

Title	Functionality for data access and data exploration (D5.4)
Work Package	WP5
Authors	Atle Alvheim, Reiner Mauer
Date	August 2009
Dissemination Level	PU (Public)

Summary/abstract

In a future CESSDA Portal, two main interconnected changes will be expected:

- a) Most data will be documented according to DDI 3.+ as the metadata standard, and;
- b) As a direct outcome of this, there will be substantial amounts of data organised as complex collections.

Technologies to explore data located via the CESSDA portal could be divided between:

- a) The ability to find and put together simple files;
- b) And the ability to keep and explore the structure of complex files, if such files are available.

Technologies require the ability to look up metadata at different levels: collection; study level; topic level; question level; variable level. More detailed suggestions:

- a) Display complex instances in two steps, show structure and content;
- b) Ability to load more than one single study via an enhanced data loader that can open several windows;
- c) Allow temporary recode of data;
- d) Ability to move freely between initial hit-list and exploration windows, with ability to re-run searches, based on synonyms, or similar.

WP5.4 Functionality for data access and data exploration

This document is written as background for planning enhancements to data access and data exploration functionalities in the new CESSDA data portal.

The present [CESSDA data catalogue](#) provides a seamless interface to data resources and variables within a large selection of datasets from social science data archives across Europe. Not all of the CESSDA data archives are as yet registered with the portal and for a variety of reasons some of the archives publish only part of their holdings in this catalogue. Each CESSDA archive maintains its own data catalogue, which generally contains information on most available sample based datasets in the archive. It is the explicit aim of the CESSDA-PPP project to extend the working of the portal to more types of data, to more complex organised data and to more varied technologies and functionalities, i.e. full DDI 3.0 functionality.

The Cessda Portal represents an implicit common catalogue for the collective data holdings of CESSDA member archives. For this implicit catalogue the portal is presenting one common data finding aid.

In the Catalogue, each dataset is described through its indexed metadata as a set of catalogue record containing bibliographic and methodological information, abstract, keywords and information on geographical and temporal coverage. When looking up data via the available search-/browse-technology these catalogue records are presented for further exploration at variable-, study-section, study- or archive repository level.

In the present portal, data may be located in several ways:

- Free Text Search;
- Browse by Topic - a CESSDA topical classification in two levels for studies;
- Browse by Keyword - the ELSST thesaurus provide the vocabulary;
- Browse by Data Publisher – a comprehensive listing of archive holdings.

The CESSDA Catalogue can be viewed in any of nine languages. The default language is dependant on the regional setting of the computer user.

Present working of the portal

The present portal represents a common catalogue of data published to the registered servers; in addition it has built in the ELSST multilingual thesaurus and a two-tier study classification scheme.

Free Text Search

A free text search can be performed on all areas of the metadata by entering a search term or phrase in a search box. Where the search terms match a concept in the multi-lingual thesaurus (ELSST), the search will be performed in all languages supported in ELSST. A list of all resources containing the search term(s) and their language equivalents will be displayed in the panel on the right hand side. The search term “Essex” is found in many studies, but not recorded in the thesaurus in any of the hierarchies or translations, therefore it generates only a list of hits, without any further functionality.

The screenshot shows the CESSDA Catalogue search interface. The search term is 'Essex', resulting in 102 hits. The results are displayed in a table with columns for Study, Section, Variable, Study, and Archive. The table lists various studies such as 'British General Election Study, 1992; Cross-Section Survey' and '[MTUSW552] Multinational Time Use Study : Release 2 of the World5.5 dataset'. The archive names include UKDA and ADP.

Study	Section	Variable	Study	Archive
British General Election Study, 1992; Cross-Section Survey				UKDA
[MTUSW552] Multinational Time Use Study : Release 2 of the World5.5 dataset				ADP
British General Election Study, 1987; Cross-Section Survey				UKDA
British General Election Study, 1997; Cross Section Survey				UKDA
British General Election Study, 2001; Cross-Section Survey				UKDA
British Election Study, 2005: Face-to-Face Survey				UKDA
[MTUSW552] Mednarodna anketa o porabi časa : Druga dopolnjena izdaja datoteke držav sveta				ADP
British Household Panel Survey; Waves 1-11, 1991-2002 : Teaching Dataset (Work, Family and Health)				UKDA
British Household Panel Survey; Waves 1-11, 1991-2002 : Teaching Dataset (Social and Political Attitudes)				UKDA
British Household Panel Survey; Waves 1-13, 1991-2004: Teaching Dataset (Time Use, Leisure and Social Memberships)				UKDA

Where a search term matches a concept in ELSST where there is a translation and a hierarchy of terms, related terms will be offered as suggestions to refine or broaden the search. Another option is to narrow the search to one specific language:

The screenshot shows the CESSDA Catalogue search interface for the term 'ABORTION (INDUCED)', resulting in 231 hits. The results are displayed in a table with columns for Section, Study, and Archive. The table lists studies such as 'Syn på abort' and 'Abort'. Below the table, there are sections for 'Related Terms', 'Top Terms', 'Broader Terms', and 'Search in Specific Language'.

Section	Study	Archive
Syn på abort	Miljøvernundersøkelsen 1995, befolkningsdelen	NSD
Syn på abort	Miljøvernundersøkelsen 1995, organisasjonsmedlemmer	NSD
Abort	Fruchtbarhetsundersøkelse - 1977	NSD Metadata
Abort	Omnibusundersøkelse mars 1995 (aktuelle samfunnsspørsmål)	NSD Metadata
Abort	ISSP 1991 Holdninger til religion, Norsk del	NSD Metadata
Abort	Undersøkelse om generasjonenes syn på arbeid, fritid og livsverdier, 1982	NSD Metadata

Related Terms:

- BIRTH CONTROL METHODS (1994 hits)
- CHILDBIRTH (420 hits)
- FAMILY PLANNING (2580 hits)
- GYNAECOLOGY (6 hits)
- LABOUR COMPLICATIONS (8 hits)
- MISCARRIAGE (431 hits)
- OBSTETRICS (40 hits)

Top Terms:

- MEDICAL SCIENCES

Broader Terms:

- OBSTETRIC SURGERY

Search in Specific Language:

- ABTREIBUNG (German) (62 hits)
- ABORTION (INDUCED) (English) (42 hits)
- ABORTO (Spanish) (2 hits)
- ABORTTI (Finnish) (32 hits)
- AVORTEMENT (French) (4 hits)
- ABORT (Norwegian) (231 hits)
- ABORT (Swedish) (231 hits)

Browse by Topic

Datasets are classified at study level by one or more of the two-tier topic classifications. The top levels can be expanded to show the second levels. By selecting a topic, a list of all resources classified with this concept in all supported languages will be displayed in the panel on the right hand side. Below the list of resources more specific terms will be offered as suggestions to refine the search, along with the option to narrow the search to one language.

The screenshot shows the CESSDA Catalogue search interface. The search term is "gender and gender roles". The results are displayed in a table with columns for "Study" and "Archive". The table lists several studies, including "Undersøkelse om familie og kjønnsroller, 1994 ISSP, Norsk del", "Ønsker om og behov for sysselsetting blant gifte kvinner 1968(Yrkesaktive)", and "Advertising and changes in society 1950-1975". The left sidebar shows a hierarchical tree of categories, with "SOCIAL STRATIFICATION AND GROUPINGS" expanded to show sub-terms like "gender and gender roles".

Browse by Keyword

The concepts from the multi-lingual thesaurus, ELSST, are displayed in multi-level hierarchies. The initial top level terms can be expanded to show the narrower terms in the hierarchy. Note that not all of the thesaurus concepts are displayed in this hierarchical tree. Only those (either directly or through a narrower term) that will lead to resources being discovered are listed.

By selecting a term a free text search will be performed across all areas of the metadata using that term and all its synonyms in all supported languages.

The "Assigned Keyword Search" displays a list of resources which have had the selected concept in any of the supported languages assigned as a controlled vocabulary keyword.

The top tabs "Study", "Section" and "Variable", which function in the **free text search** or **browse by keyword**, offer the ability to confine the search to the metadata at study, section (variable groups) or variable level.

Browse by Data Publisher

This facility allows a simple browsing of all datasets in the CESSDA Catalogue from each of the contributing archives.

How to Examine a Resource

The search results are displayed in the right hand panel. The listings function as links, when activated a Nesstar WebView window will open displaying the resource. It is possible to use the WebView interface to examine the resource metadata and depending on access and registration conditions, perform analysis online.

There is presently used a specific blue icon displayed in connection with a resource title in the hit list to signal the possibility to display a translation of selected metadata, including the topic classifications and thesaurus keywords that have been assigned to that resource.

We could outline several types of adjustments to the portal:

1. In relation to the present portal, a stricter discipline on inserting concepts and keywords where appropriate is probably the most efficient expansion possible. This could be obtained by incorporating the thesaurus into documentation/publishing tools. This action requires both tool development and work practices development.
2. It will be possible to develop the freetext-search by allowing for more complex searches, e.g. using “AND” or “OR”, etc. This could however be complicated, because terms are matched against the hierarchies of the thesaurus before searches are activated, and complex search terms are difficult to meaningfully match and translate.
3. It could be possible to better organise search output, the “hit lists” better. This has been evaluated and the conclusion so far has been that this is quite difficult, hit lists are now grouped by metadata level and further grouping variables are not obvious. One solution could be to try to implement further searches within hit lists. This has so far been implemented by using concept hierarchies and synonyms, but such a strategy is not as explicit as searches within hit-lists.
4. The best technique to help exploration would be to open more than one study or “hit” at the time and allow easier visual inspection. This is the main recommended solution.

Changes represented by the enhanced CESSDA Portal

At the Cologne workshop in October 2008 the participants discussed characteristics of a CESSDA central data portal. Some points relevant for functionality for data access and data exploration were identified:

1. Data exploration basically is an activity that works on the metadata part of a “data package”, a documented file or collection of files, both with and without standardised frequencies;
2. For exploration and comparison purposes a data-viewer should be able to load data from more than one file and more than one repository;
3. The metadata standard should build on DDI, either version 2 or 3. Dublin Core should also be supported where relevant;
4. Accessing metadata should be free of restrictions and should not require authorization;
5. Accessing data across national borders requires an explicit access policy;
6. The portal should be able to store complex data instances and to retrieve on basis of relationships;
7. A work-bench concept is not of primary importance;
8. Versioning of data should allow identification of most recent and authorised version of data;
9. It should be possible to link micro- and macro-data;
10. The whole infrastructure system needs a system of Persistent Identifiers;
11. The multilingual thesaurus and a concepts and harmonisation database should be incorporated and made available for functionality development.

In short: These discussions did not result in many suggestions for a better search interface.

From the user viewpoint, what are the basic changes represented by the enhanced portal being planned? So far the analysis has shown two distinct enhancements to be implemented in the renewed CESSDA data portal.

- a) Most data will be documented according to DDI 3+ as metadata standard, and
- b) As a direct outcome of this, there will be substantial amounts of data organised as complex collections

The majority of DDI instances that may be published and displayed in the portal as collections will be categorised as grouped by design, although various kinds of ad hoc grouping may be developed. DDI 3 allows implementation of a strong grouping possibility. The major practical problem is that presently we have no good tools for producing DDI3 complex instances, but in the suggested setup of sequence for creating DDI-files, grouping and recording of comparison data are generally reserved as the last step of the data documentation process. This probably indicates that these kinds of activities are fairly close, almost overlapping with research activities. It is to be expected that the GROUPING of single studies both will be actual as outcome of archival documentation processes and as necessary functionality under portal data exploration. Recording of comparative information / use of the COMPARATIVE module also naturally contains some information that if possible should be developed as part of archival preparation work, in particular Universe, Concept and Question relationships across national simple files in comparative collections. When documenting comparative collections, it could be a point in the best practices list that whatever may be defined as comparative by the COMPARATIVE module actually should by default have that

information inserted in the documentation, so that from the start it becomes actionable. However, it is to be expected that the COMPARATIVE module will be closer to the user perspective of working with the data and recording outcomes of such work.

The actual resulting comparative- and grouping information coming out of this data exploration process will be very complicated to lead back into a data repository as new versions. Both technically and in terms of coherent data maintenance any solutions in that direction have to be very carefully evaluated. However, for preparation of data before user download, outcomes of a comparative exploration process should naturally follow data.

Expansion of portal functionality as a tool for exploration and comparison

Science rarely stops with descriptions, science is about relationships. In contrast with this, data are often representing a specific time-point or time-interval and a specific spatial reference. Most unit-record sample based data are collected using one specific instrument: one questionnaire. This generates a need to **link and compare**, to put data together for generation of relative measurements, contrasts, trends, etc

For some data this aspect is incorporated already in the data collection or data preparation process, we collect and store complex organised data, data are measured also across third dimensions, over geographies and time. We need to keep the original complexity available through the exploration process.

Therefore the data location problem leads naturally over into the data exploration problems, both data location and data exploration are triggered by the need to find out how useful data may be for analysis of specific problems. One preparation aspect of this is to get rid of all that is of limited relevance.

Technologies to explore data located via the CESSDA portal could be divided between:

- a) The ability to find and put together/contrast simple rectangular files;
- b) And the ability to keep and explore the structure of complex files, if such files are available.

Technology requires the ability to look up metadata at different levels: collection; study level; topic level; question level; and variable level. Since the present portal only holds simple files documented according to DDI 2, only study level, topic level and variable level is displayed. In the present data model questions are linked to variables and may not be extracted as a metadata element that is hierarchically severed from the variable.

For exploration possibilities to be of any use, they have to give users the possibility to work themselves, to identify and compare. It is fairly limited what a general web-based solution can offer compared to what a user can usually do on their own equipment with more well-known technologies.

Table 1 tries to summarise the portal functionality-related questions, possibilities and problems.

1. The present CESSDA portal solution is contrasted against the ideas for an enhanced Portal solution
2. Data published by data archives are categorised as either being simple (one rectangular file) or complex (collections of several rectangular files interconnected)
3. The first portal functionality is focused on data location, via search or browse against an index
4. The outcome of the data location process is a hit-list, broken up in 3, 4 or 5 metadata levels, hit-lists may be optionally expanded by time-point and potentially sorted
5. The explore procedure loads and displays further metadata about instances. There are depicted 4 enhancements compared to the present version:
 - a) Display complex instances in two steps, show structure and content;
 - b) Ability to load more than one single study via an enhanced data loader that can open several windows;
 - c) Allow temporary recode of data;
 - d) Ability to move freely between initial hit-list and exploration windows, with the ability to rerun searches, based on synonyms, or similar.

	Present	Future				
Data		Simple	Complex instances		Model	
Metadata standard	DDI2	DDI2/DDI3	DDI3			
Search proc	Via ELSST Against index Give concepts, synonyms and translations. Support search across languages	Via index → registry	<u>GROUP</u> -module Group simple files over time, panels, geography	<u>Comparative</u> module Measure more detailed relationships between single files. May substitute for <u>GROUP</u>	Micro-macro relationships	Questions and variables to be modelled as separate elements Expand models with <u>GROUP</u> -related elements
Outcome	Hit-list presented at 3 metadata levels	Hit-list presented at 4 metadata levels	Hit-list presented at 5 metadata levels	Hit-list presented at 5 metadata levels	Hit-list	
Additional information to be indexed		Questions to be severed from variables in the hit list through enhanced modelling	Instance abstract Concepts Series-statement Coverage Archive GroupType	Dynamic or user oriented.	Application oriented: Either pre-defined files or file integration	
Explore proc	Load and display One study at the time, via Nesstar Webview	Load and display more than one study selected from the hit list	Load and display grouping elements Show structure of instance as basis for detailed selection of simple			
		Temporary recode				
Further search, sort, filter display		Comparable data	Comparative data			

The table on the previous page should be read in the following way:

1. In the present portal, all data are documented according to DDI2. The present search procedure matches search terms given as input against ELSST. If a search term is matched by an ELSST term, the term is expanded with all translations provided by ELSST. After a potential expansion of such a search term via ELSST, a search is conducted against the index. The hit-list is presented at 3 levels, study, section of study and variable. The explore procedure is to load the study into the Nesstar WebView. If the hit is on a variable, the variable will be in focus when the file is loaded. If the variable is organised in a group, then it will be shown “in context”.
2. If we shift to the suggestions for the future portal, the first column indicates that still a large part of all search results will point to simple square files. However, such files could be documented according to DDI3, which basically would not expand the information content very much. However, a more detailed metadata modelling will model a better picture of the variable/question relationship, and questions and variables could be displayed as logically separate levels of the metadata. This would give us an elementary Question Database

The basic exploration functionality that may be developed is the possible loading and display of more than one hit at a time, hits across studies in the same repository or studies located in separate repositories. By activating hits, an exploration window may be opened and make it possible to visually compare, in particular variables or questions.

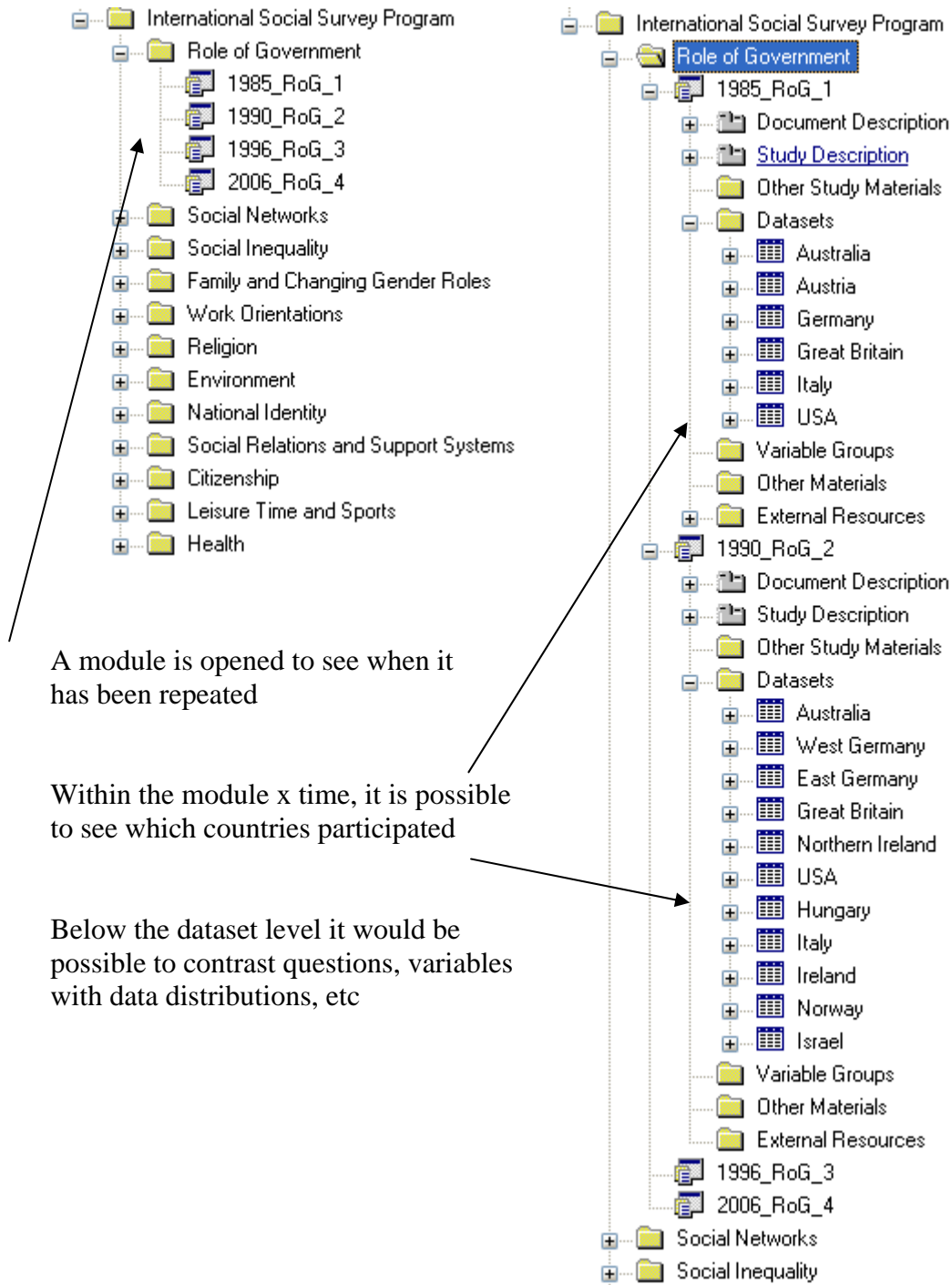
A routine for automated comparison of pieces of text should be developed.

3. We can distinguish three types of complex instances.
 - a) GROUPEd instances, over time (i.e. modules in ISSP), over panels and over geography (comparative/universes);
 - b) Comparative instances where the comparison is measured over substantive criteria, concepts, questions, variables, etc;
 - c) Multi-level files.

For GROUPEd instances, it will be of interest to present a top instance level abstract, with the possibility of displaying further information at this level, in particular additional documents stored at this logical level. Beyond that, the default view should basically give the structure of complexity as a list, where users may expand level, like shown:

ISSP	+ Module	+ Timepoint	+ Dataset
	+ Role of Gov't	1985	Australia
	+ Social Networks	1990	Austria
		1996	
		2006	

The practical visualisation of this could be presented like the illustration on the next page:



Treatment of hierarchical datasets

Hierarchical datasets are datasets with information about units at different levels. Such datasets could be created through the documentation process or through the data location and exploration process. They are often created from different sets of data being linked on key variables. E.g. one dataset could hold information at country level; another dataset holds information about individuals. One variable characterising individuals is that they live in a country. The country variable in an individual level dataset makes it possible to link in further information about the country. For many research-purposes it would be convenient if this was part of a data-exploration phase.

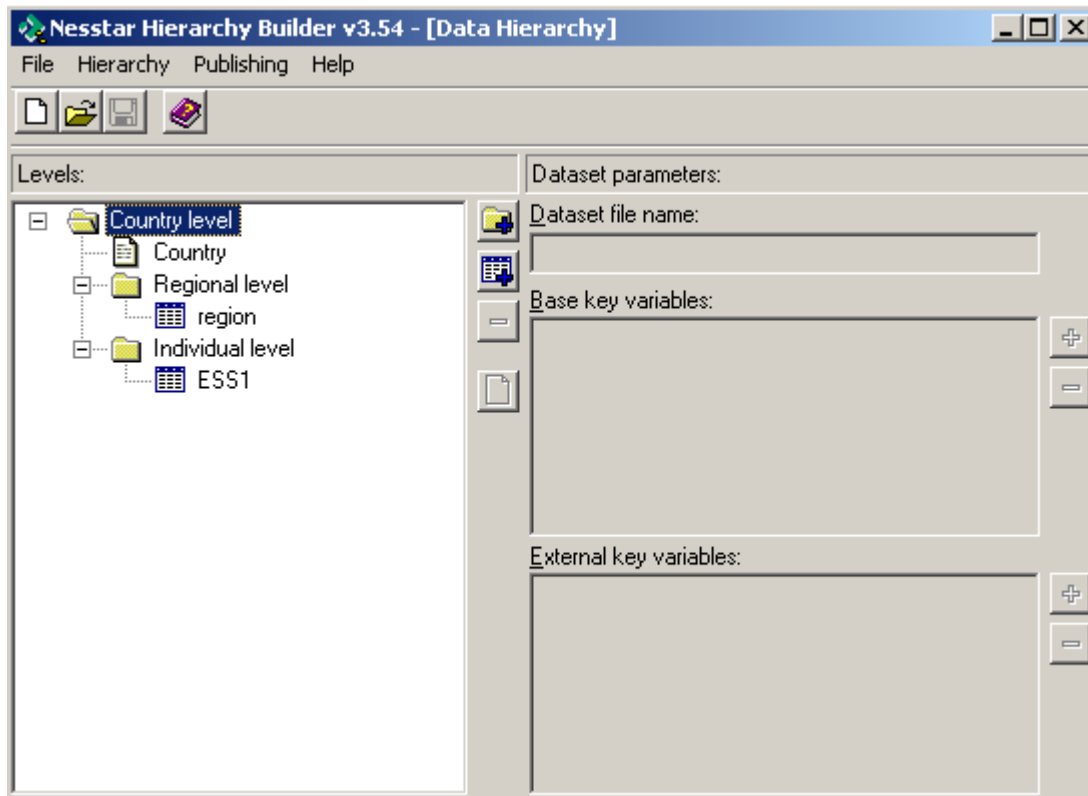
These kinds of files generally require some special documentation considerations.

Multilevel designs are of increasing importance. Therefore the CESSDA portal should make it possible to develop analytic files for such purposes. A general work procedure would be

1. Search for / identify relevant information on geographic or other relevant contextual units (could also be organisations, etc);
2. Bring together this information in one file;
3. Make sure the information is at the appropriate contextual level and that there are variables that identify the units;
4. Document original data and the processing of the original data;
5. Make a documented rectangular dataset for every contextual level (if this is geographic units, make one set for the country, one for the regions, etc). To illustrate, we here use the Nesstar Publisher;
6. Match-merge these datasets with the individual level data. Nesstar Publisher do this in a dedicated Hierarchy-builder;
7. The resulting file is to be treated as an ordinary file, although statistical analysis sometimes would require special statistical techniques, like hierarchical regression.

To illustrate: We could start from two distinct separate datasets, one with individual level data, 3.000 respondents over three countries, and one with country data for the same three countries.

The merging of the two datasets in the example below is validated by the country variable in each dataset. Country data are added to individuals.



Note that the “Study Description” of the dataset highest in the hierarchy becomes the default study description of the integrated dataset.

The individual level dataset contains information about individuals, e.g:

#	Gender	Country
1	Man	Norway
2	Man	Norway
3	Man	Germany
4	Woman	Germany
...		
3000	Woman	Sweden

The country level dataset could look like:

Country	BNP	Area
Norway	10000	Small
Germany	20000	Large
Sweden	15000	Medium

In the data loading software it should be possible to run elementary analyses selecting different types of variables. The documentation should tell what the different sources of such combined datasets are, and single variables that carry along its original descriptions.

Download:

When merging existing datasets, there is always a real danger that there will be many variables in the combined set. It is important to have an easy variable selection possibility, but identification variables should preferably be non-deletable.

