Working Paper No. 9 (Summary)

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (ii): New data release techniques

PROVIDING REMOTE ACCESS TO DATA: THE ACADEMIC PERSPECTIVE

Invited paper

Submitted by the London School of Economics, United Kingdom[1]

[1] Prepared by Tanvi Desai (t.desai@lse.ac.uk).

Providing Remote Access to Data: The Academic Perspective

Tanvi Desai: Data Manager, Research Laboratory, London School of Economics

I. Introduction: Trends in Data Provision

With the ability to supply data through the web, data providers are leaning more towards centralised control of data, making it increasingly rare for users to hold micro data locally.

This is a major concern for academics as, though there is now more easily available data than ever before, this data is of far less value. On-line systems necessarily limit data manipulation. These limitations mean that it is vital for academic research that data providers consider a remote access strategy rather than restricting their innovations to web based data provision.

II. Choosing a remote access system

The ideal for every remote access system should be to provide an environment that allows a user to feel as much as possible as if they are working on their own PC. The three major factors that contribute to this are.

- Speed: results of analyses must be returned as quickly as possible,
- Familiarity: it should not be necessary for researchers to learn new software and new programming techniques to access data,
- Flexibility: restrictions on data manipulation must be kept to a minimum

Many would also include access to metadata in this list. However, researchers are accustomed to having to spend time examining documentation, so this is not strictly necessary for a remote access system to qualify for the above definition. Therefore while access to high quality metadata is vital to good data use, it should not be the major focus of a remote access system.

Cost is also an important factor when choosing a system. The main areas where cost can be accrued are

- Hardware: how much space is necessary to operate the system?
- Software: is it necessary to write new software to provide the utilities required, or is it possible to use established software that not only reduces development costs, but also provides users with a familiar interface. For established software, what are the licensing costs?
- Data preparation: is it necessary to do a lot of work preparing the data for mounting in the system, causing delays to data release?

III. Security:

Prevention of unauthorised users: Almost all remote access systems have a password authorisation system. This is the most basic form of security, and has to be supported by a signed license agreement.

It is also possible to have a system of “trusted” computers, whereby IP addresses are registered with the access system, and only requests originating from known IP addresses are processed. This is impractical as few universities have the ability to allow researchers dedicated IP addresses.

An effective measure to support password authorisation is to restrict the delivery of output. For instance, LISSY will only return output to the user’s registered email address.

Prevention of sensitive analyses: There are a number of ways of going about this, some more practical than others.

Checking output: This is very impractical for a number of reasons. Primarily because it is very time consuming and prevents output from being returned promptly. There is also the problem of finding staff that are qualified to check output, and the difficulty of understanding other people's programs.

Blocking at source: This is the most effective method, as it can prevent users from running analyses that might expose confidential data.

Deposit papers: A commonly stated condition of data access, is that papers produced using the data are deposited with the providers. Very few data providers manage to put this into practice successfully. The only data provider I have come across who makes this work is the Luxembourg Income Study. This is primarily due to their close knit user community who are made to feel a part of the project through high levels of support and regular workshops and conferences. In addition LIS is able to enforce this policy, as the board members are all senior academics in the field who become aware of any breaches.

The best form of security is a good relationship with your users, if they feel they have a stake rather than being in the supplicant position they are more likely to act responsibly.

IV. Supporting a remote access system

High-quality support is crucial to any successful remote access system. A close relationship between users and providers not only ensures that any investment in developing the system is relevant, but also helps immeasurably with security, as users feel like they have a responsibility to the project. A good line of communication between the data collectors and distributors is also a vital component of support, in order to ensure data quality and fast response to user queries.

One of the major drawbacks of a remote access system is that researchers are not able to create their own subsets of derived variables. This is especially important for cross-national data where recoding is almost always necessary for comparability. Therefore support for remote access systems must take into account the necessity of providing users with space to store constructed subsets or a system for adding derived variables to the core dataset (or preferably both). Decisions on derived and comparable variables can only be made in conjunction with the user community; this is another reason why close links between the data providers and users is vital.

V. Conclusion

To conclude, there are two main points I would like to concentrate on.

First, a remote access system should resemble a local system as closely as possible.

Second, close links between all the stakeholders (users, collectors, distributors) is vital. Managers of remote systems should not see themselves purely as data providers, but as members of the research community. The sense of community can be encouraged through steering committees, user groups, seminar series and workshops. Good communication between all parties has a positive impact on effective allocation of resources, research quality, levels of use, and data security.