UNITED NATIONS ECONOMIC COMMISSION
FOR EUROPE (UNECE)

EUROPEAN COMMISSION

CONFERENCE OF EUROPEAN STATISTICIANS

STATISTICAL OFFICE OF THE EUROPEAN
UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (vii): Trans-border access to microdata

The Need for Networks on Data Access - Data Without Boundaries Project and the Workshop on Data Access

Prepared by Roxane Silberman, Centre National de la Recherche scientifique Réseau Quetelet,
Stefan Bender, Institute for Employment Research (IAB), Germany and
Anco Hundepool, Statistics Netherlands

The Need for Networks on Data Access - Data Without Boundaries project and the Workshop on Data Access

Roxane Silberman[*], Stefan Bender[**] and Anco Hundepool[***]

[*] Centre National de la Recherche scientifique Réseau Quetelet, email roxane.silberman@ens.fr
[**] Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nuremberg, Germany, email Stefan.Bender@iab.de
[***] Statistics Netherlands, Department of Methodology and Quality, P.O. Box 24500, 2490 HA Den Haag, The Netherlands., e-mail ahnl@cbs.nl

Abstract: There is a very active series of projects going on in Europe with respect to the access across borders to confidential micro data for scientific research and the confidentiality issues related to that. For example Eurostat is supporting several ESSNet projects with National Statistical Institutes in this field. There is also a need to involve the other actors as the Data Archives and the researchers who are the final users. In this paper we will give an overview of the 7th framework project "Data without Boundaries (DwB)" and the series of "Workshops on Data Access (WDA)". Both initiatives will help to establish an international data access. Their main goals are to exchange knowledge, define standards and find solutions for best practice.

1   Introduction

National Statistical Institutes collect very rich datasets, which serve as the basis of the production of statistics on all aspects of the modern societies in which we live. In the information era governments and decision makers require very detailed information before they can take their decisions. Also the public expects that the decisions are based on detailed evidence.

These rich dataset also could have a second life. Researchers at universities and other research institutes need detailed and precise databases for their work. They also participate to policies evaluation which is increasingly based on cutting edge research.

On the one hand side the NSIs are of course willing to meet these very respectable needs, but on the other hand they have the legal and moral obligation to respect the privacy of the respondents. They must find the right balance between the two.

As a consequence SDC procedures and solutions as "on site" access, remote execution or remote access have been set up at national level. Yet access across borders is now increasingly on the agenda as comparative research develops. Before an international data access could be established, there is a strong need for

exchanging knowledge, define standards and find solutions for best practise. There is also a need to involve all actors in that discussion, the NSIs, the researchers and other stakeholders as the Data Archives who have long experience in disseminating microdata for research.

Both the FP7 project "Data without Boundaries (DwB)" as a European initiative and the "Workshop on Data Access (WDA)" at a smaller level yet involving non European countries are working in this perspective.

In chapter 2 we will give an overview of the FP7 project Data without Boundaries, while in chapter 3 we will provide some details on the Workshops on Data Access.

2   The Data without Boundaries project

2.1 Introduction

While access to anonymised official microdata for researchers is still uneven both at national level within Europe and access to Eurostat datasets still burdensome, access to highly detailed and sensitive microdata is now increasingly on the agenda. Yet the different member states have substantially different outcomes for research access to Official data. Crossing borders is even worse as there are different legal frameworks, different institutional arrangements and different criteria for research accreditation. Providers may be different: in some countries National Statistical Institutes (NSIs) have built cooperation with the Data Archives whereas in others it is not the case. Mode of access range from no access, on site access, remote execution, bespoke tabulations to remote access. Access to foreign researchers may be forbidden. There are different views about security, anonymization, output checking. Language may be an important issue as translation is still not frequent. For a European research team requiring access to microdata from different European countries, there is no single access to find what data are available, where and how, metadata standards – if there are some - differ and they will have to deal with several applications for accreditation. Working on the different databases together may be impossible when access to detailed microdata is required. The issue is not just efficiency, but real harm to the contribution of the social sciences to democracy in an information society.

Yet there are new conditions to build in Europe. In the recent years, there have been changes in the legal framework in some countries. An increasing number of Research Data Centres (RDCs) are providing access to highly detailed microdata, in several cases also to foreign researchers. There are examples of fruitful cooperation between Data Archives and NSIs. At European level, there is a will to build a European data infrastructure both on the Archives side building on the Data Archives European network (CESSDA) long experience, and on the NSIs side building on initiatives

within the European Statistical System such as the discussions for a new EC Regulation on research access to European microdata expected about 2012.

Therefore the primary objective of Data without Boundaries (DwB) project which aims at enhancing access to official microdata in Europe is to take advantage on these existing infrastructures and prepare the essential relationships and build trust, common view and agreements on standards between the European Statistical System led by Eurostat, other stakeholders as the Central banks, CESSDA and the researchers who are the final users; from access as a postcode lottery, to an integrated model where the best solutions for access are available irrespective of national boundaries and are flexible enough to fit national arrangements. It aims at

a) discussing frameworks and proposing pilots based on volunteers for a European accreditation and a distributed remote access for confidential microdata to be expanded later to other partners, both for national and European datasets;

b) fostering discussions and promoting improvements and solutions for the entire communities through annual/bi-annual European data Forum, regional workshops, users conferences, training sessions, staff visits;

c) preparing an easy and single point of access (What data are available? How can I access them?) for the researchers, to be linked to the CESSDA portal where NSIs metadata could be harvested when not available through the CESSDA archives still providing access to official microdata;

d) immediately enhancing access to official data making European datasets more useable (metadata, routines) and supporting foreign researchers transnational access both on site and through remote access system to countries official microdata. Close coordination with the European Statistical System discussions and initiatives as well as with on-going and future related projects is developed to ensure maximum synergy and incorporation of outputs.

Close coordination with the European Statistical System discussions and initiatives as well as with on-going and future related projects is developed to ensure maximum synergy and incorporation of outputs.

There are 27 partners from 12 European countries participating the efforts, both statistical offices and data archives, also including several universities involved in methodological research. The DwB project is funded by European Commission's 7th Framework Programme for 4 years (2011-2015) The project is divided in 12 workpackages which we will highlight some important points here. Whereas WP 1 and 2 are concerned with the standard Project management, Internal and External Communication, Promotion and Dissemination, the 10 other WPs can be best presented as 3 blocks, the Access Facilities block, the Front office block, and a third block that aims at enlarging cooperation.

2.2 Access facilities objectives.

The 1ˢᵗ block about access facilities deals with 3 main issues:

- move towards a European accreditation (WP3)

- propose an architecture for a distributed remote access to highly detailed microdata (WP4)

- agree on methodological issues such as anonymization and output checking (WP11)

The work is based on overviews of the current situation, discussions and agreements between partners from both sides about standards and bests practices and implementation of pilots within the current context based on volunteers. In addition the project immediately offers transnational access through open calls for researchers to access data either on site or remotely (WP9 and WP10), and these experiences will be used as business cases for the other WPs objectives. It is also expected that changes in the European regulation will open new possibilities for access to Eurostat microdata within the time frame of the project.

Moving towards a European accreditation is essential for comparative research projects. Nowadays, researchers willing to access microdata in different European countries encounter a number of important barriers. The project will devise and promulgate a fit for purpose standard for researcher accreditation for the use of official data that NSIs (and archives) will find credible and addresses cross border issues. It aims at achieving a widely-recognised standard for accreditation which will reduce administration and costs, improve efficiency and improve confidence in custodians of official data when providing access - thus helping to remove barriers for research use of official data.

Yet, the legal frameworks for research access to official data remain an important barrier in several cases. Whereas in some countries there have been changes providing an efficient context for building secure access to highly detailed microdata for researchers, it is not the case in other parts of the European area, both at national and at European level. The project will conduct an audit and describe the legal frameworks for research access to official data in the European Research Area, present the results in a useable form as a contribution to effective policy planning and legal changes where opportunities arise and act as a definitive guide for all interested parties on the legal frameworks and accreditation processes for access to data across the ERA.

A distributed remote access architecture is a second essential condition for comparative research projects. Whereas in some cases, access across borders is allowed for foreign researchers, technical solutions differ widely. The most efficient and for researchers most convenient way of having access to microdata of NSI is remote data access. Remote data access means that an approved researcher may

access restricted micro data for her approved project via a secure internet connection (see Grim et al. 2009 and Hundepool et al. 2010). Although in many countries NSIs have started to make micro data accessible for the scientific community, national legal concerns and restrictions prevented an overall establishment of remote data access systems so far. This distributed remote access should also allow the researchers to work together and not separately with the different datasets. The project will 1) identify the challenges and solutions for building and operating a Remote Access facility in compliance with internationally recognised Information Security standards (WP3) 2) analyse a set of alternative suitable organisational architectures for a sustainable co-operational model for linking data centres, and 3) propose a suitable organisational architecture as a credible working model and if possible a proved concept (WP4).

The objective of this work package 4 is to examine how e-technology based Remote Access (RA) environments, as proved in these existing examples, may help widen and enhance data access across the European Research Area.

The WP will scope of a researcher environment in a Remote Access setting. At a minimum, the researcher environment should be interactive, dynamic, one that leverages technology, is rich in metadata, provides collaborative spaces and social networking tools and of course provides access to datasets. The concrete experiences of transnational access provided through WP9 and 10 (for a more detailed description see below) will be used as business case to define the researchers needs.

The project will also conducts an overview of the current architecture of the RDCs as a first basis for a definition of the organisational architecture of the network of RDCs / Safe Centers (SCs) and of RA which will require definition of procedures for cooperation and internal testing for a SC and RA network implementation.

Building a real RA network between some SCs will be the final objective based on volunteer partners. Secure connections between some SCs should be established operationally. With the Research Data Centre-in-Research-Data-Centre (RDC-in-RDC) approach (Bender and Heining 2011), also building in cooperation with the SDS at UKDA and the GENES RDC in France which have developed SD-Box for remote access, Workpackage 4 of DwB tries to overcome the existing legal barriers and to bring micro data access closer to the ideal perception of remote access and to start a network of RDCs, where transnational access to foreign data sets are possible for researchers. The basic idea of this approach is to allow data access from designated national and international institutions with comparable standards to the national RDC, where the data are still stored. By using – for example - a Citrix-thin-client solution, researchers can access the whole scope of micro data available for on-site use from a national RDC without travelling to it.

Agree on SDC procedures

The project general aim is to widen and enhance access to micro data by using modern data design and access techniques. Databases can be designed with different levels of disclosure risk. This ranges from no meaningful risk (i.e. safe for public use) to retaining the complete and identified original record of the data subject. There is a relationship between disclosure risk and utility. The challenge for each dataset design is to build in as much utility as possible for the least disclosure risk for a dataset.

WP11 focuses on enhancing the SDC procedures. High quality SDC is a prerequisite for taking informed decisions about giving access to microdata. Partners in this workpackage are the Statistical Offices of the Netherlands, Germany and UK and the universities of Southampton, Barcelona (technical), Tarragona (Rovira I Virgili), CSIC and Tenerife (La Laguna).

There are 5 tasks in this workpackage:

Task 1. Selection of statistical software tools for assessing disclosure risk and producing lower risk datasets and the development of an architecture document if no suitable tools are found to exist.

Task 2. Statistical methods for dataset design to maximize utility whilst keeping disclosure risks low. Research on micro-data design methodology is necessary to maximize the utility of data for a given disclosure risk. It should be noted that this is true across the spectrum of datasets, from public use through to datasets accessible only in research data centres. It is important that the modified datasets have a similar structure (correlations, etc.) to the original data. Therefore, research datasets should be generated using a good statistical disclosure control method, or, alternatively, a synthetic data generator.

Task 3. Tabular disclosure control (safe outputs) SDC remains an important task, even if the research data have already been made less disclosive. This is particularly true in the case of tabulations resulting from research on small domain data and a severe problem when the data source is full population data stemming from a census or administrative data. There are several tasks here:
- Statistical methods for flexible safe tabular output. Several methods for protecting frequency tables will be studied and eventually further developed. Some of the methods include Record swapping, Unbiased Random Rounding, Cell Perturbation Method, Invariant Post-tabular SDC method
- Methods for flexible safe tabular output based on linear optimization approaches In this task we will improve the current code of the RCTA methodology by implementing a heuristic solution.
- New versions of the cell suppression and controlled rounding. The current solutions here will be improved and rewriting using open standards, both for the code itself and the use of free open solvers.

Task 4. Testing the results in real life situations. This task will research the real life threats to confidentiality experienced by data service providers (i.e. not hypothetical, scenario-based, or mathematically modelled threats which have already been subject to substantial research work with published results). This research will feed into the other tasks of this work package to inform the research activity. We will focus on record linkage techniques here. There are two objectives here:

- Analyse the synthetic SDC methods, to find flaws and develop adhoc record linkage tools. Evaluate the risk using these adhoc algorithms and compare with the results obtained with standard record linkage algorithms. Relate the record linkage framework into the probabilistic modelling framework for disclosure risk assessment.
- Develop record linkage methods and data cleaning tools to integrate multi-databases dynamically. That is, record linkage methods will be developed so that they will be applied "on the fly" when users query the system.

Task 5: Producing standards and guidance. There is a need to have consistent guidelines for output checking. Software tool selection and design, micro-dataset design, and output checking methods have to be appropriate for all of analysis undertaken by researchers yet in aggregate adequate to manage disclosure risk. This task will research best practice in standards, guidance, and make recommendations. Output checking is the preferred model for any general-purpose RDC (see ESSNet on SDC). For international projects there is a need to have consistent guidelines. SDC methods have to be appropriate for safe-guarding data confidentiality irrelevant of the type or level of complexity of analysis undertaken by researchers. As it is impossible to predict all likely analyses in advance an SDC system must be sufficiently comprehensive and flexible to address all possible scenarios.

Immediately enhance transnational access to highly detailed official microdata

DwB will also aims at immediately facilitate and enrich transnational research that requires access to highly detailed or sensitive microdata (WP9 and WP 10). Regular calls will invite academic researchers, including PhD students from countries within the European Union or the European Free Trade Association to apply to work with data from 4 countries, France, Germany, the Netherlands and UK, where current conditions allow access for foreign researchers through the so-called Research Data Centers (RDCs). DwB will provide financial support for travel and accommodation costs and any charge to use the RDC whether for visiting the RDC prior to remote access (WP9), for onsite access (WP10) or for accreditation and training purposes when presence on site is required. Researchers will also be given special onsite support in using the RDC. Researchers teams conducting comparative cross-national research requiring simultaneous access to more than one RDC will be very welcome. RDCs partners will work together to facilitate in time accreditation and discuss any

improvement that could be possible within the current national contexts to fit the researchers needs facing different accreditation processes, different type of access, different standards for outputs checking. Conversely, these research projects will serve as "business cases" for moving toward a European accreditation (WP3) and a distributed remote access (WP4).

2.2 Front office block

Three other WPs aim at providing better and integrated information about data available, way of access and at improving the use of the datasets. Currently this information is fragmented at national as well as at European level, depending on the different providers, the NSIs and other statistical departments and agencies, also including the Central banks, depending on the relations with the Data Archives that are members of CESSDA. Metadata may be poor, NSIs and Archives do not use same metadata standards. Language issues are still important.

This block aims at :

- Discussing standards (SDMX and DDI) and develop tools to harmonize metadata. WP7 will provide an overview about the current use of metadata standards in the NSIs (mostly SDMX) and the Data Archives (DDI) and on going discussions between SDMX and DDI communities.

- Devising techniques to harvest NSI metadata by CESSDA (WP8). The CESSDA portal is essentially a discovery tool and gateway to the data holdings of the current network of CESSDA data archives. Whilst this is a tremendous resource in its own right, in relation to Official Statistics (OS) data, it is clearly incomplete. Only few archives have built strong cooperation with the NSI in their respective country, providing good information about official data. As an initial step, bringing the disparate and variable information that exists on the availability of OS research data together from across the ERA forms a key priority for WP8. It will investigate the possibilities and problems associated with harvesting NSI metadata on OS data and making them available through an enhanced CESSDA portal. The WP will take a rather different, but parallel complementary approach to the problem – current OS metadata and associated interchange formats will be evaluated and an object model applicable across the disparate OS will be developed. The information derived will inform the creation of a metadata model incorporating SDMX and DDI as well as any system-specific enrichment required to deliver extended portal functionality.

- Providing a single point of access to official data within the context of the CESSDA portal and the current CESSDA ERIC process is the final objective in the perspective of building a real European data infrastructure. . Finally, WP12 will build and implement the plans set out in WP7 and WP8 to create a more coherent system for resource discovery of official statistics. It will present a single point of information about official and other statistics from both data archives and national statistical institutes.

- It will also incorporate the work of WP5 that aims at servicing the use of the OS microdata, a) providing a comprehensive overview of available national official data in Europe, how and where to access them, and basic structured metadata on main surveys, b) providing structured metadata and user friendly routines for Eurostat data and census data. Given the complexity of microdata disseminated by Eurostat (e.g. the integrated EU-LFS data consists of over 1,200 separate files) and census data disseminated through member countries researchers invariably require more extensive documentation, improved and more flexible tools to read data into different statistical packages, and routines to build cross-sectional or longitudinal databases. The work will provide the basis for a future Service center for European OS microdata to be linked to the CESSDA portal.

2.3 Enlarging cooperation.

The DwB project gathers several CESSDA members, several NSIs, researchers and stakeholders with the aim of discussing standards that could be agreed by all communities and thus enhance access to official data across Europe. It also builds pilots for a European accreditation, for a distributed remote access for highly detailed data, for harvesting NSIs metadata. All WPs work in cooperation and aim at identifying best practices, agreeing on standards and building on volunteers while bridging the different communities (NSIs, archives, researchers).

Yet the long term success requires involving and building trust between the whole ESS, the whole CESSDA and the researchers who are the final users, also making bridges with non European partners. Other stakeholders are also interested such as the Central Banks, government agencies holding administrative data bases important for research and policy evaluations and of course Eurostat and other European and international bodies.

This objective is met through : a) A biennial European Data Access Forum and two regional workshops on data access, b) Users conferences where researchers are able to present their work based on the use of official microdata ; c) Training activities for researchers ; d) staff visits in RDCs to promote their development within Europe.

3 The Workshop on Data Access

Whereas DwB focuses on access across borders within Europe building on existing European infrastructures, Workshop on Data Access (WDA) brings together researchers who are active all over the world in promoting innovating methods to provide access to highly detailed microdata.

3.1 Introduction

All NSIs face the dilemma of how to provide microdata access to researchers while at the same time guaranteeing the confidentiality of data providers. Traditional access modalities have included public use files, licensing, and increasingly, RDCs. New innovations have recently emerged - notably the use of multiple imputation techniques to create public use files, and technological solutions to provide remote access.

3.2 Main Idea

The aims of WDA is to bring together researchers, who are active in promoting these new innovations on data access and Research Data Centers (RDCs) so that they can share knowledge, outline a future research agenda and build up a network for future activities. So, participants in WDA are reviewing access to confidential microdata for research purposes. WDA builds an expert network which could help to identify best practice and disseminate advice about effective management of research facilities.

WDA is self-organized by (a group of) RDCs and follows a "bottom up"-approach. Topics for the workshops are generated by the RDCs or researchers – mostly by presenting a relevant topic on one of the workshops. Because of the organisation "rules", WDA is highly interactive and is – in a strict sense – not a formal group of researchers/RDCs. Although some researchers/RDCs have participated in every workshop, there is not really a representation of WDA. Because of its openness representatives of RDCs and institutions/researchers, who are dealing with data access from Africa, Asia, Australia/NZ, Europe and North America joint one of the meetings.

3.3 A small history of WDA

In 2006, Julia Lane (NSF) and Stefan Bender (IAB) came up with the idea of WDA. In August 2007 the first Workshop on Data Access took place in Nuremberg, which was organized by the RDC of the German Federal Employment Agency in the IAB. Participants from Canada, Denmark, Germany, the Netherlands, New Zealand, Serbia, the UK and the US joined the workshop. Main topics were the interchange of developments in technology, the legal and statistical environments, and management issues. Special sessions on multiple imputation for public use files and metadata were organized. All participants were enthusiastic about the knowledge sharing and decided to have a second workshop.

In February 2009, the second Workshop on Data Access took place at the Office for National Statistics' head office in Newport, supported by the IAB and the German Data Forum (RatSWD) and organized by Felix Ritchie. Participants from the UK, Germany, Netherlands, Slovenia, France, Canada, the US, Japan, China and Australia reviewed developments in technology, the legal and statistical environments, and management issues across countries with a much wider range of experience. Among the highlights of the meeting was the news of developments in

the French and Japanese legal systems, greatly increasing the opportunities for microdata research in two major economies.

Two main themes arose from the meeting. The first was the increasing convergence of statistical legislation in respect of research use of data. While there is still a large variation across countries, there are common themes in recent legislation across the globe, including recognition of the value of microdata research to statistical institutes. The second theme was the growing importance of metadata, particularly as a source of analysis in its own right. Even something as notionally simple as creating micro-aggregated data comparable across countries, or describing data using international standards and definitions, can be informative about both the data and the way they can be used.

The third Workshop on Data Access left Europe and jumped to the US. On May 2009 it was organized by Felicia LeClere (ICPSR) and Maggie Levenstein (ISR) in Ann Arbor. Participants from Australia, Germany, Japan, the Netherlands, Slovenia, Sweden, the UK and the US joined the workshop. Because WDA was the first time in the US more participants from different US institutions joined the workshop and presented their new developments. From the Europeans side the ESSNet projects and Data without Boundaries were presented to show the strong need for a international coordination on data access. The participants agreed to organize WDA around the topics data access, disclosure, metadata and international data sharing.

Now, WDA will be organized within DwB, because there is a strong need to expanding the network on data access to underrepresented regions and to have a better international coordination between different activities on data access.

4 Conclusion

In the last years in an increasing number of countries access to confidential micro data were established. Research on SDC techniques like solutions on data perturbation or definitions of anonymisation risk and first steps into an automatic output control have helped to establish a culture of access for researchers to detailed micro data collected/stored by NSIs on a national level. Initiatives like WDA helped to exchange knowledge on data access. Yet access across borders is now on the agenda as cutting edge research increasingly require to access to highly detailed and sensitive data from different countries. This issue is now at the top of the European agenda.

Therefore the primary objective of DwB is to take advantage on the existing national cooperations in several countries in Europe between the NSIs, the Archives and the researchers as well as on the European infrastructures as the European Statistical System and the European network of Data Archives, CESSDA to prepare the essential relationships and build trust, common view and agreements on standards on the European level. The main goal of DwB is to have an integrated model on data

access where the best solutions are available irrespective of national boundaries and are flexible enough to fit national arrangements.

Therefore, the work in DwB is organised in three blocks. The Access Facilities block deals with establishing a European accreditation to micro data, propose an architecture for remote access and agree on SDC techniques like output checking. The Front Office block will provide better and integrated information about data available, way of access and will improve the use of micro datasets. The last block is dedicated to enlarge cooperation and to build trust between all parties by organising workshops and conference or establishing an exchange program.

DwB is just one part of a broader movement for having access to European micro data. For example a new EC Regulation on research access to European microdata is expected about 2012. Currently there are ESSNet projects on harmonising SDC and on Decentralized and Remote Access to Confidential Data going on. There are international data forums or OECD activities and national data forums like in Germany, the UK, which are dealing with access to micro data. So, there is a strong need in coordination these activities, because a broad international network of Researchers, Data Archives, NSIs and National Statistical Institutes International Stakeholders (Eurostat, OECD) have to find comparable solutions and should try to establish a (international) division of work. DwB will – and that is not only our hope – play a major role to find proper solutions for an international data access across boarders by sharing the results, coordinating activities and by building trust between NSIs, data archives, stakeholders and – of course – the final users, the researchers.

References

The Data without Boundaries project website, http://www.dwbproject.org

Bender, S., Heining, J. (2011) The research-data-centre in research-data-centre approach: A first step toward decentralised international data sharing, forthcoming, IASSIST Quarterly

Grim, R., Heus, P., Mulcahy, T., Ryssevik (2009) Secure Remote Access system for an upgrated CESSDA RI, metadata technology, Cessda ppp, http://www.cessda.org/project/doc/CESSDA_RI_SRA_FINAL.pdf

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte Nordholdt, E., Seri, G., De Wolf, P (2010) Handbook on Statistical Disclosure Control – Version 1.1. http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf (2010)