Project N°: 262608

# ACRONYM: Data without Boundaries

## DELIVERABLE D5.4

(Report and Databank Documenting Integrated EU OS Data)

## WORK PACKAGE 5

(Servicing European Researchers in the Use of OS Microdata)

| | | |
|---|---|---|
| REPORTING PERIOD: | From: Month 1 | To: Month 46 |
| PROJECT START DATE: | 1st May 2011 | DURATION: 48 Months |
| DATE OF ISSUE OF DELIVERABLE: | March 4th 2015 | |
| DOCUMENT PREPARED BY: | Partner Numbers 5, 1, 4, 8, 17, 25, 6 | Partners Names GESIS, CNRS-RQ, UL, RODA, FORS, CED, NSD |

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 (DwB - Data without Boundaries)

## Summary[1]

This report provides an overview of the outputs of the Work Package 5 Task 3 of the Data without Boundaries project. The objective of this Task was to generate metadata describing integrated European data sources. This includes on the one hand census microdata which are documented within the IECM system (http://www.iecm-project.org) and on the other hand Eurostat microdata which were documented within the MISSY system (http://www.gesis.org/missy/eu/missy-home).
One of the key developments of this work package was the establishment of a hierarchically structured metadata scheme which can adequately describe such complex data structures. Furthermore the IECM database was vastly expanded with 19 new samples being included. Finally metadata for 5 major Eurostat surveys (EU-SILC, LFS, AES, SES and CIS) has been prepared and will be published in the MISSY system by April 2015.


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

## Table of contents

# 1. Introduction

The goal of this task was to document integrated official statistics microdata. This included two different data sources: harmonized census microdata which is incorporated into the Integrated European Census Microdata (IECM) system and integrated European microdata from Eurostat which was documented in the Microdata Information System (MISSY). This data documentation should aid researchers in data exploration by providing information about the topical, geographical and temporal coverage of a data source and allow them to gauge whether a data source could be of use for a specific research project. Beyond that however the documentation produced as part of WP5 Task 3 should also provide a tool for data analysis by providing detailed information on the data collection process and sampling procedures and most importantly detailed metadata on the variable level. Thus the aim of this data documentation is to vastly reduce researchers' time investment both during data exploration and analysis. The IECM takes this one step further by disseminating the public use census microdata free of charge to certified researchers.

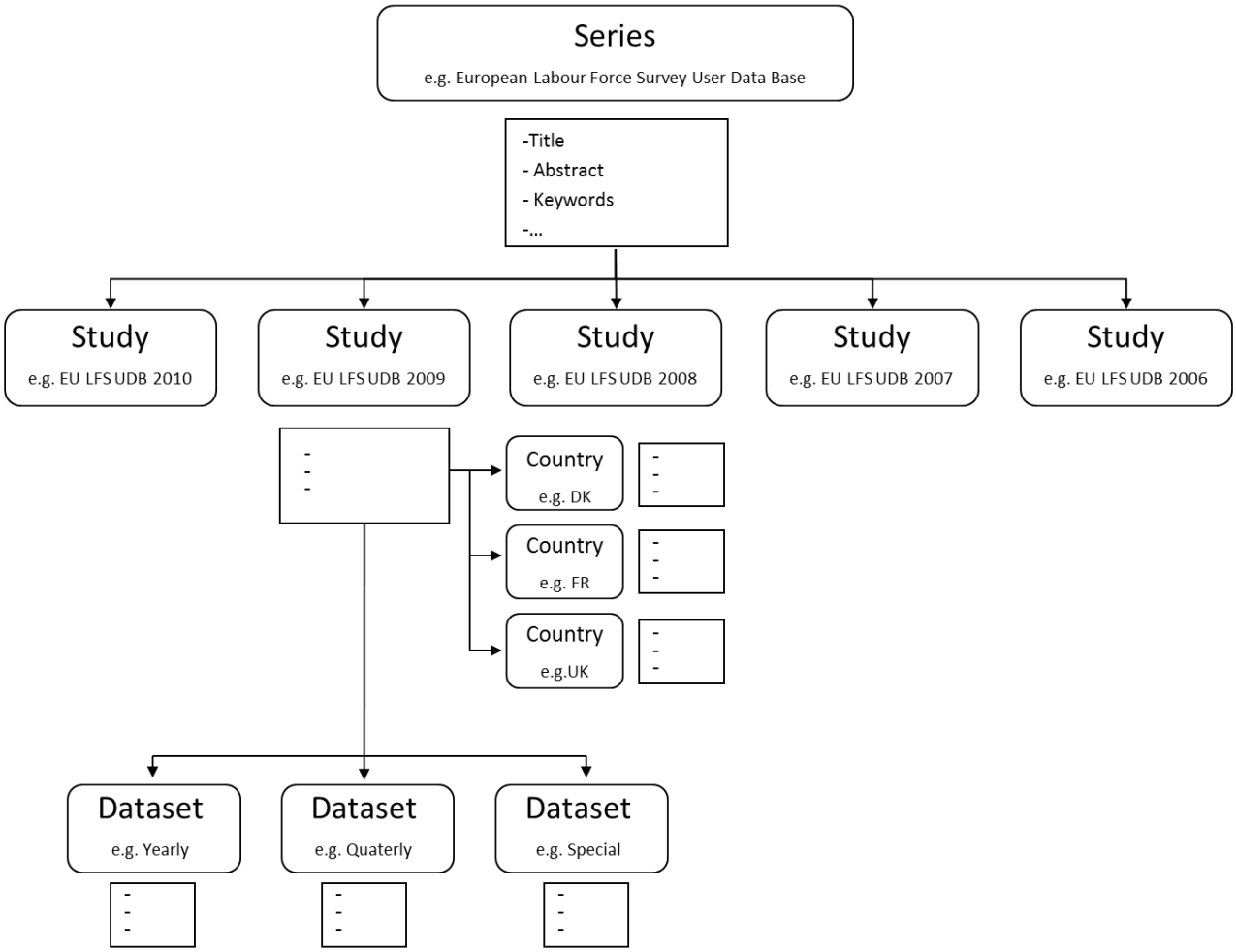
The remainder of this report will be structured as follows: Section 2 describes the metadata scheme employed for documentation of the integrated European microdata. Section 3 outlines the MISSY tool employed for preparation of the metadata. Section 4 provides an overview of the metadata produced for Eurostat datasets and Section 5 looks into how the contents of the IECM were expanded as part of DwB. Section 6 discusses how the translation budget allotted to WP5 was utilized. Section 7 looks into the coordination with DwB WP8 and 12 which were responsible for establishing the DwB portal. Section 8 concludes the report and provides an outlook.

# 2. Metadata scheme for documenting integrated Eurostat microdata

## 2.1 Study level metadata scheme

The study level metadata scheme employed in DwB WP5 Task 3 builds on that used in Task 2 in the CIMES system but expands on it. It was devised with the idea of documenting the numerous multinational and repeating survey programs carried out by Eurostat. It is comprised of three levels which follow a hierarchical structure (see Figure 1). The top level is called series and refers to a data collection program which takes place over multiple intervals in time. The main objective of the series level is to provide a brief description of the thematic content, the temporal and geographic coverage as well as the guidelines laid out by Eurostat. The second level, which we have termed study level, refers to one instance of a data collection (usually a year) and is a bit more complex as it is multi-dimensional and includes both information that is specific to a year for all countries as well as country specific information for that year. This section contains detailed information on topics such as sampling, data collection and weighting. Finally the dataset level describes different instances of a study, such as the panel and the cross sectional version of the EU-SILC. The metadata scheme developed here is compatible with requirements of the DDI metadata standard. This section will provide a short overview of each level and list all fields included in the scheme with a short description for each field.

**Figure 1** - Structure of Study Level Metadata Schema




## Series

A series is defined as a collection of studies conducted within a data generation programme. These studies share common basic contents and are carried out repeatedly.

- **Title**: Provides the title of the series in English.
- **Subtitle**: Provides the subtitle of the series in English.
- **Release Year:** The year in which the data were released regardless of when data were collected.
- **Version Number:** The version number can distinguish multiple releases from one year from another**.**
- **Data Publisher:** The legal entity responsible for authoring the data i.e. EUROSTAT.
- **Abstract**: A short summary describing the purpose, nature, and scope of the series.
- **Keywords**: Thematic coverage of the series; several keywords can be used.
- **Geographic Coverage:** Information on the geographic coverage of the series.
- **Organization:** Details the organization of the series and the responsibilities of involved parties.
- **Universe:** The group of persons or other elements that are covered within the series.
- **Sampling:** Details Eurostat's requirements regarding sampling.
- **Data Collection Methodology:** Details Eurostat's requirements regarding data collection.
- **Anonymization:** Details the anonymization procedures undertaken by Eurostat.
- **Legal Basis:** Detail the legal norms on the basis of which the series is carried out.

- **Access-Conditions:** Details the access conditions as issued by Eurostat and provides contact information.
- **Access Form:** Links to an (online) access form, with which data can be ordered.
- **Access Contact:** Contact information regarding data access.
- **Comparability over Time:** Details breaks in continuity or methodology over time.
- **Comparability over Space:** Details differences in data or data collection between countries.
- **Notes:** Provides any further pertinent information that cannot be included in other fields.
- **Citation:** Lists all sources used for the documentation at the series level and includes the following information for each source: Author, title, year of publication and link to document (if available).

**Study**

A study is defined as an individual instance of a series, which will tend to be individual years. However since some information will not only change over time but also vary between countries it becomes necessary that coding of this information is done by year and country. This is the case for the fields surmised under the subheading country specific information.

***General information***

- **Title**: Provides the title of the study in English.
- **Subtitle**: Provides the subtitle of the study.
- **Abstract**: A short summary contains information pertinent to the specific study but not the series as a whole.
- **Keywords**: These keywords are in addition to those already defined at the series level; they should be used to describe the contents of study specific content such as ad-hoc modules.
- **Geographic Coverage:** Information on the total geographic scope of the data.
- **Time Period Covered:** The time period the data refer to; this can be different points in time if for example retrospective questions are asked.
- **Notes**: Contains any additional information important for researchers which cannot be included in any other field.
- **Citation:** Lists all sources used for the documentation at the study level and includes the following information for each source: Author, title, year of publication and link to document (if available).

***Country specific information***

- **Corresponding national Study:** This field provides the alternative title of the study in the national context.
- **Producer:** The legal entity/entities responsible for carrying out the study.
- **Universe**: The group of persons or other elements that are covered within the study.
- **Sampling Units:** The sampling units employed.
- **Minimum Effective Sample Size:** The minimum sample size required for this country; usually defined as a requirement by Eurostat.
- **Actual Sample Size:** Size of the sample drawn.
- **Achieved Sample Size:** The number of interviews realized.
- **Size of Design Effect:** The design effect corrects for bias that results from applying complex sampling criteria (in comparison to a simple random sample).

- **Effective Sample Size:** Is calculated by dividing the achieved sample size through the design effect.
- **Available Data:** Provides information on whether longitudinal and/or cross sectional data are available.
- **Source of Sampling Frame:** Basis for drawing the sample; for example this might include census data or registers.
- **Sampling Design:** Includes information on whether sampling is random or systematic, stratified, one stage or multi stage.
- **Primary Sampling Units:** In the case of a multi stage sample this field provides information on which sampling units are selected in the first step.
- **Secondary Sampling Units:** If a multi stage sample was used this field provides information on which sampling units are selected in further steps.
- **Stratification Criteria:** If a stratified sample was used this field details stratification criteria.
- **Sampling Method:** Additional information regarding sampling methodology which is not provided in the other fields; for example this could include special cases such as the representativeness in regards to small subpopuations (e.g. oversea territories).
- **Number of Rotational Groups:** Number of rotational groups used in a longitudinal design; this also provides information on how often respondents are interviewed in the panel.
- **Units of Observation:** Describes the level at which data is collected.
- **Dates of Data Collection:** Defines the start date and end date of data collection.
- **Participation Mandatory:** Whether the participation in the survey is mandatory or not.
- **Type of Data Source:** Provides information on whether data are derived from register and/or survey data.
- **Interview Mode:** Provides information on what types of interviews were carried out.
- **Percentage of Proxy Interviews:** Provides information on the percentage of proxy interviews carried out.
- **Initial Design Weight Target:** The initial design weights are used to assess the representativeness of the sample in regards to the universe. This field notes what level these initial design weights refer to.
- **Initial Design Weight Method:** Method on the base of which the initial design weighting is carried out.
- **Initial Non-response Method:** Description of procedures undertaken to correct for bias from systematic non-response.
- **Initial Calibration Method:** Description of calibration procedures undertaken to increase consistency with existing population statistics.
- **Weighting Method:** Information on methodology regarding weighting.
- **Notes:** Additional important country specific information not provided in any other field.
- **Citation:** Lists all sources used for the documentation at the country level and includes the following information for each source: Author, title, year of publication and link to document (if available).

**Dataset**

This level describes specific datasets which are part of a study. Possible distinctions can include those between the cross-sectional or panel versions of a study (e.g. EU-SILC) or those between quarterly and yearly datasets (e.g. EU-LFS).

- **Title:** The name of the dataset as provided by the data producer.
- **Weighting:** Provides information on weighting (e.g. when this differs from the methodology as documented at the study level).
- **Notes:** Provides dataset specific information not listed anywhere else.
- **Citation:** Lists all sources used for the documentation at the dataset level and includes the following information for each source: Author, title, year of publication and link to document (if available).

**Documents**

Documents can be linked to any level of the metadata schema and are either uploaded or linked to. The following information is attached to a document.

- **Document Type:** A controlled vocabulary is provided.
- **Level:** The level which a document is associated with.
- **Year:** The year to which a document refers to.
- **Country:** The country a document refers to.
- **Language:** The language of the document.
- **Document Name:** The name as issued by the document producer.

## 2.2 Variable level metadata scheme

The objective of the variable level metadata scheme is to provide detailed information on individual variables. During data exploration interested researchers can learn about the availability of certain variables as well as their comparability between countries and over time. Variable level information is always associated with a specific dataset (e.g. EU-SILC 2005 Austria cross-sectional). Variables from different years of a series can be associated with each other via a thematic classification. Some variable level metadata are directly imported into the MISSY system from the microdata and do not have to be entered manually. This includes names, labels, frequencies and descriptive statistics.

The EU-SILC contains two types of auxiliary variables, flag and imputation variables. These variables are directly associated with a specific variable and marked by _F or _I suffix in the variable name. Within the MISSY system they are displayed alongside the variable they are associated with, however only variable name, statistics and frequencies are displayed.

**Variable Details:**

- **Variable Name:** The name of the variable as used in the Eurostat Data.
- **Variable Label:** Variable labels, usually those provided by Eurostat were used. However these were sometimes shortened.
- **Description of Target Variable:** Eurostat's definition of the target variable.
- **Country Specific Comments:** Country specific information regarding the collection or preparation of a variable in a country. This pertains specifically to issues that might influence comparability.
- **Other Comments:** A field which can include any additional information which does not fit other fields.

- **Thematic Classification:** Each variable must be assigned an entry in the thematic classification of the study; the distinct classification of variables builds the basis for the variable-timepoint matrix.
- **Unit of Observation:** The level at which data is collected.
- **Filter:** In the case that a variable is only collected for a subgroup of respondents this field details the rules guiding the selection process.
- **Is ad-hoc Module Variable:** Whether a variable is part of an ad hoc module or not.
- **Is derived Variable Type:** Differentiates between variables which were directly collected or those derived from a number of sources.
- **Question Wording:** Question wordings can be provided in any language. The default is English.
- **Question Text Comment:** A field for additional comments regarding the question text(s). Changes in question wording over time should be documented here.

**Values/Frequencies:**

This subsection contains descriptive statistics on variables which are derived directly from the microdata and not inputted manually. The following mutually exclusive options exist, only one can be chosen to be displayed in the MISSY system.

- **Frequencies:** Includes frequency counts, total N and missing values; should not be used for numeric variables.
- **Mean/Stddev:** Includes the total N, min/max values and standard deviation. Intended for continuous variables.
- **Min/Max:** Includes minimum and maximum values as well as the total N.
- **Total N:** Includes only the total N.

## 3. MISSY system as a tool for data documentation

The development of the MISSY III system at GESIS ran parallel to the DwB project. However it is important to emphasize that the development of the MISSY system was funded independently of Data without Boundaries. In order to account for the fact that IT staff at GESIS had to put in a large amount of additional hours to ensure that the MISSY editor could be used by the DwB partners GESIS received some additional funding as part of redistribution of funding within DwB to partially compensate for this. The MISSY Editor served as a tool for data documentation within DwB WP5 and the metadata for Eurostat microdata produced in the DwB project can now be accessed via the MISSY System. Thus a short overview of the functionality and architecture of the system shall be presented here before moving on to describe the actual process of data documentation.

The MISSY system consist of three components: the MISSY Import Application which is used to import variable level metadata from microdata in SPSS format, the MISSY Editor which is used for the manual entry of study and variable level metadata, which was implemented as a web application in order to allow the DwB partners to access the system and the MISSY Publisher which displays the structured metadata. The underlying data model was built on the basis of the DDI-RDF Discovery Vocabulary (Disco) Ontology. This also allows for metadata contained within the MISSY system to be exported in DDI-XML 3.2 to be reused in other systems such as the DwB portal developed as part of WP12.

The process of data documentation via the MISSY system within DwB WP5 was structured in five steps which will be detailed below.

***1) Setting up a Series in the MISSY Editor***

In a first step a series must be defined within the Editor, this step includes the definition of the series, and the underlying levels studies and datasets. This is done manually within the MISSY Editor by the MISSY staff at GESIS.

***2) Import of variable details from microdata***

The second step entails importing of information from the microdata in SPSS format. This microdata was generated with the help of scripts prepared as part of DwB WP5 Task 4 (see DwB Deliverable D5.3 for details). In this first step only variable names and labels are imported.

***3) Selection of variables for which no frequencies are required***

In the next step those variables where frequency counts are not needed are selected within the editor. This intermediate step is required as usually these variables are identifiers or income variables with huge numbers of values which cannot be displayed properly within the system.

***4) Import of values, value labels as well as calculation of descriptive statistics and frequencies***

In this step variable level information and statistics are imported into the system from the SPSS microdata. Now the metadata are ready to be modified manually within the editor.

***5) Manual or automated entry of metadata with the help of the editor***

The final step and the one which entails by far the greatest amount of work is the process of entering the metadata for all the fields of the schema defined in section 2. Entry of metadata can either be handled manually via the web based editor or can be imported from .csv files which are prepared in advance. Section 4 will go into more detail on the documentation of the individual series and the sources employed.

## 4. Databank on Eurostat Microdata

As part of this work package metadata for European microdata was documented. While CED was responsible for the documentation of integrated census microdata GESIS, CNRS-RQ, RODA, FORS and ADP managed the documentation of Eurostat data. NSD had initially been assigned to this task as well, however after Eurostat had not granted NSD access to the microdata NSD's responsibilities had been reassigned to ADP. While the description of work had only mentioned EU-SILC and LFS it was decided that partners had enough resources to document additional studies. In selecting additional Eurostat microdata for documentation focus was placed on studies deemed topically most pertinent to social sciences. Thus the Adult Education Survey (AES), the Structure of Earnings Survey (SES) and the Community Innovation Statistics (CIS) were documented. In order to ensure access to the relevant microdata CNRS-RQ prepared a proposal to Eurostat dissemination unit for the work package to access these studies. Table 1 provides an overview of the studies documented.

**Table 1** - Documentation of Eurostat Microdata for T5.3

| Series | Temporal Coverage | Number of countries included |
|---|---|---|
| EU-LFS | 1983-2012 (anually) | 10-31 |
| EU-SILC | 2005-2012 (annually) | 27-32 |
| SES | 2002, 2006 | 24 |
| CIS | 2002-2008 (bi-annually) | 15-16 |
| AES | 2007 | 26 |

## 4.1 EU-SILC

Documentation of EU-SILC data was handled by GESIS. The metadata includes the cross sectional data from 2005-2011. The panel is included from 2007 to 2011. As SILC panels cover 4 years and are named after the latest wave, the 2007 panel includes data for the years 2004 through 2007. The process of data documentation followed the general structure outlined above in section 3. In a first step setup files were written which can import the .csv data as delivered by Eurostat into SPSS and attach variable and value labels. The information on labels was taken from the "Description of target variables" (see D5.3 for details). Once variable level information was imported into the MISSY system additional metadata was entered manually (as detailed in section 3). The documents included at the variable level were the aforementioned "Description of target variables" as well as the "Module guidelines" for the respective Ad-hoc modules and national questionnaires from the UK, Germany and France for the question wordings. For the study level metadata national and comparative quality reports as well as the explanatory notes found on the Eurostat homepage were consulted. The latter were particularly important for the Ad-hoc modules. Overall the metadata provided for EU-SILC by Eurostat is very comprehensive and well organized making the task of documenting it comparatively easy. However the amount large amount of material to be documented meant that the task of documenting the EU-SILC was very labor intensive and GESIS committed considerable personnel resources which went beyond the funding provided by the DwB project.

## 4.2 EU-LFS

The documentation for the EU-LFS was prepared by GESIS and covers all waves from 1983 to 2011, Ad-hoc modules are documented from 1999 to 2011. The core document employed for generating the setup files and from which much of the variable level information was drawn is the "EU Labour Force Survey database User Guide". Additionally the "Methods and Definitions" which are available for 1988 to 2001 and the explanatory notes from the Eurostat homepage (which include information for the years 2003 and onward as well as information on the Ad-hoc modules) were consulted for the variable level documentation. National questionnaires from the UK, Germany and France were used for the question wordings. The study level documentation was based primarily on the series "Quality Reports of the EU LFS" and the "Main characteristics of the National Surveys" which are available as of 2003. The documentation for years prior to 2003 also relied primarily on the "Methods and Definitions". While Metadata for the newer waves (particularly as of 2003) is well organized and comprehensive the farther back in time one goes, the less detailed the metadata is. Thus the study level documentation for older survey rounds is not as comprehensive as it is for newer rounds. Much like the work on the EU-SILC the task of documenting the EU-LFS was very labor intensive and GESIS committed additional personnel resources into this task which went beyond the funding provided by the DwB project.

## 4.3 CIS

CIS metadata were prepared on four levels: variable (integrated CIS dataset for each round), national, study level and general information about the Eurostat Community Innovation Survey research. Four CIS rounds were covered: CIS 3, CIS 4, CIS 2006 and CIS 2008. The MISSY system was used to import data and enter metadata. As the data were not fully harmonized, attention was paid to presenting country specifics, especially on the variable level.

To prepare structured metadata following the metadata scheme presented above, various types of metadata documentation were used. Some documentation was provided on DVDs as accompanying materials. There was also some documentation which is publically available (synthesis of national reports, some methodological explanations, innovation activity reports etc.). However, to prepare quality national level metadata, additional documentation was required. With assistance from the Eurostat CIS department all the participating NSIs were asked to provide their CIS quality reports. Later we needed to establish direct contact with NSIs' CIS departments to receive the documentation which covered a significant part of required information about the implementation of CIS research. Part of the obtained documentation is available in addition to the CIS metadata in MISSY.

## 4.4 SES

The documentation for SES was prepared by FORS and CNRS-RQ and covers 2002 and 2006. Setup files were first created in order to import the .csv data into SPSS, including variable labels. The "Thematic classification" field (on the variable level documentation) was imported into the MISSY system.

Once variable level information was imported into the MISSY system, additional metadata were manually entered. For the study level metadata, "Quality Reports" as well as the explanatory notes found on the Eurostat website were consulted. "Implementation Arrangements" documents on the 2002 and 2006 SES were consulted for the variable level documentation.

The country-level documentation was based on information chiefly drawn from the quality reports sent by both Eurostat and the different NSIs. E-mails were sent to the persons in charge of the methodology in each NSI. The answers helped us to have access to missing information on key points which were not to be found in the quality reports. Problems arose when quality reports were either not in English or no longer available. E-mails were then sent. The given institutions were keen on delivering answers which proved efficient enough to sort things out in most cases. When information was not at our disposal, we resorted to the descriptions we found in the Synthesis Quality Report written by Eurostat.

Some countries posed major concerns owing to the lack of information in their quality reports and the absence of satisfactory answers. The quality reports from CBS (Statistics Netherlands) did not provide us with ready-to-use information within the MISSY framework. We were told that the person who used to be responsible for the SES 2002 and 2006 was no longer working for CBS. Hence, most of the information (i.e universe, sampling units, type of available data, sampling procedure) is missing. Some similarities are to be found between the Netherlands and Greece. Indeed, even though some information was to be found in the synthesis of quality reports, no description was at our disposal for the year 2002. What is more, we did not receive any answers from the Greek NSI. There is therefore no information on the target sample size, the type of data, the weighting method

and the type of interviews. As for Spain, the situation is different. We were not able to get into contact with a person who could provide us with information on the 2002 and 2006 Spanish SES. For all that, information was found in the synthesis of the quality reports produced by Eurostat. The Finnish situation differs from the previous examples. It highlights the problems MISSY faced when taking into account of all the peculiarities of the country members. Some information is yet at the users disposal in the "notes" paragraph. To do so, we copied/pasted the explanations we received from the Finnish NSI.

As previously said for the EU-LFS, the metadata for the recent waves (i.e 2006) was well-organized and comprehensive. Such was not always the case for the 2002 Structure of Earnings Survey Quality reports. All in all, the documentation highlighted the lack of homogeneity among the quality reports written by the NSIs of different countries.

### 4.5 AES

GESIS was responsible for documenting AES data. At time of the data access request only the 2007 data was available. As this was the first wave of the Adult Education Survey it was intended as a pilot study. The key source for the metadata were the national quality reports, which were designed as questionnaires on survey mode, field methods, sampling design and data quality which Eurostat sent to NSIs and which they filled in. The quality and comprehensiveness of these "reports" differed widely between countries. While overall the most important information was available for a majority of countries the comprehensiveness of study level metadata varies considerably between countries. Two further documents were employed for compiling the study level metadata: the “AES manual” and the “Synthesis Quality Report Adult Education Survey”. The variable level metadata used relied mainly on the document “Variable Definition of Anonymised AES Microdata Set”.

## 5. IECM

The IECM project is currently disseminating to the research community census microdata from 16 European countries. There are over 90 million person records in the current dataset (around 30 million households) corresponding to 54 different census samples conducted from the 1960 to the 2010 census rounds (from France 1962 to Ireland 2011). In June 2015, IECM will be disseminating 115 million records from 58 census microdata samples.

14 samples have been integrated into the IECM database during the DwB project and 4 more samples will be disseminated in June 2015 (Table 1). Data releases take place once a year in June to maximize efficiency over continuous releases throughout the year since every release requires a complete update of the entire database. During the 4 years of DwB we have integrated around 1,500 original variables into 200 harmonized variables, and we have dealt with 70 million person records.

**Table 2 -** Description of IECM database by period of integration

| | Samples | Person records | Household records | Number of harmonized variables | Number of unharmonized variables |
|---|---|---|---|---|---|
| Prior to DwB | 39 samples | 43,886,299 | 17,964,938 | 232 | 2,678 |
| During DwB | -France 2006 | 19,973,287 | 8,749,114 | | |
| | -Germany 1970 | 3,094,845 | - | | |
| | -Germany 1971 | 4,089,856 | 1,569,112 | | |
| | -Germany 1981 | 4,278,563 | 1,725,991 | | |
| | -Germany 1987 | 3,160,224 | 1,348,896 | | |
| | -Ireland 1971 | 296,878 | 83,285 | | |
| | -Ireland 1979 | 337,686 | 98,453 | | |
| | -Ireland 1981 | 344,291 | 101,890 | 151 | 1,044 |
| | -Ireland 1986 | 355,020 | 107,278 | | |
| | -Ireland 1991 | 353,149 | 112,149 | | |
| | -Ireland 1996 | 365,323 | 122,860 | | |
| | -Ireland 2002 | 410,688 | 140,040 | | |
| | -Ireland 2006 | 440,314 | 157,762 | | |
| | -Ireland 2011 | 474,535 | 175,651 | | |
| | -Ukraine 2001 | 4,889,288 | 3,623,571 | | |
| To be released (June 2015) | -Austria 2011 | 839,501 | 366,068 | 79 | 91 |
| | -France 2011 | 20,541,337 | 9,026,051 | 77 | 99 |
| | -Portugal 2011 | 528,870 | 296,707 | 68 | 99 |
| | -Spain 2011 | 4,107,465 | 1,621,643 | 63 | 95 |

In addition to microdata, census metadata are provided to the users in multiple levels. At the sample level, information is provided for 34 different fields (Table 2). At the variable level, the data extraction system allows the user to access a wide variety of information about each harmonized/unharmonized variable: (i) variable codes and their associated labels, (ii) variable description, (iii) variable comparability, that refers to the details about the comparison of the variable (among samples of the same country or among different countries) are offered, (iv) universe of the variable, which identifies the group of population that was asked, (v) availability of the variable in other samples and (vi) the questionnaire text, which reproduces the question from the enumeration text and the instructions given to the interviewer for this particular question.

**Table 3 -** IECM metadata provided at the sample level

| | | |
|---|---|---|
| Title | Sample microdata source | Dwellings |
| Census Agency | Sample design | Vacant units |
| Population | Sample unit | Households |
| Universe | Sample fraction | Individuals |
| De jure or de facto | Sample fraction (private households) | Group quarters |
| Enumeration unit | Sample fraction (institutional households) | Settled/unsettled Population |
| Census day | Sample size (person records) | Special populations |
| Field work period | Sample size (private households) | Dwellings |
| Enumeration forms used | Sample size (institutional households) | Households |
| Type of field work | Sample weights | Institutional population |
| Respondent | | Group quarters |
| Coverage | | Unsettled population |

## 6. Translation Budget

Work Package 5 of the Data without Boundaries project was allotted a budget of 50,000 € to commission translation services. The initial plan had been to translate documents produced by NSIs in the context of Task 2 however this budget was never utilized. As part of the redistribution of funds within the project 20,000 € from this budget were assigned elsewhere while the remaining 30,000€ should be used for translations of survey materials within Task 3. It was decided that translations of questionnaires for SILC and LFS would provide the greatest utility to researchers. The largest part of the budget was spent on translation of 2011 SILC questionnaires as English language SILC questionnaires were unavailable for most countries. In the case of the LFS French and German questionnaires had to be translated.

Where provided by NSIs the questionnaires also include a precise mapping of national survey questions to specific items in the EU-SILC. GESIS is planning to produce such mappings where they are missing in the future.

These translated questionnaires will be made available in the MISSY system and offer tremendous utility to researchers interested in cross country research who are looking to explore the equivalence of survey items across countries.

## 7. Coordination with WP8 &12

A substantial part of the WP5 work has focused on bringing together information and to document actual data collections, for discovery and use purposes. This work resulted in development of two databases, MISSY and CIMES, and an essential part of the work has been to develop a comprehensive and implementable metadata model for each of these products. As there are major differences in the kind of data collections covered by MISSY and CIMES, the respective metadata models are compatible but only partially overlapping.

In addition to these two focused tasks of WP5, WP8 and WP12 have developed a more general and complete metadata model (described in D8.2) aiming to cover the total data resource pool potentially available for social science research purposes through both Data Archives and NSIs. This work was divided between the packages so that WP8 has done the basic model development work, based on a thorough investigation of the data work and production processes of a wide range of potential data providers and analyzed user needs to be covered. WP12 built on the WP5 and WP8 modelling work to build a prototype implementation of a general data discovery portal.

As the WP5 databases were to be integrated into the DwB portal it has been essential to ensure compatibility between metadata models. The DwB data discovery portal will harvest metadata from a wide range of sources, including the CIMES and MISSY systems. In order to coordinate these efforts two joint meetings were organised. Additionally representatives from each WP attended the regular meetings of the other work packages so as to be up to date with the current progress. Thanks to this tight cooperation full compatibility between the metadata models for WP5 and WP8 was achieved. The technical staff from both GESIS and CNRS-RQ responsible for the MISSY and CIMES systems cooperated with WP12 to ensure the integration of metadata into the WP12 portal.

## 8. Conclusion

The objective of this task was to document integrated microdata from official statistics. This includes the integrated census microdata which was documented in the IECM system and on the other hand integrated European microdata as disseminated by Eurostat which was documented in the MISSY system. The MISSY system vastly improves the documentation of Eurostat microdata by providing a one stop shop which offers comprehensive and structured metadata for EU-SILC, EU-LFS, AES, SES and CIS and thus makes these data far more accessible for scientific enquiry. The IECM takes this one step further by providing harmonized public use microdata which accredited researchers can readily download from the web page. Both of these services could in the future become cornerstones of a European Service Center for Official Statistics. Currently both CED and GESIS plan to continue the work on IECM and MISSY, respectively and will work to keep these systems up to date and incorporate a wider range of data sources.

Furthermore it has to be emphasized that the output produced as part of this task exceeds what was proposed within the description of work. Initially it had only been planned to document EU-LFS, EU-SILC and AES. However due to dedication and commitment of all involved parties, the investment of additional institutional resources by partners as well as the reassignment of funds within the DwB project it was possible to additionally provide high quality and comprehensive metadata for CIS and SES as well.