The ESSnet-Project "Decentralised Access to EU-Microdatasets"

Maurice Brandt and Patricia Eilsberger*

* Federal Statistical Office Germany,
e-mail: maurice.brandt@destatis.de; patricia.eilsberger@destatis.de

Abstract:
In times of an increasing demand of microdata for scientific research, there is a need to discuss new and innovative ways which ease the access for researchers in the European Union. Most of the Member States already offer access to their national data; the access via a safe centre to the Community Statistics still has to be expanded. The goal of the ESSnet-Project "DECENTRALISED ACCESS TO EU MICRODATA SETS" is a recommendation to set up a network of national safe centres to ease the access to the European Household Panel (ECHP) based on the legal, technical and administrative feasibility. The paper will describe the framework, the objectives and the current status of the project. Some aspects are still in progress and some parts are almost finalised.

1. Introduction

The informational infrastructure in the national Member States (MS), as well as the infrastructure in Eurostat referring to statistical information and data, has been improved over the recent years. It is now possible for researchers to access national datasets in several MS and also European datasets at Eurostat. The request for microdata develops more and more in direction of original microdata without any applied data perturbating anonymisation procedures. There are different ways for researchers to obtain access to original or only slightly anonymised data. One feasible solution is the concept of so-called "guest researchers", who are visiting the Research Data Centre (RDC) of the National Statistical Institute (NSI) to get access to microdata in a safe centre, which is a confidential and secure environment for datasets of official statistics. To access detailed European datasets there is unfortunately only one possibility for interested researchers. They have to visit the safe centre of Eurostat in Luxembourg. On the one hand it is certainly an advantage that the access is basically possible, on the other hand due to the local constraint it is also a barrier for some researchers. Thus, the data infrastructure for European datasets has to evolve to make it easier and more feasible for researchers to use the microdata at a European level. The idea is to develop a "DECENTRALISED ACCESS TO EU MICRODATA SETS" where upon a researcher from a certain MS can use European datasets in his own MS. The advantage of the network of safe centres would be that every single MS can widen the supply of microdata in the RDC, because in addition to the national dataset the researcher can use EU microdata. The concepts of the safe centres which are realized in the MS until now

for using national datasets as well as the concept of the safe centre of Eurostat could be examples for the decentralised access to European microdata sets. The goal of the project “DECENTRALISED ACCESS TO EU MICRODATA SETS” is to prove the feasibility of an access to European microdata in safe centres (on site) of the MS exemplarily for the ECHP data. The question is, whether the implementation of a network of safe centres in the MS is possible and what kind of requirements are necessary for the procedure. According to this, a study of feasibility includes the methodology, guidelines and requirements which are essential to implement access to European microdata in safe centres of the MS. Based on these results the best practice solution for such an implementation could be focused in follow-up projects.

2. Objectives

The project procedure is described in different tasks. First it is necessary to give an overview concerning the present ways of access to microdata in the different MS. Hence, it is necessary to evaluate the actual situation of safe centres in the MS. From a broad variety of thinkable models, a shortlist of viable possibilities to create an European network of safe centres will be worked out. Then, the pros and cons of the solutions on the shortlist, including a detailed discussion about technical, legal and cost aspects of the different possibilities to develop a structure and architecture of a European network has to be conceived. Also based on these results, the construction of a guideline for European safe centres is intended. It is necessary to discuss whether the implementation of a decentralised access to EU microdata sets is practicable and which legal frames of the countries in the European Statistical System (ESS) must be regarded. Furthermore, to standardise and guarantee the anonymity of the outputs, an evaluation of the assignability of general rules of output checking to the ECHP data needs to be proved. Then, the development of a consistent guideline and documentations for researchers to use safe centres will be focused. In the context of a feasibility study a cost analysis for a future implementation is also included. To inform the non-participated NSI’s on the current status of the project, the results and also ways of contact to the national RDC’s will be offered online at www.safe-centres.eu.

3. Results

1.1 Possible network schemes

There are several NSI’s in the European Member States that offer access to national microdata for researchers in RDC’s. As mentioned above, the access to detailed microdata of the European Community Statistics such as the ECHP, is currently only possible at the safe centre of Eurostat in Luxembourg. With respect to this, by now

there was no need to implement any standardisation on dissemination, access and workflow procedures in general. By discussing possible ways to set up a network of safe centres in the EU, especially legal, technical and administrative questions need to be focused.

As for the first approach on a network recommendation, potential network schemes have been examined. The idea is to reduce the burden of Eurostat with respect to find a network solution that considers the current legal constraints (especially for submitting and granting access to the data), the use of already implemented access channels and of course the data availability. The outcome includes the main workflow of an RDC, the special needs and characteristics and so the issues that need to be addressed. Finally this presentation results in possible network schemes for an envisioned solution and summarises the criteria which have to be considered within the next steps.

1.2 Deciding among criteria

Those solutions not leading to a practical and adequate implementation have been dropped. The three possibilities on the shortlist have been proven among a list of criteria. The working progress at this time involves finally a hybrid solution where the most feasible parts of each solution on the shortlist are recognised. The recommendations that have to be considered are:

- Pilot recommendation, short term applicable: decentralised solution administratively and technically run by RDC, data host by RDCs but stored and submitted by Eurostat for each project, alternatively accessed via thin client solution
- Strategic, long term aim: remote access via thin client solution run by RDC (data storage central at Eurostat)

The criteria have been elaborated on the basis of the experiences on the workflow of an RDC (with respect to administration), the legal framework and technical aspects (regarding the IT environment already available). Each recommendation has been compared by questions that envision the simplicity of a future implementation. It was also aimed to give a brief overview of the complexity that may arise, when the solution is to be performed within the context of “setting up” a new RDC and an “ongoing” RDC (local). Considering the need of a short term recommendation that uses existing channels and agreements, and the aim to establish a “new” perspective of data-access in the future, the distinction between “pilot” and “strategic aim” has been made.

1.3 The “pilot” and the “strategic aim”

For the reason that Eurostat is allowed to transfer the ECHP to the national RDC’s (where they can also be stored), it is not necessary to reform the legal ground first. Thus, the project team finally agreed upon a “pilot” solution that varies (in data storage) and a long term recommendation for an innovative development of the informational infrastructure in Europe. The “pilot” solution aims are an easy and fast implementation of the network. Thus, it is either possible to store the data at the local RDC’s and for those who do not agree in storing (and disseminating) the data, Eurostat can implement a remote access, to which the RDCs may get access through their safe centres. The process foresees:

1) ECHP is stored at the local RDC
2) researcher requests data at local RDC
3) local RDC proves accessability of the institute and gives a recommendation to Eurostat

   3a) local RDC saves the decision at a system like "circa" to make the decisions on the institution transparent
4) Eurostat forwards recommendation to other MS (not necessary for ECHP)

   4a) for ECHP, Eurostat will decide
5) local RDC completes contracts
6) Eurostat creates datasets (countries removed who did not allow access)
7) local RDC creates user accounts
8) local RDC consults user while analysing the data
9) local RDC checks output on common guidelines first
10) peer-review process of output checking (once in a while) by other NSI or Eurostat

Variations on this solution only refer to the possibility that the data is stored at a central repository at Eurostat (which Eurostat is willing to construct). But the idea of taking the administrative burden away from Eurostat will still be realised.

Compared to the “pilot” solution, the so-called “strategic aim” includes that decision-making is also decentralised, the standardisations are more developed and approved by experience and the access will be extended for other community statistics and

national statistics. The “strategic aim” is part of a process where the MS are converging in aspects of exchanging information, standards and data. As for other community statistics, there is either currently no legal basis for Eurostat that allows submission nor the discussion on developing new legal regulations, so the remote access to the data should be considered.

4. Safety first

The intention to widen access also includes the question of standardising security aspects. Whilst the need to guarantee anonymity of the results, the legal framework includes special restrictions and conditions that have to be considered when providing access. As the study initially concentrates on implementing a network to access to the ECHP, it seems reasonable to prove whether the guidelines of Eurostat and the project partners are transferable for standardised criteria. We agree that there is a minimum legal bottom line, the specifications on allowances and restrictions have to be defined.

A safe centre is defined as a secure room in a MS/Eurostat, especially designed for researchers. It is a place where researchers can access to detailed confidential data under contractual agreements which cover the maintenance of confidentiality. The safe centre itself would consist of a secure hermetic working and data storage environment in which the confidentiality of the data for research can be ensured. Both the legal and the IT aspects of security are considered here. To ensure the security of the data, the researcher is not allowed to

- print documents
- copy data to diskettes, USB sticks, CD-ROM’s, DVDs or Zip drives
- copy data to the local hard disk,
- connect recording devices to the serial, parallel and external ports,
- connect a laptop to the network,
- use E-mail,
- make Internet connections

  • Exception: separate desktop for Internet connections. For a VPN client connection it is necessary to have internet access,
- install hardware (the PC is locked) or to take out,
- boot the PC from floppy, CD-Rom, DVD-Rom or any other media,
- access to the internal production network of the MS/Eurostat.

Within the “strategic aim”, a safe way to access data outside the environment of a RDC should be taken into account. Some MS are now using Citrix (which is more common) to set up a safe connection between the desktop (in our case it is the PC in the safe centre of the local RDC) of the researcher and a protected server of the MS (Netherlands, Italy, Denmark, Sweden). The key issue is that the microdata set remains in the controlled environment of Eurostat, while the researcher can do the analysis in the RDC. The CITRIX connection will enable the researcher to run statistical packages/programmes on the server located at Eurostat. The researcher will only see the session on his screen, which allows him to see the results on his analysis and also the microdata itself. But for this reason the researchers are in the enclosed environment of the safe centre. CITRIX will only send the pictures of the screens to the PC of the researcher, but no data is transmitted. Even copying the data from the screen to the hard disk is not possible. The MS/Eurostat has to check the output for disclosure risks and after granting the anonymity the results will be submitted.

5. Costs

Based on the experiences of already existing RDC’s it seems quite easy to calculate the costs for the hardware that allow the access to microdata either on a national server or via remote access. But the implementation of new ways of accessing community data surely leads to an increasing demand that causes an increase of staff as well. Also for NSI’s that are aiming to implement a RDC, an estimation of the occurring costs should be useful.

Thus, a cost template has been developed. The different categories are

1. staff planning rates each qualification/grade,
2. breakdown on strategy and operational costs,
3. breakdown on fixed and variable costs,
4. number of projects,
5. and IT costs.

On this basis, a cost model that gives information on the staff unit costs, the scale of operations, the operating costs per project and finally the share of costs split by categories, can be estimated.

6. Outlook

As this feasibility study ends in January 2010, the final results have not been accomplished yet. But the core of the project, the recommendation on a possible network solution, is almost available.

Compared to the objectives and the goal of the project, there are remaining items of work. Furthermore, guidelines for researchers that include useful information on the statistics should be also available. The standardisation need thus refers to a data description (which is already available at Eurostat) and a harmonised access form. The fact that there is an ongoing discussion on standardised metadata (Euro SDMX Metadata Structure (ESMS)) might have a positive benefit also.

A determining necessity for providing access to microdata on a European scale is consistency in the way each Member State checks output against disclosure of data on an individual level. A common set of guidelines is therefore needed.

The development of this set of guidelines is currently being dealt within the "guideline group on output checking", that is part of the ESSnet on Statistical Disclosure Control. The result of this guidelines project will be a set of guidelines that can be applied to all kinds of microdata (business, households, individuals).

References

European Commission (2008): *Protection of Confidential Data at Eurostat*, Luxembourg.

Eurostat (2007): *Handbook on Statistical Disclosure Control*. CENEX SDC. Luxembourg.

Gotzfried, A. & Pellegrino, M. (2008): *The Euro-SDMX Metadata Structure and Quality Indicators*, conference paper on Data Quality for International Organizations Eurostat, Luxembourg.

Hundepool, A. & De Wolf, P. (2005): *OnSite@Home: Remote Access at Statistics Netherlands*, Monographs of Official Statistics, Luxembourg.

Lenz, R., Vorgrimler, D. & Scheffler, M. (2006): *A Standard for the Release of Microdata.* RDC Germany Working Paper No. 7. Wiesbaden.