisaged on cities and urban data

In this part of the FIR, we have concentrated the description of results to a very few set of urban databases: MUAs (IGEAT), UMZ, FUAs (IGEAT) and new LUZ (OECD-Eurostat-Urban Audit) for the integration of specifications (part A), MUAs and UMZ for elaborating indicators of interoperability (part B). The aim was to describe the methodology that would be used on larger datasets that are of interest for the ESPON Program.

It is first worth noticing that, as mentioned in the Inception Report, it is important to ensure the full coverage of the work already developed under the ESPON Database Project Phase I. Accordingly, the FUA and MUA databases developed by IGEAT are now ready to be integrated in the ESPON Database Search interface (cf M4D December 2012 deliveries).

We have also asked Urban Audit for integrating the new LUZ delineations (696 new LUZ larger than 50 000 inh.) and the indicators that have been collected for these functional areas (population, population density, population growth, population share of national value, total GDP, GDP per capita, GDP growth, GDP share of national value, air pollution, CO^2 emissions per capita, urbanised area, urbanised area growth) are available for new LUZ larger than 450 000 inh. (http://www.oecd.org/). The urban nomenclature used by the ESPON project SGPTD (Secondary Growth Poles and Territorial Development in Europe) is currently under the process of integration into the ESPON DB.

I/ Reminder of the context and previous steps

Among the diversity of urban databases available at the European scale, some of them have been selected to be integrated in the ESPON Database: two of them define cities as continuous built-up areas (the MUAs from IGEAT and the UMZ from EEA) and two of them define cities as functional urban areas (the FUAs from IGEAT and the new LUZ from the consortium Eurostat/DG Regio/Urban Audit). Some other urban databases constructed or collected by ESPON projects may also be integrated in the future (for instance, see SGPTD project - Secondary Growth Poles and Territorial Development in Europe - below).

Before the data integration process, some previous steps have been followed and are reminded here:

December 2011 and June 2012: Constructing two *metadata ontology diagrams* of the selected databases (one for continuous built-up areas and one for functional urban areas). Each diagram summaries the different construction processes of the databases in order to enlighten the main similarities and differences and better understand the specificities of each one. They are integrated in the Metadata specifications available in the ESPON Database, so that each user can download the document in order to find some help for choosing one or another stored urban database, depending on its needs.

June 2012 and December 2012: Elaborating a *new nomenclature* for urban objects in the ESPON Database in order to make possible the integration of the databases and allow urban data queries on the web portal.

June 2012: Elaborating and testing *new indicators for evaluating the degree of interoperability* between databases. These indicators allow evaluating if it has some sense to compare some indicators measured for different databases or to enrich a database using the data of another one, and to prepare the elaboration of correspondence dictionaries between entities of each database.

II/ Integration of urban databases built as a collection of LAU2

Most of the selected urban databases are based on delineations that aggregate a collection of LAU2, either at the scale of continuous built-up areas (MUAs) or at the scale of functional urban areas (FUAs_IGEAT, new LUZ), or a collection of NUTS3 (SGPTD urban objects). In these cases, the data integration process is relatively simple as the correspondence with LAU2 is very easy to construct. We have summarized in Table 1 the different fields that have been uploaded or that will be in the next weeks.

Table 1: Fields of the urban databases built as a collection of LAU2 that are integrated in the ESPON Database

	Geometries	Attributes
MUAs_IGEAT	Shape files	Population 2000
FUAs_IGEAT	Shape files	Population 2000 and 2006 Population growth 2000 - 2006 Surface 2000 Density 2000 Gross domestic product (GDP) at current market prices 2006 Gross domestic product (GDP) at current market prices per capita 2006 Percent of persons in unemployment 2006 Share of the unit population in the MUA 2006 NACE (Nomenclature of economic activities; established by the European Community 2006)
SGPTD entities	Shape files	Gross domestic product (GDP) in Purchasing Power Standards (PPS) at current market prices 2000 to 2007 Gross domestic product (GDP) per capita in Purchasing Power Standards (PPS) at current market prices 2000 to 2007 Gross domestic product (GDP) per person employed in Purchasing Power Standards (PPS) at current market prices 2007 Population 2000 to 2007 Persons in employment 2000 to 2007 Share of persons in employment, employed in Financial Intermediation, Real Estate, Renting & Business Activities 2000 to 2007 Employment Rate 2008 Unemployment rate 2000 to 2009 Crude Net Migration Rate 2007 Share of 25-64s educated to tertiary level 2008 European Patent Office Patent Applications 2006

		Potential Accessibility by Air 2006 Total Gross Domestic Product Annual Average Real Percentage Change (2000/2007 – 2007/2009) Gross Domestic Product Per Capita Annual Average Real Percentage Change (2000/2007) Gross Domestic Product Per Person Employed Annual Average Real Percentage Change (2000/2007)
New LUZ	Work in progress (shape files received in December 2012)	Attributes will be required when the work on geometries will be finished

III/ Integration of UMZ database and construction of a correspondence dictionary

The Urban Morphological Zone database has been uploaded in the ESPON Database a few months ago. However, two different tasks were necessary in order to ensure a full and friendly utilisation of the DB by ESPON users.

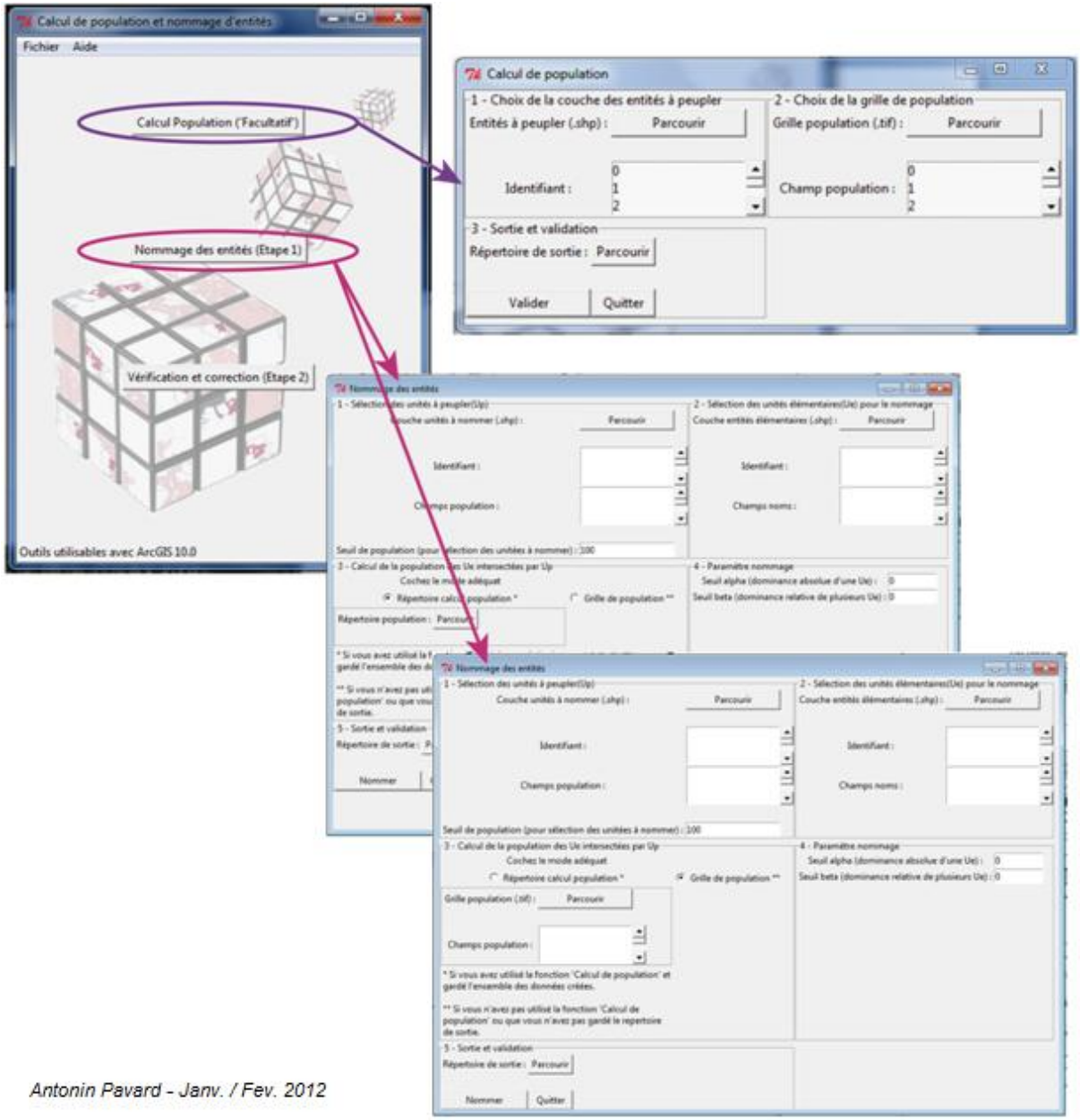
3.1 Improving the methodology for naming UMZ

We just give here some insights summarising the improvements that will be described with more details in a revised version of the 2011 Technical Report "Naming U.M.Z.: a database now operational for urban studies".

First, the different parameters used to calibrate the choice of a simple name or a multiple name have been improved. These parameters consist in population thresholds and are computed in areas that intersect UMZ and reference administrative units (LAU2, LAU1, etc.). Until now, the very simple rule based on the threshold of 50% of the population concentrated in the selected LAU2, had been used in the naming procedure, as in the original method used by the French census board INSEE. We have tested different thresholds, from 30% to 70% and examined systematically the results for a set of well known conurbations, where multiple names are expected. The 40% gives the best results and has been retained for the updated version of the database that is currently available on the ESPON Database. For further details, see the revised version of the technical report, shortcoming.

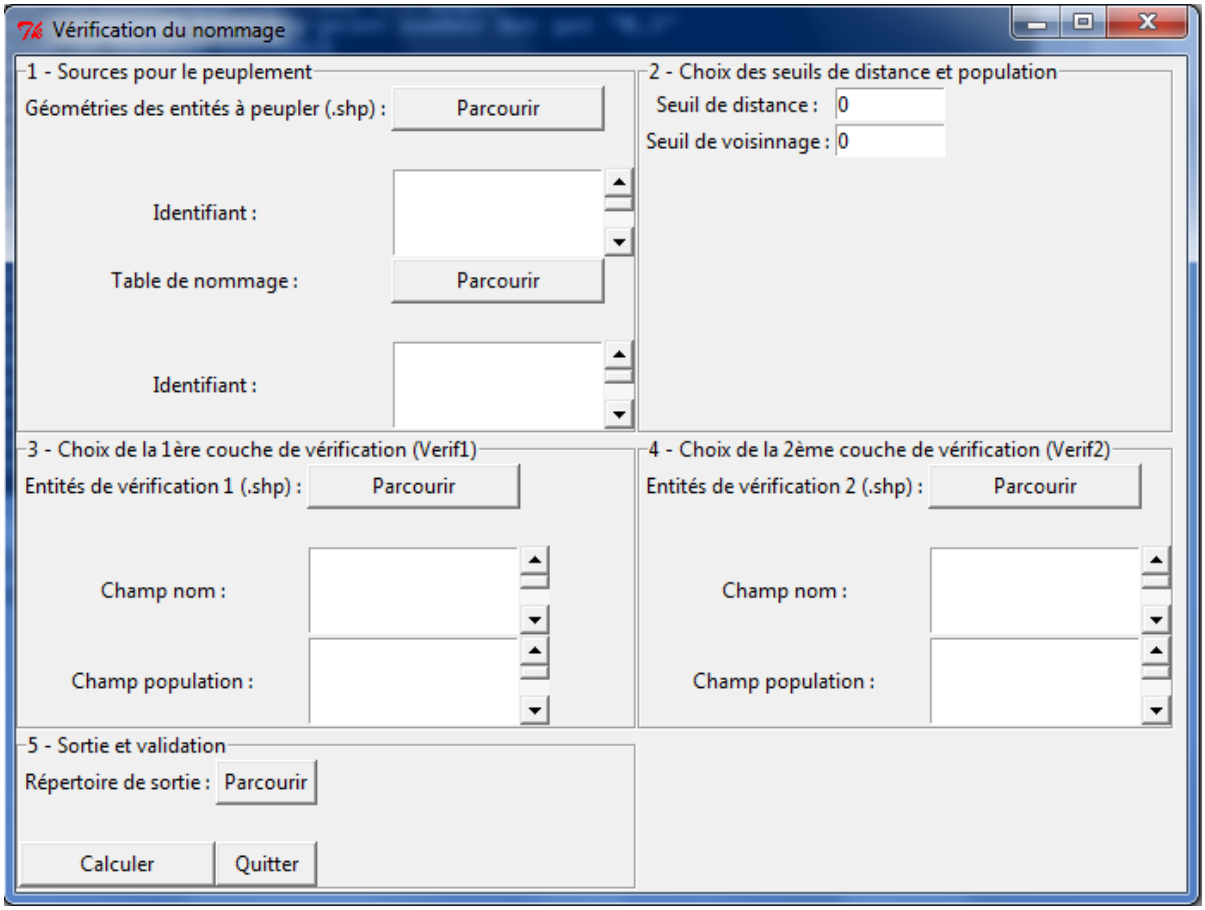
Secondly, the automatic algorithm for giving the names as well as the procedure for checking the results have been improved, in four ways. Instead of using different computing languages and environments (like Python, Post Gray and Post Gis), the procedure only uses now the Python language and the Arc Gis tool. Furthermore, we have simplified the procedure by creating a graphical interface (Figure 1), that allows choosing the geometric layers, the parameters, the directory files etc. In particular, the choice of a specific version of LAU (for example LAU2 2001 or LAU2 2006) may bring very different results in final names for countries that have dramatically reform their administrative boundaries, such as Poland.

Figure 1: The graphical interface for the UMZ naming algorithm



Lastly, the procedure for checking results by comparing the names with Eurostat database (*Geographical names: Settlements*) and Geopolis database (F. Moriconi-Ebrard 1994) has also been automated and has been simplified by the construction of a graphical interface (Figure 2). Two principles are used for the data checking, the research of the closest cities around each named UMZ (the distance thresholds are chosen by the interface user) and the comparison between the closest cities names and the UMZ names (the Levenstein distance is used and consists in calculating the number of modifications – replacement, removal or displacing of letters – necessary to go from the UMZ names to the Eurostat or Geopolis ones).

Figure 2: The graphical interface for checking the UMZ names



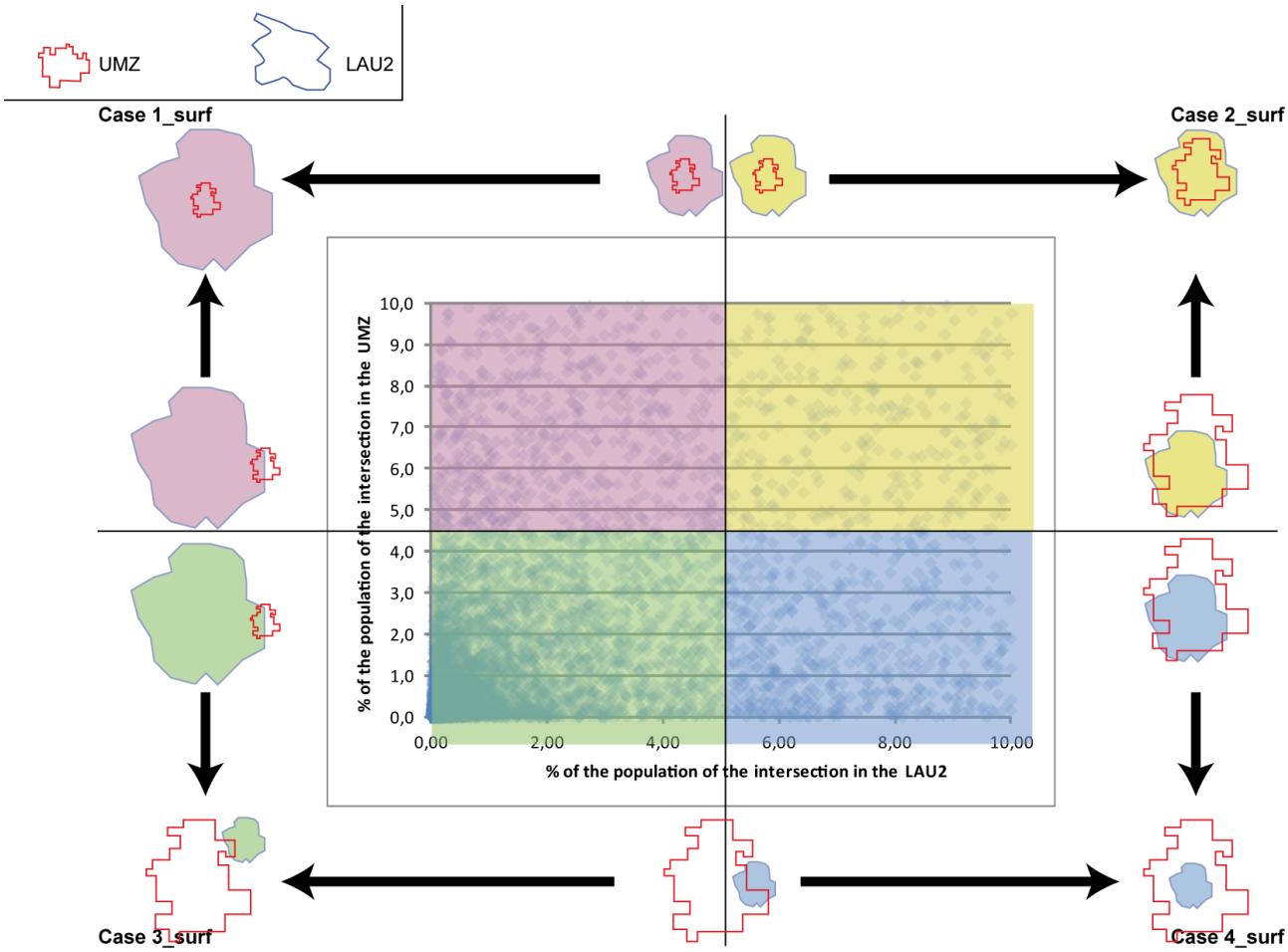
Finally, the whole chain of procedures necessary to give a name and check the results only takes now a couple of hours, instead of two days. This huge reduction of time makes this tool very useful and efficient for the future versions of UMZ database (for example the 2006 UMZ) or for naming UMZ smaller than 10 000 inhabitants (which is the current minimal threshold of population in the ESPON Database). Let us also notice that these tools have been conceived as generic and could be used for naming other types of urban objects. The different menus can be translated in English easily.

3.2 Creating a dictionary of correspondence with LAU2

The UMZ delineations are not based on a collection of LAU2 but on a collection of pixels, as they are created from CORINE Land cover. In order to facilitate the use of socio-economic indicators based on LAU2 (for instance the SIRE database, Eurostat or national census), we have worked on a methodology to create a dictionary of LAU2. This dictionary allows creating a simplified version of UMZ database, called UMZ_LAU2, that gives, for each UMZ, the collection of LAU2 whose aggregation contains this UMZ: one single LAU2 if it is an "isolated city", several LAU2 if it is a "multi-LAU2 city". The LAU2 version that has been used here is "Comm_census_2001", from Gisco 2001. In order to simplify the procedure, we have aggregated all the UMZ inside one LAU2 when they do not cross other LAU2. The total number of UMZ is then reduced from 4185 to 4070.

The first step consisted to intersect the geometries of UMZ and LAU2 and compute, for each resulting area, the population and surface. A total of 23257 intersections have been considered for the UMZ larger than 10 000 inhabitants. Figure 3 and Figure 4 enlighten the variability of the resulting configurations (share of UMZ surface versus share of LAU2 surface, in Figure 3, and share of UMZ population versus share of LAU2 population, in Figure 4). Let us precise that surface approach is just given here as an illustration of the nature of intersections between UMZ and LAU2. The following dictionaries are all based on population qualification of intersection and not on surface ones.

Figure 3: Qualifying the intersections between UMZ and LAU2 in terms of surfaces



Legend:

Case 1: UMZ surface much smaller than LAU2 surface

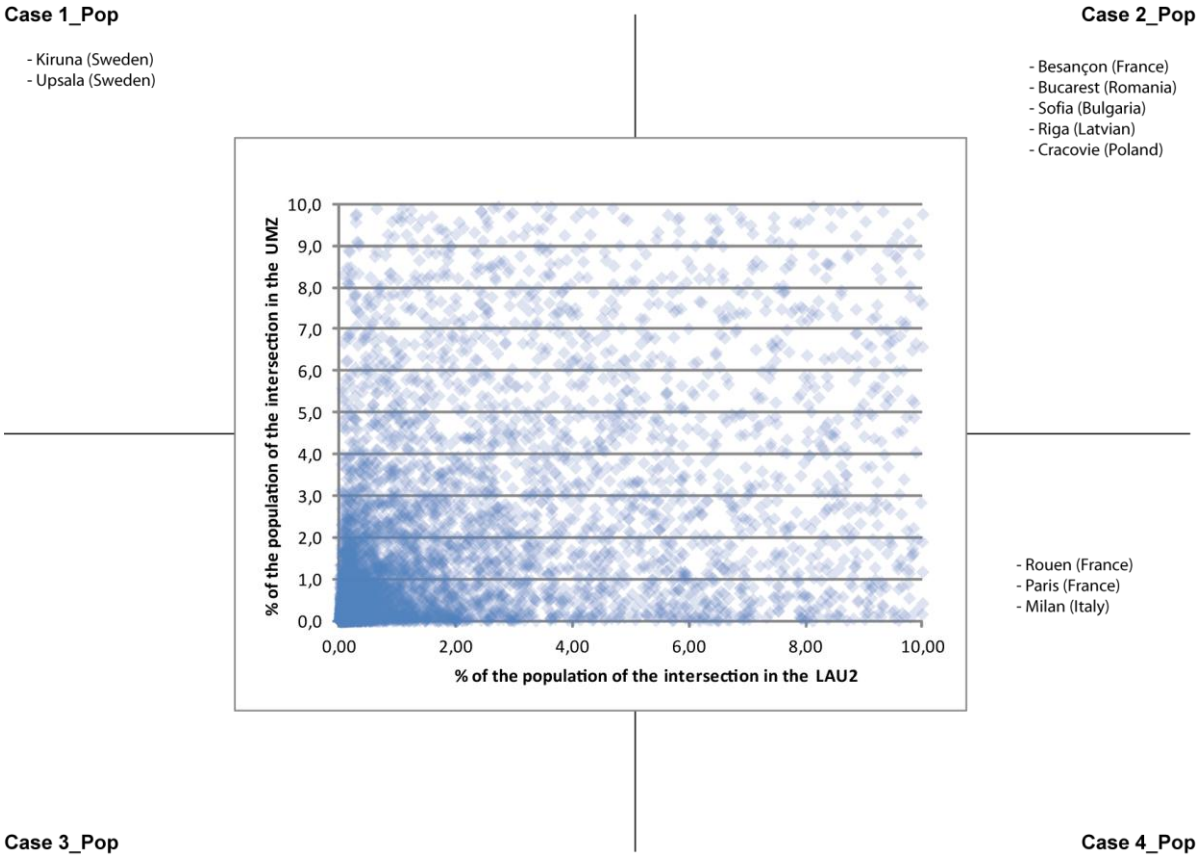
Case 2: UMZ surface very close to LAU 2 surface

Case 3: The intersected surface between UMZ(s) and LAU2(s) is small

Case 4: The intersected surface between UMZ(s) and LAU2(s) is large

The interoperability is high for case 1 and 2, whereas it is more complex in case 3 and 4.

Figure 4: Qualifying the intersections between UMZ and LAU2 in terms of populations



Legend:

Case 1: UMZ population much smaller than LAU2 population

Case 2: UMZ population very close to LAU 2 population

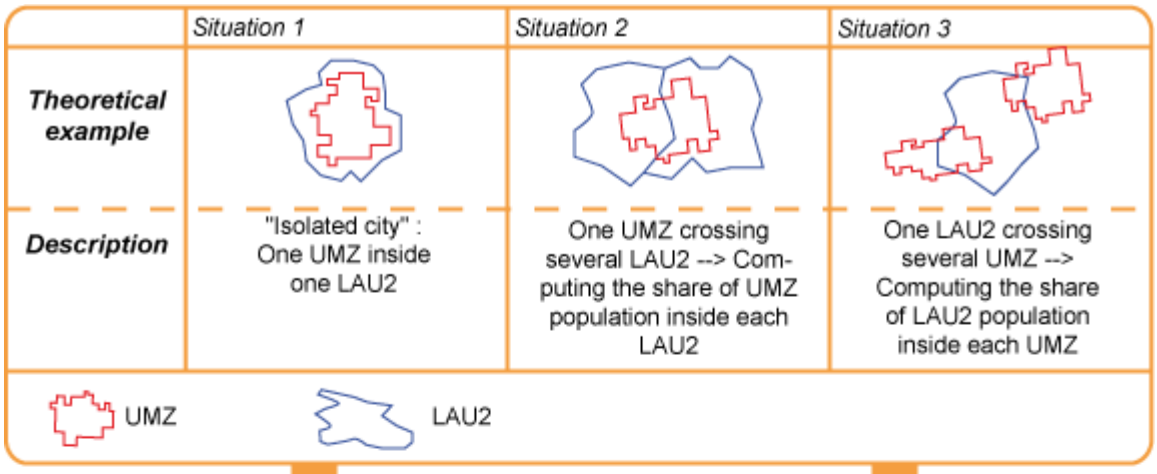
Case 3: The intersected population between UMZ(s) and LAU2(s) is small

Case 4: The intersected population between UMZ(s) and LAU2(s) is large

The interoperability is high for case 1 and 2, whereas it is more complex in case 3 and 4

Facing the complexity of the cases 3 and 4 has leaded us to propose two different final products. The first one, called **UMZ_LAU2_Restricted**, is constructed in a very simple way and its use is very easy, but it is relatively poor regarding the question of harmonization as it does not give the possibility for the user to change any thresholds. The second one, called **UMZ_LAU2_Complete**, gives much more fields and possibilities to the user, who can, for example, vary himself the thresholds in order to adapt them to the average size of LAU2 per country, or in order to study the different configurations of intersections that characterize, for instance, polycentric areas or conurbations. The different variables of these two databases can be separated into three main fields (Figure 5).

Figure 5: Spatial configurations and main situations in the UMZ_LAU2 database



Two dictionaries are produced

Simple : UMZ_LAU2_Restricted *(selection of LAU2 for which at least 50% of the population is laying on the UMZ intersection + allocation of one LAU2 to the major UMZ when it crosses several UMZ)*

ID_UMZ_Agr	Country	NAME_Up	ID_LAU2	Type_UMZ	P_Pop_Int_Lau2 (>50%)	Nb_LAU	Flag_Inf_30
1	AT	Ansfelden	AT41002	0	95,25	1	0
2	AT	Bad Ischl	AT40703	0	92,39	2	0
2	AT	Bad Ischl	AT50336	0	56,05	2	0
3	AT	Bad Voslau // Kottingbrunn	AT30603	0	88,41	4	0
3	AT	Bad Voslau // Kottingbrunn	AT30618	0	95,40	4	0
3	AT	Bad Voslau // Kottingbrunn	AT30620	0	92,84	4	0
3	AT	Bad Voslau // Kottingbrunn	AT30635	0	92,89	4	0

ID_UMZ_Agr : *ID of the aggregated UMZ*
Country : *Country code of the UMZ*
NAME_Up : *Name of the most populated UMZ among the aggregated one*
ID_LAU2 : *ID of the LAU2 (COMM_Census2001)*
Type_UMZ : *Type of the UMZ (Isolated city or multi-LAU2)*

P_Pop_Int_Lau2 : *Share of the intersection population among the total LAU2 population*
Nb_LAU2 : *Number of LAU2 intersected by the UMZ*
Flag_Inf_30 : *Boolean information based on the field "P_Pop_Int_LAU2" (If value is < 30% : flag = 1 ; if value is >=30% : flag = 0). Concerns isolated cities.*

Elaborate : UMZ_LAU2_Complete

ID_UMZ_Agr	Country	NAME_Up	ID_LAU2	Type_UMZ	P_Pop_Int_Lau2	P_Pop_Int_UMZ	Nb_LAU2	Flag_Inf_30	Classif_P_Pop	RankInf50	Flag_Sup50	RankSup50	Rank_LAU2_Sim
1	AT	Ansfel	AT41002	0	95,25	97,74	3	0	1	1	1	1	1
1	AT	Ansfel	AT40101	0	0,18	2,27	3	1	10	2	0		2
1	AT	Ansfel	AT41019	0	0,00	0,00	3	1	10	3	0		1

P_Pop_Int_UMZ : *Share of the intersection population among the total UMZ population*
Classif_P_Pop : *Classification of the intersections according the field "P_Pop_Int_LAU2"*
RankInf50 : *Rank for each intersection in the UMZ according the field "P_Pop_Int_UMZ"*

Flag_Sup50 : *Boolean information based on the field " P_Pop_Int_LAU2" (If value is < 50% : flag = "" ; if value is >=50% : flag = 1).*
RankSup50 : *Rank for the intersections in the UMZ according the field "P_Pop_Int_UMZ" and if the Sup50 = 1*
RankLAU2_Sim : *Rank for UMZ crossing one LAU2 relative to the share of population*

3.2.1 The "isolated cities"

The first situation of Figure 5 corresponds to the cases 1 and 2 of Figure 4 (one UMZ is laying inside one LAU2). It concerns a total of 1091 UMZ that can be called "isolated cities". Of course, this result depends mainly on the mean size of LAU2, by country (see the difference between France and Great Britain and Sweden, Germany or Baltic countries on Table 2 or Figure 7) or even inside one country (for instance, the LAU2 are very different in size in Italy, very large along the coasts and smaller in the central regions). In the database UMZ_LAU2, there is a column indicating if the UMZ is an "isolated city" or a "multi-LAU2 city".

Figure 6: UMZ laying inside one LAU2 ("isolated cities")

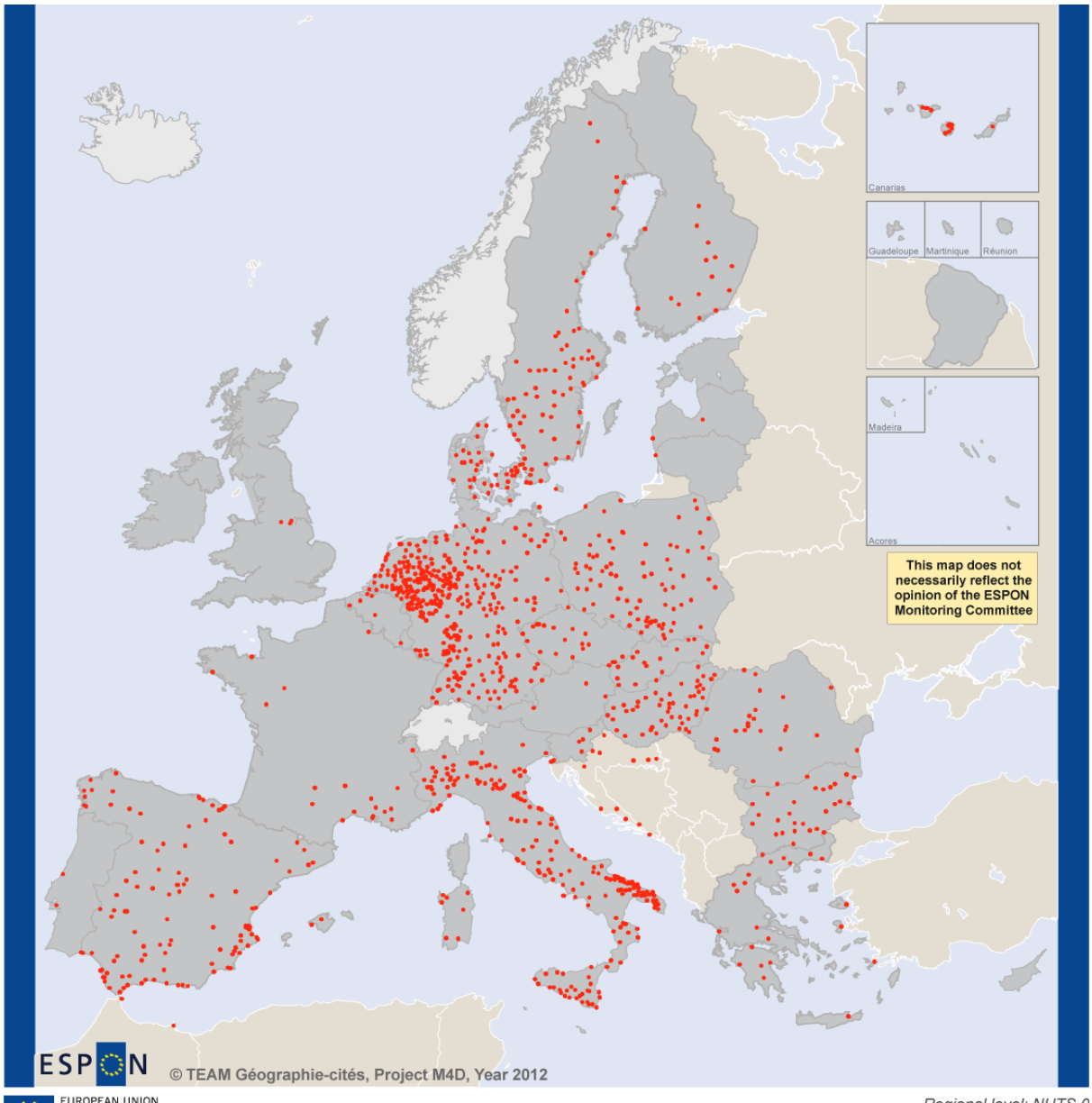


Figure 7: Number of "isolated cities" and average size of LAU2 per country

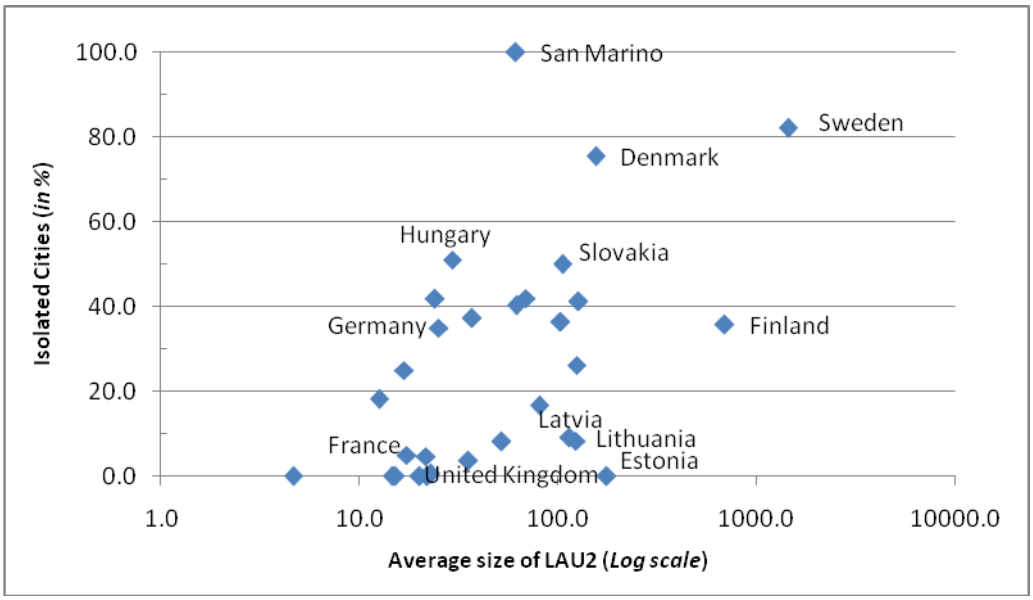


Table 2 : Number of « isolated cities » and average size of LAU2 per country

Ct	Av_Nb_LAU2_By_UMZ	Nb_Isolated_Cities	Ave_Size_LAU2	Nb_UMZ	P_Isolated_Cities
SE	2,9	69	**1444,1**	84	**82,1**
FI	3,4	15	693,2	42	35,7
EE	3,6	0	175,4	13	0,0
DK	7,8	37	154,8	49	75,5
GR	33,2	21	126,9	51	41,2
PL	5,3	81	124,2	310	26,1
LT	6,1	1	**123,3**	12	**8,3**
LV	4,2	2	**114,3**	22	**9,1**
SI	2,3	6	**105,5**	12	**50,0**
HR	3,3	12	103,1	33	36,4
RO	5,3	25	80,7	150	16,7
NL	6,0	78	69,1	186	41,9
ES	7,4	125	61,6	311	40,2
SM	1,0	1	61,0	1	100,0
BE	81,9	4	52,0	49	8,2
IT	49,8	199	37,0	535	37,2
AT	12,4	2	35,4	56	3,6
HU	7,6	49	**29,4**	96	**51,0**
DE	7,4	276	25,2	789	35,0
BG	4,1	28	23,9	67	41,8
UK	149,0	3	22,8	502	0,6
LU	10,0	0	22,0	2	0,0
PT	51,0	3	21,7	64	4,7
IE	182,3	0	20,1	24	0,0

FR	76,2	19	17,4	380	5,0
SK	5,4	16	16,7	64	25,0
CY	12,0	0	15,1	5	0,0
LI	7,0	0	14,6	1	0,0
CZ	5,7	19	12,6	105	18,1
MT	47,0	0	4,6	1	0,0

3.2.2 One UMZ intersects several LAU2

When one UMZ is crossing different LAU2s (Figure 5 situation 2), different choices are possible. The most simple consists in identifying only one LAU2, i.e. the one that contains the major part of the UMZ. We have not retained this method as it prevents the study of interesting urban forms, like conurbations for example (several urban cores historically well differentiated and laying in different LAU2 but latterly joined by urban sprawl into one single morphological patch). A more sophisticated solution consists in selecting several LAU2s until we exceed a certain share of the UMZ population, for example 50% or 75%. We have not retained this second method as it may lead to select some LAU2 that represent a very few part of the UMZ.

We have then adopted a third solution, which is currently used, for example in the cores of the new LUZ, or for the French urban areas ("unités urbaines"). It consists in selecting the different LAU2s for which more than a certain threshold of population (generally 50%) lay inside the UMZ. For the UMZ_LAU2_Restricted database, we have adopted the threshold of 50%. But for the UMZ_LAU2_Complete database, and considering the large heterogeneity of the average size of LAU2 per country, we have decided not to determine ourselves a certain threshold but to let the possibility to the user to choose this threshold. For instance, a high threshold such as 50% is more restrictive for the selection of LAU2 than a low threshold such as 30% (201 UMZ would be excluded with this high threshold, especially in Germany, Italy, Netherlands, Poland, Spain and Sweden). It would thus create more cities linked to only one LAU2 and, by this way, there would be more "rural" LAU2 in Sweden. The user can then choose to modify the threshold according to the average size of LAU2 in each country. Another possibility is to adopt the same threshold, but the user should then keep in mind this heterogeneity when interpreting, for instance, results coming from socio-economic indicators: if the UMZ only represents 30% of the LAU2 population, the use of indicators collected for this LAU2 should be done with caution. In the database, we have added a column with a flag in case the threshold in lower than 30%. We have also added a column with a discretisation of the shares in ten classes (from 0 to 10%, from 10 to 20% etc.).

3.2.2 One LAU2 intersects several UMZ

When one LAU2 is crossing different UMZ (Figure 5, situation 3), we have a particular case of the situation 2 that raises the question of "LAU2 multiple belonging" (one LAU2 that may belong to two or more different UMZ_LAU2). This case is not very frequent (it concerns less about 7% of the total number of LAU2 intersecting UMZ) but is very concentrated in certain areas in Europe, characterized by large LAU2 and/or dense urban settlement patterns, like Germany, Belgium, Netherlands, Italy or Spain. In each of these countries, some LAU2 may intersect 5, 6 or 7 different UMZ. Different choices are possible for managing this case. The most simple, selected for the UMZ_LAU2_Restricted database, consists in retaining only the UMZ that concentrates the largest share of LAU2 population, which relates directly to the situation 2 (if the share of LAU2 population inside this UMZ intersection is larger than 50%, the LAU2 is retained). For the

UMZ_LAU2_Complete database, we have prepared three indicators that allow the user to make its own choices. First, the different UMZ that intersect the LAU2 are ranked according to the share of intersected population, so that the user may choose only the the main UMZ (as in the UMZ_LAU2_Restricted version) or retain, for example, the two largest intersecting UMZ etc. Secondly, we make apparent the particular threshold of 50% as an under-category of the ranking (Rank1_>50% or Rank1_<50%) in order to allow the user selecting the largest intersecting UMZ whether its share is larger than 50% (Rank1_>50%) or not (Rank1_<50%).

In order to help the user choosing the most appropriate version of the database (restricted or complete) and understand easily the different variables and their use, we will prepare a technical report on the UMZ_LAU2 database, with different examples and illustrations, and will make it available on the ESPON Database platform.

Annex 6: Clarifications on the Statistics section

- *"The introduction to the Statistics section clearly explains the work done for this work package and the progress made and the instances (result of applying the methodologies for a particular dataset) that will result from the various checks. However, the actual application within the Database could be more clearly explained. What will be done with these instances? How will they be used for the datasets and the Database?" (ESPON CU, Nov. 2012)*

The Outlier Check is an important component of the process of creating and maintaining the Database. The **input** to this component is a candidate data file downloaded from the Database Portal. The file will have already undergone both a Semantic Check and a Syntactic Check. This means that the indicators in the file itself will form a coherent and useful addition to the database, and also the data relationship between the metadata and the indicators are also coherent. This second condition is important for the outlier check.

The **output** from the outlier check will be a **report**, in PDF form, which will contain a summary of the findings of the outlier check itself. It may also contain any of the printed text from the outlier check software (the R code), and any of the diagnostic graphics which are also created as during the outlier check. We say 'may' here, since the output from the outlier check software can be voluminous – we do wish for the staff on the projects not to feel intimidated by the outputs. Some of the graphical outputs may be in an unfamiliar form (we have been experimenting with cartograms, for example). One of the experiences of teaching statistics for many years to both undergraduate and postgraduate students is that they require guidance in interpreting the outputs from statistical software, and there are those who feel intimidated by the detail. The summary will highlight all the data items that need to be checked.

The **outlier check report** will be uploaded back to the Database Portal. This will allow it to be caught by the tracking tool and the relevant project partner (and the CU) advised. We seek ***confirmation*** that a data item is ***unusual*** or that it has been flagged as ***missing***. The final identification of unusual data will require judgement. There are three types of test: (i) aspatial univariate, (ii) aspatial multivariate, and (iii) spatial univariate. The confirmation we seek is "this data value is correct" or "this data value is not correct". If a data value is incorrect, we will ask for corrected values.

The aspatial univariate tests will be most easy for researchers to comprehend - a value either very high or very low compared with the others in its distribution. The reasons can be many and varied. There is an interesting paradox with aspatial/spatial outliers - these sometimes cluster spatially (that is, groups of adjacent NUTS or LAU regions can have unusually high or low values). Such clustered outliers are unlikely to be spatial outliers (neighbouring values are similar, not very different). The interpretation of the tests does require judgement at the NCG end.

A question has arisen as to the treatment of ***missing data***: that is, data for which a code has been supplied to indicate that no value is available for the given indicator and the given region. Typically a missing value code of -999 is used. We will highlight the NUTS regions with these codes and seek confirmation that the data are indeed missing. For the multivariate tests, we encounter the ***curse of dimensionality***. To compute a multivariate test we need to have all data present for all regions. If a single value for an indicator is missing, then the entire region has to be dropped from the analysis - this is known as ***case wise*** deletion. Hence the importance of seeking confirmation that missing data are actually missing.

The project partner who supplied the data reports back the confirmations through the Database Portal to the team at RIATE. If there a need to update any data, the outlier check cycle will need to be repeated. If all of the data are correct, and the unusual values confirmed, then the data can be uploaded into the Database.