Luxembourg Income Study Working Paper Series

Working Paper No. 12

The Luxembourg Income Study: The Use of Telecommunications in the Social Sciences

Lee Rainwater and Timothy Smeeding

May 1987

(scanned copy)

Luxembourg Income Study (LIS), asbl

The Luxembourg Income Study (LIS): The Use of International Telecommunications in the Social Sciences*

Lee Rainwater
Professor of Sociology
Harvard University
and Research Director, LIS

Timothy Smeeding
Professor of Public Policy and Economics
Vanderbilt University
and Project Director, LIS

May 1987

forthcoming: THE ANNALS of the American Academy of Political and Social Science

*The authors are grateful for the support of the Luxembourg Income Study project which helped provide funding for part of this paper. The opinions of the authors should, however, be attributed only to them and not to their respective institutions or supporters.

ABSTRACT: The purpose of this article is to explain how a computerized telecommunications network (the IBM EARN-BITNET system) has contributed to the institution of social science research by making it possible to efficiently access the Luxembourg Income Study database stored in Luxembourg from over 400 major universities worldwide. The article outlines the practical and traditional difficulties encountered in cooperative comparative international research projects and the way that LIS seeks to overcome these problems. The article explains the advantages and disadvantages of housing a household income dataset such as LIS in one place. The "one place" advantages of building an expert staff of consultants who thoroughly understand the database, and the privacy and confidentiality guarantees necessary to access household income databases from several countries, are contrasted with the basic disadvantages of time, cost and user distance from the dataset. The paper goes on to explain how the IBM EARN-BITNET telecommunications system plus the LIS user package (sample file, documentation, software package, etc.) solves this dilemma by allowing the user at any EARN-BITNET sites to easily access the database. Finally we argue that the challenge for realizing the promise of distributed databases and colleagueship lies in payment mechanisms and funding consortiums which facilitate rather than frustrate international collaboration.

The practical and organizational difficulties of time, distance and cost have for several years plagued cooperative international research projects in the social and in other sciences. While international research societies in all of the social sciences regularly convene meetings, conferences and congresses at which papers are presented, the papers normally remain parochial in scope, limiting their analyses to the experience of one country concerning a given social issue. The basic physical sciences can most often deal with scientific "laboratory" problems which transcend social, economic and political boundaries. However, these boundaries are often the essence of comparative social science research because of the opportunity which they present for the researcher to investigate the ways in which different societies cope with similar social problems.

Up until a few years ago, the most common and pragmatic solution for social scientists interested in comparative research was to call together a group of country experts who would meet once, discuss the project, and then write papers about how their own countries were dealing with a given substantive research issue: e.g., poverty or public health problems. The papers were then assembled into a single volume with an introduction, several country chapters, and a summary paper which attempted, usually in vain, to unify the volume. The results of such endeavors were often perplexing to those who sought to learn from such studies. In particular, when the

country papers dealt with quantitative data of differential quality, coverage, definition and scope, the "results" which emerged were frustratingly hard to compare across countries. Experiences of the "apples and oranges" variety such as these have for decades discouraged true comparative research in the social sciences.

The rapidly evolving technology of computerized databanks provides a challenging opportunity to assemble multinational databases that provide a common foundation upon which teams of social scientists can build truly long term and comparative international research programs. These databases provide the opportunity to define a range of theoretical and substantive problems and to combine analyses of several countries into a single paper or book. This is in fact what the Luxembourg Income Study or "LIS" project does.

The next section of the paper describes LIS: its nature and objectives. However, even with LIS in place, important issues of cost, distance and time must be overcome to create long term research projects which do not unduly penalize the researcher by forcing her/him to be away from their normal place of work for inordinate periods of time at high cost. One potential solution to this problem is the dissemination of "public use" datasets to individual researchers. However, due to respondent privacy and confidentiality problems created by large datasets, this solution is not always possible. The second section of the paper discusses telecommunication as

essential to a solution to these problems, using LIS as an example.

We describe our use of the BITNET (Because It's Time Network)-EARN (European Academic Research Network) interuniversity telecommunications system. We argue that because of the BITNET-EARN system, LIS is able to overcome not only distance, time and cost, but also the technological and political/administrative barriers which it faces. Here we find that rather than posing an intrusive threat to individual privacy, modern telescience has offered researchers the opportunity to begin to explore at relatively low cost an entirely new and exciting realm of scientific inquiry which would otherwise be lost.

The final section of the paper mentions the long run implications of such systems for comparative research, both the potential and the difficulties which it affords. We conclude by pointing to the challenge of meeting the short and long term costs of maintaining and expanding both the comparative database and the telecommunications network upon which it relies as the key element in fostering future collaborative social science research.

I. The Luxembourg Income Study - LIS

The Luxembourg Income Study (LIS) database experiment began in April 1983. Its purpose is to gather in one central location, the Center for Population, Poverty, and Policy Studies

(C.E.P.S.) in Luxembourg, sophisticated microdata sets that contain comprehensive measures of income and economic well-being for a set of industrialized welfare states. Because of the breadth and flexibility afforded by microdata, each researcher is free to make several choices such as definition of unit (family, household, etc.); measure of income; population to be studied (e.g., males, females, urban families, elderly households). This truly comparable microdata creates a potentially rich resource for human resource and related policy research.

The LIS databank currently contains datasets from Australia, Britain, Canada, Germany, Israel, Norway, Sweden, Switzerland, and the United States. Datasets from Holland, Denmark, Finland, France and Spain will likely to added in 1988. Table 1 gives an overview of these datasets: country, dataset name and size, income year, data sampling frame, and representativeness of the population.

- - - - - - - - - - -

Table 1 about here

- - - - - - - - - - -

The database consists of income microdata sets prepared to a common plan, based on common definitions of income (by source), taxes, and family and household composition and characteristics. Spouses' earnings and average annual wage rates (earnings divided by hours worked) are separately recorded as well. This resource has already proved extremely

Table 1
An Overview of LIS Datasets

Country	Dataset Name, Income Year (and Size[1])	Population Coverage	Basis of Household Sampling Frame[8]
Australia	Income and Housing Survey, 1981-82 (45,000)	97.5[4]	Dicennial Census
Canada	Survey of Consumer Finances, 1981 (37,900)	97.5[4]	Dicennial Census
Germany	Transfer Survey, 1981[2] (2,800)	91.5[7]	Electoral Register & Census
Israel	Family Expenditure Survey, 1979 (2,300)	89.0[5]	Electoral Register
Norway	Norwegian Tax Files, 1979 (10,400)	98.5[4]	Tax Records
Sweden	Swedish Income Distribution Survey, 1981 (9,600)	98.0[4]	Population Register
Switzerland	Income & Wealth Survey, 1982 (7,036)	95.5[9]	Electoral Register and Central Register for Foreigners
U.K.	Family Expenditure Survey,[2] 1979 (6,800)	96.5[6]	Electoral Register
U.S.A.	Current Population Survey, 1979 (65,000)	97.5[4]	Dicennial Census

[1]Dataset size is the number of actual household units surveyed.

[2]The U.K. and German surveys collect subannual income data which is normalized to annual income levels.

[3]As a percent of total national population.

[4]Excludes institutionalized and homeless populations. Also some far northern rural residents (inuits, eskimos, laps, etc.) may be undersampled.

[5]Excludes rural population (those living in places of 2000 or less), institutionalized, homeless, people in kibbutzum and guest workers.

[6]Excludes those not on the electoral register, the homeless, and the institutionalized.

[7]Excludes foreign-born heads of households, the institutionalized, and the homeless.

[8]Sampling Frame indicates the overall base from which the relevant household population sample was drawn. Actual sample may be drawn on a stratified probability basis, e.g., by area or age.

[9]Excludes nonresident foreigners and the institutionalized, but includes foreign residents.

useful in both basic and applied social and economic research concerned with such human resource issues as:

1. The distribution of household income and the relative income positions of the old and the young; urban and rural residents, and other groups of policy interest, e.g., single parents.
2. The distribution of earnings for both men and women, their change over the worker's life cycle, including the transition to retirement.
3. Comparative studies of the workings of the welfare state and its policies towards the elderly, the disabled, and the unemployed.

The LIS database has been used to study income poverty, the relative economic status of one-parent families, of children, and of the elderly, and the overall distribution of government cash transfers vs. direct taxes (see Appendix). Projects to add noncash income and to explore the role of women's earnings in family income are currently underway.

LIS has now moved beyond the initial experimental stage to provide a databank that can be perpetually updated and expanded to include the most recent data available for any and all nations with high quality income microdata sets that choose to participate. The datasets will be updated during 1988, adding 1984-85 cross-section datasets and the initial waves from several new European household panel studies.

The LIS project and dataset are permanently housed at the CEPS Research Center in Luxembourg. The data are stored on the government of Luxembourg's central computers which are accessed via several computer terminals at the Institute, under the strict rules of the government of Luxembourg's Data Access and Privacy laws.

Once research papers or reports are prepared from LIS, the researcher is required to make the results available as a LIS-CEPS Working Paper. In this way we can document previous LIS research from those interested in furthering the use of our network and provide for a statistical review of results by LIS member country central statistical offices. While there is no charge to reasonable use of the LIS data by member countries who have joined financial forces to underwrite the maintenance and renewal of the database, minimal user charges must be levied on researchers from nonmember countries and international research organizations to pay for data preparation: programmer salaries, data set computer maintenance charges, and transaction costs. Cost estimates depend on expected use and difficulty/ease of proposed manipulations.

II. Access to LIS

In order to use the LIS dataset, a system of communication between the researcher and the dataset is a necessity. Either the researcher must travel to the data, or the data must be

transported to the user. Excluding the possibility of computerized telecommunications, researchers could access the LIS database in Luxembourg by either traveling to Luxembourg or by using traditional telephone and mail linkages. These alternatives are both costly and time consuming, especially for transoceanic access to the database. Research funding organizations are extremely suspicious of international travel due to its high cost and the supposed personal consumption (i.e., "tourist") flavor of such endeavors. Moreover, such travel is disruptive to the researcher, forcing interruption of normal job duties, home life and acclimation to work in a foreign environment. On the other hand, there are advantages to on-site access, in particular the ability to interact with the expert staff who thoroughly understand the database and its nuances. Because there is usually no perfect substitute for interactive face-to-face discourse with such experts, especially when first using a dataset, remote access systems face the challenge of developing user-friendly modes of discourse as a substitute for this verbal interaction.

The usual and preferred alternative to travel researcher is for the data center to create a "public use" datafile which can be exported to the user for minimal cost. However, because of the strict privacy restrictions and confidentiality assurances under which some foreign central statistical offices have loaned LIS copies of their datasets, LIS public use tapes are at present impossible to provide. Despite the incredulity of

American reseachers who fail to understand why "public use" files cannot be created for LIS, foreign governments can make a strong case for data access limitations based on political concerns about confidentiality threats. Censuses of the population scheduled for 1981 in the Netherlands and 1983 in West Germany had to be postponed due to public concern about privacy, confidentiality and access to data. Similar privacy issues in Sweden have led to severe criticism of Statistics Sweden and, much to the dismay of all persons involved, a substantial increase in refusal to participate in Swedish income and labor force participation studies. Due to such concerns as these, the restrictions placed on access to the LIS datafiles are severe. Direct access is restricted to the staff of the LIS project center in Luxembourg under the supervision of Brigitte Buhmann and Gunther Schmaus, the LIS technical team who have sworn to uphold the Luxembourg government privacy restrictions.

In summary, the dual problems of limited direct access due to cost and distance, and restrictions on secondary public use distribution of the datasets have created a severe technical problem for LIS. Fortunately, a solution to both problems is available with computer networking which ties together geographically distant researchers with the centrally located datset and its technical support staff. The BITNET-EARN telecommunications system is therefore the key to the technical success of LIS.

III. BITNET-EARN and LIS

The BITNET-EARN system is an electronic mail and file transfer network available at some 400 or more academic and research centers in the United States, Japan, Canada, Great Britain, Europe, Scandinavia and the Middle East. The diffusion of this network has been extremely rapid. By the turn of the decade, virtually all major universities and social science research centers should be connected to the system. This rapid spread is due to two basic features: low cost and ease of use. The system is highly subsidized by IBM, especially the EARN linkages which are being provided free until approximately 1989. A LIS user with a BITNET "address" (log-on name and computer node) can type in a "message" (letter, program, paper, etc.) which is then transferred by leased lines from university to university computer until it reaches its final destination in the United States or a so-called "gateway" which connects to the EARN network via satellite, The process then repeats to complete the route within Europe ending up at the Luxembourg Computer Center. The message is held at the center until it is retrieved by a LIS staff member. Once the message is received, the Luxembourg staff either replies to the message, or sends a data request to the central computer on which LIS is stored.

In order to facilitate the procedure of data access to LIS, two separate and independent steps are necessary. First, the user, or potential LIS researcher, must have enough information to be able to efficiently request the output which they

desire. This requires a complete and user friendly package to serve as an introduction to LIS. Second, the LIS staff must be able to quickly and efficiently process the request and return the output to the user being sure to protect the confidentiality of the file.

Once a potential LIS user has contacted the project, the first step is to send the person a brief document, similar to section I above, which describes LIS. Researchers who wish to proceed request the complete User Package to LIS. This package includes:

a. A Technical Description of each country datafile which goes into LIS, including sampling frame, expected sampling and nonsampling errors, and other pertinent information.

b. A Definition of Variables list which explains in detail the exact income components from the raw country datafile which went into each LIS variable. For demographic variables, this includes the exact wording and codes for such variables as occupation, education, marital status, etc. for each country. For LIS income variables the maximum and minimum values, mean, median, and percentage of population receiving each type of income are included along with the name of the income components from the country dataset which have been included in the LIS variable.

c. An Institutional Information Codebook which includes a basic description of those income components which are social transfer programs: history, overall outlays, eligibility rules and bibliographic sources for additional information on each such income source in each country.

d. A list of Standard Recodes of LIS income definitions (e.g., pre-tax income; disposable income, etc.) and other recodes (e.g., marital status; one parent families, etc.) for those who wish to compare their results to earlier LIS analyses using these same concepts.

e. A Sample Datafile containing a random sample of about 200 records from each country. This sample is used to test data runs to ensure computer software commands and correct specifications.

f. A package of Technical Request Information, including available software packages and EARN-BITNET technical conventions for sending requests.

A fee of about $20.00 (U.S.) is paid for this package which provides the potential user with answers to most questions which one could initially ask about LIS. Armed with such a package, it is relatively easy to insure that timely, non-duplicative and non-wasteful output requests are sent and returned to the LIS datacenter with a minimum of turnaround delay. Moreover, this package substantially reduces the amount

of "up front" investment which the researcher needs to make in order to both understand and access the datafile. Given test output from the sample datafile, the researcher will have the wherewithal to both debug the software used to request the data, and some sense of the sensibility and utility of the output itself.

Once the job is sent via BITNET-EARN and received in Luxembourg, specially designed software reads through the submitted jobs, verifies their consistency, and sends the job to the main computer in Luxembourg. At present job requests can only be processed using the SPSSX software package. Eventually, SAS, LIMDEP, and other widely used social science packages will also be available. The finished output is returned from the main computer to the LIS center where it undergoes data and confidentiality protection review via software which maintains minimum allowable cell output restrictions, and checks that raw data is not being transmitted, before sending the output file back to the researcher using the BITNET-EARN network.

The two key portals of request submission and output retrieval are under the control of the LIS center staff only. A distant researcher can neither directly access the dataset via job submission nor directly receive output with positive action on their part. In this way, job input and output can be screened to prevent the violation of data protection and confidentiality laws. While turnaround is not instantaneous, we

hope that overnight job submission and output return will become the norm in cases where the researcher realizes not only the local time at which they submit the job, e.g., for those in the United States, but also the local time at which the message will reach the LIS center, e.g., six to nine hours later than the local time in the USA.

While systems somewhat similar to this are functioning within the United States, the Survey of Income and Program Participation (SIPP) Center at the University of Wisconsin-Madison for instance, the LIS system is unique due to its virtual worldwide access. Other systems such as the SIPP system are in place mainly to facilitate user understanding of a very complex datafile such as SIPP, for which public use files are also available. While the SIPP center does allow the user an opportunity to more easily be introduced to the dataset, user friendliness is their primary advantage. Maintenance of privacy and confidentiality, and overcoming the costs of distance are the added features of the LIS EARN-BITNET access system.

IV. Long Run Implications

Just as workstations in office and home have replaced working in a central on-campus computer center, in long run working with data at geographically remote locations may prove more economical than each investigator starting from scratch by installing raw data at his homebase--at least with large,

complex databases, as the SIPP experiment has shown.

However, as LIS is beginning to demonstrate, remote telecommunications access coupled with input and output screens, also provides the central dataset operators with the means to provide some protection of privacy and confidentiality to the suppliers of the input datasets. Without these protections, there would be no LIS because there would not be the input datasets which are its essence.

However, while the uniqueness of the LIS dataset and of EARN-BITNET are the strengths of our project, they in turn are also its weaknesses. Comparative social policy research is in its infancy. In its second year of operation LIS is at approximately the same stage that longitudinal household panel data research was in 1969, the second year of the Panel Study of Income Dynamics (PSID). While the PSID has gone on to create fresh new and exciting opportunities and methodologies for social science research, it took a decade or more for the tools, strengths and weaknesses of panel data analysis to sufficiently permeate the social sciences before it really caught on. Similarly, comparative cross-national social science research is a virtual unknown for most potential LIS users. The enormity of the enterprise of learning to think cross-nationally is one which few have bothered to undertake and even fewer have mastered. The LIS collaborators struggle with this challenge almost daily. Regardless of the "user friendliness" of the technical process for accessing the LIS

dataset, like most highly rewarding life endeavors meaningful interpretation of the results is something which requires time, effort and commitment on the part of the researcher. In this sense, the real job of using LIS is just beginning.

Our second and more immediate concern is that of ensuring funding mechanisms which would provide the basic "public goods" which LIS requires: maintenance and upkeep of both the central LIS database and the EARN-BITNET telecommunications system. By early 1988 we expect that at a reasonable annual cost, the initial nine LIS countries will have joined in a cooperative funding consortium which underwrites the basic cost of LIS datafile maintenance and renewal for five years. This funding will allow us to offer reasonable usage of LIS at no cost to researchers in all nine countries.

At the moment, use of the BITNET-EARN network is also free to member universities with appropriate equipment. The key to the continued use of the system is the charge which BITNET-EARN will eventually levy on user universities. To the extent that universities are willing to fund the up front costs of capital, maintenance and message transmission, BITNET-EARN will also remain free to users. The combination of zero money cost for both data and telecommunication linkage will allow LIS users to achieve maximum economies of scale in sharing of distributed databases.

Under similar regimes, we can expect international collaborative research projects and colleagueship to reach

their full potential. Under alternative higher cost regimes, especially for telecommunications, we are less sanguine about comparative research reaching its full potential.