Working Paper No. 7
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

LISSY Remote Access System

Invited paper

Submitted by Statistics Netherlands and the Luxembourg Income Study[1]

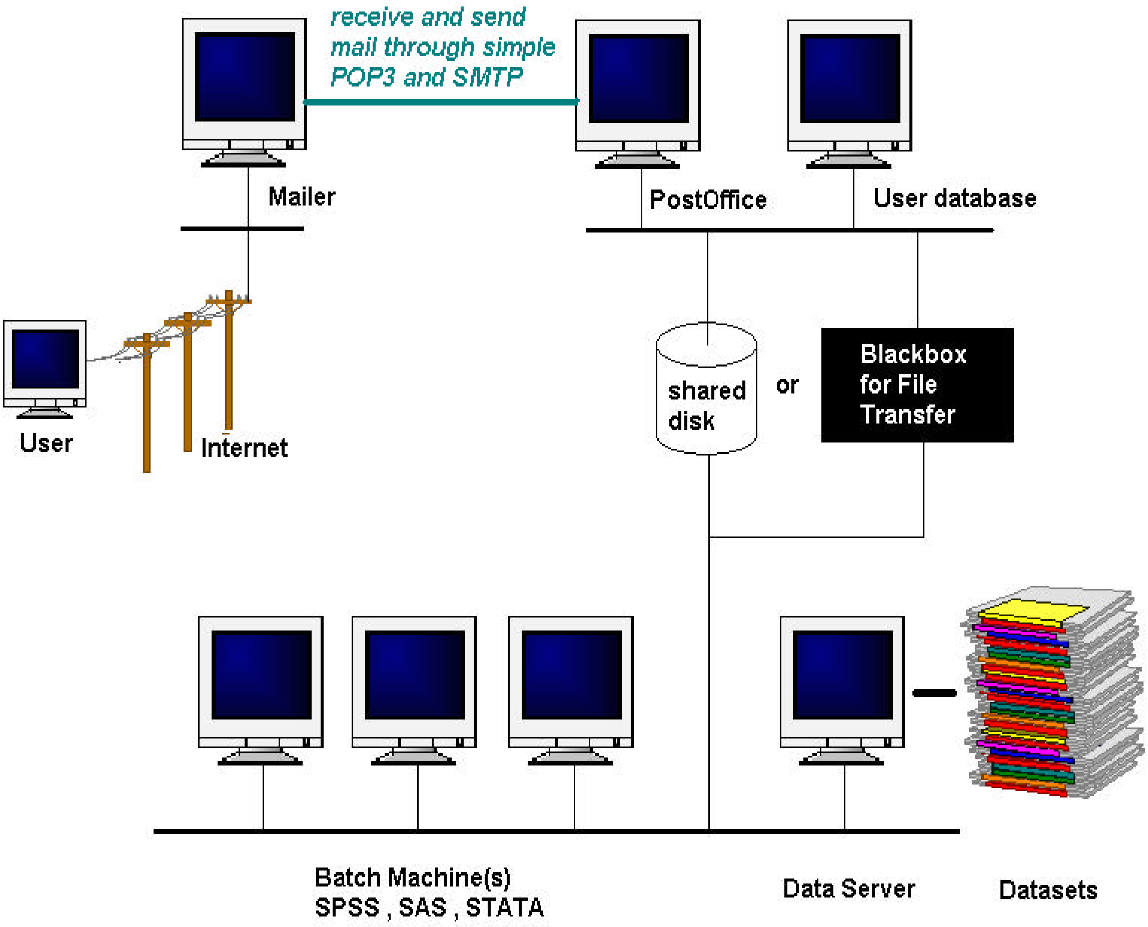
[1] Prepared by John Coder (Statistics Netherlands) and Marc Cigrang, Luxembourg Income Study (LIS) (marc.cigrang@pt.lu).

LISSY Remote Access System
By John Coder and Marc Cigrang

Lissy has been developed to provide easy remote access to statistical datasets in the form of request submitted through simple email messages. At the moment Lissy provides users with the choice of SAS, SPSS, and STATA as statistical packages. The email requests contain the "code" created by the user for the specific statistical package used.

The processing system automatically accepts these requests, processes them, and then returns the results in the form of email to the person making the request. This is a fully automated system capable of running 24 hours a day, 7 days a week. It has been programmed in the JAVA programming language. Therefor it can be run on almost any recent operating platform.

The LISSY operating system consists of a series of software components connected through on one or more networks. These components work together to receive, process, and return statistical requests.



The LISSY system components can be installed on a minimum of one computer or multiple machines linked across a network depending on the needs and the type of a project. An ideal configuration would consist of 4 main system machines and one or more job execution machines according to the numbers of jobs to be processed. These computers communicate with each other using or shared system resources like standard email, shared disks or some blackbox mechanism for forwarding files in order to provide an automated processing system.

Mail Server

The mail server is the connection to the outer world, the internet and is the only part of the system that is visible from outside. It receives all the emails addressed to a precise mail account specified as the repository for a precise project. Users wishing to access the system must address their emails containing the program code needed to access the data to that precise mail account. Their mail also must contain all necessary information needed to clearly identify a user. We will see later how such an email request looks like.

As is normal, mail arriving on the mail server waits for the user, in this case the "system job control or PostOffice" computer to receive it.

Any Pop3/SMTP compliant mail server can be used.

For security reasons the mail server machine should not be on the same logical network then the user database or the data itself as it is visible from the outside world, meaning internet.

The database systems employed in the LISSY system are Oracle or DB2 or any JDBC compliant database engine.

A certain number of tables are used to manage the system. One of them, the USER table contains all information provided by when someone registers to access the database.

Another table, the job table, gives exact information about the jobs.

Both these tables among some others allow to get very precise statistics about usage and performance of the system.

System Job Control or PostOffice

The heart of the access system is the system job control component or PostOffice. It plays the role of the “traffic cop” if you will which manages the entire LIS access mechanism.

At the speed of five seconds interval the PostOffice accomplishes the following tasks :

- It retrieves the email requests from the mail server
- It prepares these request for processing by checking for all security issues like clearly identifying the user, checking the use of not allowed statistical commands, check for the usage of sequences of commands or variables or any other combinations not allowed
- It distributes the requests to the batch processor computers
- It returns the statistical results to the proper user email addresses
- Finally It maintains critical databases needed for the overall operation.

The PostOffice does not need to know anything about the batch processor machines as these operate in a semi-independent way. A batch processing computer can be added and started without any notification to the PostOffice. The PostOffice also does not require any knowledge of the location of the microdata since it does not use the microdata.

Step one in the sequence of the PostOffice routine is a query of the mail server to determine if any requests have been received. Once received, the request is scanned for the mandatory information located at the beginning of the request. The mandatory information includes the requestor’s userid, the requestor’s password, the statistical package that is being used and finally the project he wants to work on. The statistical package can be one of three, SAS, SPSS, or STATA.

```
*USER = marc;
*PASSWORD = mypassword;
*PACKAGE = stata;
*PROJECT = projectname

 stata code follows here ……….
```

The user identification and password are checked against the list of registered users.

If this is ok, then a check is made to determine if this user’s access to the database is still active.

Following the access check, the syntax of the request is examined to determine any violations regarding the types of statistical procedures used and sequences or combinations of words.

Any requests that appear to be violating the rules established for a precis project are or set aside in the review area where they are examined manually by the project staff or send back to the email address contained in the header of the mail message with a notification of the error to the sender.

Upon examination, the project staff can then permit the request to continue or contact the sender and discuss the problem.

A copy of the request is saved to the archive so that a complete list of all jobs ever submitted is maintained. This archive not only provides a backup in case a request has been lost but also provides evidence of misuse of the database.

Finally the file containing the request is moved to an area for further processing by the Batch machines. Note that this area need not be on the PostOffice computer but anywhere on the local or remote network where the batch processing computers can have access.

Requests are processed by the batch processing computers and the results are packaged and returned to the common area accessible to the PostOffice. The PostOffice queries that directory at specified intervals to see if any result files are waiting to be returned to users. If it finds a request file waiting to be sent back to a user, it initiates an examination of the file size and contents. If any security violations are found, the file is moved directly to the review area where the contents are examined manually by the staff.

Output that is judged to be acceptable under the project rules is returned to the sender automatically following the check as an email. The results are sent back using the email address stored in the user database and not to the address of the sender of the request which could be different.

For all these operations the database of job submissions is updated and can be used to provide numerous statistics concerning system usage.

Data Server

The data server computer is a very simply computer that acts as the repository for all of the datasets available for access. Separate directories are maintained for each of the three file formats required for the three statistical packages available. In addition, the SAS data directory includes one SAS format file for each dataset.

The centralization of datasets on the server assures that all requests are accessing the same data.

The directories containing these data files are write protected so that a user could not accidentally change them as part of their job submission.

For obvious security reasons there should definitely be no direct link between this machine and the mail server.

Batch Processors

Each time a user request has been accepted, a batch processor computer ultimately fulfills that request by executing the program code sent by the user. The batch processors are computers equipped with one or more of the statistical packages offered by the project. Each batch processor acts independently without any knowledge of the number, function, or status of other batch processors that may or may not exist on the system. This independence permits the addition or deletion of batch processors at any time. This allows to balance workload by adding or deleting batch processors at any time in order to cope with the workload.

At 5-second intervals, the batch program looks into the common area where requests are placed by the PostOffice. If it finds a job, it gets it for execution and clears it out of the queue. After execution, it moves the results to the output queue where the PostOffice can find it for further processing. Then again, after five seconds it checks for further jobs to be executed.