

Coder, John F.

**Working Paper**

## User Services and Data Management in the Luxembourg Income Study

LIS Working Paper Series, No. 247

**Provided in Cooperation with:**

Luxembourg Income Study (LIS)

*Suggested Citation:* Coder, John F. (2000) : User Services and Data Management in the Luxembourg Income Study, LIS Working Paper Series, No. 247, Luxembourg Income Study (LIS), Luxembourg

This Version is available at:

<https://hdl.handle.net/10419/160919>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

**Luxembourg Income Study Working Paper No. 247**

**USER SERVICES AND DATA MANAGEMENT  
IN THE LUXEMBOURG INCOME STUDY**

**John Coder**

**September 2000**

# **User Services and Data Management in the Luxembourg Income Study**

## **I. Introduction**

A major goal of the Luxembourg Income Study (LIS) has been to provide researchers with access to information about the social and economic characteristics of households and families for countries throughout the world. Since this information is in the form of data files containing actual observations from household surveys, the researcher does not need to rely on statistical summaries that others have previously created. They can create tabulations and statistical analyses from the survey data that precisely fit their needs and update them easily when new data become available. Further, a system to access these data has been developed which permits researchers to generate statistics from remote locations using email as a mechanism to transmit requests and forward results. Access to these data is, in fact, restricted to this system of remote access because much of the data available through the project has been provided by the member countries with the restriction that they not be redistributed or otherwise copied.

This introduction describes the evolution of the techniques and methods used to maintain the databases of household microdata and to provide remote access services to users throughout the world. It covers a period that began in 1988 when the first successful remote access system was put in place. While technology has changed dramatically since then, the basic principals that govern use and access to the data have remained the same. The discussion begins with an overview of the system. Topics covered include:

- management of databases
- statistical tools
- formulation of access requests
- registration and management of users
- details of the automated request processing system
- data available to users and how it is managed and maintained, the workload
- plans for LIS in the 21<sup>st</sup> century.

The final chapter presents results of a recent survey of LIS users. This survey was designed to provide feedback on how the system is being used and what areas may need improvement.

## **II. Evolution of the Access System**

More than 10 years have passed since the first LIS system was launched for providing remote access to the databases maintained by the project. The design of the first access system was built around the emergence of email (in those days known as EARN/BITNET). Email provided a fast and easy mechanism for any individual at any location to send a request for services to the project (although the performance of email in its early stages was not always fast and reliable). The statistical tool chosen at that time for generating tabulations, statistics, and other analyses based on the survey data was the SPSS (Statistical Package for the Social Sciences) which, at that time, was available for use on mainframe computers. SPSS was chosen because it was widely used, especially in Europe, and it was provided without cost by the Centre Informatique de l'Etat of Luxembourg whose computing resources and email connection provided both of these essential resources.

Requests for statistics were made by embedding the SPSS code needed to generate the statistics within the body of the email message. The system required that the SPSS code include several additional comment lines (using SPSS comment line syntax) that provided the identification of the requestor, a password, and the databases on which the code should be executed. The email messages were received at the LIS headquarters by manually checking and receiving the email several times a day. Each email was examined and the SPSS code extracted, again manually. This code was then executed per the user request. The results were then repackaged within the body of an email and returned to the requestor. Staff intervention was required at every step in the earliest versions of the system in order to process each request. It was not unusual to require several days to complete the request and return the results to the user. Since manual intervention was needed, access was not available on weekends and holidays.

In the early days, most of the requests for access came directly from those persons involved closely with the project. That included the LIS staff in Luxembourg and persons in member countries who were instrumental in obtaining and providing the household survey data needed to make the project a success. Data from 10 countries were available within one year of the initialization of service. These included:

**Table 1. Initial Household Survey Data in LIS Database circa 1988**

Country	Reference Year for Income
Australia	1981
Germany	1984
Israel	1979
Norway	1979
Sweden	1981
Canada	1981
Netherlands	1983
Switzerland	1982
United Kingdom	1979
United States	1979

These early data were household or family-based. That is, the income information was based on the sum of annual income received by all members of the household.

Demographic variables were also summarized at the household level. For example, these early data sets included the age of the head and spouse if present, age of youngest child, number of household members, number of children under age 18, etc. There was no separate data available for individuals within the household, i.e. a person-level database.

While the computing resources available were substantial for that point in time, they were limited to the point that subsampling of sample cases for the large surveys was required. For example, the large sample sizes for the United States and Canada presented

unacceptable loads on both the disk storage and on execution times. In order to reduce the burden imposed by these larger data sets, only a subsample of the cases were taken.

By the early part of 1989 it was clear that the manual method for receiving requests and returning results required more human resources than were available to the project. Additional programming resources were made available to attempt development of a more automated method for handling the increasing number of statistical requests. This effort resulted in the implementation of a second mainframe-based processing system using the IBM's REXX command language. In this system the survey data were stored in IBM's DB2 relational database system. Each time a request was processed the variables needed for the request were unloaded from the database to create an ASCII data file. This file was then transformed into an SPSS system file, and the execution of the request was made. Also, for the first time, information on the registered users of the system was stored in a relational database and checks of user identification and passwords were made automatically.

More importantly, the email requests were received in an automated rather than a manual way. Once received, the contents are analyzed by the program, the requests were processed, and the results mailed back with very little human intervention. This system was designed to meet our goal of being operational 24 hours a day. This was far from true, however, as scheduled maintenance, backups, and other problems on the mainframe computer frequently interrupted service.

Revisions to the content and detail of the data from the surveys were also undertaken in 1989. It decided that, where possible, the data for new, incoming data sets should include a person-level dimension. This meant that the amount of data to be processed and stored would increase dramatically as the new data sets arrived.

We had not gone far into the decade of the 1990's before we began to look for yet other ways of improving service to the growing number of users. Our inability to control the computing resources was a major concern as we were totally dependent on a large

government organization where our project was of little importance. Also of concern was the now need for more storage and faster turnaround. Efforts to add the data for new countries had paid off and pressure to provide updates of the original data sets grew. By the end of 1990, the number of country data sets had increased to 16. The number of requests for statistics was growing as well as more and more users learned about the access through the workshops and the number of persons using email grew. In addition, a significant number of requests for data access using the SAS statistical system were being made by the expanding user community, a service that was not available from the provider of computing resources.

The onset of these events was accompanied by the introduction of faster and faster personal computers and the expansion of both the SPSS and SAS packages to the personal computing environment. For the first time, we were presented with a viable alternative to a mainframe system that was becoming increasingly difficult to use.

A decision was made to migrate the entire access services and data to an environment based on personal computers toward the end of 1990. The IBM OS/2 operating system was chosen as it was the first true multitasking system available running on a personal computer. Both SPSS and SAS were available under this operating system so that this choice permitted the addition of SAS as a choice for our users. The migration to the personal computing environment also required that a network be established in order to provide enough resources to handle the workload and to provide access using both SPSS and SAS.

Access to email through EARN/BITNET continued to be a problem. A dedicated telephone line was established to provide communication to the Centre Informatique de l'Etat, which remained our only connection to the EARN/BITNET. This connection was made via a dialup telephone modem. The modem connection, as they were at that time, was difficult to maintain and frequent problems resulted in continued untimely disruptions of service.

Nonetheless, the shift to the personal computer environment was a great success. The addition of SAS services was key to bringing in a large number of new users who were not familiar with SPSS. In addition, we were able to actually reduce the time required to process requests and return statistical output.

There were very few significant changes that occurred during the period between 1991 and 1995 in the system for processing user requests. The personal computing environment was stable and working well. It was easy and relatively inexpensive to add storage as the number of data sets expanded during this period. The speed of access improved as the stability of the connection on the dialup telephone line improved.

The end of this quiet period in terms of system operations came in February 1996 when the offices of the LIS project were moved to new location. A reevaluation of the system at that time resulted in a decision to migrate from the OS/2 operating system to Windows NT. While the OS/2 system performed very well, SPSS and SAS development on that platform was halted due to the increased popularity of Windows NT. At this same moment, a direct email connection was established which removed the need for the telephone connection to the Centre Informatique. For the first time since the project began it was no longer dependent in any way with the government of Luxembourg's computing resources.

A number of revisions to the system have been implemented since the major overhaul undertaken in 1996 in the move to Windows NT. First, the computing infrastructure has been redesigned to permit additional computers to be added to the network to process user requests. In this design, a new powerful computer can be setup to accept user requests within one hour. More details on this design are covered later. Second, a third statistical package, STATA, was installed to provide users with another option for data analysis. Third, a relational database table was added to provide details on each and every job that was processed. This database provides information concerning the workload overall, by user, by statistical package, etc. Fourth, a website was established



for the project which provides extensive technical documentation regarding the data and access requirements.

### III. Computing Details of the Access System

As outlined above, the LIS provides remote access to its data sets using email to transmit requests for statistical summaries and other analytical output. This system is currently employing a total of seven personal computers linked across a Windows NT network. These computers communicate with each other and use shared system resource (system disks, etc) to provide an automated processing system running 24 hours a day, 7 days per week. The computers are listed by function in Table 2 below.

**Table 2. System Resources and Functions**

<b>Computer</b>	<b>Functions Provided</b>
Mail Server	Receives user requests at specified email address as email
System Job Control	Retrieves email requests, prepares request for processing, sends requests to batch processors, returns statistical results, maintains critical databases, houses critical system parameters, houses the batch processor executable, houses computational routines
Data Server	Houses all microdata files as system files applicable to each statistical package
Batch Processor 1	Processes statistical requests and returns output to System Post Office
Batch Processor 2	Processes statistical requests and returns output to System Post Office
Batch Processor 3	Processes statistical requests and returns output to System Post Office
Web Server	Provides technical documentation

## **Mail Server**

The mail server is the connection to the internet. It receives all email addressed to the mail account specified as the repository for requests to access the system. The current email address used is [postbox@lissy.ceps.lu](mailto:postbox@lissy.ceps.lu). Users wishing to access the system must address their emails containing the program code needed to access the data. As is normal, mail arriving on the mail server waits for the user, in this case the “system job control” computer to open and read it. The mail server program is InterMail’s “Post.Office” mail server from Software.com

## **System Job Control**

The heart of the access system is the system job control computer. It is a window’s based program written in the C++ programming language. It is the “traffic cop” if you will which manages the entire LIS access mechanism. Functions include retrieval of the email requests, preparation of the request for processing, distributing access requests to the batch processor computers, returning statistical results to the proper user email addresses, maintaining critical databases, and housing critical system parameters.

Once this program is started, it repeats a series of tasks at five-second intervals. Startup of the program begins with the establishment of password protected linkages to the user and job databases. User identification codes, passwords, and email addresses are read from the database and stored in memory to speed processing. Startup also includes reading of an initialization file that contains information concerning the location of critical resources and files on the network. These include:

- the directory paths for the locations where the files containing the access requests and statistical results are to be stored
- the directory path defining the location to which problem requests are diverted
- the IP (network) address and email address of the mail server
- the directory path indicating where the job archive is located
- parameters governing security checks
- file name of a file used to indicate the system status to the batch processors (i.e. open or closed)
- a series of text messages returned to the user in an email message when an error is found in the submissions required syntax

Note that the system control computer 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 system control computer. The system control also does not require any knowledge of the location of the microdata since it does not use the microdata.

**Receiving a Request** - Step one in the sequence of the system control computers routine is a query of the mail server to determine if any requests have been received. This query is handled programmatically, that is, a direct connection (socket) is established with the server and the requests are downloaded and stored in a file accessible to the system control program. Once received, the request is scanned for the mandatory information located at the beginning of the request. The mandatory information includes the requestor's user identification code, made up of a series of alphanumeric characters chosen by user or the LIS staff (usually the initials of the name or part of the name), the requestor's password (up to 8 characters), and the statistical package that is being used. The statistical package can be one of three, SAS, SPSS, or STATA. The syntax for the mandatory information for a SAS request would be as follows:

```
*USER = lisuser;  
*PASSWORD = mypass;  
*PACKAGE = sas;  
  
sas code follows here .....
```

The user identification and password are checked against the list of registered users. If the identification and password pair matches a registered user the request is processed further. If a user corresponding to the identification/password can not be found, an email is prepared noting that an error has occurred in the request. Since an error in the required syntax prevents retrieval of the appropriate email address from the LIS user database (information provided by that user at registration), the email address contained in the header of the mail message is used to return notification of the error to the sender.

After establishing that the request has been received from a registered user, a check is made to determine if this user's access to the database is still active. This check reflects the option available to the LIS staff to disable or prevent access to the database for a specific registered user. To deny access the LIS staff must enter the user database and change the access code. Denial of access would only occur if it had been determined that that particular user had failed to comply with the LIS rules of operation.

The database also contains a mechanism for assigning priority to requests once they have been checked in and are waiting for processing by the batch processing computers. The priority code for a user is stored in the database and can be changed from the default to something higher. If a higher priority code has been set, those requests with the highest priority will be processed first.

Following the access check, the syntax of the request is examined to determine what types of statistical procedures are being used. Any requests that appear to be printing or copying individual records are set aside in the review area where they are examined by the LIS staff. Upon examination, the LIS staff will then permit the request to continue or contact the sender and discuss the problem. The system control program provides the graphical interfaces needed to perform these checking operations in an efficient, "point and click" manner.

A special check is also required to determine if the request specifies use of any data for the United Kingdom (UK). If references to the UK data are found, the user database is queried to determine if the user has signed to UK use agreement. If not, the request is returned to the requestor with a message indicating that a UK use agreement must be completed before these data can be used.

After all checks have been completed the request is nearly ready for submission. Three tasks have yet to be completed. First, the file containing the statistical code must be uniquely named as all mail content retrieved by the mail server is placed in a file with a standard check-in naming convention. The file naming convention used to assign file

names is quite important as the name provides a means to control actions taken on the file. The file name assigned is composed of separate elements related to the statistical package, the job priority, the job number, and the user's identification code. These elements can be examined to determine details about the request without actually opening the file and looking at its contents. This can save time during time of peak loads on the system. Next 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 should such activity be uncovered. Finally the file containing the request is moved to a directory on the system job control computer that is known to all elements of the system as the queue where all requests are stored pending processing. The location of this directory is specified in the initialization files for both the system control and batch processing machines. Note that directory need not be on the system control computer but anywhere on the network where the batch processing computers can have access.

The master job number is incremented when a job is moved to the job queue for processing. This number is maintained in a database table along with a number of other items such as the number of users (incremented when a new user is entered into the system). In addition, the job is entered into the job database by number and date and time of entry.

**Recent Email Content Problems** - The proliferation of mail client software and free email accounts have caused some problems recently in the checkin of requests. The problem arises first because the content of email can now be in formats other than simple text. For example, the content may be in rich text format, something that our current checkin system cannot identify and correctly process. Second, those persons using free email accounts are subject to the insertion of advertisements with the body of the email by the vendor providing the free email service. The sender of the email has no control over the insertion and this insertion, depending on its location, can result in errors in the submission or failure at checkin.

**Returning Statistical Output** -Requests are processed by the batch processing computers and the results (text files) are packaged and returned to a common directory accessible to the system control computer. The system control 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 the size of the file in bytes is larger than the threshold specified in the initialization parameters the file is automatically moved to the directory reserved for output that must be examined by the LIS staff. In addition the contents of the output are examined in order to reveal any operations, listings, etc. that indicate the user is attempting to copy individual records or subsets of variables from individual records from the microdata file. If any occurrences indicating the possible file copying are found, the file is moved directly to the review area where the contents are examined by the staff.

Output that is judged to be acceptable under the LIS rules is returned to the sender automatically following the check. The results are package up as the body of an email message and sent back using the email address stored in the user database for that user. The connection to the mail server is again made programmatically by opening direct communication to the server using a socket. Note that the source email address (email address from where the request was made) is not used to return results. If a registered user is working at a different email address than the one they provided when they registered, they must contact the LIS staff and have the new email address entered into the database. This mechanism provides added security and helps discourage a registered user from giving their user identification and password to a non-registered user since the results can only be returned to the registered email address.

Return of the job is accompanied by the addition of a record to the job database. This database includes the following information for each job:

- user identification of user who submitted the job
- number of bytes (characters) of output
- execution time in seconds required to complete the statistical operations

- statistical package used
- date and time the request was received
- date and time the request was sent back
- job number

This database of job submissions can be used to provide statistics concerning system usage by various characteristics, such as individual users, groups of users, country of request, statistical package, etc. A later section provides some breakdowns regarding these statistics.

**Critical Databases-**The databases system employed in the LIS system is Oracle. Oracle was chosen because it is an extremely reliable system capable of handling very large and diverse data tables and because it can be totally integrated into C++ programs, a feature that was absolutely required for this system.

A total of four database tables are used to manage the system. These are the JOB table discussed previously, the USER table, the GENERAL table, and the COUNTRY table. The USER table contains all information provided by when someone registers to access the database. It includes the following information:

- Title
- Name
- Address
- Institute or organization
- Telephone number
- Country of residence (code)
- User identification used in submission of requests
- Password used in submission of requests
- Email address to which output is returned
- Priority (default is normal or can be set higher by the LIS staff)
- Access permission switch to United Kingdom data
- User number (sequential number used as primary key)
- User status (enable or disable access)

Each time a new person registers for access to the database the information provided is entered via a custom interface designed specifically for LIS. This interface permits new users to be entered as well as modifications to existing information. This database

interface is another windows program that maintains programmatic linkages to the database. It is necessary to “logon” to the database through this program in order to add a user or make modifications. Passwords are not changed unless the user makes a request to the staff and at no time do users have any access to make changes themselves.

Most items in the list above are self-explanatory. One, however, the access permission indicator to the United Kingdom (UK) data, requires explanation. Those providing the LIS data for the UK, the Archive at Essex University, have determined that persons using these data must sign a specific access agreement in order to access their data. Each user wishing to use the UK data must sign the special use agreement provided by Essex and available from the LIS staff. If this agreement is signed, the switch in the database that grants use is turned to the “on” position.

The remaining two tables, GENERAL and COUNTRY, are small but important. One is very important. The GENERAL table contains only two items. These are the count of the number of requests processed and the count of the number of registered users. Each time a request is processed the count in the database is incremented. Each time a new user is registered the new user number (not user id) is created by incrementing the current database number. The COUNTRY table also contains only two items, the name of every country as text and a country code as a number. This number code is used to identify the country of residence in the user database. If the actual country name is needed the user and country databases must be joined.

**Critical System Parameter Files** – The design of the LIS processing system loosely integrates the system control computer with batch processing computers across the network. Global type parameter files needed by the batch processing computers (system parameters and files that are the same regardless of batch processor) are stored on the system control computer and are accessible to all batch processors. This is to assure that all batch processors are operating based on the same set of rules. Files used by the batch processor include:



- Header text inserted at the beginning of all results files returned to users (See Appendix A). This text reminds users of the basic rules and regulations governing access to the LIS database.
- SAS runtime parameter file (a SAS autoexec file). A large file containing macro variable names for all SAS data sets available, macro variable references to SAS formats, names of the SAS data libraries containing the LIS data, form characters used to define default separators used to delimit output, etc. This file is read each time SAS is executed.
- SPSS alias parameter file. Another large file containing a alias names for the SPSS datasets. This file is also read by SPSS at execution.
- STATA alias file. A “.do” file containing aliases for the datasets

All of these files are text file that can be generated or updated using any text editor program. The alias files must be updated by the staff each time a new dataset is added to the project.

**Batch Processor Application (executable) File** – Perhaps the most important file contained on the system control computer is the executable program for the batch processors. This executable is maintained on the system control machine so that any updates to the program are immediate and do not require updating of the program on whatever batch machines exist at that time.

**Computational Routine Files** – Over the years many users of the LIS dataset have developed valuable routines for computing and manipulating the data to generate many important statistics. The program code for these routines have been donated by the LIS community and made available to all via files stored on the system control computer. Documentation for these routines is located on the LIS website. The routines are utilized by including them within the body of the program submission.

### **Data Server**

The data server computer is very simply a 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 now available. In addition, the

SAS data directory includes one SAS format file for each dataset. This SAS format file is copied to a directory on the batch processor computer at execution time.

The centralization of datasets on the server assures that all requests are accessing the same data. While it would be possible to maintain separate copies of the data on each batch processor, managing that process would be difficult as it would likely be a manual operation that would likely be tedious to monitor. Purchasing larger disk drives to provide sufficient storage on each batch processor would also increase costs. To date we have not experienced any serious network bottlenecks or slowdowns using this centralized approach.

In general the data server needs to be a basic computer with a large, high performance disk drive. SCSI-type disk drives are used to enhance the speed at which the datasets are delivered to the batch processors.

The data server manages a very large number of data files. The directories containing these data files are write protected so that a user could not accidentally change them as part of their job submission. Without this kind of protection it is very likely that some unintended error in specifying the program code would result in destruction or modification of the data.

### **The Survey Microdata**

The heart of the LIS project is the large amount of survey data that has been made available to the research community. These data have been donated by participating countries. When the first data became available in the late 1980's all datasets were at the household (family) level. As the second round of survey data became available a decision was made to expand the database to include individual level data for the members of the households. In this revised scheme separate datasets were added for adult members (age 15 and over) and child members (under age 15). The division

between adults and children was made in order to save computer storage space since income and related data are not applicable to children.

For the most part then, there are three datasets for each country observation for each year, one for households, one for adults, and one for children (the separation of adult and child datasets was not possible in a small number of surveys). These datasets can be linked together in user requests by referencing a unique household identification number present on each. Since users have a choice of accessing the data using SPSS, SAS, and STATA, each of the datasets must be made available in each of the formats required by these statistical packages. This makes a total of nine files for each survey.

### **Batch Processors**

Each time a user sends a request for access to the LIS system 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. Not only do the batch processors act independent of each other, they are as mentioned earlier, for the most part independent of the system control computer. The batch computers obtain some of their operating instructions through parameter files housed at system control, but the existence of a batch computer is not known to system control.

**The Batch Processor Application** - Processing at the batch level is managed by a program written especially for this function. It is a windows-based program written in the C++ programming language. This program is invoked by clicking on a desktop icon representing the program. Recall that the batch processor executable program actually resides on the system control computer so that any computer on the network can be transformed into a batch processor by making the correct mapping of network drives.

The batch processing application will appear as a “shortcut” icon on the batch machine desktop.

At startup the first task performed by the batch processor is to read an initialization file stored in the “c:\winnt” directory of the batch computer itself. This initialization file provides details about the location of directories and files for proper operation of that machine. Among the these initialization variables are:

- network path to the directory where request files can be found
- network path to the directory where results of executions are to be returned
- local directory where executions take place
- local path to the SAS executable
- local path to the SPSS executable
- local path to the STATA executable
- local paths to directory where SAS print and log files will be placed
- network directory where SAS autoexec file is located
- parameter permitting SAS programs for execution
- parameter permitting SPSS programs for execution
- parameter permitting STATA programs for execution

The initialization file is a text file that is can be modified as necessary by the staff. By using an initialization file, the operation of the batch machine can easily be modified without changing the program itself. Because the program is driven by the initialization file, it is not necessary to place the executables for the statistical packages in the same location in all batch machines. Computers with very different configurations regarding their hard drives, drive letters, etc. can be configured as a batch processor very easily. It is also easy to limit the types of requests a machine can process by changing the parameters that control execution of each statistical package. If a computer does not have SAS installed or there is some problem with the program, for example, the parameters can be changed to exclude processing of SAS requests.

After initialization the batch program begins to look for user requests. At 5-second intervals, the batch program looks into the directory where requests are placed by the system control computer. A single directory is used for all types of requests (SAS, SPSS,

STATA). When a request is found, the filename of the request is parsed into logical elements and examined to determine the priority level and the type of statistical package used. The highest priority request is selected of those requests that are permitted for execution on that batch processor (based on statistical package parameter values).

The file containing the selected request is removed from the queue of requests awaiting processing and placed in the local execution directory on the batch processing machine. Once this file is in place, the execution of the request can take place. The execution is started as a “child process” of the controlling batch processor application. A timer is initiated simultaneously so that the execution time of the request can be captured. After the child process has begun execution, the batch processor “waits” until the child process is completed. The timer is stopped and the execution time is computed. At that point the batch processor regains control. For SPSS and STATA job requests it is a simple matter to modify the name of the output file created by the statistical package. The output file name is transformed to include the execution time in milliseconds and the file is then moved to the network directory where output files wait to be returned by the system control computer. For SAS requests, two files can be created by the request, a “.log” file and a “.lst” (listing) file. If both of these files exist following execution (a successful request will create both) they are combined into a single file. The file is then renamed and moved to the output directory on the network.

There is a complication that must be overcome with regard to STATA executions as child processes in the LIS system. For reasons not completely understood, the termination of the STATA program is not “clean”. That is to say, it is, programmatically it is difficult to know when STATA has completely finished its work. Substantial effort was required in programming the LIS system to develop the program code that correctly detects the end of the STATA execution.

**Getting the Data for Execution** - Each request for data access contains one or more references to data sets residing on the data server computer somewhere on the network. When a data set is referenced, for example in SPSS through a “get file” command or in

SAS through a “set” specification, the data are sent over the network to the batch processor where the system or working file is needed. The transfer of data across the network is minimized by encouraging users to limit the number of variables requested since only those variables requested are transferred for the execution.

As these statistical programs can generate various output files in the process of executing the request, the execution directory must be cleared after each request is completed. This is the final task in the job processing cycle. The batch program then returns to look for a new request.

**Installing a Batch Processor** – An unlimited number of batch processors can be installed on the LIS system. Current workloads can be handled adequately using three processors. Under the current system design the installation of a new batch machine is quite a simple process. The first step is mapping the required network resources. Included are the location of the data on the data server and the location of the directories containing parameter files on the system control machine. The next step is installation of the statistical software packages that will be accessible on this processor. These can be placed in any convenient local directory. Next a “shortcut” to the batch program’s executable located on the system control computer must be created and copied to the desktop of the new batch processor. Once these two steps have been taken the initialization file can be created. Initialization files present on any existing batch processors can be used as templates.

An additional step is required to enable SPSS batch processing on the machine. A “dummy” production file must be created using the SPSS production facility. This file is created in the execution directory of the batch machine so that when SPSS starts it executes the specified “batch” run in that directory. What in fact happens is that the batch processor replaces this file in that directory with the user’s request each time an SPSS job is executed. The user’s request must be renamed to the same name specified in the production facility setup.

**Technical Requirements for Batch Processors** - At today's prices it is very inexpensive to purchase a computer with the components needed to provide efficient processing of LIS user requests. A list of minimal requirements based on experiences over the past year would include the items below.

- 500 + MgHz CPU
- 10 + gigabyte disk drive with ULTRA DMA capability
- 128 mb of memory
- Windows NT or Windows 2000 operating system
- network card
- motherboard with 66mb per second or higher data bus

By today's standards this is not a top of the line computer with the exception of the operating system. It is perhaps close to the standard machine offered by many of the low price vendors. Given the nature and complexity of the LIS system, it is not possible to operate batch machines using the Windows 98 system.

STATA has presented some special problems that are valuable to mention here. As STATA attempts to load the entire data set into memory and as many STATA users submit the most computationally intense requests, a system processing large numbers of STATA requests should perhaps provide at least one more powerful batch machine to expedite processing of this work. Providing a somewhat faster CPU and 256 mb of memory could provide substantial benefit. Additional discussion regarding STATA and the system can be found later.

**Practical Problems for Batch Processors** - The batch processors are fully automated systems that are capable of operating with little human intervention. From the practical standpoint, however, our experience shows that there are certain occurrences that are difficult to manage programmatically. These fall into three main categories. The first is catastrophic error termination of the statistical software. In these situations an error message is generally posted on the batch machine monitor indicating the error condition. Such failures have been observed in both SAS and SPSS executions. SPSS terminates more often than we would like in an "error 91" condition. SAS catastrophic failures are

much less frequent. The cause of such failures is not well understood but the consequences are twofold. First, the batch processor is blocked until the message is cleared and the system is reset. Second, the user's request is lost.

The second practical problem associated with batch processors is user requests that lock or block execution, most of the time for reasons that are not well understood. These problems are exclusively associated with SPSS requests. Because no error message appears on the monitor, they are difficult to detect, as a casual glance at the monitor does not reveal that a problem has occurred. Currently, the only means to detect this situation are to examine the list of requests displayed on the monitor to check on the arrival time of the request currently executing or to examine resources used by the program using the task manager.

A third practical problem is user requests that require extremely large amounts of computing time to complete. These requests tend to be STATA request, however, they can occur under SAS and SPSS as well. In some instances, the request has been correctly specified in the program code, but requires a long time to finish (define a long time as 30 minutes or more). In other cases, the request may be poorly specified. For example, requests that fail to specify a subset of variables for use in forming the working or system file on the batch computer cause the full dataset to be transported over the network to the batch processor. This condition, coupled with other types of merges, matches, table lookups, etc. can virtually halt processing on that machine.

**Practical Solutions** - We have found no real way to prevent the occurrence of these practical problems. That is not to say that solutions to some of them do not exist. Presently, however, the best way to minimize disruptions in service caused by these problems is to place a large number of batch processors on line so that the loss of one or two processors does not stop processing altogether. This partial solution combined with periodic monitoring by the staff works very well. The risk of service disruptions does go up considerably when the staff is not present to check on the operation. Fortunately, the low cost of computers today permit the addition of batch machines almost as needed.



## **Web Server**

The web server and LIS website are relatively new additions to the overall system providing services and data access. The web server computer provides access to the information at <http://www.lis.ceps.lu> and is now a critical element in the project.

Included within the website are detailed instructions regarding:

- the syntax of job submission
- dataset lists and identifiers
- details concerning dataset construction and content
- variable descriptions
- computational routines
- test files for testing programs on the user's computer
- institutional documentation
- pretabulated income and poverty measures

Virtually everything needed to guide a user in the access of the LIS data can be found on the website.

The web server used for the LIS project is the freeware "Apache" for windows. This is an excellent web server and is used in more than half of all web server installations worldwide.

The entire web site was designed, implemented, and now maintained by the LIS staff. It represents a very large amount of work but in today's electronic world it is absolutely essential. Once established it certainly saves an enormous amount of staff time as nearly all information requests can be met by the web site.

## **System Workload**

It is imperative that an access system be able to chronicle the usage of its resources. The job submission data table (and associated tables) provides this capability within the LIS system. Presented here are summaries of information contained in these data tables covering the period from January 1, 2000 to June 30, 2000. This is an arbitrary period, but one that we believe is representative of system usage. During this period a total of 30,082 requests for data access were processed in the LIS system (this count does not include jobs that failed the check-in process). This is an average of 166 request per day including weekends.

These 30,082 requests were submitted by 153 different users each of whom submitted at least one request during the 6-month period ending June 30, 2000. Of this total 3, users had submitted 1,000 requests or more and one user had submitted a total of 3,824 requests. Fifty-five users submitted at least 100 requests and more than half of this group had submitted 300 jobs or more.

**Table 3. Number of Requests by Statistical Software Package for the Period January 1, 2000 to June 31, 2000**

<b>Statistical Software Package</b>	<b>Number of Requests Processed</b>
Total	30,082(100%)
SPSS	14,323 (48%)
SAS	10,116 (34%)
STATA	5,643 (18%)

As indicted by Table 3, nearly half of all requests were based in the SPSS package while 34 percent used SAS and 18 percent chose to access the data using STATA.

**Table 4. Mean Execution Time in Seconds by Statistical Software Package for the Period January 1, 2000 to June 31, 2000**

<b>Statistical Software Package</b>	<b>Mean Execution Time in Seconds</b>
Total	101
SPSS	84
SAS	119
STATA	110

On average, the time required to process requests is quite low for all three statistical packages. This average includes jobs that were processed but whose code contained errors that caused the request to terminate normally but with errors. It is not possible to generate an average based on jobs that were totally successful.

Based on these figures, it would appear that the LIS processing system as now configured is quite adequately with three batch processors. This is only partly true however, as these statistics do not reveal how often service may be disrupted as the result of the practical problem discussed earlier. It also does not reveal peak workload periods such as those experienced during the summer workshop. For example, the participants in the 2000 workshop submitted nearly 4,000 requests during the five-day period while attending the workshop. During the workshop period, this group of inexperienced users often submitted requests that resulting in batch machine blockages or long running jobs that caused delays in the processing of requests submitted by others. The system, therefore, cannot be designed by observing the averages. It must be designed to comfortably handle peak load situations. The addition of several more batch-processing machines would help to improve the turnaround in job submissions.

### **System Backups**

One very critical element in a system such as LIS is a method for creating backups of the data and programs that make it all work smoothly. The LIS project includes a backup system that assures no data or programs are lost due to a failure in any system devices.

### **Programming Resources**

The system for processing user requests is based mainly on three computer programs. These programs are custom windows-based applications written in the C++ programming language with direct linkages to Oracle databases. Historically, programming resources have been provided by a team that has been associated with the project since the early stages of the project. This team worked for the LIS project first as employees of the project and later as external consultants. Programming of systems such as that used by LIS would ordinarily be very expensive, however, the costs here have been far below what would be required if a private firm were asked to develop and maintain such as system. Total cost for programming and consulting over the past 2 years have been less than \$10,000. This cost included a full upgrade of the system and software maintenance.

## **Data Protection Issues**

We have touched on data security and protection at several earlier points, but a more direct discussion of this topic is required. Most of the data provided to the LIS project were provided with the restriction that the data files must be maintained solely at the LIS project headquarters facility and that access be granted only through the statistical packages offered by the project.

There are two main data security and protection issues in the LIS. One is the risk that a user could identify a specific individual, that is the linkage of the survey data with a name of the sample person. Here, data protection begins with the data files themselves. None of the data received by the LIS project contain any personal identifying information such as name, address, telephone number, etc. thus all information has been anonymized by the participating country. Second, the data provided to LIS are based on relatively small samples of households taken from relatively large populations. While some of the variables contained on these files, such as occupation, might be used in some way in attempts to identify a particular individual, the chances of doing so are extremely remote. Third, response variance in surveys is well documented. This response variance makes it even more difficult to be sure that the sample observation in the survey is precisely a person known by name in the overall population. In short, there is virtually no way in which confidentiality could be broken for an individual.

The second main data protection issue is unlawful copying of the datasets as the LIS rules of operation forbid this action. It would be extremely difficult and tedious, but not impossible, for a clever and motivated individual to copy portions of the data sets for all observations in the dataset. The LIS systems security procedures perform numerous checks on the job submissions and results files, but it is not possible to anticipate every clever method that might be used. For example, someone could create statistical tables within the same or different requests that, when combined, isolate a variable value for a single observation. This process would then be repeated until a value for that variable

was isolated for all sample cases. Then the values would need to be extracted from the statistical output files and combined. For most of the LIS datasets this would require thousands of job submissions to complete the task.

It must be emphasized that no files are ever returned to users except as listings containing the output from the executions of the requests using the SPSS, SAS, and STATA statistical packages.

A third issue that is related to data security is the purpose for which the data are used. It is forbidden to use the data for commercial or profit ventures. That is the data are provided to users whose intention is to do research. To control the use of the data each person wishing to access the data must register. The registration collects detailed information about the individual, including their place of employment and a statement outlining the purpose for accessing the data. They must sign a pledge in which they agree to obey the LIS rules of access. Finally, a warning stating the rules of use is inserted at the beginning of all statistical output returned to users.

#### **IV. The Survey of LIS Users**

The first survey of LIS users was conducted during the summer of 2000. This survey was designed to provide feedback on how the system is being used and what areas may need improvement. The sample of LIS users consisted of all persons submitting at least one request to the LIS system during the period from January 1, 2000 to July 1, 2000. This yielded a total of 133 sample cases. The survey was conducted by sending an email to each sample case to notify them that they had been selected and directing them to a web site where the questionnaire could be completed. Of the 133 sample cases a total of 76 responded (57 percent). This relatively low response rate and the fact that this is the first time this survey was undertaken makes some of the responses difficult to interpret so care should be taken when drawing conclusions about the results presented here.

**Table 1. Field of Formal Training** (more than one field could be marked)

Economics	47
Political Science	9
Social Policy	22
Sociology	12
Philosophy	1
Law	1
General Public Policy Analysis	1
Gerontology	1
Statistics	2
Social Psychology	1
Labor Studies	1
Total Responses	98

Table 1 shows a breakdown of users by field of study. As might be expected more than half of the users have or were being trained or trained in the economics field (46 percent). A significant proportion classified themselves as being in the social policy field (21 percent).

**Table 2. Country of Access for Users** (where LIS is accessed)

USA	26
Finland	7
Germany	6
Netherlands	3
France	3
United Kingdom	5
Italy	6
Belgium	2
Ireland	2
Canada	1
Spain	5
Luxembourg	1
Austria	2
Norway	1
Denmark	2
Switzerland	1
Israel	1
Australia	1
Sweden	1
Total Responses	76

The LIS users during the first half of 2000 who responded to the survey were distributed over at least 19 different countries as indicated by Table 2. Of the 76 users responding to the survey, 26, or 34 percent resided in the United States. Finland and Germany contributed 7 and 6 users respectively. As only 57 percent of the users responded to the survey, counts by country are clearly biased downward. It is not known if the distribution of users by country is also biased downward although queries of the LIS database could provide a more accurate depiction of use by country.

Table 3 reveals user reasons for using the LIS database. Of the 91 responses, 50 responses or 56 percent use the database for research projects while 24 percent of the users accessed the data for either their master's thesis or Ph.D. dissertations.

**Table 3. Reasons for Using the LIS Database** (more than one reason could be marked)

Own research project	50
Master's Thesis	7
Ph.D. Dissertation	15
Expert opinion for others/research in commission	7
Teaching (as part of a course)	3
No response	9
Total Responses	91

Table 4 shows user responses to a question asking about satisfaction with the LIS website that provides a wide variety of project documentation. Ninety-six percent of users were either very satisfied or satisfied with the website.

**Table 4. Satisfaction With the Website**

Very Satisfied	21
Satisfied	52
Not very satisfied	1
Unsatisfied	0
I do not consult the LIS Website	1
No response	1
Total Responses	76

Table 5 shows turnaround times for jobs submitted to the LIS system as reported by users. Half reported that jobs were normally returned within 5 minutes and 84 percent reported receiving output in 15 minutes or less.

**Table 5. Turnaround Time on Requests**

Less than 5 minutes	38
5 to 10 minutes	15
10 to 15 minutes	11
15 to 30 minutes	3
More than 30 minutes	3
No response	6
Total Responses	76

Other findings from the survey indicate the following:

- 70 percent of users plan to use the database in the future
- 83 percent give the project an overall rating of 7 or higher on a scale of 10
- More than half reported the turnaround time for LIS was just as fast or faster as the computer system at their institution or place of work
- 88 percent were satisfied with the technical documentation
- Nearly 30 percent were not able to use the LIS database as planned
- About half of the users experienced comparability problems
- One in four users had accessed the self-teaching program on the website
- About half of the users had attended the LIS summer workshop

### **V. LIS System Redesign for 21<sup>st</sup> Century**

The current strategy of providing access to the LIS database using email has served the project well for more than 12 years, but serious out of date in terms of today's technology. Consideration is now underway that would transform the access system to one that provides expanded services through an interactive interface accessed using the power of web browser technology.

The details regarding the transformation from an access system based in email to one based on web browser technology have not been established. We anticipate that major elements of the system would include:



- Entry into the interface will require userid and password validation
- Access to the data will continue to be absolutely limited to statistical packages and that access will remain indirect (no interactive use of the statistical packages)
- Users will submit requests by entering the program code into an edit box provided by the system
- Users will be provided with a list of result files currently waiting for return so they are responsible for choosing those to view, delete, or receive
- System status reports will be provided to inform users waiting for results as to the status of all pending requests
- JAVA applets will be used to provide the interface (Java is a programming language particularly well-suited to systems accessed using browsers)
- The email access system will continue to be available for those users who prefer that method of access

The consultant group now responsible for providing programming resources has already converted the current email system to the JAVA programming language and the final bugs in this system are now being worked out. This move to JAVA permits the LIS access system to be deployed on computers operating Windows, UNIX, LINUX, or Sun Solaris. This flexibility could prove very useful and help reduce computing costs as the LINUX operating system can be obtained for free. It is an important step in the direction of an access system utilizing web browsers.

Development of a plan for the web access system will take place during the period between September 2000 and May 2001. During this time the LIS staff will consider the many possibilities to provide better service to the user community while at the same time making sure that the system adheres to the LIS projects access regulations. The LIS board of director will be presented with this proposal prior to the biennial meeting in 2001. Their approval is required before any work can begin.

## **Appendix A. System Software and Components Requirements List**

### **Software in Use in the LIS System**

- Microsoft Visual C++ 6.0 to develop system applications
- SPSS for Windows for request processing
- SAS for Windows for request processing
- STATA for Windows for request processing
- Windows NT 4.0 Operating System for network and system operation
- DiskKeeper to defragment system disks
- Oracle for maintaining and managing databases
- Apache Web Server for website operation
- Microsoft Front Page for website development
- Postoffice mail server for receipt and return of database access requests
- Microsoft Office for general staff use
- Adobe for creation of files in PDF format

### **Hardware in Use in the LIS System**

#### Computers On Windows NT 4.0 Network

- Mail Server – running Post Office mail server software
- Web Server – running Apache web server software
- System Control – running LIS system custom application and Oracle database
- Batch Processor 1 – running LIS system custom application and SAS, SPSS, and STATA statistical packages
- Batch Processor 2 – running LIS system custom application and SAS, SPSS, and STATA statistical packages
- Batch Processor 3 – running LIS system custom application and SAS, SPSS, and STATA statistical packages
- Data Server – network computer housing all country datasets

## **Appendix B. Data Protection and Confidentiality**

### **User Pledge**

I, \_\_\_\_\_, have completed the LIS/LES Project Information Form and hereby promise to use these materials only for purposes of academic research or teaching as specified in the attached application. I further promise to act at all times so as to preserve the confidentiality of individuals and institutions whose information is recorded in these materials. In particular I undertake not to use or attempt to use these materials alone or in combination with any other data to derive information relating specifically to an identified individual or institution nor to claim to have done so. I understand that attempts to make copies of the data, in whole or in part, stored in the LIS database, or any violation of the above clauses, may be subject to censure, fine or imprisonment. I understand that it is my responsibility that any research papers written or assisted by me and based on LIS must be entered into the LIS working paper series before they are published elsewhere.

### **Statement on Return of Statistical Results**

NOTICE TO USERS USE OF THE DATA IN THE LUXEMBOURG INCOME STUDY AND THE LUXEMBOURG EMPLOYMENT STUDY DATABASE IS GOVERNED BY REGULATIONS WHICH DO NOT ALLOW COPYING OR FURTHER DISTRIBUTION OF THE SURVEY MICRODATA. ANYONE VIOLATING THESE REGULATIONS WILL LOSE ALL PRIVILEGES TO THE DATABASE AND MAY BE SUBJECT TO PROSECUTION UNDER THE LAW. IN ADDITION, ANY ATTEMPT TO CIRCUMVENT THE LIS PROCESSING SYSTEM OR UNAUTHOR-ENTRY INTO THE LIS COMPUTING SYSTEM IN LUXEMBURG WILL RESULT IN PROSECUTION ALL PAPERS WRITTEN USING THE LIS DATABASE MUST BE SUBMITTED FOR ENTRY INTO THE LIS WORKING PAPER SERIES PLEASE CONSULT OUR WEB SITE FOR MORE INFORMATION at <http://www.lis.ceps.lu>