Project N°: 262608

ACRONYM: Data without Boundaries

DELIVERABLE D5.3

Report on Developed Routines

WORK PACKAGE 5

Servicing European Researchers in the use of OS Microdata

REPORTING PERIOD:	**From: Month 37**	**To: Month 48**
PROJECT START DATE:	**1st May 2011**	**DURATION: 48 Months**
DATE OF ISSUE OF DELIVERABLE:	**31 March 2015**	
DOCUMENT PREPARED BY:	**Partners 5, 1, 4, 8, 17**	**GESIS, CNRS-RQ, UL, RODA, FORS**

Combination of CP & CSA project funded by the European Community
Under the programme "FP7 - SP4 Capacities"
Priority 1.1.3: European Social Science Data Archives and remote access to Official Statistics

“The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 262608”

Summary[1]

This report provides a summary of the outputs of Work Package 5 Task 4 of the Data without Boundaries project. The core objective of this task was to provide routines which aid researchers in data preparation and analysis when conducting research with integrated European microdata. The key outputs of this work package can be summarized as follows.

As part of the work package routines were written to aid users in importing Eurostat microdata, which are delivered in .csv format, into statistical analysis software. These, 'setup files' as we term them, assist researchers in labeling variables and values and provide them with ready to use data files. Saving them time and easing their work with official statistics microdata.

As part of the efforts to generate such setup files an R-tool was developed, which allows users to generate such setup files for SPSS, Stata, SAS or R. This tool is available in the CRAN repository (http://cran.r-project.org/web/packages/DDIwR/index.html) and will generate files from a list of variables and labels.

Additionally tools were written which aid researchers in data preparation and analysis when working with European official statistics microdata. This includes for example a tool to calculate innovation concepts for the CIS or a tool which will calculate poverty thresholds for the longitudinal version of the EU-SILC.

[1] This report was prepared by Alexander Mack, Christof Wolf, Adrian Dusa, Janez Stebe, Sebastian Kocar, Andreas Perret, Alexia Ricard and Alexandre Diallo.

TABLE OF CONTENTS

Summary 4

1. Introduction 6

2. Setup files to read Eurostat data into statistical packages 6

2.1 EU-SILC 7

2.2 EU-LFS 7

2.3 CIS 8

2.4 SES 8

2.5 AES 9

2.6 R-Script for generating setup files 9

3. Tools and Routines 10

3.1 Integrating, harmonizing and processing routines for CIS 10

3.2 Innovation concepts for CIS 11

3.3 Technical report on harmonization of income data in EU-SILC 11

3.4 At risk of poverty thresholds for longitudinal EU-SILC 11

3.5 Indicators for SES 12

4. Conclusion 13

1. Introduction

The objective of this task was to assist users of integrated European microdata by developing routines for data preparation and analysis. The most important exercise here was to prepare what we term “setup files” for Eurostat data. Theses routines allow users to import Eurostat microdata, which are commonly delivered to researchers in comma separated value format, into the statistical package of their choice ready with standardized variable and value labels attached thus saving data users hours of preparatory work. Additionally a number of “microdata tools” were developed. These tools assist researchers by implementing social science scales, standardizing variables over time or between countries or aiding in data analysis by providing code and instruction for preparing datasets. These tools will often involve a more elaborate documentation or in some cases take the form of a technical report. Furthermore as part of this workpackage the DDIRW package for the R language was developed with which setup files can be generated automatically.

This report is structured as follows: section 2 details the process of generating setup files individually for EU-SILC, EU-LFS, CIS, SES and AES. Each subsection details the process of generating these files, which data documentation was employed and whether any problems were discovered with the Eurostat data. This section is concluded by a description of the DDIRW tool which was developed by RODA. Section 3 covers tools and routines and provides a short description of each developed routine. Section 4 provides a short conclusion.

2. Setup files to read Eurostat data into statistical packages

The highest priority for WP 5 Task 4 was to produce setup files for all OS microdata included in the documentation of WP Task 3 as these routines provide an invaluable service for all researchers employing these data. These datasets are rather complex and require a large initial time investment before one can begin with data analysis. A brief description of how Eurostat data are currently distributed is necessary to illustrate the benefits of such a service. Microdata for these datasets are distributed by Eurostat in .csv format. These files consist of a number of records separated by commas. Each record consists of fields or variables. Once the data file is opened, researchers need to consult the codebook in order to understand the meaning of variables and their labels. The codebook is provided in PDF format and it is left to the researcher to create routines which will label variables and values. The setup files handle this tedious task for researchers thus saving researchers hours of time which can instead be spent on data analysis. Additionally missing values and labels are harmonized thus easing comparability over time and between countries. In order to automate the process of generating these setup files for a wide range of statistical packages, all of which use different code, RODA developed a tool for the R programming language which can generate setup files from DDI based metadata. This package has been uploaded to the CRAN repository[2] and is thus freely available to the research community.

[2] http://cran.r-project.org/web/packages/DDIwR/index.html

2.1 EU-SILC

The EU-SILC User Database is delivered to researchers on DVD in .csv format. There are two data releases per year, one in spring which includes the newest cross sectional data and one in autumn which includes the most recent longitudinal data. Updated revisions of older data with error corrections are delivered on these DVDs as well. Both the longitudinal and cross sectional data contain four separate data files, two household and two person level files.

The most important document used for preparing the EU-SILC syntax files are the Guidelines from which variable and value labels are derived. Additionally the national quality reports were consulted for data checking, the document "Differences between Data Collected and UDB" provides information on anonymization, while the document "Problems and Modifications" provided by Eurostat on the DVDs was also consulted.

The setups were initially prepared manually in SPSS as well as Stata and were then updated for every new survey year as well as every new data release from Eurostat. They attach variable and value labels to all variables, recode alphanumeric variables to numeric format where necessary and where possible correct data errors. GESIS offers separate setup files for each year of the survey from 2005 through 2012 which retain the original data structure as provided by Eurostat. Also as of 2010 setup files are provided for all data revisions.

2.2 EU-LFS

EU-LFS data are disseminated by Eurostat as .csv files. Researchers granted access to data receive a DVD which includes all survey rounds to date, for the most recent 2013 release this includes data from 1983 to 2012. For each survey round the following data are included per country: 1 yearly file, 1 to 4 quarterly files, 1 ad-hoc file (for 1999 and for every year as of 2002), special household files (DK, FI and SE only). Additionally Eurostat delivers SAS macros which import .csv data into SAS and attach variable labels.

The core document employed in preparing setup files was the User Guide which includes information on variable and particularly value labels as well as information on anonymization of the scientific use files. Information on the ad-hoc modules was gathered from the respective Statistics Explained pages at Eurostat.

The GESIS setup files were written for SPSS on the basis of the SAS macros provided by Eurostat. The setup files provided by GESIS expanded on these by including value labels, implementing a common coding scheme for missing values which corresponds to social science standards and recoding alphanumeric variables to numeric. Additionally the setup files correct some minor errors in the data where possible. With every new release from Eurostat these setup files are then updated to incorporate the newest year of the survey as well as retroactive changes made to older data by Eurostat, such as the inclusion of additional countries, variables or data corrections. All setup files are now available for both SPSS and Stata.

Separate syntax files are provided for yearly and quarterly files as well as for files from ad-hoc modules. An integrated label file covers all variables from all these files for all years of the survey and a readme document is provided with instructions as well as information on any data errors GESIS has

come across. The setup files stick to the data structure provided by Eurostat and generate separate files for each country but also include information for users on how they can merge multiple country files into a single file for ease of use.

2.3 CIS

The CIS User Database is delivered to researchers on DVDs in two different formats: .csv format (CIS 2006, CIS 2008 and partially CIS 4) and Excel 97-2003 .xls format (CIS 3 and CIS 4). The data are cross sectional only, issued once every two years. The number of participating countries is between 14 (CIS 2006) and 16 (CIS 2008) and there is one dataset available per country per round.

The core documents used for preparing the CIS syntax files were harmonized questionnaires and anonymization method summaries. Due to country specific cases, additional documentation was used if required: individual country questionnaires, data structure for transmission to Eurostat and file description documents. The majority of listed documentation is received from Eurostat on DVDs with the microdata.

The SPSS setup files for CIS 3 and CIS 2008 were prepared both manually and automatically by using an R package called DDIwR, developed by Adrian Dusa (at RODA) for the DwB project. This R package can produce setup files for a number of statistical packages like SPSS, STATA, SAS and R.

Individual setup files were prepared for each country dataset (where available) in each of the four waves (CIS 3, CIS 4, CIS 2006 and CIS 2008), and additionally one master setup file was created, to obtain one integrated, cross-country dataset per wave.

For the individual country datasets, the setup files are able to import the raw data, assign variable and value labels, and even declare missing values where appropriate. Due to the different treatment of string variables in different statistical software, the setup files uniformly recode those variables to numeric variables and assign labels to those values.

The harmonization of data was slightly challenging, mostly due to the fact that Eurostat allows NSIs to deliver different formats of datasets with different anonymization levels and various sets of variables with different numeric and alphanumeric values of specific variables (e.g. NACE), and in some cases even different variable names. As the harmonization had not been done by Eurostat before the distribution to researchers on DVDs, the prepared setup files are meant to attempt solving this problem as well.

2.4 SES

FORS has provided two scripts that will import 2002 and 2006 SES data from the csv format under which they have been made available to the DwB WP5 into two separate SPSS files. The scripts will load the data, define the variable labels, the value labels as well as missing values.

The main challenge of this task is due to the fact that the files intended for dissemination have been subjected to several anonymization measures by each data provider. Although these procedures tend to follow general rules, data providers were granted autonomy in the way in which they have insured the confidentiality of the responding enterprises.

As the national files have been prepared independently, there are discrepancies in the filenames, variable formats, definition of missing or optional values. The import scripts are therefore customized for each year and country file.

The resulting files allow some degree of comparison between countries, with some caveats. As the sampling is very heterogeneous, great care is necessary when weighting records. Since several variables are regrouped differently among national files, researchers will have to create broader categories if they wish to use common measures. Furthermore, there is some inconsistency in the variable naming scheme between 2002 and 2006 data (variables with identical names describing different dimensions, variables with different names describing identical dimensions).

2.5 AES

Data for the 2007 AES are disseminated by Eurostat as .csv files. The document "Variable Definition of Anonymized AES microdata" included on the DVD includes variable and value labels for all variables, however these labels were not fully comprehensive thus the AES Manual had to be consulted and in some cases the responsible unit at Eurostat was contacted as well.

The GESIS setup files were written for SPSS and follow the same logic as those for the LFS and EU-SILC. The setup files provided import the .csv data to SPSS and add variable and value labels and recode alphanumeric variables to numeric. As the coding of missing values was not consistent within the .csv data delivered by Eurostat they were harmonized to a common coding scheme.

2.6 R-Script for generating setup files

As of 30 July 2014, a new package called DDIwR has been submitted to the CRAN - Comprehensive R Archive Network. The main purpose of this package, developed specifically for the work in the Work Package 5 of DwB, is to generate setup files to import .csv files in four of the most widely used statistical software packages, namely SPSS, Stata, SAS and R.

This package is written in the R language, and the main function called setupfile() expects as its first argument an object containing the metadata information for a particular .csv file. This metadata can be constructed directly into an R list object, of read from a DDI metadata file which are normally stored in the XML format.

Since R has a seamless way to import XML files, DDI metadata information can be thus served as the first argument in the setupfile() function, which has some additional useful arguments to read the .csv file, and also to deal with the missing information in the imported dataset.

The function can be used on a file by file basis, but it can also work in batch mode, by providing the path to the directory containing the metadata and the path to the directory containing the .csv files. The final output is saved in a (newly created if not existing) directory called "Setup files", normally in the default working directory but the user can specify (or set) another working directory of interest, mainly to ensure there are reading and writing permissions for those Operating Systems which have a stricter policy. Within the "Setup files" directory, the function will create a subdirectory for each of the four setup file types, which can be selected via the argument "type".

Individual variables are formatted upon importing the .csv file (for example SPSS has such options), and the function reads the .csv file to first detect whether a variable is a string or numeric, and if numeric it also detects the number of decimals. Special delimiters can be specified via the argument "sep", but otherwise the function can automatically detect three commonly used column separators like comma (","), semicolon (";") and tab separators ("\t").

For the missing data information, the function has three dedicated arguments called "miss", "trymiss" and "uniqueid". The first one accepts a vector of missing values (like -1, -2, -3 etc.) or even commonly used missing labels (like "Don't know", "Not answered", "Can't say" etc.). It will then identify all categorical variables which contain such values (or value labels) and add the relevant setup file commands to deal with these cases. For the R setup file, the function creates a very unique (and universal) way of dealing with the missing information, by storing variable level attributes containing the particular missing types (whether intentional or missing by design etc.), the exact values that are found in the dataset for each missing type, and also the unique IDs of the cases which contain those particular missing values.

Special care was dedicated to the target Operating System where the setup files will be used, because there are different eol (end of line) characters between Linux, Windows and Mac families.

3. Tools and Routines

In addition to the setup-files and the R-tool which were described in section two a number of routines and tools which are aimed at aiding researchers in data analysis were also developed as part of this work package. On a whole the amount of routines developed was not as large as we would have hoped for at the onset of the project. However this was largely due to the fact that the time investment in writing the setup files, which were the top priority of this task and benefit all researchers interested in employing these data, was far larger than originally anticipated. This was in no small part due to the problems encountered in the CIS and SES data and issues resulting from a lack of harmonization in the data. Below we have provided short descriptions of the routines developed as part of the task all of which can be downloaded from the DwB homepage as well as the MISSY system.

3.1 Integrating, harmonizing and processing routines for CIS

This routine was provided for all four waves of the CIS and consists of the following three elements.

PART 1: Merging and harmonizing databases (integrates and harmonizes all CIS 3 individual country datasets, possible renaming and recoding when needed)

PART 2: Adding weighting variables (adding a weighting factor setting to 5000 cases per country)

PART 3: NACE harmonization and conversion (creating harmonized NACE variables, developing a converter between NACE Rev. 1 and Rev. 2, specially addressing CIS coding specifics)

3.2 Innovation concepts for CIS

The innovation concepts routines added innovation concepts variables, derived from the existing CIS variables. Innovation concepts are based on up to six Innovation Union Scoreboard 2013 indicators and one Oslo Manual (3rd edition) indicator. The innovation indicators derived were:

1) Innovation or R&D intensity
2) Non-R&D innovation expenditures (% of turnover)
3) SMEs innovating in-house (% of SMEs)
4) Innovative SMEs co-operating with others (% of all SMEs)
5) SMEs introducing product or process innovations (% of SMEs)
6) SMEs introducing marketing or organizational innovations (% of SMEs)
7) Sales of new-to-market and new-to-firm innovations as % of turnover

The innovation concepts routines also include a routine which creates an integrated harmonized CIS dataset to be later processed (aggregation) and merged with economic indicators, publicly available on the World Bank website. This routine could be used for making innovation comparisons across time and space; it introduces newly derived innovation activity indicators based on CIS data.

3.3 Technical report on harmonization of income data in EU-SILC

This report describes how income data is collected in the EU-SILC and outlines different procedures to harmonize these data over time and between countries such as purchasing power parities, inflation weighting or income quintiles. After a short introductory section on each subtopic the report provides syntax for Stata and SPSS which will handle the necessary data transformations. Additionally an accompanying file provides PPPs and inflation factors for user convenience.

3.4 At risk of poverty thresholds for longitudinal EU-SILC

This Stata command syntax file adds information on at risk of poverty thresholds to EU-SILC longitudinal data of year T. The longitudinal UDB does not contain information on poverty thresholds. This information cannot be computed out of the panel data file itself given the non-representativeness of the longitudinal data. Instead, the dofile makes use of the representativeness of the cross-sectional EU-SILC data (year T, T-1, T-2, and T-3) and extracts the necessary information for calculating the prevalent at-risk-of-poverty thresholds from the cross sectional data (40, 50, 60, and 70% of the median equivalised total disposable household income). This results in a macro data file containing only the variables Year, Country, and the different poverty thresholds. This makes it linkable to the longitudinal data unlike the micro data, where cross-sectional and longitudinal data are not linkable.

3.5 Indicators for SES

Within the framework of the Structure of Earnings Survey, this SAS command adds information for the analysis of the relationships between earnings, individual characteristics of employees and their employers.

The calculation of indicators will be based on the median earnings. Following publications by INSEE and ONS, we opted for the median earnings rather than the mean.

1) Inter-decile ratio (D9/D1)

The first indicator is the inter-decile ratio. This latter gives evidence of the difference between the top and bottom of the distribution. It is one of the chief measurements of the inequality of the distribution.

2) Ratios measuring earnings depending on individual characteristics

We created routines so as to help SES users to be able to calculate the median earnings depending on key individual characteristics. The indicators are:

a. Earnings depending on the sex and the age of the employees;
b. Earnings according to the highest completed level of education (and training);
c. Earnings depending on the occupation of the employees, sex and age;
d. Earnings depending on the length of service in enterprises;
e. Earnings according to the type of employment contract (types...);
f. Earnings depending on full-time or part time on the level of remuneration.

3) Ratios measuring earnings depending on the characteristics of local units

The goal of the second set of indicators is to study the relationship between characteristics of the local units and the level of remuneration (i.e. earnings).

a) The first indicator of that series measures the level of remuneration depending on the Nace identification.

b) the second indicator measures the level of remuneration depending on the economic and financial control.

c) a third indicator calculates the mean remuneration depending on the total number of employees in the local unit.

4) Regression

We show an example of an analysis of the relationship between earnings and some individual characteristics.

To begin with, we choose three key individual characteristics. The previous tables showed the importance of sex, age, and the highest completed level of education (and training).

We use a multiple regression model. The total earnings per year will be the dependent variable. Age, sex and the highest completed level of education (written "Education") will serve as explanatory variables. Ui will be the dummy variable. β 1, β2, β3 and β4 are the parameters of the equation.

$$\text{Ln}(\textit{Wage}) = \beta 1 + \beta 2\ \textit{Sex} + \beta 3\ \textit{Education} + \beta 4\ \textit{Age}\ ui$$

4. Conclusion

As part of this work package a number of key tools have been developed to aid users of integrated European microdata. These tools complement the metadata generated in WP 5 Task 3 and make official statistics microdata far more accessible to researchers. Not only can researchers now easily inform themselves on the contents and coverage of data but the tedious labor of preparing data for analysis is largely done for them by the provided setup files. Furthermore the tools and routines developed as part of this work package can provide assistance to researchers during data analysis. Together these tools provide an invaluable service to researchers looking to work with European official statistics microdata.

For the future further work on these types of tools is in planning. The provision of setup routines is a central element of GESIS services for Eurostat microdata and will be continued for EU-SILC and LFS based on the current staffing for the foreseeable future. These tools will be offered to researchers within the MISSY system. Additionally further routines, tools and technical reports to aid researchers in data preparation and analysis of European official statistics microdata are currently in planning.

A dedicated section of the DwB website will be created under the section "DwB Services" (see http://dwbproject.org/service/) to make these tools available to the wider public. It will provide the list of developed routines and make them downloadable in a zipped folder together with the corresponding documentation text file; while pointing toward the Setup files that will be hosted on MISSY and the R-script for generating setup files (see section 2.6 above).