Eurostat Workshop on Remote Access

Luxembourg, 19 June 2009

Joseph Bech Building (BECH), Room Quetelet

Starting: 9.00 a.m.

Item 2

Remote access to microdata sets from official statistics for scientific purposes – Issue paper
Remote access to microdata sets from official statistics for scientific purposes

Table of Contents

1. OBJECTIVES AND BACKGROUND OF THE WORKSHOP ........................................... 3

2. REMOTE ACCESS AT NATIONAL LEVEL – EXPERIENCES OF MEMBER STATES ........................................... 5

3. REMOTE ACCESS AT THE EUROPEAN LEVEL ........................................... 5
   3.1. Objectives and expected benefits ................................................................. 5
   3.2. Possible architectures of the remote access system ....................................... 5
   3.3. Costs and constraints .................................................................................... 6
       3.3.1. Risks and costs of not proceeding ............................................................ 9
   3.4. Possible collaborative projects and strategies .............................................. 9
       3.4.1. Partners and their roles, coordination with other projects ....................... 9
       3.4.2. Required resources and workload ........................................................... 10
   3.5. Action plan .................................................................................................. 11
1. OBJECTIVES AND BACKGROUND OF THE WORKSHOP

The workshop aims at discussing different options of remote access to microdata. The following objectives are set:

- Discussing existing remote access solutions;
- Sharing knowledge between Member States and identifying the best practices;
- Enabling participants to learn from each other's experience;
- Prioritizing the ways to progress in providing remote access to microdata for researchers;
- Defining a relevant action plan and possible cooperative actions.

In 2007 two meetings of the Expert Group on Remote Access were organized. The outcomes of these discussions are enclosed in the roadmap document entitled: Integrated European wide access to microdata sets from official statistics for scientific purposes. Since then progress is to be noted in some actions foreseen in this document, notably:

- A second ESSnet on SDC has been launched in 2007 with expected results on output checking guidelines

- In 2007 a first call was opened for 7th Framework programme (INFRA-2008-1.1.2.27 Promoting European wide access to microdata sets of official statistics for research and leading to a European statistical system open to researchers). Although the project submitted has not been chosen for realisation, recommendations have been issued to improve the proposal. The second call that will be opened in July 2009 (deadline December 2009).

- An ESSnet on decentralized access to EU microdata has been launched in 2008 with the objective to study the feasibility of setting up a network of safe centres allowing researchers to access confidential EU microdata sets. The outcomes of this project will permit to evaluate different models of access to EU microdata via safe centres network in the EU. Remote access to data collections located in one safe centre from another safe centre is one of the options envisaged.

- The legal, administrative and methodological issues have been discussed at the workshop on microdata access organized together with ONS and CESSDA in December 2009 in Luxembourg.
In the roadmap document a stepwise approach to achieve objectives has been recommended. The following steps have been recommended:

1. Launch of the two pilot projects: on ECHP and data archive

2. Prioritisation of 2 access options:
   - Remote access: data can be seen on the computer screen, manual confidentiality check after submission of final output
   - Remote execution: data not seen on the computer screen, query submitted by researchers, first automatic confidentiality check of the query, final output checked automatically, manual checks on the random basis.

3. Testing of remote access to the microdata stored at the safe centres located in the Member States via national safe centres.

4. Developing and testing remote access from the workstation of the researcher.

At the moment step 1 and 3 are being realized within the ESSnet on decentralized access to EU microdata. The main issues that are still open include:

- Prioritization of remote access or remote execution?
- Can we define a cross-border access project?

The discussion at the workshop will be thus focused on the 3 outstanding issues:

1. existing remote access solutions and possible re-applications in other countries
2. remote access versus remote executions: weaknesses and strengths
3. possible collaborative projects.

This document aims at presenting different aspects of remote access and at initiating the discussion on the specific issues. It is seen as a working document that can and should be completed and or modified on the basis of the comments of the participants and discussion at the workshop. The participants of the workshop are welcomed to make any remarks to the document they find appropriate.

The document is based on the following assumptions:

- the sharing of practices regarding remote access facilities constitutes an important part of the workshop, however the document focuses on the possible setup of the cross-border project
- there is a willingness to develop cross-border access to EU microdata
- provision has been made to launch a pilot project with some countries
- there is a need to define modalities of such project
2. **REMOTE ACCESS AT NATIONAL LEVEL – EXPERIENCES OF MEMBER STATES**

Contributions from NL, UK, FR

3. **REMOTE ACCESS AT THE EUROPEAN LEVEL**

3.1. Objectives and expected benefits

The experiences of countries providing remote access facilities show that this is a preferred way of access to microdata. It offers the users the possibility to work with the non-anonymised data, on their own computers and using the most preferred tools. The advantages of such access are incontestable and well described in the available literature. It is also important to note that the remote access is complementary to other ways of access such as anonymised microdata sets or on-site access.

The remote access to EU microdata offers additional advantage of access to the microdata from different countries allowing for comparative analysis of the detailed statistics.

It is not possible to evaluate exact benefits of such a facility. However, it is evident that the remote access allows numerous research projects to be realized and therefore contributes to the increase of public good. For example in the Netherlands the remote access facility allows 200 researchers to work on around 90 projects yearly.\(^1\)

3.2. Possible architectures of the remote access system

For the testing, 2 remote access architectures are discussed here. Remote execution can be envisaged separately as well.

The difference between 2 remote access options is mainly the involvement of Eurostat. In the first option (scenario) the data are sent to Eurostat and Eurostat manages the access by researchers (on the basis of the Regulation 831/2002).

In the second scenario the data remains at the NSIs and access is managed by one of them (leading NSI). The advantage of this solution is that it can be based on the infrastructure already existing in the MS. The possible drawback is that the administrative and legal barriers might be difficult to overcome. This option can be seen as a fall back solution, in case Eurostat is not in position to open the secured environment of the Commission in the short term.\(^2\)

The schemes of 2 possible architectures are illustrated below. A more detailed analysis of possible architectures is expected as an output of the ESSnet on decentralized access to EU microdata sets.

---

\(^1\) F. Hoeve, Microdata access in the Netherlands, document presented at the ABS/OECD Workshop on remote access, 5-7 May 2009.

\(^2\) The intermediate solution, in between the 2 above mentioned "extreme" scenarios is that the access is provided in national safe centres and managed by the Regulation 831/2002 (art. 8(3). The standard procedures of access would then apply (contract etc).
The third option - remote execution – could be based on scenario 1. In this option only access to EU microdata would be possible.

3.3. Costs and constraints

We have mentioned above that the benefits of the remote access are difficult to evaluate. This evaluation is however easier for the costs. They can be divided in 2 types:

- The costs of the establishment of the remote access (execution) facility
- The costs of the maintenance of the system

The establishment of the system include: design of the system architecture and IT related costs: hardware and software.

The maintenance costs are staff costs and relate to: preparation of the data, output checking, IT support and administration (see more details in item 3.4.2). Depending on the solution chosen it may also be an equipment to be installed separately for each user (e.g. user identification tool).

The following constraints should be addressed:

1) Legal
2) Technological
3) Methodological
4) Resource based/organizational
5) Other
The different considerations related to the identified constraints are analysed in details below.

1) **Legal**

According to Article 23 of Regulation (EC) No 223/2009 on European Statistics, access to confidential microdata can be given to researchers, by Eurostat or the national authorities within their spheres of competence, at Community level and national level, respectively. At Community level, these rules are laid down in Regulation (EC) No 831/2002. At the national level, different provisions exist. The differences are focused on the type of users allowed to have access to microdata and access means. In some countries the access is restricted to researchers only, in others, some research departments of ministries can also work with microdata (under certain conditions). There are countries where access to microdata for scientific purposes is not practicable at all. In many countries foreign researchers’ access is forbidden. It is a crucial obstacle in many cases for the provision of microdata over the border.

If we consider scenario 1 the provision of microdata is regulated by Regulation (EC) No 831/2002. However the Regulation does not foresee the remote access to microdata. This will have to be taken into consideration when Regulation 831/2002 is revised.

In the second option, when Eurostat is not involved in the provision of microdata, the access is regulated by national law. The data sets which could be accessed is likewise determined by national law. In case a breach of confidentiality occurs, the rules of international private law will have to determine the national legislation applicable. International private law will for instance allow identifying where the violation of confidentiality of data occurred and according to which national legislation the possible sanctions, penal or others, should be determined.

The experience of some countries shows, however, that the most efficient sanction for a researcher is that he is prevented from access to microdata for a limited or unlimited period of time.

In case of remote execution the access is managed on the basis of the Regulation on European Statistics. As the user does not have real access to confidential data, the Regulation 831 does not apply.

2) **Technological aspects**

The technological solutions for remote access to EU microdata have to be carefully discussed. Currently, the remote access architecture is to be developed for the Euro Group Register (EGR). Although the system for EGR is designed for restricted EG Register users in NSIs only (and ESCB later on; researchers' access is not foreseen), general security features developed under this framework can be used to build the facilities for scientific purposes access as well.

For remote execution the solution that can be envisaged is LISSY system or similar. The advantage of LISSY system is that it is already installed in Eurostat and provides access to SES microdata for London School of Economics.

3) **Methodological aspects—output checking, metadata standard**

The methodological aspects of remote access to microdata refer mostly to output checking rules. These rules (guidelines) are being currently developed with an ESSnet on SDC. The results of this work are expected at the end of the year 2009. However, these guidelines will
concern mainly the manual output checking rules on queries either run by identified researchers or on anonymised data. The extension of output checking guidelines (more automation and/or wider public) may require additional research (pattern recognition, protocol to maintain anonymity in a multiple query systems, etc.).

The other methodological aspect that should be considered for remote access/execution is a metadata standard. This problem will be also discussed within an ESSnet on decentralized access to EU microdata. The outcomes of this project will allow to assess the potential metadata standards and to decide on particular solution.

4) **Resource based/organizational**

These aspects are probably the most difficult to decide on. The main question is: who is managing the remote access facility, i.e. who prepares the data, who checks the output, who consults the MS etc. Several scenarios can be envisaged according to the architecture model chosen.

If the scenario 1 is considered, the management of access could be done centrally in Eurostat:

1. The data are prepared and sent to Eurostat by the NSIs participating in the project
2. The data is listed in the Regulation (EC) No 831/2002
3. Eurostat compiles and prepares the data file with the data of MS participating in the project
4. Data are stored at the separate secure server in Eurostat (further technical and security features to be further developed)
5. The access request is dealt with by the Eurostat Unit D4 Dissemination (former B6) as in case of other means of access (anonymised microdata sets and access in the safe centre in Eurostat); two options may be envisaged:
   a) First option: the access request is dealt with as in case of anonymised microdata sets:
      - Only the admissible bodies falling into the definition laid down in the Regulation (EC) No 831/2002 can require remote access to microdata
      - The researcher apply for access in the same way as in case of anonymised microdata sets (detailed research project is required)
      - The MS (participating in the project) are consulted as in case of anonymised microdata sets
   b) Second option:
      - In the test phase only the restricted kinds of bodies are authorized to have remote access to microdata (e.g. selected universities only); the countries may propose themselves a list of "trusted researchers" or "trusted institutions" that would participate in the project
— The researcher (if authorized to have access) apply for access in the same way as in case of anonymised microdata sets (detailed research project is required)
— There is no consultation of MS, Eurostat verifies the relevance of the project.

(6) The contract is signed with researchers’ organization

(7) The access is opened for a specific period of time (defined in the contract)

(8) The final output is submitted for checking, three options are envisaged:

(a) First option: each MS checks the output separately for its own data (not always possible)

(b) Second options: Eurostat production units checks the output (very difficult as there are limited resources)

(c) Third option: output checking is outsourced to data archives or NSI or other body (restricted call for proposal)

(9) The publications produced on the basis of the remote access facilities are publicly available

In the second scenario the management work would have to be carried out by one of the NSIS participating in the project.

If we consider the remote execution scenario the workflow would be similar as described for scenario 1.

3.3.1. Risks and costs of not proceeding

The remote access facilities can be provided by other organisations, either privet or public as the needs for such access are high.

3.4. Possible collaborative projects and strategies

3.4.1. Partners and their roles, coordination with other projects

The countries participating in the meeting are welcomed to express their interest in participation in collaborative project. The possible scenarios will be further discussed and analysed in more details. Eurostat would assure general management and coordination, as there are quite a lot of different projects and initiatives on remote access ongoing.

First of all, the two ESSnets projects mentioned in the introductory part will deliver important outcomes, notably on possible architecture of the remote access system and output checking guidelines.

Secondly, the remote access project for EGR will allow to design the security requirements for Eurostat that could be further elaborated (separately from EGR project which is restricted for EGR domain managers in the ESS) for the remote access to microdata for scientific purposes.
We can also expect the development of the infrastructure for remote access in the ESS under the 7th framework programme. The new project can be submitted under the call 6: FP7-INFRASTRUCTURES-2010-1, entitled: European Social Science Data Archives and remote access to Official Statistics (INFRA-2010-1.1.3)3. The publication of this call is expected in July 2009.

An important link needs to be maintained within other initiatives such as the ABS/OECD project, assembling some EU countries, Eurostat, Australian Bureau of Statistics, Statistics New Zealand, Statistics Canada and some others aiming at sharing of knowledge, experience and developing standards like common terminology and metadata structure.

In order to continue the work of the ESSnets partners in the domain of remote access and further explore specific solutions developed within ESSnet, it is planned to launch the special ESSnet on remote access in 2010. The aim of this ESSnet will be to implement and test the solutions developed within previous projects. It will also assure the necessary continuation between current actions and future FP7 project (if chosen for realisation).

3.4.2. Required resources and workload

At Eurostat side we assume that the implementation of the project would require an important mobilisation of resources. Unit B2 Methodology and research will be responsible for overall management and coordination. The preparation of the data and maybe output checking would rely within production Units. Unit D4 Dissemination would assure the administrative side of the undertaking and IT staff the maintenance of the system.

At MS’ side there will be similar resources necessary. The actual workload related to the project will depend on the number of users. We may estimate that for each research project the following resources will need to be activated (from Eurostat and MS altogether):

- 5 man-days for administration of the project (including research project consultations)
- 2 man-days for IT related issues (setting up a secure connection, maintenance)
- 10 man-days for preparation of the data and output checking (Eurostat and MS)

This estimation can be roughly applied to all three architecture models discussed here.

The resource estimation allows to calculate the fees that would have to be charged on researchers using the facilities. It does not include the costs of the data itself as this would not be in line with Eurostat's (and some others countries') policy of the free dissemination of statistics. The fees would therefore cover (a part?) of the staff costs related to provision of the data.

3 A project under this topic should aim at a further improvement of the consistency of a European system of Social Science Data Archives and of the researcher's access to official statistics. It could also address new technologies for data collection.
3.5. **Action plan**

The following actions are foreseen:

(1) **2009**

   (a) Continuation of the 2 ESSnets according to relevant timetable (until 31 December 2009 and 31 January 2010)

   (b) Preparation of the project for FP7

   (c) First meeting of the High Level Working Group on Statistical Confidentiality (October) – overall microdata access strategy to be discussed

(2) **2010**

   (a) Launch of the ESSnet on remote access on the basis of the results of the previous projects