Working Paper No. 9
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

Introduction: Trends in Data Provision

As technological advances in hardware, software, data documentation and the Web make access to, and analysis of micro data more and more practical and desirable, researchers and students now expect a wide range of micro data to support their research. In particular, the growing interest in cross-national comparative research with increasing demand for international data, means that bodies such as Eurostat have a vital role to play in the future of data provision.

Recently we have seen trends in data provision coming full circle. Ten years ago, the only practical way to access data was via a main frame, often via remote access to the data providers. This was necessary as very few data users had the processing power to deal with large datasets on their PCs.

With advances in desktop computing, researchers became able to manipulate and analyse large micro data sets locally, and the trend was to supply CDRoms containing the data. This is the most practical format for academics as it gives the most control over data manipulation. However, with the advent of the Web we now have a move towards on-line data provision, and following from this, more centralised control of data and a renewed interest in remote access. Data providers are becoming reluctant to allow researchers to hold copies of micro datasets. This is a major concern from an academic point of view. There are now many more easily accessible data sources available through the Web, but the value of this data is far lower. This is because providing access on-line necessarily limits what you can do with the data in terms of analyses and exploration.

This is a major reason why it is vital from an academic point of view that data providers consider a remote access strategy rather than restricting their innovations to the Web.

Choosing a remote access system

1. **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, while access to high quality metadata is vital to good data use, researchers are still accustomed to having to spend time examining documentation, so metadata provision should not be the major focus of a remote access system. In fact, it is often more practical to present metadata through a website.

Speed: The speed with which output is returned is obviously very important when trying to recreate the feeling of working locally. Delays are a particular disadvantage when introducing new users. If the relatively straightforward exploratory programs that are necessary to get to know the data take a long time to run researchers are likely to be discouraged from using the system.

Familiarity: If researchers have to invest a lot of time learning new software, they are less likely to use the data. With familiar software they can start working as soon as they gain access, thus strengthening the impression that the data is local. Using established software also reduces costs as it is means that there is no need to develop new software.

Flexibility: Of all the three factors mentioned above flexibility is the most important. The balance between securing the data and allowing meaningful analyses is delicate. If the security restrictions mean that the data cannot be examined properly or the analyses allowed are too few to provide meaningful results, then the data become useless. One of the primary reasons for restricting analyses when providing access to data is to prevent the identification of individual cases, either by repeated refining of a selection, or by reporting cells with too few observations. From an academic point of view these restrictions are almost always unnecessary. Academics have neither the inclination nor the time to identify individuals. Cells with a low number of observations are not statistically significant, and to report them would only invite the censure of colleagues for poor research. In addition, since all users will have signed a legal document in the form of a license agreement stating that they will not perform any of these restricted actions, they are unlikely to endanger their data access, and thus their research project by breaking this agreement for no academic return.

2. Graphics:

The production of graphics is a common problem faced by users of remote access systems, as very few have the provision to return graphics (for instance part of the LISSY[*] security prevents any file that is not text format from entering the system, this means that graphics cannot be returned). LIS gets around this, by allowing users to submit jobs, which are then run locally by support staff, and returned as attachments to email. There are some remote access packages available commercially that produce beautiful graphics, but these tend to be aimed towards the private sector, and users whose priority is producing high quality graphics from relatively simple analyses to form part of commercial presentations. Therefore the range of analysis techniques can be too limited for academic use (this software also tends to be very very expensive).

3. Cost:

Cost is also an important factor when choosing a system. The main areas where cost need to be considered are

- Hardware: how much space is needed for storage, and to allow users to run jobs remotely? Also, would a PC or UNIX based operating system be most suitable to the software?
- 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 on the system, causing delays to data release and using up staffing resources?

Security

1. Prevention of unauthorised users:

The most common form of security for remote access systems is a password authorisation system supported by a signed license agreement.

[*] for an outline of the PiEP and LIS projects and the LISSY software mentioned in this paper, please see the APPENDIX.

License agreements are usually between the individual user and the data provider, however some data providers insist on an agreement with the institution where the data will be held. This is a serious concern, as a corollary of this is that only employees of the institution can have access to the data, this marginalises students as they are rarely employed by the institution at which they are studying. Considering the shortage of European researchers with high level methodological skills, any system that prevents students, in particular PhD students, from accessing international data is going to contribute to the problem.

An additional security measure that is often considered is a system of "trusted" computers, whereby IP addresses are registered with the access system, and only requests originating from known IP addresses are processed. As it is impossible for every researcher to have a unique IP address, the only way to implement this is to have a "trusted" server. Remote Desktop Access software would allow a number of researchers to link to this server from their desktops, thus providing them with a "virtual trusted" computer. However, this method needs a dedicated server, an expense which individual users and even many institutions would be unable to manage.

Another effective measure to support password authorisation is to restrict the delivery of output. For instance, the LISSY system will only return output to the user's registered email address. This means that no unauthorised user can access output, as even if they have managed to get hold of a username and password, the results are still returned to the registered user.

Hacking: Another concern is unauthorised users hacking the data server or "sniffing" the data in transfer. To prevent "sniffing" it is possible to encrypt data as it is being transferred. The LISSY system does not use encryption as no micro data nor the results of any confidential analyses are transmitted. Therefore it was decided that encryption would only slow the system down unnecessarily and speed was one of our priorities.

As is standard practice, data files should always be stored with non-descriptive names providing additional protection against anyone who gains direct access to the data server due to a network security breach. Users are given aliases with which to access the data.

2. Prevention of confidential analyses:

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

Checking output: Checking the output generated before it is returned to the user 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 who are qualified to check output. The person who is responsible for checking output has to be a statistician of equal skill to the most sophisticated user of the system. In addition they must have an understanding of the data users' fields of research so that they can see how the results will be used and whether this will be sensitive, and a good knowledge of the security needs of the individual countries since confidentiality concerns vary across nations. What chance is there that someone who has this range of knowledge will be content just checking other users output?

There is also the difficulty of understanding the structure of other people's programs. Therefore, if it was decided that checking output was necessary, it would be advisable to provide users with a template to encourage them to submit annotated jobs in a standard format.

Blocking at source: This is the most effective method we have found for preventing sensitive analyses. In this method a string search is run on the text of all programs submitted. If any strings are identified that might represent confidential analyses the program is not sent to the data server, but returned to the user with an error message. As any combination of strings can be specified, this system offers a lot of flexibility. It allows data providers to define and block problems particular to their national data.

Different blocks can also be set depending on the user. Thus, we have a method that can tailor security to the individual case, and block sensitive analyses before they gain access to the data.

3. 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 close links with their user community and the involvement of senior academics who know the publications in their field and soon become aware of any rare breach. I will discuss ways of maintaining these links below.

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.

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, as mentioned above, also helps immeasurably with security, as users feel that 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.

1. Human Resources

Dedicated personnel are necessary to efficiently support a remote access system. Their assistance is vital in

Data preparation: Support staff are needed to prepare data before it is mounted on the system. Ideally the minimum of data preparation should be necessary, partly to preserve as much of the original structure of the data as possible, but also to minimise use of staff resources. However, it is almost always necessary to do some preparation, even if it is just naming and labelling variables.

Technical support: IT staff are needed to maintain the systems, hardware and software, and ensure that downtime is kept to a minimum.

Research support: It is exceedingly rare for a dataset to be perfect. Often inaccuracies are not noticed until the data is being used, when analysis produces unexpected or suspicious results. There are also likely to be methodological queries that arise about the exact definition of variables, how value ranges are selected etc. Support staff can only answer these questions if they are familiar with the data. They must also be in close contact with representatives of the data providers, so that they can quickly obtain answers to queries that cannot be solved without referring to the original dataset.

If researchers are to work effectively, it is vital that any queries are answered promptly. Again, if a user's questions are answered promptly and intelligently by people who seem interested in their research they become far more well disposed to the data providers and feel more of a responsibility to safeguard the data. Long delays cause resentment and encourage researchers to find alternative data sources.

Derived Variables: 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.

2. Networking and Partnerships

Mail Lists: All active users should be included in a mail list so that they can be informed of any changes to the data, technical problems with the system and other issues that arise.

It is also a good idea to have a user group mail list where researchers can discuss issues amongst themselves, as researchers are often best equipped to answer each others queries. This also reduces the workload on support staff, as many researchers will query their colleagues before contacting support staff.

Website: A good website is vital to attract new users, reduce the pressure on support staff, facilitate data use and enhance security.

These days many users become aware of new data sources through the web. A website acts as an introduction to the service attracting new users, and allowing them to decide whether the data is likely to be useful. Documentation, high quality metadata and answers to frequently asked questions can be mounted on the website reducing the number of individual questions support staff have to deal with. Access to metadata such as details of variable coverage and the question texts can save a researcher time by minimising the need to examine the data interactively before starting analysis, and good methodological information minimises errors producing a higher standard of data use. Sample exercises and programs can also be provided as training aids for new users.

Another very important function of a website is to provide information on registration procedures. A straightforward, quickly implemented registration procedure is vital as it makes it significantly less likely that researchers will try to find illegal methods of gaining access to the data.

Conferences: One of the key ways in which the Luxembourg Income Study develops and maintains such a good relationship with their users, apart from the superb level of support offered, is through a regular program of workshops and conferences.

LIS run an annual workshop for new users. Here they are introduced to the data and access system in a supported environment, where problems can be discussed with the technical team. Experienced users of the data are also invited to give seminars on their work and to guide new users in the possibilities offered by the dataset. This is not only a great way of forming relationships between data users and providers, but is also a way to reduce the problems experienced by new users, and thus the time needed to support them remotely.

In addition to the annual workshop, LIS also organise a conference once every two years, recent and future topics include Child Poverty, and Immigration. Conferences have an important part to play in developing research networks and encouraging good data use. They also provide support staff with an opportunity to gain a more in depth knowledge of what the data is being used for.

Steering Committee: An active steering committee made up of data providers and data users is vital to provide a forum for discussion of topics such as security, data quality and system development. In addition, individual researchers often don't have the breadth of international knowledge necessary to create comparable cross-national variables, therefore a steering committee is invaluable when taking decisions on derived variables. Members of the steering committee will not only be aware of which derived variables it would be useful to add to a dataset, but will also be able to decide on the methodology for constructing these variables and can pool their national expertise to decide how to derive the variable accurately for each country.

Conclusion

The development of remote access systems is vital to the future of academic research if data providers continue to become increasingly reluctant to allow users to hold data locally.

Working with a remote access system should resemble working on your local PC as closely as possible. This is affected by the speed at which the output is returned, the familiarity of the statistical software, and the restrictions placed on analyses.

Security measures that are too severe render the data useless.

Blocking confidential analyses before the program is delivered to the data server is an effective way of preventing confidential data being released.

It is crucial that the system administrators provide a high quality support network to enable users to make effective use of the data. A well designed website makes a vital contribution to remote access support.

Finally, close links between all the stakeholders (users, collectors, distributors) are 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.

APPENDIX:

PiEP: The Pay Inequalities and Economic Performance Project is conducted by an international team of academic researchers with support from the European Commission, and in close collaboration with Eurostat and the national statistical institutes.

The project makes use of the 1995 Structure of Earnings Survey microdata for 6 countries (Belgium, Denmark, Ireland, Italy, Spain, UK). This data, which is held at Eurostat in Luxembourg, is accessed via a remote system managed by the London School of Economics in the UK. The access system is an adaptation of the LISSY software commonly referred to as PiEP-LISSY. http://cep.lse.ac.uk/piep/

Tanvi Desai is the Data Manager and System Administrator to the PiEP project, as well as being Data Manager for the LSE Research Laboratory. http://rlab.lse.ac.uk/

LIS: The Luxembourg Income Study provides remote access to a collection of household income surveys for 25 countries on 4 continents. The LISSY remote access software was originally developed for this project, and has been running successfully for 20 years.
http://www.lisproject.org/

LISSY: The LISSY system is a remote data access system developed by HAL Consulting. It provides secure access to micro data through email, allowing users to send programs in any of three commonly used statistical software packages (SPSS, SAS, STATA). The PiEP version of LISSY has the added ability to block any strings or combination of strings that might provide access to confidential information.