Project N°: 262608

Acronym: Data without Boundaries

DELIVERABLE D5.5

Final report & recommendations for the continuation of services for European OS Microdata

WORK PACKAGE 5

Servicing European Researchers in the use of OS Microdata

REPORTING PERIOD:	From: Month 18	TO: Month 48
PROJECT START DATE:	1st May 2011	DURATION: 48 Months
DATE OF ISSUE OF DELIVERABLE:	15 April 2015	
DOCUMENT PREPARED BY:	Partners 5, 1, 4, 8, 17, 25, 6, 25	GESIS, CNRS-RQ, UL, RODA, FORS, CED, NSD, CED

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]

The overarching objective of Work Package 5 (WP5) of Data without Boundaries (DwB) was to facilitate the use of official statistics (OS) microdata in Europe by collecting, structuring and disseminating information on existing microdata in Europe. To this aim the work package has provided structured data documentation, developed routines to facilitate data use and increased the census material in the Integrated European Census Microdata Database (IECM). The work package was subdivided into four Tasks which will be shortly outlined here.

Task 5.1 was intended as a conceptual thought exercise and the key output was a report on a virtual service center. The report detailed how such a European Service Center for Official Statistics (ESCOS) could be implemented and which services it could provide on- and offline and how such a service center could be integrated into the CESSDA-ERIC. In order to incorporate feedback received from within the Data without Boundaries project as well as from external partners this report was revised and resubmitted. It provides suggestions on how the services developed within this work package could be incorporated into ongoing activities such as the CESSDA-ERIC and thus will also inform this report.

The objective of Task 5.2 was to collect, structure and code information on available microdata from official statistics at the national level for all countries in Europe in order to enable researchers to engage in cumulative and comparative research in Europe. The original intent of this data documentation effort was to integrate these metadata into an existing metadata system. However as no currently available systems were able to adequately display the envisioned hierarchical structure of the metadata CNRS-RQ developed the CIMES system for entering and displaying these metadata.

Task 5.3 was responsible for documenting integrated European microdata from official statistics held by Eurostat. The metadata prepared as part of this task includes the European Labour Force Survey (EU-LFS), the European Statistics on Income and Living Conditions (EU-SILC), the Structure of Earnings Survey (SES), the Community Innovation Statistics (CIS) as well as the Adult Education Survey (AES). The metadata was entered with the MISSY system which was independently developed by GESIS.

Task 5.4 is also aimed at assisting users of integrated European microdata from Eurostat and consisted of preparing routines and microdata tools for data preparation and analysis. The most important part of this task was to prepare what we term setup routines for these data. These routines allow users to import a Eurostat dataset, which are commonly delivered to researchers as comma separated values (csv), into the statistical package of their choice with standardized variable and value labels. Additionally microdata tools were developed which aid in data analysis by providing code and instruction for preparing datasets.

CED had a special position within the work package as their prime responsibility was to expand and enhance the services offered by the Integrated European Census Microdata (IECM) database. The goals of the IECM project are to preserve the microdata and metadata of population censuses conducted since 1960, to harmonize the original information into variables following a common coding scheme that enhances comparability and to disseminate theses data to the research community.

[1] This report was prepared by Alexander Mack, Christof Wolf, Albert Esteve, Antonio Lopez-Gay and Roxane Silberman.

TABLE OF CONTENTS

1. INTRODUCTION

The objective of Work Package 5 (WP5) of Data without Boundaries (DwB) was to facilitate the use of official statistics (OS) microdata in Europe by collecting and disseminating information on existing microdata in Europe, providing structured data documentation, developing routines to facilitate data use, increasing the census material in the Integrated European Census Microdata Database (IECM). The work package was subdivided into four Tasks each of which will be described in section 2 which will highlight the concrete task and the relevant outputs. Section 3 then discusses how the outputs of each task can be maintained in the future and how they could be integrated into the CESSDA workplan. Section four provides a short summary.

2. WORK PROGRAM AND OUTPUTS

2.1 - Task 5.1

The aim of this task was to develop a concept for a European Service Center for Official Statistics (ESCOS), to discuss which kinds of services it could provide on- and offline and how such a service center could be integrated into the CESSDA-ERIC. The sole deliverable for Task 5.1 was a report on a service center for official statistics which was later revised[2]. These reports were the product of extensive discussions and feedback from partners within the DwB project as well as the CESSDA and ESS communities.

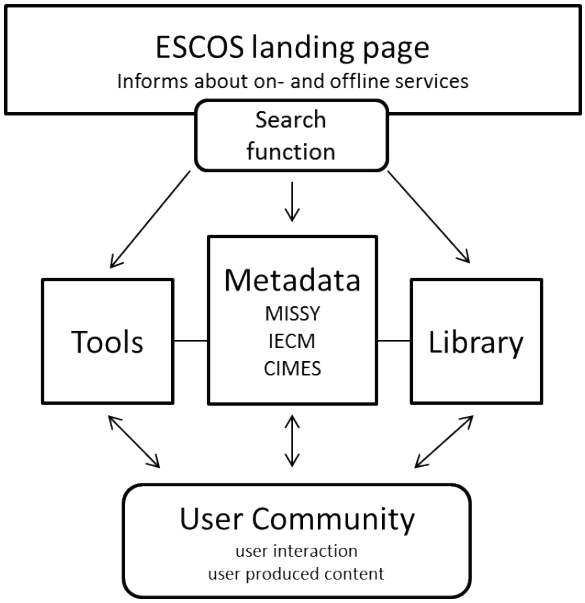
We will outline here shortly the structure of the report and its key recommendations. The report developed a concept for a European Service Centre for Official Statistics which should function as a research infrastructure for European official statistics microdata. Such a center should ideally be established on the basis of the existing CESSDA network of European data archives and cooperate tightly with the European Statistical System (ESS) and specifically Eurostat. The underlying idea is that services provided by the ESCOS should benefit the research community by providing it with services tailored to their needs, as well as the ESS by disburdening them of work which is not part of their core responsibilities. The ESCOS primary objective should be to promote the scientific use of European official statistics microdata by providing services for researchers such as comprehensive metadata and data access infrastructures and by working towards harmonization of data access for scientific purposes throughout Europe.

The report outlined the goals and objectives to be reached by a European Service Center for Official Statistics as well as the underlying motivation and scope of the services to be offered. It also detailed the specific tasks which should be carried out by a European Service Center for Official Statistics. First and foremost the establishment of an online platform (see Figure 1) which provides researchers with structured metadata, adjoining documents and routines as well as a platform for feedback, discussion and user input. Further envisioned services include the provision of training courses on the use of OS microdata, organization of scientific conferences for users of OS microdata and coordination of researcher accreditation and cross border access. Furthermore such a service center could also serve as a mediator between the European research community and the ESS and could

[2] http://www.dwbproject.org/export/sites/default/about/public_deliveraples/d5_1_european_service_centre_report.pdf

lobby for harmonization of data access conditions throughout Europe. The report makes suggestions for how such a service center could be implemented, ideally as a subunit of the CESSDA-ERIC, and outlines stages for a stepwise implementation.

Figure 1 - Structure of the ESCOS Online Platform



2.2 - Task 5.2

The objective of Task 5.2 was to facilitate the scientific use of national OS microdata for comparative research by providing comprehensive study level metadata in English language for every dataset available to researchers from European NSIs and NSAs in one centralized database. To this aim CNRS-RQ together with its partner CASD/GENES developed the CIMES system, which was primarily funded by institutional resources.

Currently acquiring information on restricted data and accreditation procedures, in particular for microdata for countries other than one's country of residence, is rather burdensome to researchers and sometimes simply impossible. Conducting comparative research requires an in-depth search through the websites of each national statistical agency to discover which data are available while understanding how each producer documents their own data. While most providers of OS microdata host information on data and access conditions online, detailed information is usually available only in the native language. Making these data easier to find, more comprehensible and more usable, requires two improvements to the descriptive metadata. Firstly a centralized resource needs to be made available and secondly data documentation needs to be structured, standardized and presented in English language. In order to ensure that the produced metadata is standardized and thus can also be reused by others (specifically the DwB WP 12 portal) the standard of the Data Documentation Initative (DDI) was employed.

The metadata scheme employed within Task 5.2 uses a hierarchical structure and is composed of three levels. The series level describes a data collection program which is carried out over time (e.g. German Microcensus). The study level then describes an individual instance of this study program, usually a year in which a study was carried out (e.g. German Microcensus 2007). The dataset level then describes different versions of this study which are issued by the data producer. For the

purpose of Task 5.2 files with varying degrees of anonymization and accessibility such as Secure Use Files, Scientific or Public Use Files are differentiated here.

Table 1 - Documentation of national OS microdata in Task 5.2

Country	Number of series	Number of studies	Number of datasets
Austria	5	29	51
Belgium	7	42	42
Bulgaria	0	0	0
Croatia	4	37	91
Cyprus	0	0	0
Czech Republic	5	35	125
Denmark	20	18	16
Estonia	3	34	40
Finland	22	17	11
France	18	134	156
Germany	9	87	131
Greece	4	17	30
Hungary	0	0	0
Iceland	1	0	0
Ireland	6	6	0
Italy	10	49	49
Latvia	6	29	27
Lithuania	4	10	6
Luxembourg	1	6	6
Malta	0	0	0
Netherland	13	171	0
Norway	8	137	231
Poland	4	14	13
Portugal	13	82	85
Romania	15	145	143
Slovakia	5	26	27
Slovenia	19	35	67
Spain	21	227	290
Sweden	2	8	0
Switzerland	15	151	160
United Kingdom	8	24	24
Total	248	1570	1821

CNRS was responsible for Task 2 and headed the coordination of duties as well as the development of a data editor which was named CIMES (Centralising and Integrating Metadata from European Statistics). The other partners were assigned to document data from a number of countries to which they likely had easy access either due to existing language skills or ties to respective NSIs. Metadata was collected mainly from the webpages of data providers and entered into the metadata scheme manually online into the CIMES tool. Responsibilities for countries were assigned as follows:

CNRS – Denmark, Estonia, Finland, France, Greece, Latvia, Lithuania, Poland
CED – Spain, Portugal, Italy
GESIS – Germany, Austria, Netherlands
RODA – Romania, Hungary
UL – Slovenia, Slovakia, Czech Republic, Croatia
FORS – Switzerland, Belgium, Luxembourg

Data were documented with the CIMES tool which CNRS-RQ developed independently and largely from its own budget for the purpose of this data collection. CIMES is a web-based application which allows different partners to produce metadata simultaneously and to store it in the same database (a MySQL relational database). The CIMES system has been made available for public use as of March 2015 and can be accessed at http://cimes.casd.eu. Furthermore the metadata contained within was successfully harvested by the WP12 portal.

In sum over 1500 studies from 27 countries in Europe were documented (see Table 1 for details). While this is an impressive amount of metadata, coverage is not complete as there are some countries for which no studies were documented. Furthermore the coverage of the database is far from complete. The project team was rather surprised by the amount of studies from official statistics available throughout Europe which are produced by a wide number of institutions such as ministries, central banks and so forth. Thus selection criteria were developed for selecting the most relevant studies. The criteria employed here were: relevance to social sciences, cross national comparability, broad topical coverage.

After completion of Task 2 Réseau Quetelet, in cooperation with its partners, CASD/GENES, CNRS is currently continuing the development of CIMES and adding to the documentation efforts in DwB by including additional datasets. Latest developments of the tool include making CIMES open to the public, adding a few additional functionalities to make CIMES more user friendly when going through the list of countries and datasets, adding a field for the European integrated micro data and integrating at the country level the fact sheets DwB WP3 have produced for each country regarding the legal framework and the accreditation procedure in general for their National Statistical Institute.

Réseau Quetelet has also started to document more datasets under its own funding, with a first objective of completing the list for some countries. At the writing of this report 32 datasets have been added for Ireland and 24 for UK. Under the field of the European integrated microdata, the list includes a) the Eurostat microdata providing the link to MISSY for those currently documented in MISSY as well as the link with Eurostat for access; b) other integrated micro data such as those provided by Eurofound, the DG ECFIN business and consumer tendency surveys and the Eurobarometers with links to the providers. The results of this work package are summarized in greater detail in the DwB Deliverable 5.2[3].

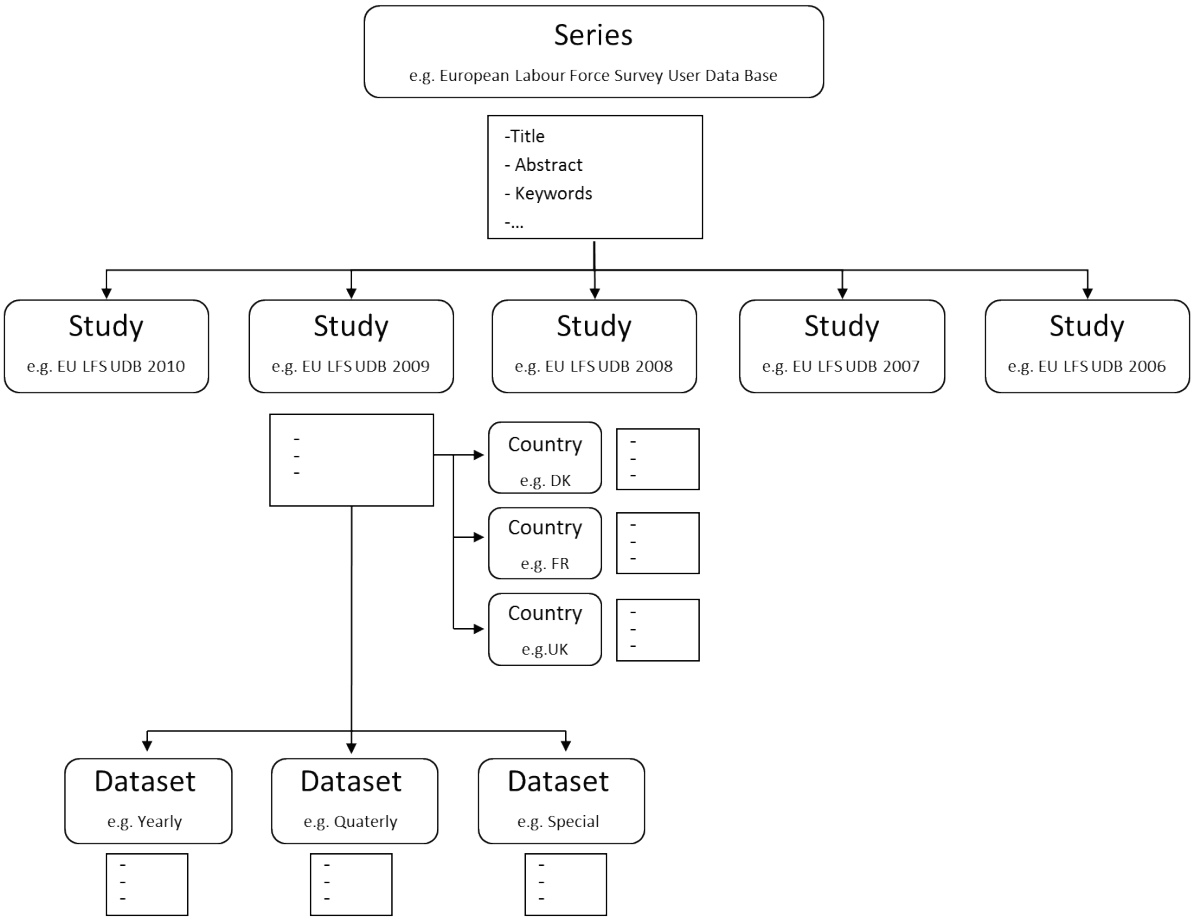
2.3 - Task 5.3

The goal of this task was to document integrated European official statistics microdata. This included two different data sources: census microdata which was incorporated into the IECM system and microdata from Eurostat which was entered into the MISSY system. While the objective of the data

[3] http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d5-2_databank-national-survey_report_final2.pdf

documentation in Task 2 was to provide a broad database this task aimed at providing in depth information for a far smaller amount of studies. This data documentation should likewise aid researchers in data exploration and allow them to learn about the topical, geographical and temporal coverage of a data source in order to gauge whether it is a useful basis for a specific research project. Beyond that however the documentation produced as part of Task 5.3 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. While data was entered into two different databases great care was taken to ensure compatibility of metadata and seamless integration into the WP12 portal.

Figure 2 - Task 5.3 Study Level Metadata Scheme



The metadata schema used for the documentation of Eurostat microdata uses the same general structure as that employed in Task 5.2 but is more detailed and slightly more complex (for a schematic representation see Figure 3). It is also structured hierarchically with a series, study and dataset level. The series level describes a data collection which usually spans over time and over multiple countries (e.g. EU-SILC). This level includes a number of general metadata items describing the coverage and intent of the series. The study level then goes on to describe a specific instance of a series (e.g. EU-SILC 2009). The metadata scheme used in Task 5.2 had to be expanded to include a subsection on country specific information which details specifics of data collection and sampling in each country. The dataset level is then linked to the study level and describes different instances of a study; in the case of the EU-SILC for example this would include the cross sectional and the longitudinal dataset (see Figure 2 for a schematic representation).

Entry of metadata was handled via the MISSY Editor. A web application which allows for entry of metadata either manually or via import of structured Excel files. The MISSY editor was developed by GESIS as part of an independent project but extra efforts were made to ensure that the partners of DwB could access the editor. Data documentation was handled as a two-step process. In a first step variable level metadata is imported from an SPSS system file into the MISSY Editor, this file has to be generated by the respective partners. To this aim CNRS had coordinated a data access request to Eurostat so that all partners would have access to the microdata. The metadata imported from the system file was then complemented manually by information drawn from documentation provided by Eurostat and respective NSIs. The data model underlying MISSY is based on the DDI-DISCO data model which made it easy for the WP12 team to harvest metadata from the system.

The MISSY system went online in January 2015 and can be fully accessed by the public at http://www.gesis.org/missy/eu/missy-home and to date includes the following integrated European microdata series: European Labor Force Survey, European Statistics on Income and Living Conditions, Structure of Earnings Survey, Community Innovation Statistics and Adult Education Survey (Table 2 provides an overview of the included series and the temporal and geographical coverage of the metadata).

Table 2 - Documentation of Eurostat Microdata for Task 5.3

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

The IECM project is currently disseminating to the research community census microdata from 16 European countries. There are a 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 3). 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 around 1,500 original variables have been integrated into 200 harmonized variables. This includes 70 million person records. The IECM can be accessed at http://www.iecm-project.org.

Task 5.3 has developed a wide array of comprehensive and structured metadata for different microdata from official statistics. A more comprehensive account of the outputs of this task can be found in DwB Deliverable 5.4[4].

[4] http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d5-4_databank-documenting-integrated-eu-data_report.pdf

Table 3 - 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

2.4 - Task 5.4

The prime objective of Task 5.4 was to produce routines for use with common statistical packages (i.e. SPSS, SAS, Stata and R) which aid users of Eurostat data. Two different types of services have been provided here: routines which assist users in data preparation, these “setup files” and routines which operationalize common social scientific concepts, scales or indicators or which assist researchers in restructuring data files, we term such files “microdata tools”. These tools will often involve a more elaborate documentation or in some cases take the form of a technical report.

The highest priority for Task 5.4 was to produce setup files for all OS microdata included in the documentation of Task 5.3 as these setup routines for use with integrated European OS microdata provide an invaluable service for researchers since these datasets are rather complex and require a large initial time investment before one can begin with data analysis. A brief description on 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 lines. 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 produced as part of Task 5.4 handle this for researchers thus saving them hours of time, which can instead be spent on actual data analysis, and increasing the comparability of research outputs produced with these data. 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[5] and is thus freely available to the research community.

Additionally a number of tools and routines were developed to assist researchers working with Eurostat microdata, this includes for example a report on Income harmonization in the EU-SILC or a tool which calculates a number of innvation concepts on the basis of the CIS.

The setups and microdata tools provided within Task 5.4 complement the metadata generated in Task 5.3 and make official statistics microdata far more accessible to researchers. Together these services can become a valuable tool to explore the contents of the data. This will guarantee that users obtain a better sense of the contents and quality of data and have an easier time in doing their actual data analysis. A more comprehensive overview of the outputs of this task can be found in DwB Deliverable 5.3[6].

[5] http://cran.at.r-project.org/web/packages/DDIwR/index.html
[6] http://www.dwbproject.org/about/public_deliveraples/dwb_d5-3_routines-eu-data_report_final.pdf

3. OUTLOOK

Deliverable 5.1 contains a first outline of how we envision the long term integration of services developed as part of this work package into the CESSDA-ERIC. As mentioned in section 2.1 metadata on national and integrated OS microdata should become the cornerstone of an online service center which is further complimented by tools and routines which help researchers in the process of data preparation and analysis. While the report also provides suggestions for steps to be taken toward that direction the overall focus of Deliverable 5.1 is more conceptual while the report at hand focuses on the concrete tasks and their technical implementation. In this report we have provided a summary of the outputs of Work Package 5 and will provide an assessment of the current state of affairs and whether and how these services can be maintained, expanded and updated.

The IECM database has been vastly expanded as part of the work conducted within the Data without Boundaries project. CED plans to continue this project and include new samples of census data continuously. For the future it is CED's objective to intensify cooperation with CESSDA and participate in plans for a European Service Center. Meanwhile, IECM will continue participating in EU-wide level research infrastructures. IECM is a partner of the Integrating Expertise in Inclusive Growth project (InGRID), funded by the European Union's Seventh Framework Programme. Within this project, CED plans to develop poverty mappings based on IECM data.

The database produced as part of Task 5.2 has been integrated into the CIMES system. After some additional work which was put into the system by CNRS-RQ it is now up and running and can be accessed by the public. Sustainability and future development of CIMES after the end of DwB is currently envisioned in three steps.

1) In a first step, Reseau Quetelet will continue with its own funds to document additional data sources for some countries, including more data sources which can be used for comparative research. A particular focus will be placed on data available in the Research Data Centres for confidentiality reasons such as the business files and employer/employees datasets. Furthermore Reseau Quetelet will revise and complete the section on type of files (PUF, CUF, SUF, ScUF with the links to the data providers) when necessary for some of the datasets currently documented.

2) In a second step, it is envisioned that partners from DwB who have participated in Task 5.2 could agree via a Memorandum of Understanding to continue to work on data sources from countries they had worked on prior within DwB.

3) In parallel, it is envisioned that two processes will go on to ensure long term sustainability and development of CIMES. One process will be to discuss with the respective teams the opportunity to develop the tool at the variable level, using MISSY thus going further in the integration of the two tools used within DwB Work Package 5. The Horizon 2020 calls can provide an opportunity for such developments. The other process is linked to CESSDA which is currently setting up a procedure intended to assess the output of projects conducted by its member institutions (e.g. DwB, DASISH) and decide whether and how they can be integrated in the CESSDA work plan. In the current context, where few CESSDA members have developed cooperation with their respective NSI, and where most NSIs and other government bodies producing official microdata do not provide metadata that could be harvested by the new CESSDA portal to be set up according the CESSDA work plan, CIMES provides a basis for such development in the area of official statistics microdata and can be seen as a

driver for more cooperation at the national level between the members of CESSDA and the NSIs. With this perspective, steps 1 and 2 are seen as intermediary steps and a contribution for step 3 and full integration into the CESSDA work plan.

As an important requirement stated in the description of work was to ensure that the database can be integrated into the WP12 portal the necessary steps have been undertaken by the IT staff at CNRS-RQ on the one hand and NSD and Metadata Technology on the other to ensure that metadata from CIMES can be harvested and displayed within the WP12 portal.

The metadata on Eurostat data collected as part of Task 5.3 has been successfully entered into the MISSY system. The system was launched in January 2015 with SILC and LFS metadata and the documentation of AES, SES and CIS was completed in April 2015. Due to the tight cooperation with the WP 8 and 12 it has been made possible to harvest this metadata into the WP12 portal.

The MISSY system is a cornerstone in the services provided by GESIS for official statistics microdata and will be maintained. Based on the current staffing the updating of metadata on EU-SILC, EU-LFS and AES should be secured for the foreseeable future, however currently GESIS has no capacity for continuing the documentation of SES or CIS. While the technical infrastructure which would enable external partners to enter metadata remains in place, none of the DwB partners have the personnel resources to do so without additional funding. However the technical infrastructure is in place to enable similar cooperation in the future and could be a fruitful possibility for future cooperation in the CESSDA context and a possibility for CESSDA partners besides GESIS to develop expertise on Eurostat Microdata. Such expertise is not only important in providing metadata but also for training and consultancy of researchers and data producers.

As an important requirement stated in the description of work was to ensure that the database can be integrated into the WP12 portal the necessary steps have been undertaken by the IT staff at GESIS on one side and NSD and Metadata Technology on the other side to ensure that metadata from MISSY can be harvested and displayed within the WP12 portal.

As to integrating the services presented here into the CESSDA work program first steps have been undertaken at the time of this writing. CESSDA has started to evaluate DwB outputs and to discuss which of these should be maintained as CESSDA services in the future. The participants of Work Package 5 will continue to lobby for the idea of a European Service Center for Official Statistics in order to ensure the sustainability of the developed services.

4. CONCLUSION

Work package 5 of the Data without Boundaries project has produced tools and services to aid the scientific usage of official statistics microdata from Europe. The availability of centrally accessible and structured metadata as well as a wide range of routines will ease the tasks of data exploration and analysis for researchers using European OS microdata. These services benefit not only the research community but are also a considerable aid to data providers. On the one hand the task of structuring and translating metadata is handled for them on the other hand they can also benefit from the increased scientific applicability of their data.

As has been argued for in Deliverable 5.1 we recommend that these services should become an integral part of CESSDA-ERIC ideally by establishing a subunit which is responsible for OS microdata and coordinates, maintains and advances these services.