TECHNICAL REPORT

ESPON

The ESPON Database Application

CONTENT

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

ESPON 2013 DATABASE

MARCH 2011

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

38 pages

LIST OF AUTHORS

Bogdan Moisuc, LIG Steamer
Jérôme Gensel, LIG Steamer
Anton Telechev, LIG Steamer
Benoit Le Rubrus, LIG Steamer

Contact

moisuc@imag.fr

jerome.gensel@imag.fr

anton.telechev@imag.fr

benoit.le-rubrus@imag.fr

.

LIG Grenoble

Bâtiment ENSIMAG D

681 rue de la Passerelle

38400 Saint Martin d'Hères

(+33) 4 76 82 72 11

TABLE OF CONTENTS

Introduction

This technical report describes the main features of the ESPON Database Application. The ESPON 2013 Database Application is a complex information system dedicated to the management of statistical data about the European territory, spanning over a long period of time. The part that is visible to the users is the Web interface of the application. Equally important is the hidden part, relying on two databases: one is used for storing ontology data, and the other, called the ESPON Database, is meant to be queried by end-users.

The ESPON 2013 is the path followed by both data and metadata from the moment they are entered in the ESPON DB Application, until they are output as answers to queries expressed by end-users (see ***Figure 1***). Four phases are identified along this data flow:

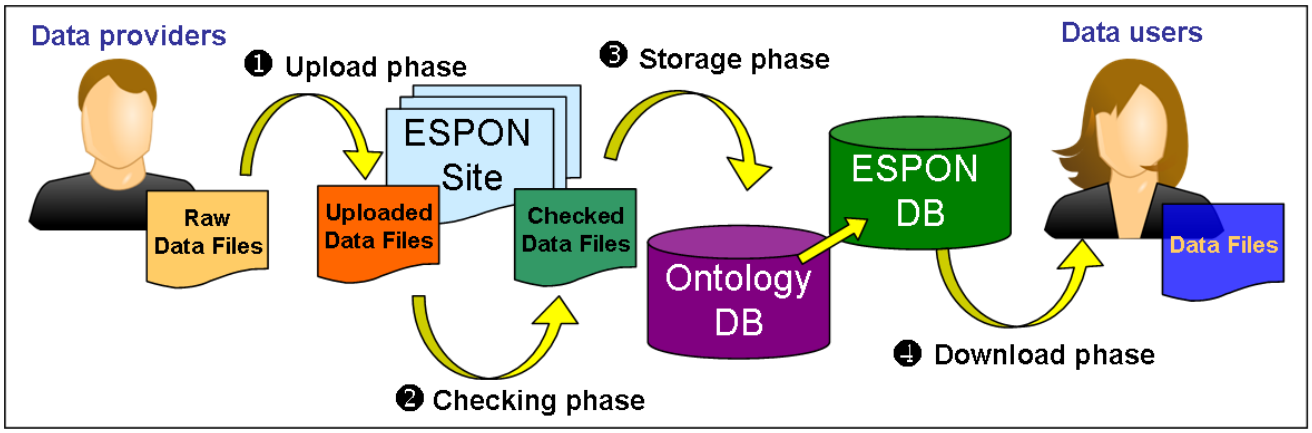


Figure 1: *The ESPON 2013 DB data flow.*

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

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

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

4. During the download phase, end-users of the ESPON DB Application are invited to explore, search and retrieve both data and metadata through a

Web interface. Free data and metadata can be accessed and downloaded by any end-user, while data and metadata subject to copyright restrictions are made available for authorized and registered users only.

The first section of the document shows the screens and features of the Web application interface version 1.0, also known as the ESPON 2013 Database Web extraction tool. This gives a *vademecum* for the application by following a typical user session, from the authentication to the download of query results.

The second section of this document contains a description of the databases upon which this application is based: the ontology base and the ESPON database, and of the software that allows filling them.

1 ESPON DB Web Interface

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

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

The underlying subsections contain a description of the different pages that can be displayed by the application. The order of the pages follows the logical order of the navigation performed by a common user. The starting page of the application is the search page (see section 1.5), but any user willing to use the full functionality of the ESPON DB Application will have to register first (see section 1.1). A registered user can login to the application (see section 1.2), update his registration information (see section 1.3), see the help page (see section 1.4), use the application to search (see section 1.5) and download data (see section 1.6). If users are data providers, they can also upload their own information into the ESPON Database (see section 1.7). Privileged users (administrators of the application) can also change the application settings and manage user registration requests (see section 1.8).

1.1 Registration page

Figure 2: *The registration page of the ESPON DB Web Application.*

The registration form contains a small number of fields required for establishing the identity of the user (first name, family name and organization) their contact coordinates (telephone number and email address) and their relationships with ESPON (the ESPON Project which they are members of and their function within this project).
Once the users submit the form, they receive a confirmation email. In the same time, the database administrator is notified for the registration request. The administrator can check the information entered by the user and validate or not the registration request, via the administration interface (see section 1.8 for the administration page). Once the administrator has validated the user registration, the user receives via emailAdministration page (administrators only)their login and password for the application.

1.2 Login page

When typing the URL of the ESPON 2013 Database Web application in his/her browser address bar, any user is first invited to choose between both following types of login:

- *anonymous* browsing;
- *registered account* login.

Figure 3: *The login page of the ESPON DB Web Application.*

The differences between these hese two ways of browsing essentially consist in the features the user will be offered by the application. Roughly speaking, a registered user will be allowed to access the whole set of pages that can be delivered by the application; an anonymous user will be allowed to search and download results from only a subset of available data. Also, anonymous users cannot upload data or edit metadata. Besides the login fields (*Login* and *Password*), ***Figure 3*** shows that the login page displays a link in order to ask for a new login or password, if the users have lost the old ones.

On the login page, there a link that redirects the user towards a registration form. The same form can be accessed by clicking on the menu item "Register", which is only visible to unregistered users. Upon clicking on one of these links, the user arrives on the registration form page.

Although the layout of the pages is identical for both types of session (logged in or anonymous), the displayed menu bar on the header of pages is different. ***Figure 4*** shows the menu bar for an anonymous session, ***Figure 5*** shows the menu bar for a registered user. The main difference between both menus is the availability of the upload page for a registered user (described in section 1.7) and of the account page (described in section 1.3).

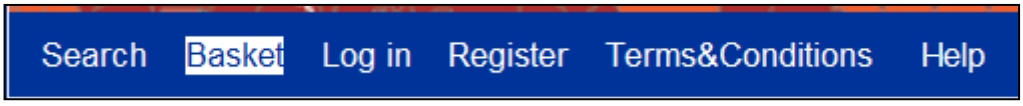


Figure 4: *Menu bar for an anonymous session.*

Figure 5: *Menu bar for a registered user.*

For registered users, the authentication leads them to the ESPON Database Web Interface Home Page. Figure 5 shows that this menu bar gives shortcuts the available features of the Web application:

- the home page;
- a page to update one's account;
- the search page;
- the basket page;
- the upload page;
- a link to quit the session;
- the on-line help.

1.3 Account page (registered users only)

The account page allows users to update and complete their personal information and to change their password for accessing the application. In order to update any profile element, the users are required to enter their old password.

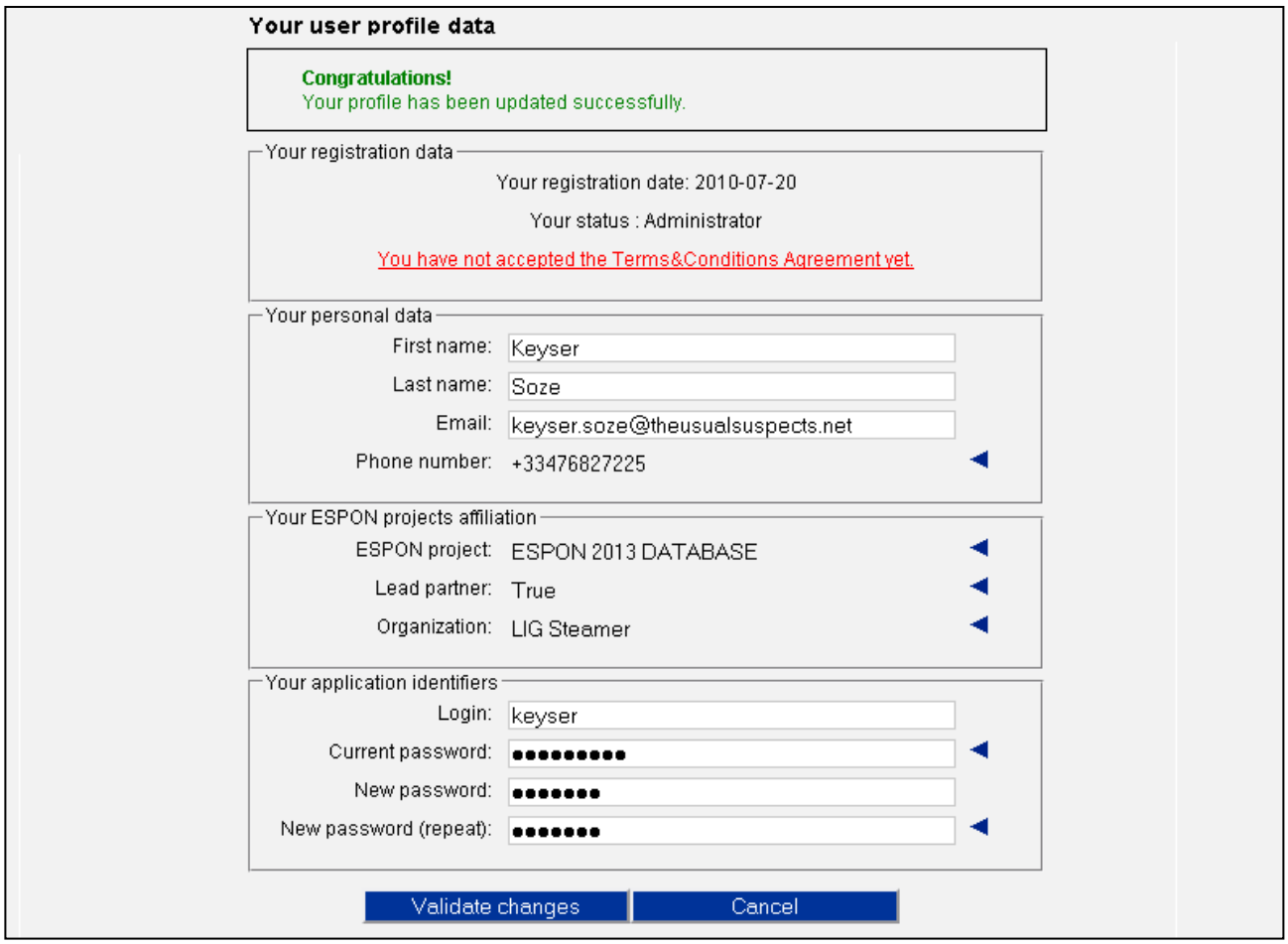


Figure 6: *The account page of the ESPON DB Web Application.*

From the account page, users can also see if they have agreed with the terms and conditions of the ESPON Database Application. As they are already registered, they only need to agree with the T&C once. Unregistered users have to accept the T&C at each session.

1.4 Help page

Figure 7: *The help page of the ESPON DB Web Application.*

The help page of the ESPON 2013 Database Application is aimed at helping users navigate easier through the pages of the ESPON 2013 Database Application (see ***Figure 7***). The help contains mainly the same *vademecum* as the first chapter of the present document, that is, the logical succession of pages (with their explanations) navigated in a typical user session.
An index allows the users to visit the six categories of the online help:
1. Registration page
2. Login page
3. Profile page (registered users)
4. Search page
5. Basket page
6. Upload page (registered users)
7. Help page

From each of these pages, a link to the help content (index) is provided.

1.5 Search page

The ESPON 2013 Database Application offers currently two types of searches: a search by project (the data providers) and a search by theme (according to the ESPON 2013 Database thematic classification).
For each type of search, two modes of research can be performed:

- simple search;

- advanced search.

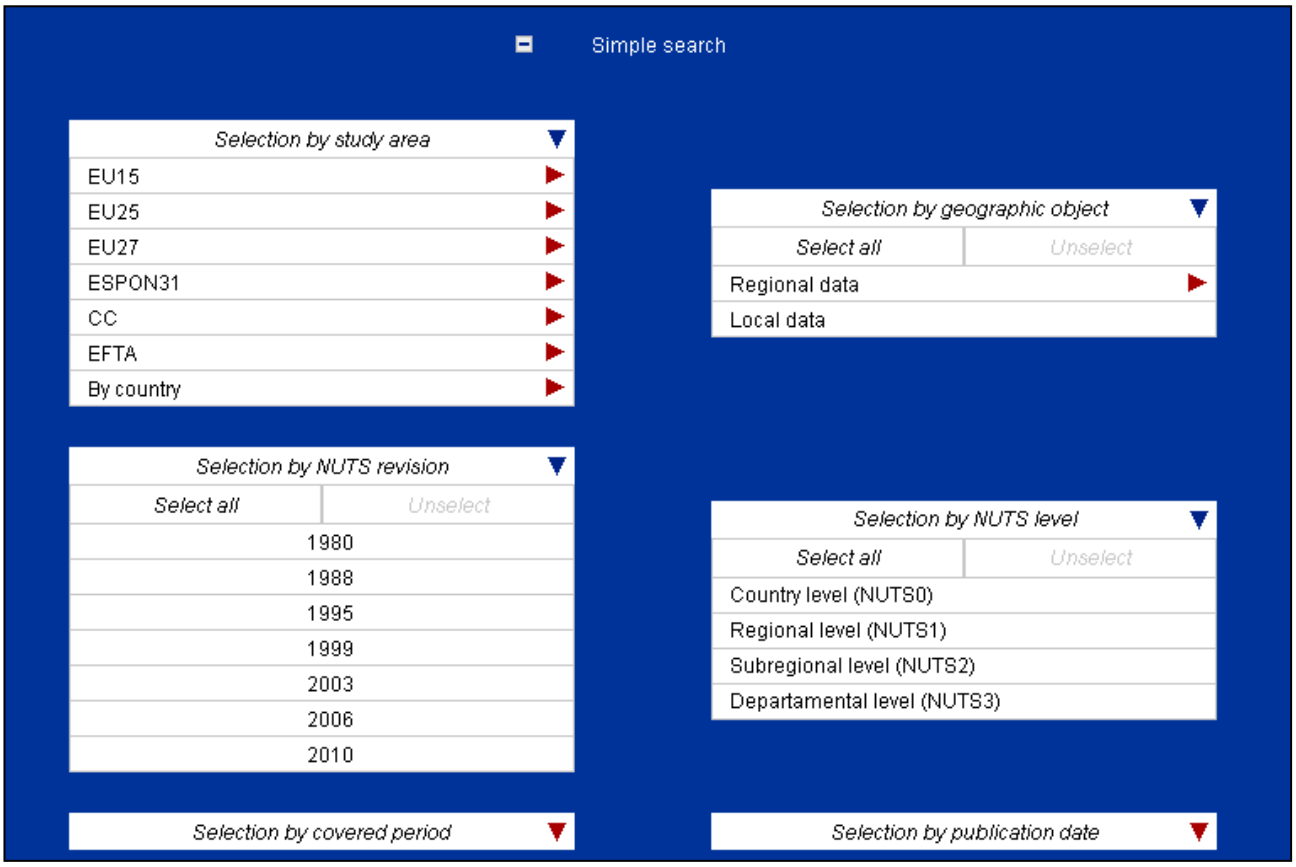


Figure 8: *The advanced search criteria.*

The user may switch from one mode to the other mode by clicking a button situated in the middle of the screen. In advanced mode (see ***Figure 8***), a series of additional criteria can be added to a search in order to narrow down the results:

- study area (which means choosing either one of the groups of countries – EU15, EEFTA, etc., or choosing countries one by one);
- geographic object type (regional data, world data, local data etc.);
- covered period;
- publication date.

Once the user has chose a type of geographic object, two new search criteria appear:

- nomenclature revision (e.g. NUTS 2003, NUTS 2006, etc.)
- nomenclature level (e.g. NUTS 2, NUTS 3, etc.)

For each type of search, the user must select at least one mandatory field in order for the query to return any result.
For the search by project (see ***Figure 9***), the user has to select at least one (possibly all) project in order to perform a query. When moving the mouse over one project, a sublist with all the available datasets for that project becomes visible, allowing users to refine their search. Multiple datasets and/or projects can be selected. Once a dataset is clicked, all the indicators available

in it are displayed in the indicator list displayed on the right. The window with the indicator list can be maximized by clicking on the small button below.

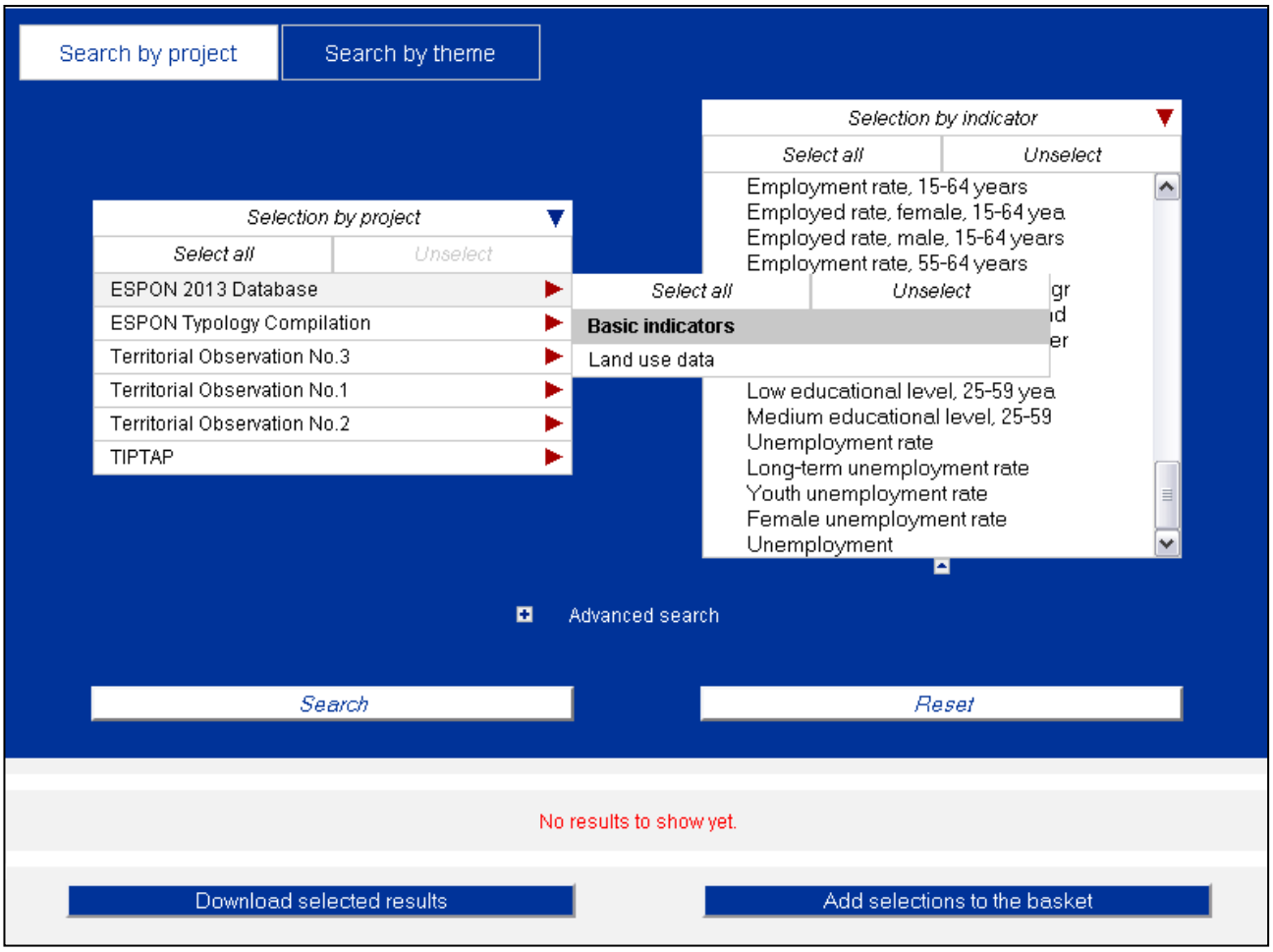


Figure 9: *The search by project.*

For the search by theme (see ***Figure 10Figure 9***), the user has to select at least one (possibly all) theme in order to perform a query. When hovering over one theme, a sublist with all the available subthemes for that theme becomes visible, allowing users to refine their search. Once a theme or a subtheme is clicked, all the indicators available in it are displayed in the indicator list on the right. Multiple themes and/or subthemes can be selected.

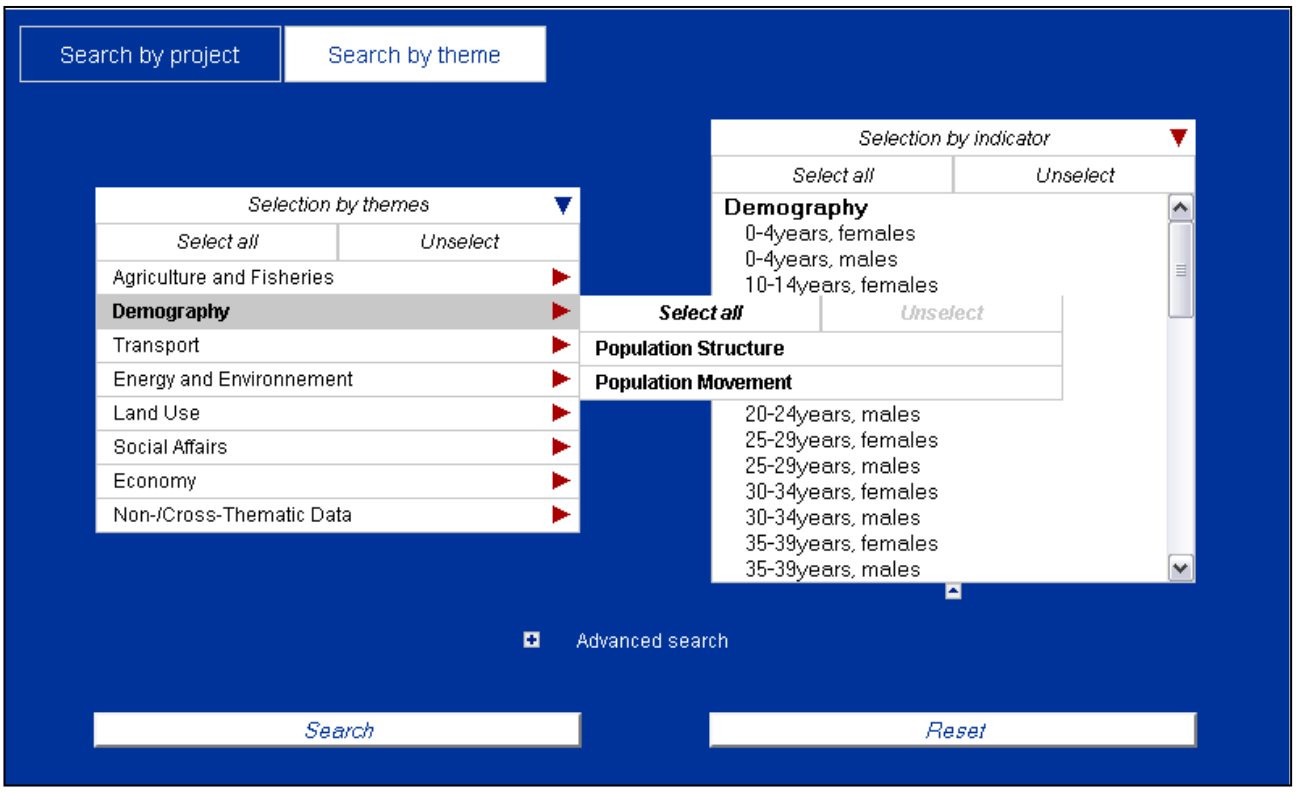


Figure 10: *The search by theme.*

As previously mentioned, the query is performed only if the user has selected at least one criteria (theme, subtheme, project or dataset). If this requirement is not met, clicking the "Search" button will return a warning message. Otherwise, the application performs the query and the results are displayed in the same page, under the form of a result table, as shown on ***Figure 11:*** *The result table, displayed by datasets.*.

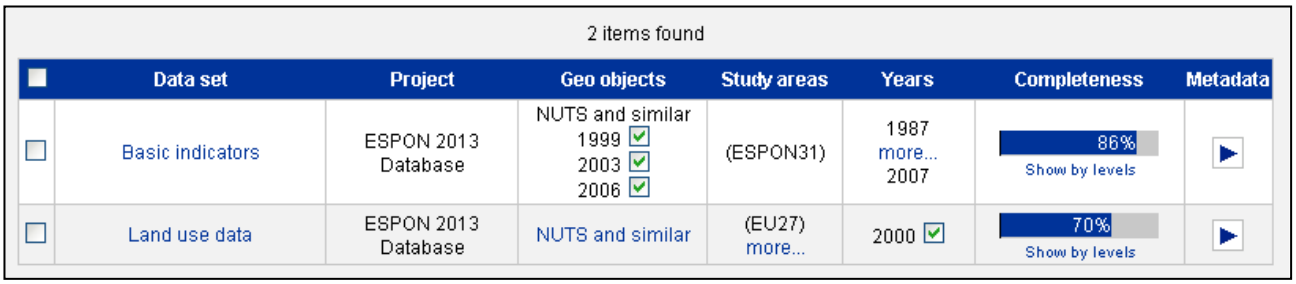


Figure 11: *The result table, displayed by datasets.*

The screenshot above shows that the query returned two results (two datasets). Each result is displayed on one line. Considering this table of results, the displayed information for each result is respectively made of the following columns:

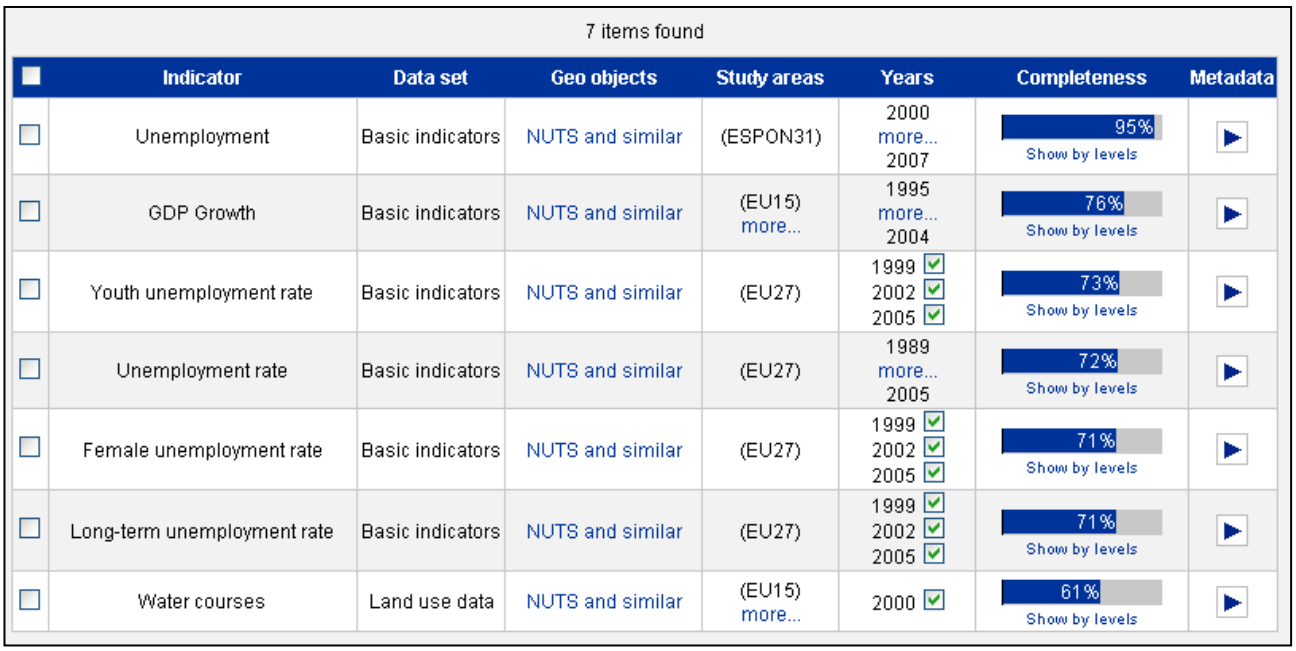
1. First column (no header title): the checkbox allows to select the current result in order to perform two possible actions:
    - adding this result to the basket;
    - immediately downloading this result.

The basket is roughly speaking a temporary area where the user can save multiple search results while he/she performs various queries. This basket functionality is further described in section 1.6.

2. Dataset column: displays the name of the dataset which is concerned by this result item. Clicking on this name displays all the indicators that have been delivered within this dataset. The indicator names can be selected and unselected for inclusion into the basket or in the direct download. The  icon is displayed if the result concerns data with copyright constraints attached. This case implies both behaviors:

    - if the user is anonymous, the checkbox in the second column is disabled: though he/she can see this result and the associated metadata, he/she is not allowed to download it;
    - if the user has logged in with a registered account, the checkbox is enabled. Nevertheless, he/she is warned about the confidential status of data for this result.

3. Project: this field shows the project that provided the data of this result.
4. Geographic object column: the current version of the application only manages NUTS. Clicking this name displays all the nomenclature versions for which data have been delivered within this dataset.
5. Study area: this field shows the spatial coverage of the current result. Whenever the countries present in the result can be aggregated by groups, they are aggregated to the largest group possible, when aggregation is no longer possible, the remaining countries are aggregated as individual countries (e.g. “EU 15 + IS, NO”).
6. Covered period: this field shows the temporal coverage of the current result. When data range over a period of more than 3 years, a clickable “more” label allows the users to see the complete list of years. When there are three years or less, they are all visible by default.
7. Completeness: the percentage of the global completeness of the current result is represented by a colored bar. Long blue bar: high rate; short blue bar: low rate. A clickable “Show by levels” text allows having a more refined view of the completeness, per nomenclature level.



7 items found

Indicator	Data set	Geo objects	Study areas	Years	Completeness	Metadata
Unemployment	Basic indicators	NUTS and similar	(ESPON31)	2000 more... 2007	95% Show by levels	
GDP Growth	Basic indicators	NUTS and similar	(EU15) more...	1995 more... 2004	76% Show by levels	
Youth unemployment rate	Basic indicators	NUTS and similar	(EU27)	1999 2002 2005	73% Show by levels	
Unemployment rate	Basic indicators	NUTS and similar	(EU27)	1989 more... 2005	72% Show by levels	
Female unemployment rate	Basic indicators	NUTS and similar	(EU27)	1999 2002 2005	71% Show by levels	
Long-term unemployment rate	Basic indicators	NUTS and similar	(EU27)	1999 2002 2005	71% Show by levels	
Water courses	Land use data	NUTS and similar	(EU15) more...	2000	61% Show by levels	

Figure 12: *The result table, displayed by indicators.*

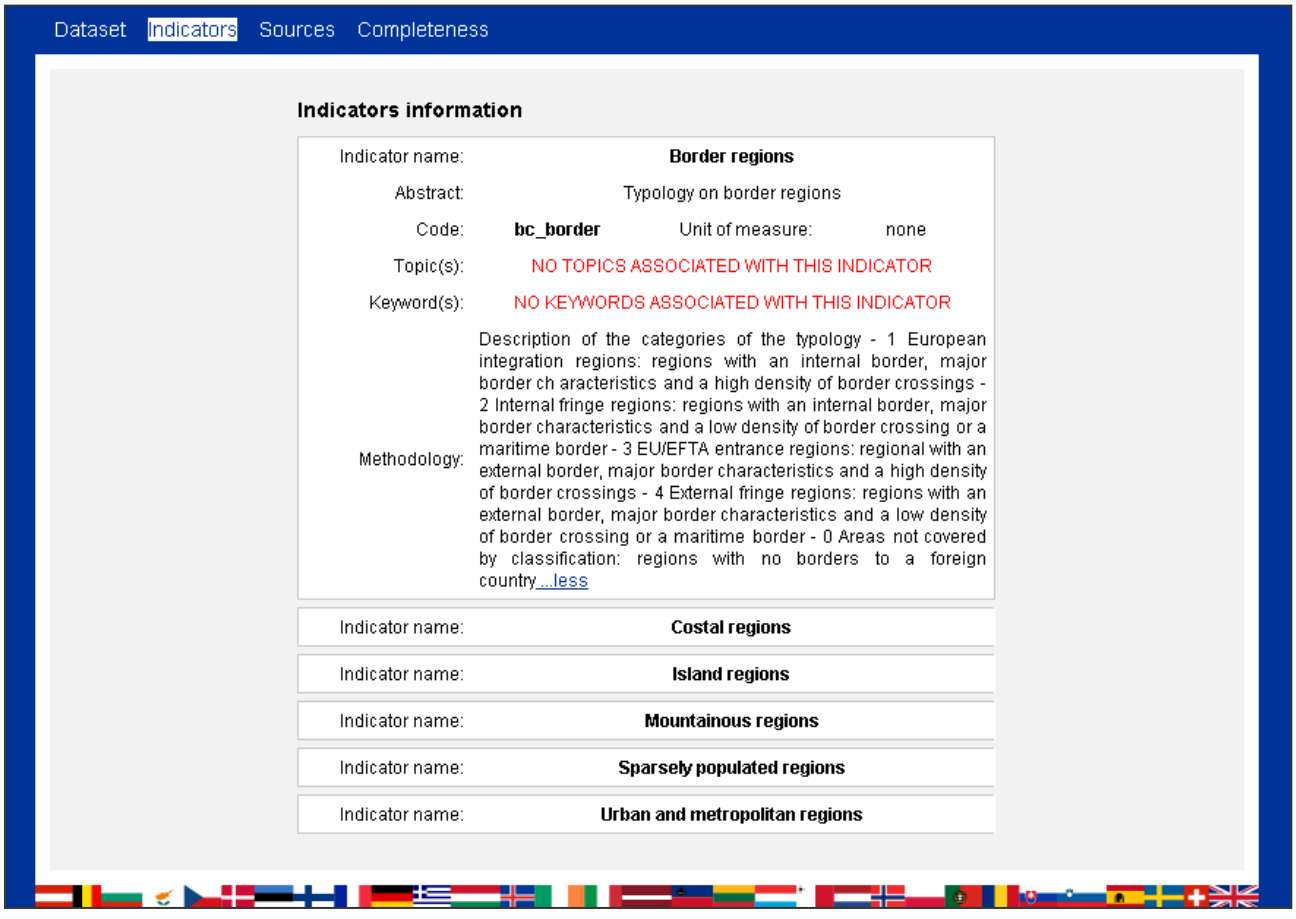


Figure 13: *The detailed metadata page.*

Metadata: by clicking the icon, the application opens a popup window where the user can consult further details about the metadata of this current result (see ***Figure 13***). Four different types of information are available in this metadata page:

- identification of the dataset;
- the list of indicators for this dataset;
- the data lineage (the data sources for all the values) of this dataset;
- the map of Europe showing the rate of completeness of data for each country which is concerned by the dataset. Figure 9 describes this feature.

A cartographic representation of the level of completeness of the data for the chosen dataset is also displayed in the metadata section. The completeness described the result as a whole. If more indicators or more years or more NUTS revisions have been chosen from the search criteria, then the displayed completeness is averaged for all these criteria. The data completeness can be seen at every available nomenclature level.

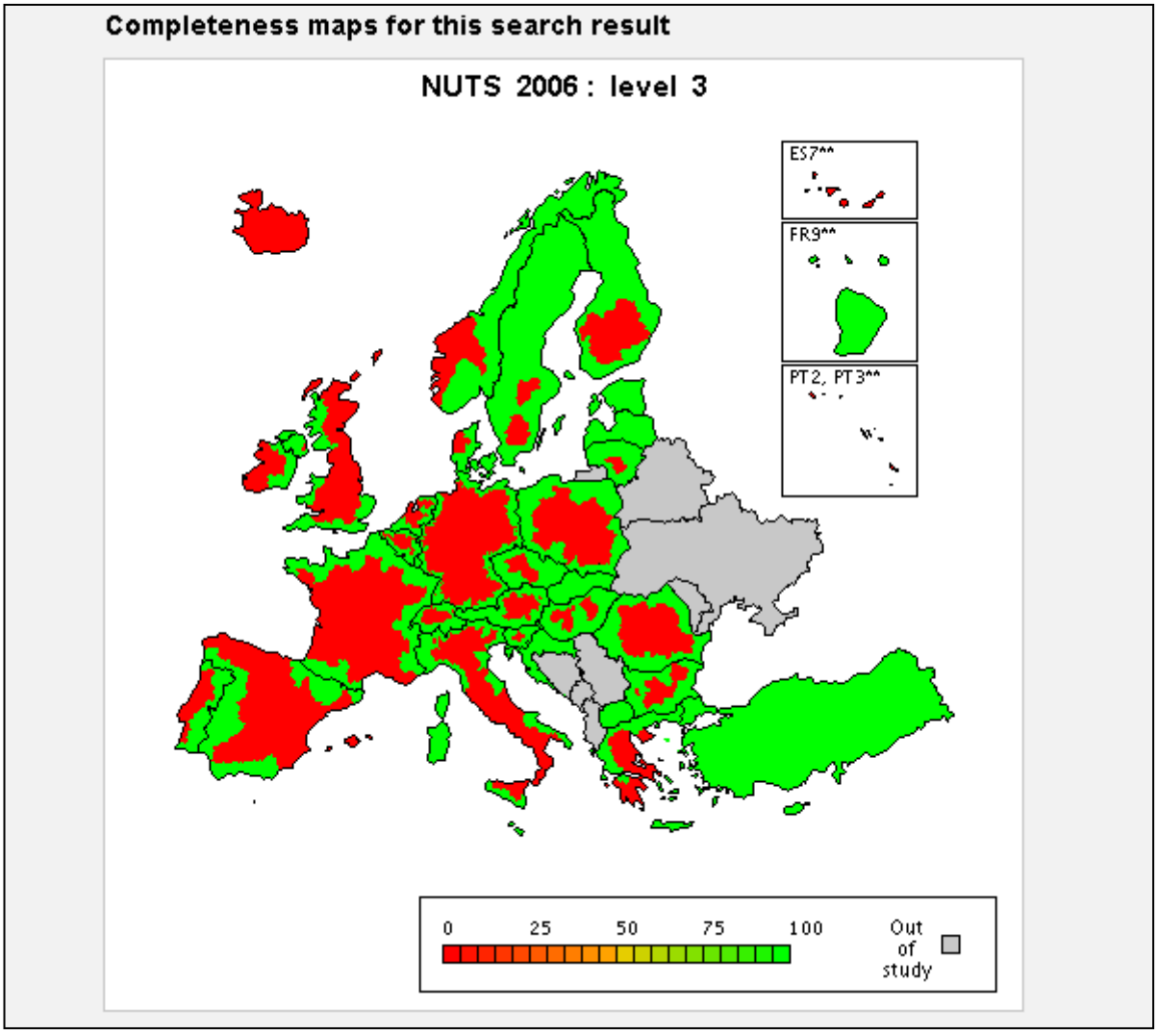


Figure 14: *The data completeness metadata screen.*

Finally, from this search result table, the user can perform the two following actions on selected items (by checking checkboxes on each line):

- adding selected results to his/her basket;
- downloading selected items.

The "basket" feature allows to temporary save search result items. Thus, the users can keep their attention on their queries, the basket allows them to drill their research process.
However, the basket is not a mandatory step for downloading results: if the user has found the expected results, they may directly use the "Direct download" functionality. This "direct download" can be considered as a short-cut to the basket service, which is further described in the following section.

1.6 Basket page

Figure 15 shows the basket page: excepting the project name (which disappears) and the dataset name (which is renamed to "Item"), the displayed information is similar to the search results table on the search page: the basket is filled with search result items. Once the user completed his/her research, he/she can now refine the selection of search results items that he/she wants to download.

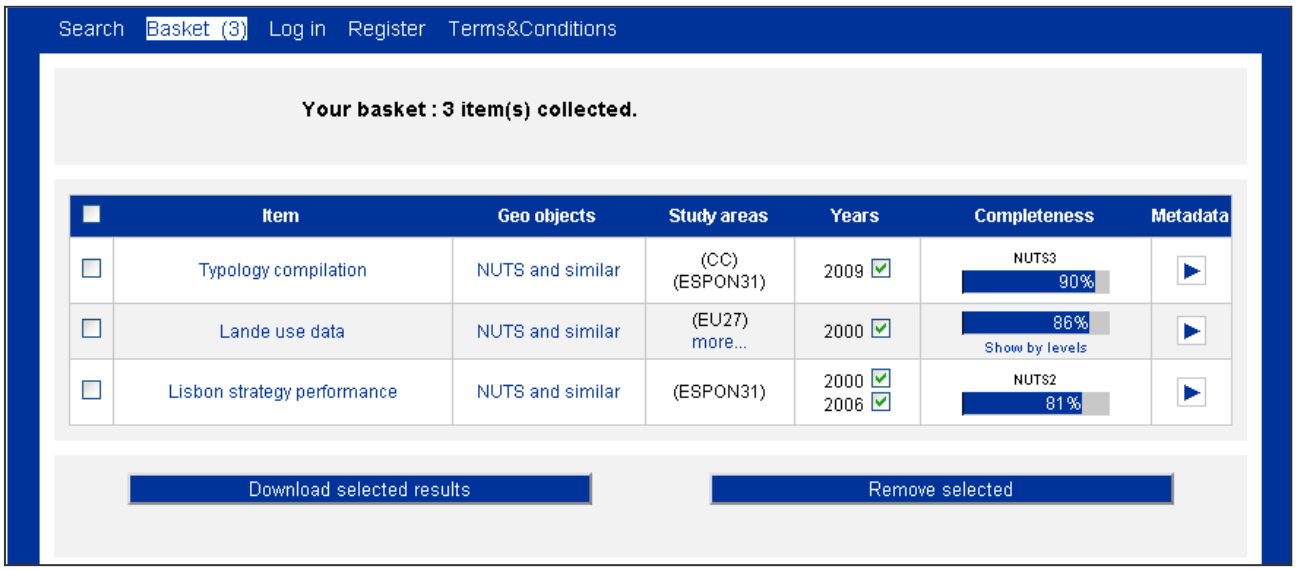


Figure 15: *The basket page.*

Through the checkboxes on the left side of each line, the user is invited to select items on which he/she can perform both following actions:

- a deletion;
- a download.

The deletion consists in removing the selected items from the basket. Note that in such a case, the user is asked to confirm before processing.

The download action triggers the following set of tasks on the server-side:a Microsoft Excel file is built for each selected item which is expected to be downloaded. This spreadsheet is composed of at least four sheets:

1. the first sheet is untitled "Dataset", it provides general information about the current dataset (name, contact, etc.);
2. the second sheet is untitled "Indicators", it provides metadata information about included indicators in this dataset: a description, the unit, the methodology, etc.;
3. the third sheet is untitled "Lineage", it provides metadata information about the lineage of the dataset: validity start, methodology, etc.
4. the following sheets display the values of the dataset, broken down by nomenclature revision and nomenclature level, there will be one sheet for each revision/level combination existing in the dataset (e.g. one sheet for NUTS0 2006, one for NUTS1 2006, one for NUTS0 2003, etc.).

If the user has selected several items to be downloaded, a zip archive is built, gathering the built xls files (one file per result item) by the previous step.
The application finally returns the built file to the user as an attachment (an xls file or a zip archive, depending on the number of selected items).

On the client-side, the users are prompted by their browser to open or to save the built file on their disk.
In the case of one selected item, the proposed filename for the downloadable file will be **ESPON_data.xls**. In the case of several selected items to be downloaded, the default filename for the proposed downloadable file will be **ESPON_data.zip**.
Once downloaded on their local disk, the content of the zip archive can be extracted with a standard archiving tool, the included files are simply named **ESPON_*number of result_dataset title*.xls**, and there will be as many files as there are selected items in the search results or in the basket (see ***Figure 16***).

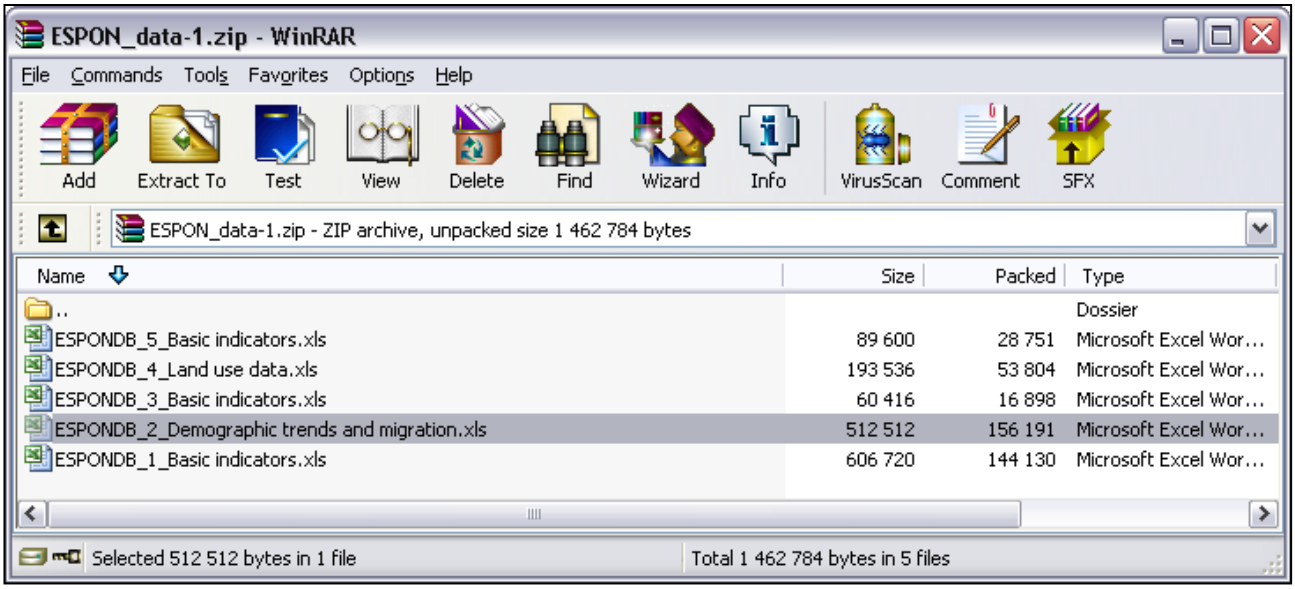


Figure 16: *An example of downloaded zip file with 5 search results.*

Caution: note that the basket is emptied when the user logs out. This is the reason why the user is asked to confirm his/her wish when he/she clicks the "Log out" button on the menu bar.

1.7 Upload page (registered users only)

The upload page aims at allowing ESPON project members to upload their data deliveries in order to fill the ESPON 2013 Database. The ESPON 2013 Database Application allows the users to upload 3 types of files to the ESPON server:

- metadata files;
- data files;
- additional documentation files (technical reports, etc., describing in more detail the different methodologies used for calculating indicators or for estimating missing data).

This is typically done in a 4 step process.

In the first step users are invited to upload a metadata file on the server. However, the step can be much more complex. When loading a metadata file within the application, the users are prompted the existing errors in their document. Errors can be either syntactical errors, or simply missing information for mandatory fields. Fields with errors are highlighted in red and a textual description of the error is also given in the header of the metadata editor (see an example in ***Figure 17***).

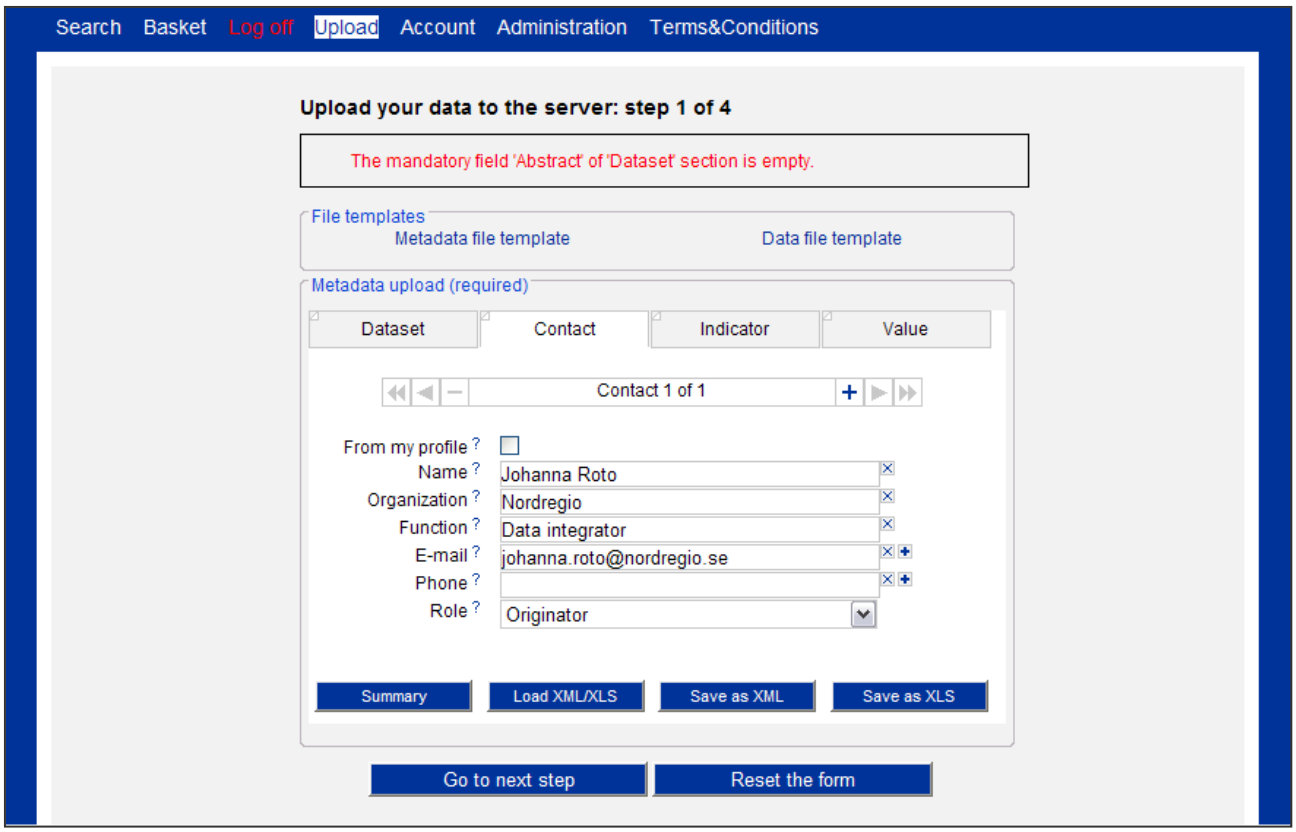


Figure 17: *The upload page, step 1.*

The metadata profile of the ESPON 2013 Database has been previously defined as an Excel template. The Web editor contains largely the same fields as the Excel template, but allows filling them in a more user friendly manner. The interface defines the same three levels for filling the metadata: description of the dataset, of the indicator and of the record (for lineage). It includes all the information required for compliance with the INSPIRE directives. However, in order to simplify the task of the data providers, the application fills in automatically all the data that can be derived from the dataset. For instance, the data providers are not required to describe the spatial or the temporal extent of their dataset, this information will be derived by the ESPON 2013 Database Application and filled in automatically.

The first level defines the main characteristics of the whole dataset. It includes all the common points for all the elements inside the dataset and, also, the metadata description, like the name of the delivery, the authors, etc.

Information	Description
Dataset name	Name of the data delivery
Abstract	Short description of the dataset content
Date	Date of creation of data and metadata
Contact Point	Name, email, etc of the author of this metadata

The second level of metadata information is based on the characteristics of each socio-economic indicator. Inside a dataset, each indicator has its particular specifications and characteristics. The selection of appropriate themes and keywords is a key point, at this level, to obtain a good description of the resource.

In order to ease the edition of indicators and to harmonize definitions and codes inside the database, the editor provides a list of well known indicators (that are already present in the database).

Information	Description
Code	Code of the indicator (harmonized with existing nomenclature)
Name	Name of the indicator
Abstract	Full text description of the indicator
Unit of measure	Can be meter / square, thousands, etc: free text
Classification (themes, keywords)	Various themes and keywords found into thesauri to describe the resource. Multiple themes and keywords can be attached to one indicator, if needed.

The value level is based on the characteristics and description of the indicator values and it is linked with the "category column" of the data files. This is the most specific level because it defines the main characteristics of the value of a specific indicator that belongs to one dataset. The lineage and reliability contribute to give information about the quality, the procedures for data collection, the sources and the methods used for delivering and transforming data.

The information for each value is grouped by "scopes". A scope defines a subset of data which have the same lineage. For example, if a dataset for the EU 27 contains official data coming from Eurostat, from the Romanian and from the Bulgarian national statistical institutes, the values metadata will be grouped in three scopes, one for EU 25 values (source EUROSTAT), one for Ro values (source: Romanian national statistical institute) and one for Bg values (source: Bulgarian national statistical institute).

Information		Description
Scope		
	Label	Label in front of each value in dataset, that allows to map metadata information with data information
Lineage		
	Provider, provider URL	The statistical institute or ESPON project that provides the values and a link to its website
	Date	The date of extraction of the data
	Source , source URL	The document name (report or article) and its URL, if available on the Web
	Methodology	A free text description of the methodology used for estimating, calculating or correcting the data in this scope This is required whenever modifications were applied by the provider to the original data, for instance when data are complex indicators that were built from less complex indicators, when data are result of corrections or estimations of the original data, etc.
Reliability		
	Estimation	***true*** if the value comes from an estimation
	Quality	A level of confidence to the reliability of the value: ***low***, ***medium***, ***high***.
Constraints		
	Public data access	***true***: the data is public which means all users, without any restriction, can view and download the data ***false***:means that only members of ESPON projects will

		be able to download those datasets
	Public metadata access	If true, users can see what exists in the database, whatever the data access's value. If false, nobody can see the data stored inside the database. It is a special issue for purchased data that can only be used on the demand of an expert.
	Copyrights	A text area that allows users to communicate who owns the rights for the data within this scope

The users have a lot of flexibility in editing and storing their metadata files. They have the possibility to edit their metadata online, via an interactive process, or to simply edit it offline (in Excel, for instance), and then load it in the metadata editor. They can save it either in Excel or XML format and reload it at a later time. However, an important point to note is that metadata file upload cannot be performed if the user has not corrected all the errors and filled in all the mandatory fields. Once the file is error free, clicking on the "Go to the next step" brings them to the second step.

In the second step, the user is required to upload a data file to the ESPON server (see ***Figure 18***). The data file must correspond to the previously uploaded metadata file, that is, all the indicators and scopes present in the data must have been described in the metadata file.

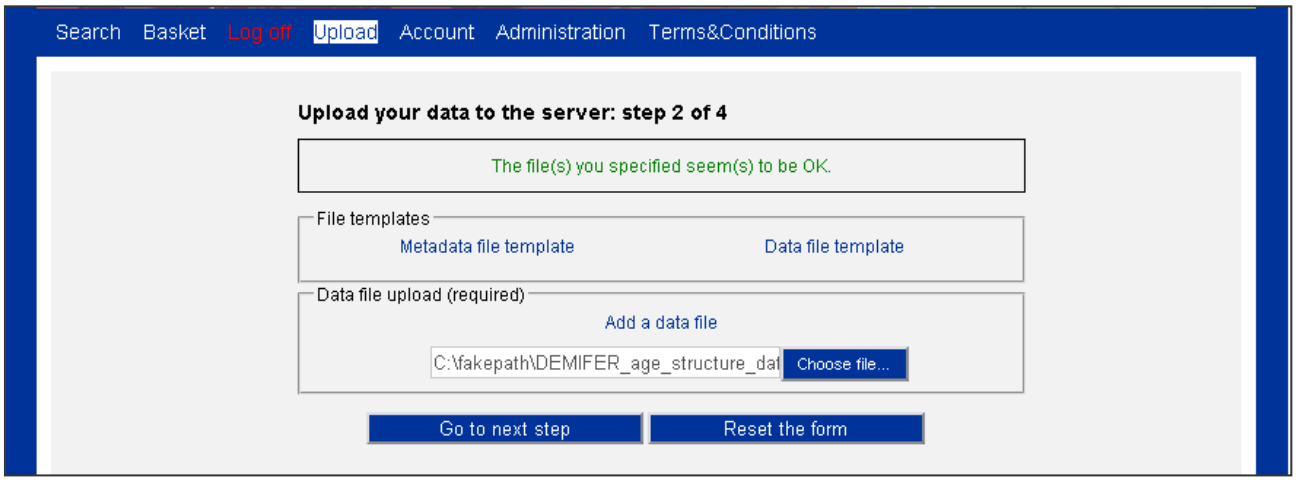


Figure 18: *The upload page, step 2.*

Note that the uploaded files are obviously not directly integrated to the ESPON Database, this upload step has to be followed by several processes: verification, harmonization, etc.
If no correspondence errors are detected, clicking on the "Go to the next step" brings the users to the third upload step.

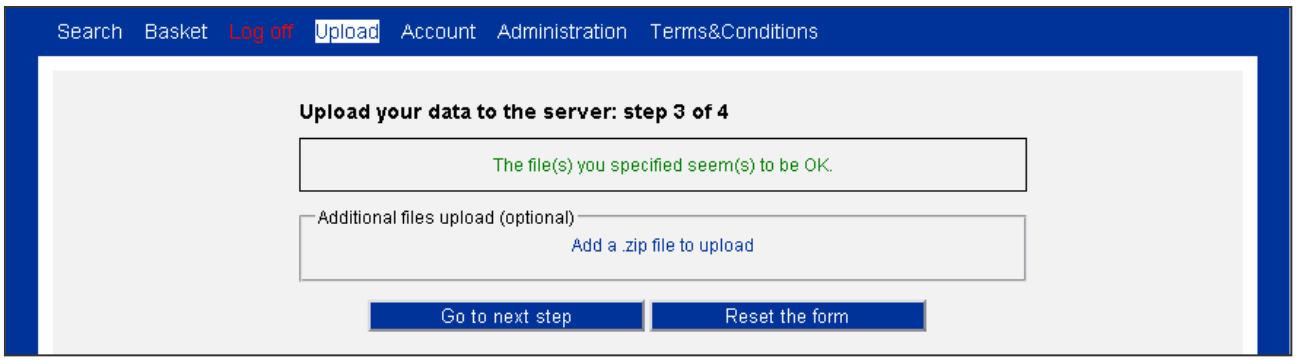


Figure 19: *The upload page, step 3.*

In this step one or more reports can be inserted, including intermediate data in Excel sheets in order to illustrate more clearly the methodology. All the files must be archived together in one .zip file. This step is optional, so, if there are no additional documentation files, the user can go directly to the fourth and last upload step.

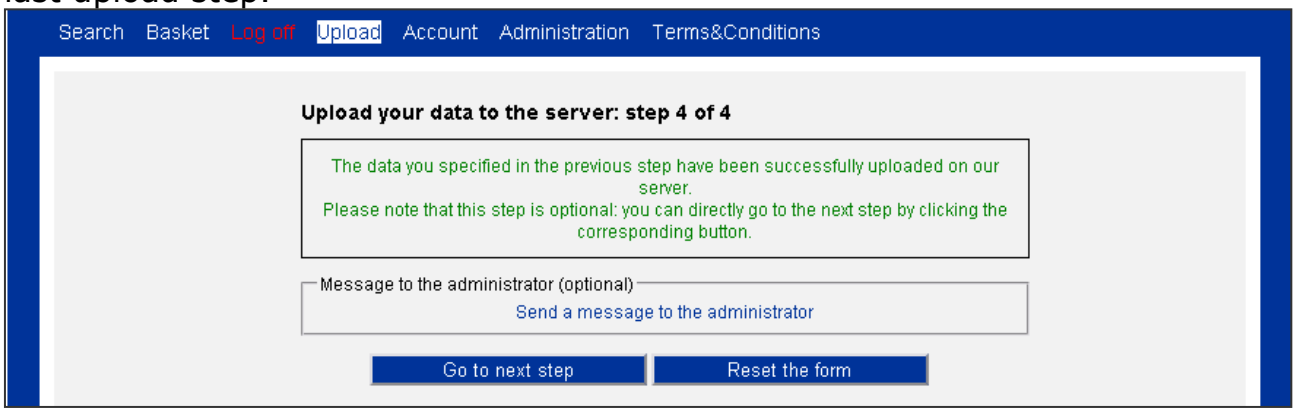


Figure 20: *The upload page, step 4.*

In this step, users can send a message to the application administrator. If an error was committed, the upload process can be restarted by clicking the button "Reset this form".

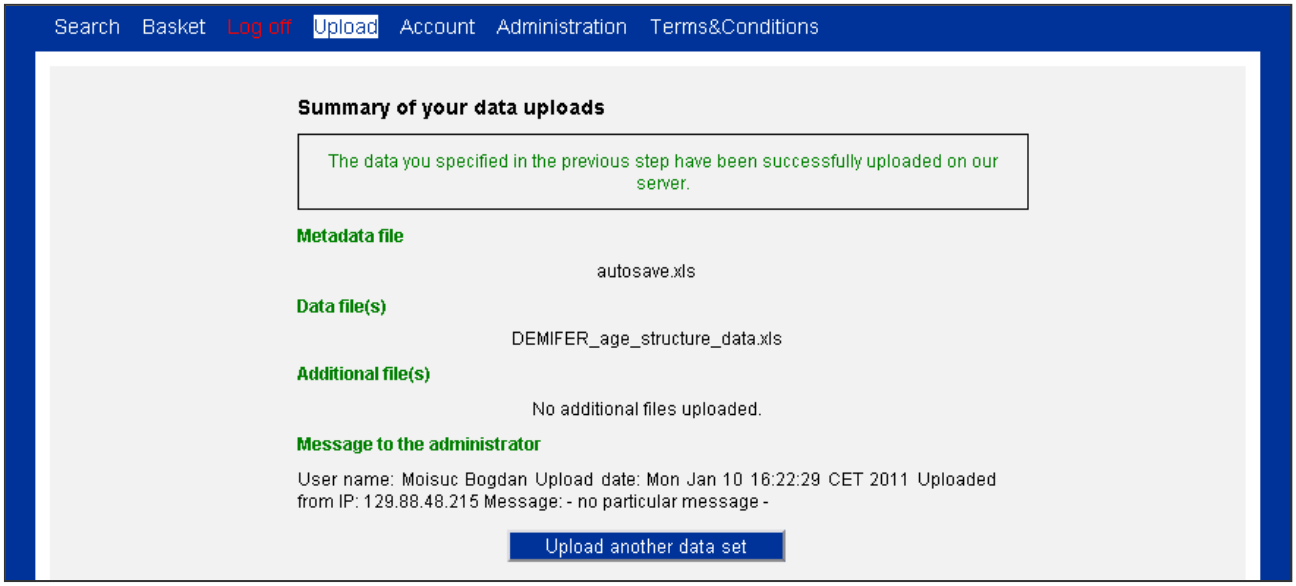


Figure 21: *Successful upload message.*

In this step, users can send a message to the application administrator. If an error was committed, the upload process can be restarted by clicking the button "Reset this form".

1.8 Administration page (administrators only)

The administration page is aimed at helping ESPON Application administrators with the database maintenance and update tasks. There are three main categories of tasks in the administration interface: application management, users management and upload (data) management.
Application management tasks are related to the well-functioning of the application as a software. This comprises 3 categories:

- database connection properties, containing all the descriptors necessary for the application in order to have physical access to the ESPON Database;
- mailing properties, containing all the notification properties for the application (who gets notified when new uploads occur or when new users register);
- other properties include the formats and the templates for data and metadata.

User management tasks are related to the maintenance a list of application users with their respective privileges. They are also divided into 3 categories:

- users registration management allows administrators to have a fast overview of new registration requests and to validate or invalidate them;
- users data management allows updating information about registered users in a quick manner;
- user removal, for users who are no longer active members of the ESPON Database community.

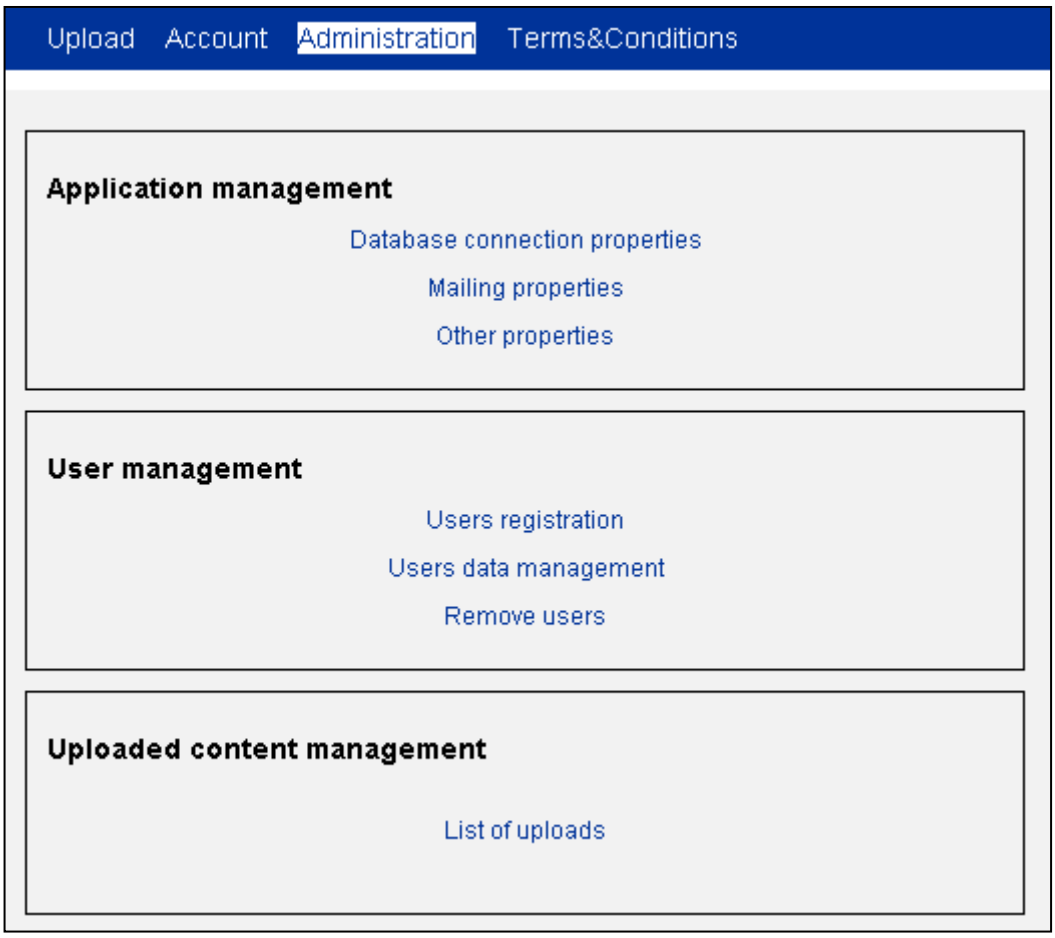


Figure 22: *The administration page.*

The third type of tasks is aimed at dealing with new data uploads performed by ESPON Database data providers. The interface allows administrators to see what new uploads exist and to retrieve the files in order to perform data checking and acquisition into the ESPON Database.

2 The Databases

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

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

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

2.1 The ontology database

Due to the size of the ontology database model, we split its presentation into several parts. Each of these parts is equivalent to one dimension of the data. From a conceptual point of view, the ontology database can be viewed as a hypercube with 5 dimensions to any statistical data:

1. The spatial dimension identifies precisely the spatial unit described by the indicator value;
2. The thematic dimension identifies precisely what type of statistical indicator is described by a given indicator value;
3. The lineage dimension identifies precisely the source of the indicator value (the database or the organization that published the value), as well as the transformations (corrections, estimations, etc.) that were applied to the value until its actual state;
4. The temporal dimension defines the moment in time or the period described by the indicator value;
5. The dataset dimension describes the publication (name, author, etc.) of which the indicator value is part of.

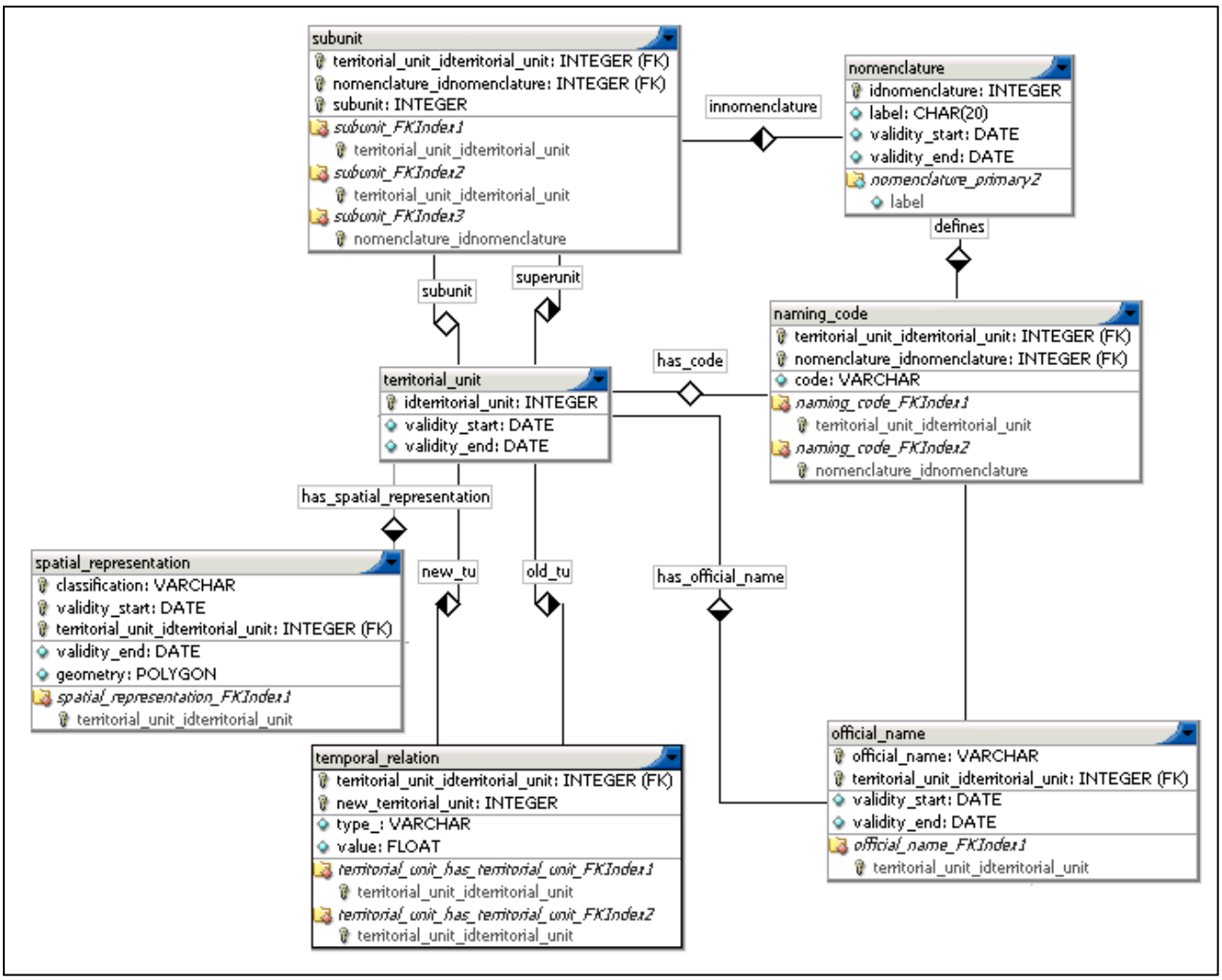


Figure 23: *The spatial dimension in the ontology base.*

The spatial description of the data is given via a spatiotemporal ontology. This ontology describes a complete list of the administrative units, stored in the territorial_unit table.

Territorial units are described as parts of a nomenclature. A nomenclature (see table *nomenclature* in ***Figure 23***) defines two main properties for territorial units: their code, which is an abstract, standard and unambiguous name based on conventions (see table *naming_code* in ***Figure 23***), and their hierarchical inclusion relations (see table *subunit* in ***Figure 23***). Nomenclatures can evolve in time (see attributes *validity_start* and *validity_end* in the table nomenclature), and, as such, units codes can also change.

Territorial units also have official names that can change in time (see table *official_name* in ***Figure 23***).

The temporal relations between territorial units are depicted via the table *temporal_relation*, which contains an attribute (see *type* attribute in ***Figure 23***) that allows describing the specific type of relation linking two territorial units.

Each territorial unit can also have several spatial representations (see table *spatial_representation* in ***Figure 23***), different geometries being fit for different purposes (e.g. more detailed geometries for applying geometrical operations and simplified geometries for display on screen). The spatial representation of a territorial unit can also evolve in time.

The thematic description of the indicator values is given via a thematic ontology. The table *theme* stores the ontology of themes and subthemes used in order to class the indicators in the ESPON Database (see ***Figure 24***). Each indicator must belong to one or more themes and subthemes.
Keywords (see table *keyword* in ***Figure 24***) are used in order to add more thematic information to indicators. Although keywords are not used by the current version of the ESPON 2013 Database Application, future versions may use keywords in order to make faster and more precise querying possible
The table *units* is required in order to keep track of the different measure units for the same indicator. This covers conversions between different measure systems and also multiples. For instance, the same indicator, total population, can be expressed in one dataset in thousands of inhabitants and, in another dataset, in inhabitants.
The table *indicator_relation* is a generic table (can be used for a specific type of relation by changing the type attribute) designed to store vertical and horizontal relations between indicators.

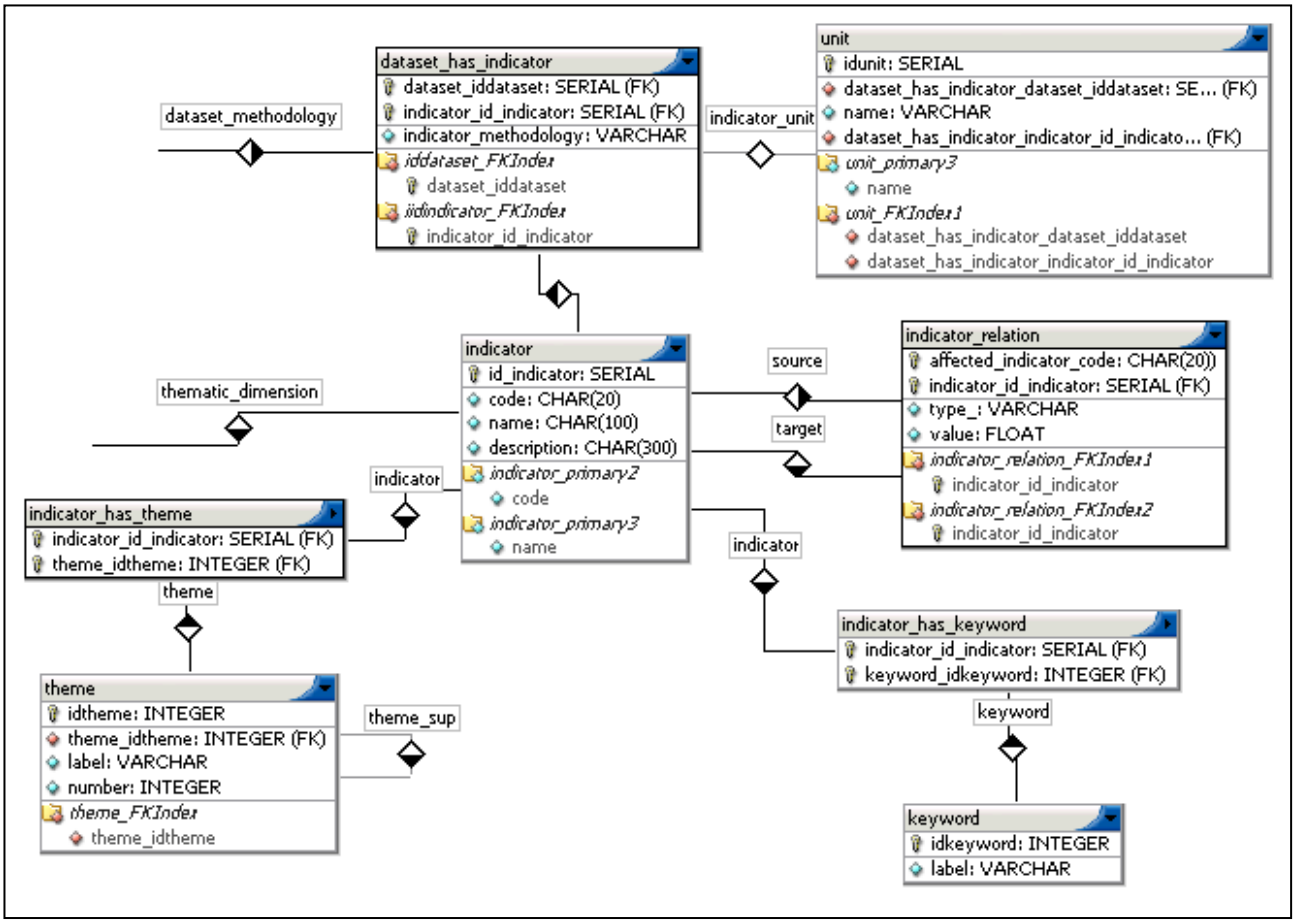


Figure 24: *The thematic dimension in the ontology base.*

The description of the lineage of an indicator value is given in the *lineage* table (see ***Figure 25***). The lineage of an indicator value is described by three elements: the original data provider (table *provider*), the methodology for the transformations applied to the data (table *methodology*), and the copyright constraints attached to the indicator value (table *constraints*).

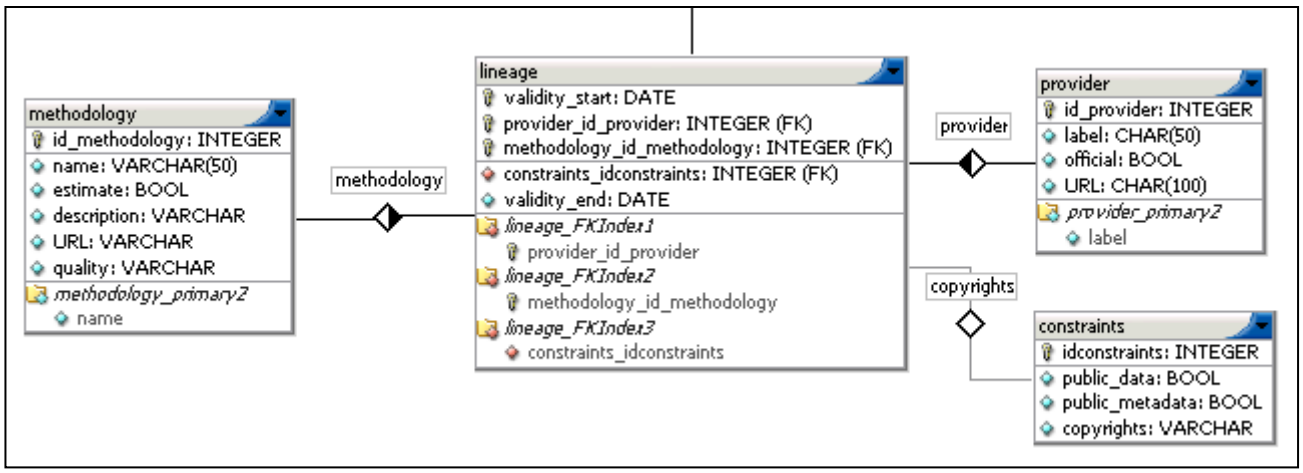


Figure 25: *The lineage dimension in the ontology base.*

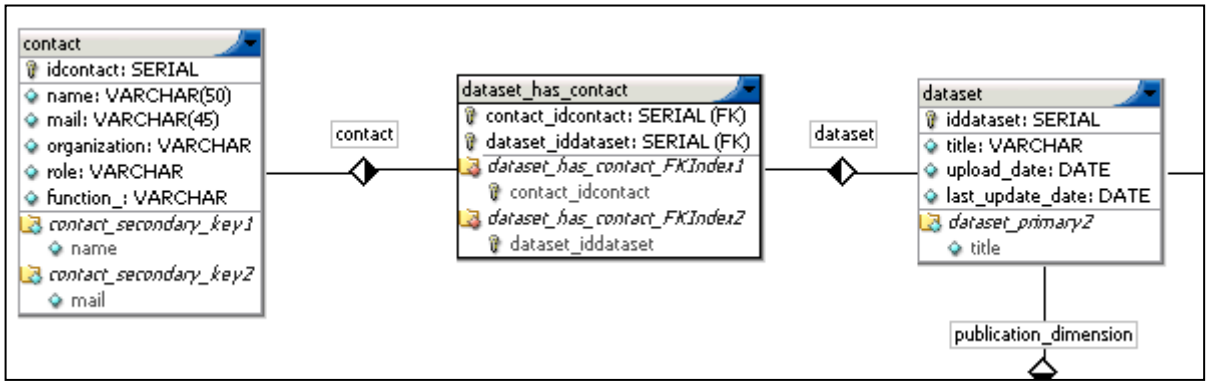


Figure 26: *The publication dimension in the ontology base.*

A publication is described in the ontology database through the table *dataset* (see ***Figure 26***). Another type of information about the dataset maintained in the ontology database is about the person or persons who submitted the publication (see table *contact*). For one dataset, there can be more than one author, and the same author can work on several datasets. Besides keeping a trace of data deliveries from ESPON projects, the coordinates of the contact person from projects are kept with a double purpose, in order to help with the fast filling of the metadata in the Web metadata editor (where, through a connection to the ontology database a list of contacts will be displayed) and in order to manage user access to the ESPON database (as most of the data providers will be members of ESPON projects that will also need privileged access to the ESPON database).

In order to fill the ontology base with data from the data and metadata files uploaded by data providers, a custom data acquisition software has been developed. The main concept behind this software is that of mapping. A generic, open source software package allows describing mappings between documents and databases. This allows taking values described in text documents and inserting them in a database schema (here, in the ontology database), at the right place. However, before actually inserting the data in the database, a preformatting of the excel file is needed (this is done directly in the Excel files, via *Visual Basic* scripts).

After preformatting, metadata and data files are input in the Java desktop application and checked for errors. There are three categories of errors to be corrected: metadata errors, data errors and data/metadata correspondence

errors (e.g. when indicators described in the metadata are not present in the data or *viceversa*). The errors are pointed out to the application administrator, who must correct them before entering the data and metadata in the ontology database.

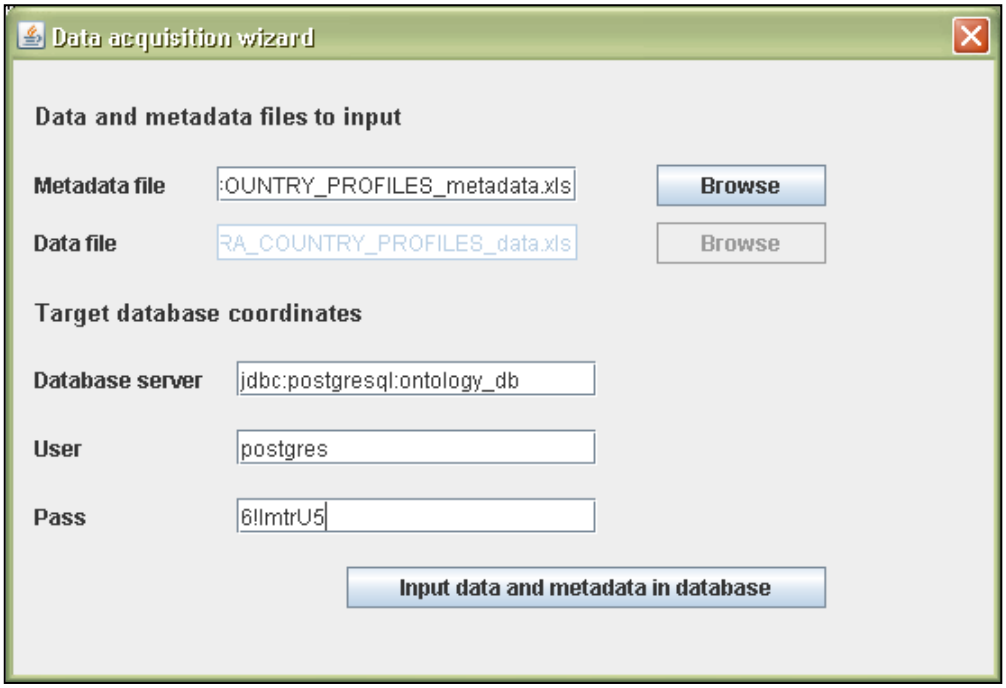


Figure 27: *The data acquisition application.*

Once data are in the ontology base, they are ready to be processed and enriched via automatic harmonization processes. Although automatic harmonization methods were not developed yet, in future phases of the ESPON 2013 Database Application they are expected to be.
After the harmonization process, the data and metadata are ready to be transferred to the ESPON Database, in order to be queried and downloaded by the final users. The transfer between the ontology database and the ESPON Database is done via an SQL script, which is used in order to make the conversion between the two schemas.

2.2 The Espon database

The ESPON 2013 Database uses a complex data structure representation that is based on the principles of reflecting the main features of the ESPON metadata scheme and introducing some conceptual entities that allow to optimize data access in a Web application context.
The present paper describes the core elements of the ESPON 2013 database and introduces the algorithm that is used to search statistical data in the database and to produce the information about the completeness of the existing data.

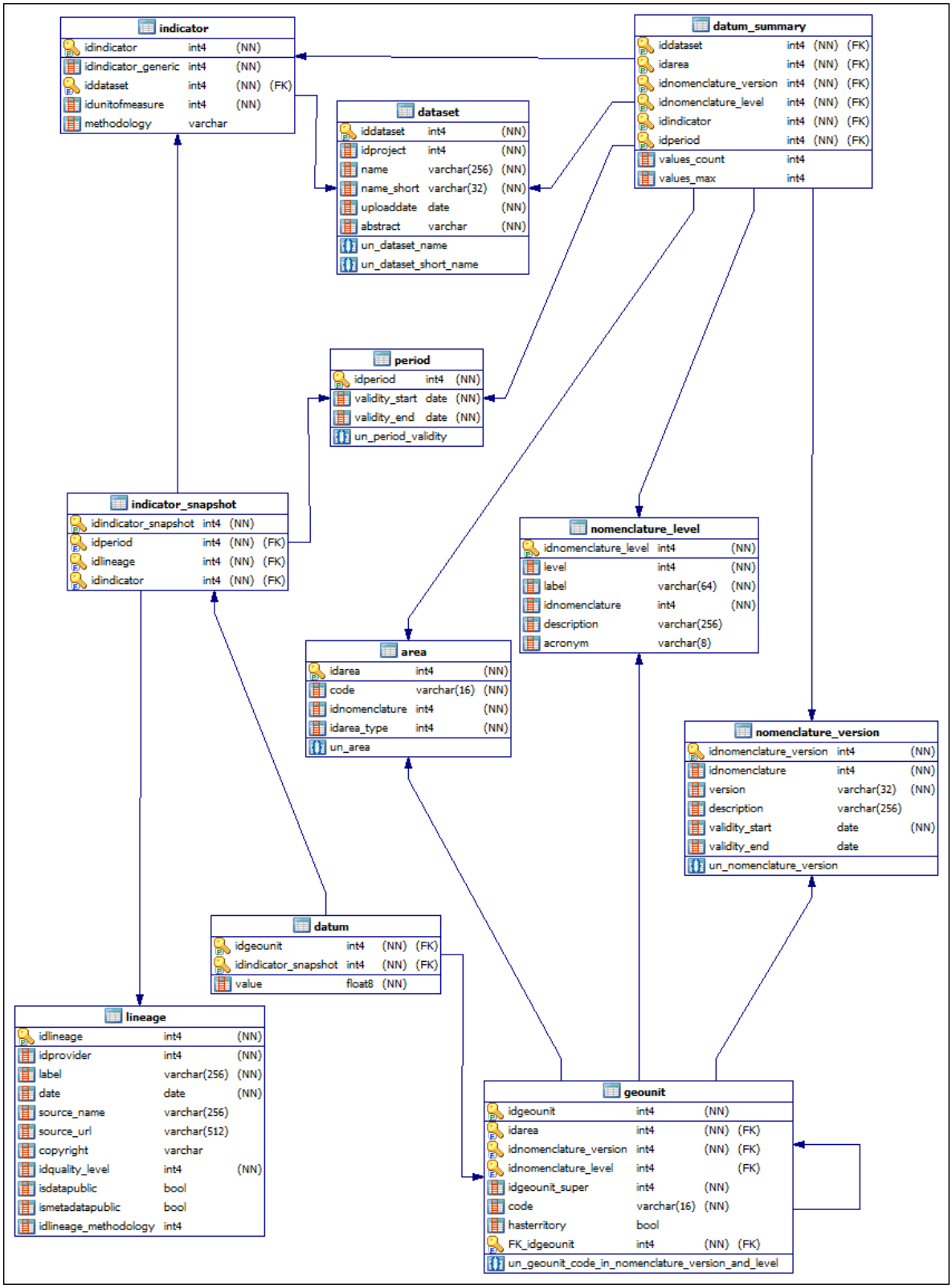
2.2.1 Core data structures in ESPON 2013 Database

Most of the elements of the essential ESPON metadata structure are reflected directly in the ESPON 2013 database. For example, a dataset, the main element of the metadata structure, is represented in the database as an entry in the table called DATASET. The columns of this table reference the mandatory attributes of a dataset in terms of the metadata: the name of the dataset, its creation (upload) date and its abstract. Some additional attributes have been also created: each dataset table entry has a serial ID that facilitates search processing, and a short version of the name, that is used in the Web application user interface to harmonize the view aspect of the search form items. This principle of almost direct reflection of the metadata structure concerns multiple entities used to describe the ESPON statistical data: contacts, indicators, thematic classification and keywords, lineages and other elements are represented using this principle.
Another side of the ESPON 2013 database scheme is the representation of territorial units used for statistics. They are not described by the ESPON metadata, but are directly taken from the official documentation on territorial units for statistics. The datasets already processed for use in the Web application reference different versions of the NUTS nomenclature. So, the part of the ESPON database describing territorial (geographic) units contains a library of nomenclature and geographic units data for the NUTS of versions 1995, 1999, 2003 and 2006. This library was created during the database building process and is based on official documents fully covering the corresponding nomenclature versions.
The ESPON statistical data represent a link between metadata information and a territorial unit. For example, the total population (an ESPON indicator) is measured for different territorial units. The link between them is an indicator value stored in the table called DATUM. In the context of the ESPON metadata structure, this association becomes complex because ESPON statistical indicators have two dimensions. The first one is the period of validity of the indicator value. The total population indicator changes in time, it varies according to the demographic, social and economic trends characterizing the region to which belongs the territorial unit. This means that the database scheme must take into account the fact that there may exist as many associations between the indicator and the territorial unit, as there are time periods for which the value was calculated. The second dimension is the source reference for the indicator value, called lineage. The same indicator for the same period of validity may be calculated by different statistical organizations or research actors. Each indicator value must be associated with its source of production in order to be valid.
The amount of statistical data produced by ESPON projects is very important: the smallest datasets already processed contain thousands of indicator values; the hugest ones can contain hundreds of indicators, coupled with periods and lineages references they give hundreds of thousands of values registered. These quantities of data are difficult to store and manipulate in a complex table, so, while designing the database structure, we decided to make the DATUM table as simple as possible. To obtain a simplified data structure, we

introduced the notion of the indicator snapshot, which associates an indicator with its time and source dimensions. These snapshots are stored in the table called INDICATOR_SNAPSHOT. Thus, the DATUM table can contain no more than three columns: a territorial unit reference, an indicator snapshot reference and a numeric value corresponding to the association between them. This approach facilitates the data search and significantly optimizes the access to the database table.

Even if the main statistical data table can be considered as simplified and optimized for access, it remains very inefficient for data search when general criteria are applied, because of the quantities of data present in it. In fact, all the search requests that can be made through the Web application interface concern higher-level entities and never descend to indicator values level or to small territorial units: the search can be made by dataset, theme, country, nomenclature level or version and covered period. This important detail gave us the idea to design a particular table in the database, that contains a generalized view of the low-level data registered. That table was called DATUM_SUMMARY; it groups the existing statistical values by the main search criteria of the application and contains about ten times less entries than the DATUM table. It does not provide the statistical values, but the information about the values that exist. Thus, it contains six foreign-key columns that reference the dataset, the geographic area, the nomenclature version and level, the indicator and the period. Two other columns contain the number of indicator values existing in the database for the combination of the foreign keys and the maximum number of these values possible for the area and the level referenced by the table entry. The DATUM_SUMMARY table allows to process search requests without accessing the DATUM table and to extract all the necessary information to produce search results. This table ensures the rapidity of search in the Web application context. It is also necessary to add that DATUM_SUMMARY table is maintained automatically, without any interference from the application administrator's side: the entries of this table are calculated and stored by internal database requests when the application is executing its initialization sequence.

The following diagram presents the relationships between the core tables of the ESPON 2013 database.



2.2.2 Data search algorithm

The search algorithm used in the ESPON 2013 Web application consists of two phases.

During the first phase, the search is done using the simple search criteria of the search form (search by dataset, theme or indicators). This search is done independently of the type of search requested by the user. If he requests a simple search, the result of this phase is already the final search result that will appear in the search results table. If the user requests an advanced search, the simple search algorithm is executed with the aim of restricting the amount of data to search in, because it helps to optimize the advanced search algorithm and to make it more rapid.
The design of the application takes into account that the objects retrievable from the database are relatively constant, they do not evolve after being placed into the database unless corrections are made by the application maintainers in the data. That is why we found that one of the ways to improve the application performance is to define a set of objects that are cached in the application memory and that can be directly accessed without the necessity to query the database. This collection of objects is built upon the application start up or when the database connection parameters have changed. Objects like datasets, indicators, thematic data and some others are stored in this collection. The simple search algorithm finds the objects cached in this collection without querying the database and produces the search result. For example, if the user selects a project in the simple search form and submits a search request, the search algorithm simply collects all the datasets that correspond to the selected project and produces the search result. The same sequence is applied for the other simple criteria - topics and indicators selections.
The second phase of the search algorithm is executed only in case of an advanced search request. If the user has made at least one selection in the simple search form, the simple search algorithm is applied before this phase. If there were no selections in the simple search form, but only in the advanced form, an additional database request is made to retrieve candidate dataset and indicator identifiers that can restrict the field of the main advanced search request. That search for candidates is made on the DATUM_SUMMARY without significant joins on other tables and is usually very quick. If the simple search made previously or the candidate search give at least one result, we may conclude that there is a high probability of finding data for the final result and we build a collection of preliminary results using the data retrieved. If both the results are empty, the advanced search algorithm does not continue and the application outputs an empty search results table.
The next step of the advanced search phase is to filter the candidate results according to the search criteria imposed by the user and to retrieve all the necessary information to fill the search results table. The dynamic SQL request in this case is more complex than in the candidates search step, it uses the DATUM_SUMMARY table joined with other tables that contain data for building the results: NOMENCALTURE_VERSION, PERIOD and DATASET. It is parametrized using the indicator values of the candidate search result and thus executes very quickly. If the response for this database request contains at least one row of data, that means that the information found corresponds to the search criteria specified by the user and can be transformed into a search result object. For each search result object, the algorithm defines the nomenclature versions and levels, the periods and the geographic areas found.

It also calculates the priority of the search result: how many criteria specified by the user were found for this object. The priority value is used to sort the entries in the search results table. Finally, after executing all the advanced search results, the remaining part of the algorithm calculates the completeness values of each result object.
The complete search algorithm used by the application can be summarized in the following pseudo-code listing:

```
Find the cached objects that correspond to the user's search
criteria.

Produce simple search results (collection CS).

If (it's a simple search)
    output the search results table with CS and exit the
algorithm.

If (CS is empty)
    Build the candidates search query and execute it to produce CS
with elements.

If (CS is still empty)
    Output an empty search results table and exit the algorithm.

Build the advanced search query and execute it with the data of
each element of CS (collection CA).

For each element in CA
    calculate the completeness of the data found.

Sort by priority the elements of CA, if any.

Output the search results table with the elements of CA, if any.

End of the algorithm.
```

2.2.3 Data completeness calculation

In the ESPON 2013 Database Web application context, the completeness of data can be defined as the percentage of effectively present indicator values against the maximum number of values possible for the given indicators, time periods and study areas. A dataset has the 100% completeness if all its indicators for all time periods have a value associated with each territorial unit, on any level of nomenclature hierarchy. The completeness value is a constant for a dataset taken entirely, but it can vary significantly between dataset subsets of indicators, time periods and territorial units. The completeness does not show the quality of data; it means only how much data is really present in

the selected range of criteria against the maximum amount of data that could be present for this selection if all combinations of spatial, temporal and thematic criteria had a value for these criteria.
During the application development, we considered two ways of calculating the completeness value. The first one was to calculate the completeness on-the-fly, counting each value present and each value absent, then producing the percentage. This a very exact, but at the same time a very expensive method of calculation.
The way that we finally chose to produce completeness values for search results is to use the DATUM_SUMMARY table. The contents of this table is always up to date with the one of the DATUM table. Two columns of DATUM_SUMMARY, values_count and values_max, contain information about the effective and maximum number of indicator values present for the given indicator, time period, study area, nomenclature version and level in the DATUM table.
The calculation of completeness values is made when search result objects are already produced. They contain information about the indicators, periods, study areas and nomenclature details found. A simple SQL query allows to summarize the number of existing values for each search result and the maximum number of values possible. This sum is calculated for each level of nomenclature separately: that allows to show the completeness details per levels in the search results table. The sum of the completeness values by levels give the general completeness of the search result. The percentage value is calculated using the very simple formula:

```
Completeness = values_count * 100 / values_max
```

The SQL query built to retrieve the completeness data may be illustrated by the following example:

```
SELECT idnomenclature_level, SUM(values_count), SUM(values_max)
FROM datum_summary
WHERE idarea IN ({list or area IDs})
AND idnomenclature_version IN ({list of nomenclature versions
IDs})
AND idnomenclature_level IN ({list of nomenclature levels IDs})
AND idindicator IN ({list of indicator IDs})
AND idperiod IN ({list of period IDs})
GROUP BY idnomenclature_level;
```

The result of this query is produced in a very short time and lets to calculate the exact values of search result data completeness without demanding additional algorithmic procedures.

Conclusion and future works

The specific achievements for the current phase of the ESPON 2013 Database Project from an application point of view can be divided into three categories: metadata related achievements, data related achievements and interface related achievements.
From a metadata point of view, the outputs of the ESPON 2013 Database Project include:

- a metadata profile specifically dedicated to statistical territorial data, compliant with existing standards (ISO 19115, the INSPIRE directive);
- an Excel template allowing ESPON 2013 DB data providers to fill compliant metadata offline;
- a Web application allowing ESPON 2013 DB data providers to fill compliant metadata online, to import metadata files in XLS and XML formats, to correct syntactical errors and to export metadata, again, in either XLS or XML format. The Web application is completely integrated within the ESPON DB data flow and within the ESPON 2013 DB Web Application.

From a data point of view, the outputs of the ESPON 2013 Database Project include:

- two operational database schemas optimized for data harmonization and, respectively, for fast querying;
- a software suite allowing importing data from XLS files into the ontology database and transferring them to the ESPON database, after harmonization;
- a complete ontology of territorial changes for the NUTS nomenclature between 1995-2006;
- a coding scheme and a thematic ontology for indicators;
- an operational relational database filled with regional (NUTS) data supplied by ESPON Projects, complying with the ESPON DB format (over 1 million records);
- an operational file database filled with all the remaining data delivered by ESPON projects that do not fit in the ESPON database yet (local data, cities data, world data, environmental data, etc.).

From an interface point of view, the outputs of the ESPON 2013 Database Project include:

- an operational Web application, allowing to fulfill the complete ESPON DB data flow, from user registration, data upload to data search and download;
- a database administration interface, integrated in the same Web application, allowing administrators to reconfigure the application on the fly.

Future improvements to be brought to the ESPON 2013 Database Application can also be divided into the same three categories.

From a metadata point of view, in future phases of the ESPON 2013 Database Project the TPG will have to ensure that the metadata profile is coherent with the data formats used for including new types of data in the ESPON database, i.e. continuous spatial data (usually referred to as environmental data) and flow data (i.e. statistical information describing not one territorial unit, but a relation between two territorial units).
From a data point of view, future outputs of the ESPON 2013 Database Project should include:

- extending the ESPON DB schema in order to accommodate discrete flow data (i.e. occuring between territorial units, like regions or cities);
- designing and implementing some automatic data harmonization and estimation methods;
- designing and implementing a solution for storing spatially continuous data (environmental data) and for easy conversion of this data towards a discrete format (i.e., the actual model of the Espon database);
- building spatial ontologies for world territorial units and, as much as possible, for local units and cities (completeness is hard to achieve within this scope, due to constant changes);
- continuously filling the ESPON DB with spatially discrete statistical data (regional, local, world and cities) data, flow data and spatially continuous data.

From an interface point of view, future outputs of the ESPON 2013 Database Project should include:

- extending the ESPON DB Web Application with new types of search (i.e. keyword search);
- extending the ESPON DB Web Application with new search criteria (i.e. spatial typology criteria);
- making the search interface more flexible (i.e. allowing spatial selection of other types of units than countries).