CONSORTIUM FOR INTERNATIONAL EARTH SCIENCE INFORMATION NETWORK

Sharing Data through the Information Cooperative

Second CIESIN Users Workshop

October 11-13, 1993

Atlanta, Georgia

*

DEVELOPING COMPARATIVE DATABASES

by

Gaston SCHABER

CEPS/INSTEAD

DEVELOPING COMPARATIVE DATABASES

Prof. Gaston Schaber, President of CEPS/INSTEAD (Center for Population, Poverty and Public Policy Studies/International Networks for Studies in Technology, Environment, Alternatives, Development).

ABSTRACT

* The paper gives a short presentation of the Center and of the comparative socio-economic databases the Center develops :

- the Luxembourg Income Study, with more than 40 micro-datasets on income distribution from 20 industrialized countries;

- the Panel Comparability Project (PACO), setting up a common database for longitudinal studies on households, presently integrating Western national panels and preparing for including nascent panels of Eastern countries;

- a similar comparative database project for longitudinal studies on firms, set up at a regional level.

Information is included on data availability, data accessibility and data protection, data documentation, institutional documentation and electronic communication networks.

Referring to these projects, the paper deals with basic issues such as

- production and control of comparability (ex post/ex ante),
- longitudinal data production and quality control,
- comparative analysis as well as the corresponding necessity of having
- training programs for young economic and social science researchers (in order to turn the enormous national investments absorbed by continuous data production into transnationally comparable, scientifically grounded and cumulative knowledge, hopefully usable for policy analysis.

* In relation to CIESIN's explicit concern to address to take into account the human dimensions of global change, and to initiate collaborative ventures for identifying, acquiring and harmonizing socioeconomic and environmental data, the paper also reflects on the conceptual prerequisites for

- transnational and interregional pilot-projects geared to develop
- integrated databases with selected demographic, economic, social, and environmental data, in a
- comparative and geo-referenced information system to be tested.

Pilot-projects of this type need the support of an appropriate infrastructure, to be offered by multinationally located research centers operating together as a large-scale network facility.

PRESENTING THE CENTER

CEPS/INSTEAD is both a CENTER (small) and a NETWORK (inter-continental).

RESEARCH POLICY OF CEPS/INSTEAD

At different levels, **national, trans-national, inter-regional and inter-continental,**

CEPS/INSTEAD carries out micro-economic and micro-social **studies** and creates micro-economic and micro-social **DATA BASES**, with the aim of developing **INSTRUMENTS** for analyzing, programming and simulating socio-economic **POLICIES**. These studies and data bases are geared to either:

- produce **innovative information,** or
- add value to conventional, classical data by creating **compatibility** and **comparability,** or
- develop **innovative methodology**, or
- develop new **information instruments**, useful either for monitoring **policies** or for **technology transfer** ...

DEVELOPMENT POLICY OF CEPS/INSTEAD

CEPS/INSTEAD develops and consolidates its networks of researchers and of research through the **joint execution of trans-national projects** under contract.

At present, these networks function mainly with social sciences, but the Center is now by bringing in exact sciences and technology, so that the networks may make stronger contributions to the economic and social development and/or re-development and re-investment in regions.

PRIORITIES are:

- development of information systems focusing on transnational comparability,
- development of a system of spatialized data bases interactive with classical data bases, for
- integrative studies on regional structural change, particularly in the Greater Region.

GENERAL POLICY OF CEPS/INSTEAD IN RELATION TO INFORMATION POLICY

CEPS/INSTEAD creates through its integrating networks compatibility between the data and systems of its partners to guarantee comparability at a regional, inter-regional and international level. Standardization is a key for compatibility and comparability.

PRESENT RESEARCH ACTIVITIES AT CEPS/INSTEAD

The **national studies** are conceived in a way that should make them valuable as **prototypes** for other countries.
The **international or inter-regional** studies have among their main objectives to produce **comparability**.

*** Longitudinal** study of **living conditions of households** in the Grand-Duchy of Luxembourg. Sample representative nationally. 2.000 households with 6.000 individuals. Per individual 800 variables. Study being used as prototype in other countries.

*** Longitudinal** study on **firms** in the Grand-Duchy of Luxembourg. Representative sample of over 400 firms. 300 variables per firm. Study being used as prototype in other countries.

***** The **Luxembourg Income Study**, an international comparative study on Income Distribution. Countries involved: Australia, Austria, Belgium, Canada, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Japan, Luxembourg, the Netherlands, New Zealand, Norway, Poland, Portugal, Spain, Sweden, Switzerland, United Kingdom, United States of America.

*** The Panel Comparability Project (PACO)**: for the **international comparison** of the already existing **longitudinal studies on living conditions** of households (and a similar project on firms). Since May 1990, CEPS/INSTEAD is, within the framework of the EUROPEAN SCIENCE FOUNDATION, the Center of the Network for Panel Studies of Households. Countries involved up to now: Belgium, France, Germany, Greece, Ireland, Luxembourg, the Netherlands, Sweden, United Kingdom. Observers: USA, Hungary, Spain, Poland.

***** Development of an **information system on environment** - with focus on transnational comparability.

***** Development of a **system of spatialized data bases** interactive with classical data bases.

CENTERS & NETWORKS IN PREPARATION WITHIN THE FRAMEWORK OF CEPS/INSTEAD

(1) Risk Data Bases and Risk Communication Networks.
(2) Comparative Center on Health Economics
(3) Technology Transfer and Information Broking.
(4) Networks for International Development.

*** * ***

Working Modes:

*** Inter-disciplinary:** with permanent collaborators in the fields of psychology, pedagogy, sociology, economics, econometry, mathematics, computer sciences and more recently also in the fields of natural and exact sciences and technology;

*** Inter-university:** with long term collaborators from the universities of Liège, Nancy, Brussels, Antwerp, Tilburg, Frankfurt, Bath, Bristol, Paris, Louvain, Ann Arbor (Michigan), Clark University, Harvard, (Massachusetts), Syracuse (New York), Pittsburgh University (Pennsylvania). And on a long term cooperation agreement with Florida Atlantic University, Boca Raton.

*** Inter-regional:** Luxembourg, Nancy, Liège, Louvain, Saarbrücken, Aachen, Trier, Maastricht, ...

*** * ***

Status:

A former nonprofit organization, CEPS/INSTEAD is now a Public Establishment managed according to private law under the supervision of the Prime Minister (Loi du 10 novembre 1989, Mémorial A N°72). It has scientific, administrative and financial autonomy.

Mission as defined by law:

1) conduct and organize cross-sectional and longitudinal studies of populations, poverty and socio-economic policy;

2) create, manage and utilize data bases with national and international comparative scientific data;

3) develop instruments of analysis, modeling and simulation for socio-economic policy;

4) develop and improve data-processing tools, within the spheres and subject matter envisaged by the present Article;

5) create and maintain interregional and international research and information networks with regard to technologies, the environment, development and alternative kinds of development;

6) organize, at the post-graduate level, training relevant to the proposed research.

The Center may be charged with research on any other matter deemed appropriate by the Government.

*

See following page for synoptic view on databases.

*

PRESENTING THE MAJOR COMPARATIVE DATABASES AND STUDIES

THE LUXEMBOURG INCOME STUDY (LIS)

INTRODUCTION AND OVERVIEW OF LIS

The Luxembourg Income Study (LIS) project began in 1983 under the joint sponsorship of the government of Luxembourg and the Center for Population, Poverty and Policy Studies (CEPS) in Walferdange. It is now funded on a continuing basis by CEPS/INSTEAD and by the national science foundations of its member countries.

LIS has the following objectives :

1. To test the feasibility for creating a database containing social and economic data collected in household surveys from different countries;

2. To provide a method which allows researchers to use the data under restrictions required by the countries providing the data;

3. To create a system that would allow research requests to be received from and returned to users at remote locations;

4. To promote comparative research on the social and economic status of various populations and subgroups in different countries.

Since its beginning in 1983, the experiment has grown into a cooperative research project with a membership that includes countries in Europe, North America, and Australia. The database now contains information for more than 20 countries for one or more years (see third page). Negotiations are in process to add data from additional countries including Korea, Taiwan, Finland, Portugal and Spain.

The LIS databank has a total of over 45 datasets covering the period 1968 to 1989. In 1993 and 1994, additional surveys will be added to represent the period of the early 1990's, for most of the nations (see wave III on page 3). The dataset is accessed globally via electronic mail networks. A new operating system for our remote access network is scheduled to go on line early next year. It will provide additional user options including SAS. Extensive documentation concerning technical aspects of the survey data, and the social institutions of income provision in member countries is also available to users. This work is being supported by the U.S. National Institute of Aging, the Statistical Office of the European Community, the OECD and our member nations.

Reports by participants in the LIS project have appeared in several books, articles and dissertations. Each completed study is published in the LIS working paper series, which currently numbers more than 90 papers. The project conducts annual summer workshops to introduce researchers to the database, and to give scholars experience in cross-analysis of social policy issues related to income distribution. Over 110 students attended the 1988 thru 1992 sessions; 30 more are expected for the 1993 workshop. In December 1993, our first "overseas" workshop will be held in Washington D.C. A LIS Newsletter is published twice yearly and mailed to over 1300 scholars in 25 nations.

The LIS project is supervised by Timothy M. Smeeding (Project Director), Lee Rainwater (Research Director) and Gaston Schaber (President, CEPS/INSTEAD). Further information is available from Caroline de Tombeur at the LIS address on the cover page or Timothy M. Smeeding, Metropolitan Studies Program, 400 Maxwell Hall, Syracuse University, Syracuse, N.Y. 13244-1090). Telephone (315) 443-9045, fax (315) 443-1081, BITNET : SMEEDING@SUVM).

I. The Main Components of the LIS micro data

LIS is based on micro data. The user can access the database to get individual information. At the same time the researcher is free to define his own **observation unit** according to his research question:

* individuals and groups of persons,
* families and types of families,
* households and categories of households.

Because of the relational structure of the database the user can request information at all available levels. He can also link variables from different observation units without losing information.

The micro data also inform on the sources of individual incomes, so that the study can differentiate the different types of incomes. Detailed **individual income variables** exist for:

* wages and salaries,
* social transfers,
* private transfers,
* retirement benefits and pensions,
* near-cash and non-cash benefits, and
* direct taxes and contributions.

All income variables cover a one year period and are counted in the original country currencies.

On the household/family level **demographic variables** give information on:

* number of members,
* number of children,
* number of earners,
* type of unit,
* geography and region,
* etc.

Demographics on the person level are available for:

* gender,
* age,
* relation to head of household,
* level of education,
* type of work,
* industry,
* etc.

These variables are stored in the database for the head of the household, the spouse and three other adults. For three children LIS has variables about gender, relation to the head and age.

The **LIS-variables** allow the user to compare the different countries as far as incomes are concerned. These variables are calculated for every country according the same rules and they measure the same facts for each country:

* self-employment income,
* earnings,
* social insurance,
* social insurance transfers,
* means-tested benefits,
* social transfers,
* private transfers,
* transfer incomes,
* mandatory payroll taxes,
* occupational pensions,
* factor income,
* market income,
* gross income, and
* disposable income after taxes.

II. The Technical Documentation

Since 1990 LIS works on a comprehensive set of documents for each country survey. The purpose of this project is to give LIS users an overview on

* the source of the stored data,
* the quality of the income measures and
* the history and the characteristics of the original survey.

Each country survey is covered by one chapter, which contains information on the following topics:

* general information,
* population and sample size, sampling methods,
* measures of data quality,
* data collection and acquisition,
* weighing procedures,
* determination of survey unit membership,
* children and spouses,
* availability of basic social and demographic information,
* availability of labor market information,
* availability of geographical information,
* sources and amounts of cash incomes, taxes,
* variable list, description of variables, and other variable specific information, and
* a bibliography of important publications based on the survey.

A country comparisons summary follows the same headings and allows the reader to make a quick comparison of the different LIS databases.
The technical documentation is available on diskette.

III. The Institutional Database

Beside the statistical procedure output and the information about the quality of the data, the user also needs information about the social and economic context of the different countries to interpret the number from LIS. For this we offer our institutional database.

The first section contains aggregate, time-series statistics, macro-economic indicators, social expenditure and revenue figures, labor force and economic statistics.

For each country the second part consist of a set of qualitative and quantitative descriptions of the major tax and cash transfer programs. The tax section describes the basic rules underlying employer and employee contribution and the income tax system. The transfer chapter is organized to correspond exactly to the LIS income variables and it describes the basic program rules for:

* sick pay,
* accident pay,
* disability pay,
* social retirement benefits,
* child or family allowances,
* unemployment compensation,
* maternity allowance,
* military, vet, war benefits,
* other social insurances,
* means-tested cash benefits,
* all near cash benefits,
* private pensions,
* public sector pensions, and
* alimony or child support.

For each of the programs information is provided on the first and current laws, the beneficiaries, qualifying conditions, benefit levels including indexation provisions and taxation requirements, financing, and provisions for dependents or survivors.

Each country chapter includes also a brief text giving a historical synopsis of the social protection system in that country, a bibliography, a very brief description of other major programs and politics that effect wages and/or incomes, and a section that assist LIS researchers to link this institutional database to the technical documentation as well as to the corresponding micro variables.

Partial list of variables:

V1	GROSS WAGES AND SALARIES	D6	NUMBER OF EARNERS
V4	FARM SELF-EMPLOYMENT INCOME	D7	GEOGRAPHIC LOCATION
V5	NONFARM SELF-EMPLOYMENT INC.	D22	TENURE (OWNED OR RENTED)
V8	CASH PROPERTY INCOME	D27	NUMBER OF CHILDREN
V10	MARKET VALUE OF RESIDENCE	D28	AGE OF YOUNGEST CHILD
V11	INCOME TAXES	D1	AGE OF FAMILY HEAD
V16	SICK PAY	D2	AGE OF SPOUSE OF HEAD
V17	ACCIDENT PAY	D3	SEX OF FAMILY HEAD
V18	DISABILITY PAY	D8	ETHNICITY/RACE HEAD
V19	SOCIAL RETIREMENT BENEFITS	D10	HEAD LEVEL OF EDUCATION
V20	CHILD OR FAMILY ALLOWANCES	D11	SPOUSE LEVEL OF EDUCATION
V21	UNEMPLOYMENT COMPENSATION	D14	HEAD'S OCCUPATION
V22	MATERNITY ALLOWANCES	D15	SPOUSE'S OCCUPATION
V23	MILITARY/VET/WAR BENEFITS	D16	HEAD INDUSTRYC CLASSIFIC.
V25	MEANS-TESTED CASH BENEFITS	D17	SPOUSE INDUSTRY CLASSIFIC.
V26	ALL NEAR CASH BENEFITS	D18	STATUS OF WORKER HEAD
V32	PRIVATE PENSIONS	D19	STATUS OF WORKER SPOUSE
V33	PUBLIC SECTOR PENSIONS	D21	MARITAL STATUS FAMILY HEAD
V34	ALIMONY OR CHILD SUPPORT	D25	HEAD DISABILITY STATUS
V39	GROSS WAGES/SALARY HEAD	D26	SPOUSE DISABILITY STATUS
V40	HOURLY WAGE RATE HEAD	LFSHD	LABOUR FORCE STATUS HEAD
V41	GROSS WAGES/SALARY SPOUSE	LFTSP	LABOUR FORCE STATUS SP.
V42	HOURLY WAGE RATE SPOUSE	HRSHD	HOURS WORKED/WEEK HEAD
D4	NUMBER OF PERSONS IN FAMILY	YTAXHD	INCOME TAX HEAD
D5	FAMILY STRUCTURE	YTAXSP	INCOME TAX SPOUSE

IV

August 1, 1993

LIS DATABASE LIST **

COUNTRY	HISTORICAL DATABASES		WAVE I	WAVE II	WAVE III (3)
Australia			1982	1986	1990*
Austria				1987	1992*
Belgium				1985*	1988*/1992*
Canada	1971	1975	1981	1987	1991*
Czechoslovakia		1976		1988	1992*
Finland				1987*	1991*
France (1)		1974	1979 1984	1989*/1990*	
Germany (2)	1973	1978	1981/83	1984*	1989*/1991*
Hungary			1983	1987*	1992*
Ireland				1987	
Israel			1979	1987	1992*
Italy				1986	1991*
Luxembourg				1985	1991*
Netherlands			1983	1987	1991*
Norway			1979	1986	1991*
Poland				1986	1992*
Spain			1980-81*		1990-91*
Sweden	1968	1975	1981	1987	1991*
Switzerland			1982		1992*
United Kingdom	1969	1974	1979	1986	1991*
United States	1971	1975	1979	1986	1991*
Yugoslavia				1987	1992*

Notes :

(1) France has an income survey (1979, 1984, 1990*) and a budget survey (1984*, 1989*)

(2) Germany has three different databases, an income and consumption survey (1973, 1978, 1983), a transfer income survey (1981) , and three cross-sections from the socio-economic panel (1984, 1989*, 1991)

(3) We are also in negotiation with Korea (1993), Portugal (1980, 1989), Republic of China (Taiwan) (1990), Turkey (1987) and Greece (1990). Denmark, Japan and New Zealand are unable or unwilling to join at this time.

* Available after August 1, 1993
Year = year that data applies to (reference year), not the year data was collected.

THE PANEL COMPARABILITY PROJECT FOR LONGITUDINAL HOUSEHOLD STUDIES (PACO)

PACO, initiated by Gaston Schaber[1], Gert G. Wagner[2], Günther Schmaus[3], is a project, whose purpose is to produce a public good according to the rules of the scientific community: the creation of a comparative microdatabase, with national and regional panel data, accessible to independent researchers and analysts, under legal and physical conditions which guarantee anonymization and data protection.

1. The Problem

Panel analysis puts a heavy demand on researchers: they have to spend a large amount of time to become familiar with the panel's data organization and with the procedures for its exploitation. This holds true already for one single panel - but difficulties increase sharply for researchers who set out to **compare** two or more of these complex studies.

So up to now panel analysts have focused their efforts mainly on single panels, and in most cases on the panel of their own country. As a rule, only single country data and country related problems have been treated. Little is yet known about differences and similarities between countries.

At the present stage, internationally comparative studies on panel data are feasible only by research teams who manage to secure active participation of staff from the respective national panels. Single researchers or policy analysts are not in a position to do comparative work without help from or close contact with the corresponding panel teams. Even then, the research process is burdensome.

Cross-national research with data sets from national panels is so difficult right from the beginning because each set is organized in a different manner and uses a different format. In short there are:

- no common variable names,
- no common format,
- no common storage system, e.g. as SPSSX/SAS system files,
- no common software.

And there is no central database for hosting the various national datasets.

Without a central or common databank, it is practically impossible to cope **systematically** with the tasks to be undertaken to standardize each of the variables of each of the panels, in order to work out in detail the concepts and the definitions needed for harmonized analyses.

---

[1] Prof. Dr. Dr.h.c. Gaston Schaber, President of CEPS/INSTEAD, Luxembourg; University of Liege, Belgium; Clark University, Massachusetts

[2] Prof Dr. Gert G. Wagner, Ruhr-University Bochum, Germany and German Institute for Economic Research (DIW,Berlin);

[3] Günther Schmaus, Senior Researcher at CEPS/INSTEAD and Associate Project Director of PACO; Luxembourg

2. The Solution

In order to overcome these problems, CEPS/INSTEAD[4] is creating in Luxembourg - in partnership with DIW[5] Berlin - a comparative database.

The PACO research **network** comprises already partners from Belgium, France, Germany, Ireland, Luxembourg, Spain, the United Kingdom and the United States. They cooperate in setting up **progressively** a base containing microdata and complementary information from the various national household panels, **starting** with data from Belgium, Germany, Lorraine/France, Luxembourg, the United Kingdom, Luxembourg and the United States. The PACO Network is in the process of involving research partners in Eastern countries (Poland, Hungary and the Czech Republic) as well as in South Korea.

By building up the Database, the PACO network facilitates comparative cross-national research on policy issues such as labor force participation, income distribution, household formation and dissolution, poverty, problems of the elderly ...

2.1 The very **first step** towards achieving comparability is to establish **data archive files** of available panel data **without harmonizing the existing variables**, and to store the data **in a consistent manner** and **document** the data sets. The variables of each file are to be stored under their original variables' names. These data archives represent a first intermediate step on the way to a fully harmonized panel databank.

Characteristics of the Panel Data Archive: the files

- contain all original variables, and
- use original variable names, but
- use common format,
- use a relational database structure,
- which is accessible as SPSSX system files,
- and offer the possibility of raw data output.

These data archives may be used to analyze panel data separately in a cross-national perspective. These files are not comparable in a strict sense, but they represent a real improvement for international research because they are stored on the same software, use similar file organization and the same computer hardware.

2.2 In a **second and more difficult step**, PACO is adding value to the original panel data by creating **COMPATIBILITY** and **COMPARABILITY**. This means that the PACO Database at this level contains harmonized and standardized variables, both at the cross-sectional **and** at the longitudinal level: with identical variable names, corresponding to a common plan established for defining and recoding variables.

---

[4]) CEPS/INSTEAD: Centre d'Etude de Population, de pauvreté et de Politique Socio-Economiques, International Networks for Studies in Technology, Environment, Alternatives, Development

[5]) DIW: Deutsches Institut für Wirtschaftsforschung

The PACO Database will contain **as many comparable variables as possible**. Each panel carries a set of questions which are identical from wave to wave. These **core questions** are the first candidates for variables to be standardized.

In comparison to the cross-sectional datasets that make up the LIS Database, the PACO Base will contain for all countries the same (or very similar) variables as LIS, but present **in addition** a much richer list of variables in the following respects:

a) it will offer more detailed information on

- demography,
- education,
- income,
- labor force,

b) and definitely more sufficient information on

* cross-sectional and longitudinal aspects of
  - housing,
  - health,
  - unemployment,
  - calendar activities,

* history of individuals concerning
  - family background,
  - education history,
  - employment history,
  - marriage history,
  - fertility history,

* household formation and dissolution

* relationship and links
  - links between partners,
  - links between children and parents,
  - links to former members of the household.

Each national panel has or may have - in addition to the core issues and questions - specific components which do not appear in many of the other panels or which do appear only in one or in a limited number of its own waves. Such components are poor candidates for harmonization and will be stored only in their original form.

The **PACO RESULT FILES** will contain all the variables which can be standardized. **MOREOVER** the user of the result file will have the possibility of **accessing** those **ORIGINAL VARIABLES** in the panel studies which have not been made comparable for some reason. This procedure allows researchers to simultaneously access original and harmonized variables.

Characteristics of the PACO Database:

- access to harmonized panel variables,
- access to LIS variables,
- possibility to access original variables,
- standardized variables names,
- common format,
- common software,
- storage in a **relational** database **structure,**
- i.e. storage as SPSSX system files,
- possibility of raw data output.

2.3 The database will be expanded in a **third step** by a documentation system (**META-DATABANK)**. All necessary information on original and standardized variables will be integrated into the documentation system (on PC) which CEPS/INSTEAD has developed for its own Household Panel. Additional documentation about the newly created comparable variables, in machine readable and in written form, will be prepared. But in the first stage we are only able to collect the original user manuals of each panel study and have them available for PACO users.

We are going to standardize the variables, the file structures and the access system in such a form that the analysis of different panel studies in a cross-national and longitudinal context will be possible with a **MINIMUM NEED FOR MODIFICATION OF THE PROGRAMS** which have been written for the respective country panels. This will be the case at least for standard tabulations and standard analyses. More complex analyses could probably not be standardized in such a way, but will be efficiently supported by the data organization.

At the beginning not all required harmonized variables are already available. Therefore, a researcher may have to create some harmonized variables from the non-standardized data archives and to match these variables with those from the harmonized panel database. This will not be difficult because unique identifiers will allow the matching of both types of files.

2.4 Working with the PACO Database:

* **Data protection**: names, addresses, birthdays and detailed geographical information do not exist on the data files, nor any other variable allowing identification of individuals and families. A further step in anonymization: subsampling of datasets before any possible distribution.

* **Database structure**: the base contains micro-data on individuals and households stored according to relational principles as shown in figure below.

* **Data availability**: country datasets available as SPSS system files/ export files for mainframe and for PC (see database list below).

* **User groups**: in the very initial phases of PACO, for reasons of privacy constraints, access will be restricted to the researchers/institutes who are involved in providing the datasets and in setting up the structures and devices which shall both guarantee the protection of the data and, as soon as feasible, their availability for use by the larger scientific community. - The user system will be modeled after the one developed by LIS.

* **Communication networks**: While the microdata stay protected from direct access, the directory of the Database, the data documentation and the meta-databank will be directly accessible via electronic network. Network facilities will be used for communication between users and the PACO staff.

3. Advantages of the PACO Approach for Cross-National Research

a) Identical analyses with longitudinal data for different countries will be possible.

b) The analyst can simultaneously access the original and the harmonized and standardized variables.

c) Relevant parts of the documentation on the national panels will be translated into one common language (English).

d) Researchers are free to concentrate on analysis without wasting time on data problems.

e) Researchers can access the data within 'their' statistical package, they do not have to learn the retrieval language of a Database Management system.

f) The PACO Database will have a user friendly interface: the user will have to specify only the data he wants to work on, without having to specify the exact procedure for retrieval.

g) The PACO database will have a relational data structure. This data model helps reducing the complexity of panel data.

4. The Future

a) Interpretation of results from cross-national research with panel surveys requires adequate information on the countries' systems of social security, taxation, schooling, etc. After having set up the micro-database, we will have to develop an **institutional database.**

b) The Database will be updated with the new waves coming in from the respective panel studies.

c) In some countries (United Kingdom, Belgium, Hungary, South Korea) new household panel studies have been started. These panel data will be added as soon as the datasets become available.

d) In regard to the Eastern countries where partners are starting panel studies, resources are urgently needed for common workshops, common training session for young researchers as well as for exchange of experienced scholars - in order to create at an early stage the best possible conditions for a good comparative approach in **training** and **learning panel design** as well as **panel analysis of substantial issues** such as labor economics, social policy etc.

In March 1993 the Directorate General XII (Science, Research and Development) of the C.E.C. has informed the PACO Project Coordinator that the project belongs to one of the eligible programs for the EAST and invited him to look for appropriate partners in the eligible countries. - DG XII has included the PACO Project in the information package it circulates in Central and Eastern Europe to allow prospective candidates to contact the Project Coordinator.

*

Having presented the Center and its major comparative studies, we may now **address explicitly the underlying questions and (our) options in matters of RESEARCH POLICY.**

I will do this **from the standpoint of data producers and comparability producers**, knowing well that the audience may consist of a majority of **data users** ... **facing thus a situation that has the potential for misunderstanding and/or mutual stimulation**...

*

BLUEPRINT FOR DEVELOPING COMPARATIVE RESEARCH AND THE CORRESPONDING NETWORK FACILITIES

I

THE INITIAL PROPOSAL OF "MAKING DATA EUROPEAN"

II

TOWARDS SCIENTIFIC ENTERPRISES

III

TOWARDS LARGE-SCALE NETWORK FACILITIES
FOR ECONOMIC, SOCIAL AND HUMAN SCIENCES
INCLUDING THE ENVIRONMENTAL DIMENSION

This section draws on the orientation paper the author has given at the planning conference on "Making Data European: Integrating the Social Sciences Data Base", organized jointly by the EUROPEAN SCIENCE FOUNDATION and CEPS/INSTEAD, April 15-17, 1993, in WALFERDANGE/LUXEMBOURG. - The initial proposal for this conference had been submitted to ESF by a group of scholars: Anthony Atkinson, Robert Erikson, Gosta Esping-Andersen, Jon Eivind Kolberg, Walter Müller, Lee Rainwater, Gaston Schaber and Christophe Starzec.

- The first paragraphs of the present section reproduce the original European perspective, where the need for comparable data is particularly evident, but they apply as well in a much broader perspective. **The basic issue is not making data 'European' but making data 'comparable' across as many countries as feasible:**

* Research on social and economic aspects of European society and institutions is seriously hampered by the **unavailability to researchers of Europe-wide data**:

(a) in many cases the data are available **only at the national** level - with serious obstacles to combining datasets in a way that would make them **representative of Europe**;

(b) in some (very rare) cases European micro databases do exist, but are **not available** for researchers due to policies prohibiting access because of data protection concerns.

But

ad (a): Over the last ten years the experience gathered in developing the **Luxembourg Income Study** has demonstrated that **it is possible** to combine datasets from many nations into a micro-database that is **both** protected against intrusion **and** publically usable for analyses.

ad (b): The recent research on the German microcensus by Walter Müller and colleagues demonstrates that realistic concerns about data protection can be met by the **effective anonymization of data before their release for public use**.

* There is an urgent need to foster in the near future

- the **development of a policy encouraging public use of microdata representative of the European Community and the European Free Trade Area**... and

- the formation of collaborative groups of scholars and of public officials who will focus on shaping policies and institutional approaches which will put into practice at the European level the lessons drawn from a variety of public use facilities.

*** We will have to foster** already existing and nascent

DATA ENTERPRISES

which move forward social and economic sciences
in giving them a European dimension

Bringing existing data sources together, in such a way that Europe and not only its separate nations can be studied, should be the aim.

* Many different kinds of surveys are carried out in most of the countries of Europe (as in other advanced industrial countries). **They can be combined to represent Europe** (and allow at a second stage comparison beyond Europe).

* Microdata collected by national statistical offices and other public institutions represent one of the richest and most reliable **data sources for a wide variety of comparative research problems**, such as

- population structure and demographic processes,
- family and household structures,
- educational and qualification developments,
- labor market processes,
- structures of social inequality and
- other fields of economic and social studies.

Reference is to be made here to the pioneering efforts and accomplishments of (a) the European Community institutions with the Euro-Barometer and the European Labor Force Survey and to (b) the Council of European Social Data Archives (CESSDA), the European Consortium of Political Research (ESPR), the European Consortium of Sociological Research (ECSR) for developing support and services to researchers.

But whatever data are produced and wherever they are stored, **on the European level access to such data is at present extremely difficult if not impossible**.

* The main keys to turning the enormous national investments in gathering social and economic statistics into scientifically useful and cumulative knowledge are

- to develop mechanisms and facilities for making microdata available to researchers and

- to provide explicitly in the training programs for the next generation of economic and social science researchers adequate opportunities for working with such databases in a comparative perspective.

*** Various ways for making micro databases available to scientists**:

In many countries with developed social science communities, national statistical data are made available as microdata to the scientists. **This is done under conditions imposed by a necessary concern about data protection.** The ways in which data are made available and data are protected vary from country to country. They should be analyzed in their respective national contexts and compared for their relative advantages and disadvantages - in order to come to a **generally acceptable understanding and standard in Europe** in regard to **both availability and protection**.

*** The need for developing a common public use policy**

At the level of the E.C. and at the level of the countries a public use policy should be worked out that would be responsive to the need **for a broader range of analyses** than can be done by any single national or community institution.

It is the **diversity of analyses** of public use data **by researchers who vary** in their policy interests and orientation and who vary in their options for applied versus basic social science, which produces the kind of social knowledge that is most useful for policy making.

*** The need for organizing publicly available knowledge and documentation about the complexities of available data.**

Moving from domestic to foreign, from national to European data (and from European to Transcontinental) is extremely difficult in any particular area: the diversity of languages, currencies, distribution systems, social institutions, governmental programs among countries is such, that without adequate and systematically organized documentation any single scholar or research group has to make investments which are in most cases out of proportion to their means and which do not necessarily guarantee the quality of the outcome.

*** Some mechanisms are needed** whereby specialists can organize knowledge in ways that can be made readily accessible to researchers working in a particular area.

One way of achieving this goal would be to **develop institutional arrangements** whereby responsibility for particular kinds of data are assigned to particular institutions which bring together the data, organize it so that others can work with it, and document the data and the institutional background for the data in a readily mastered way. With the development of academic computer communications networks **it would then be possible for widely dispersed researchers to work together or alone on common databases, much as many physicists do**. Such a way of organizing social science data should produce both more accurate research and research that could be conducted more economically because of a reduction in the waste of resources that comes from individual research

groups having to repeat tasks that have been completed already by other research groups. The Luxembourg Income Study project is a pioneering example of this kind of "data utility".

So far, in our common proposal, we had in mind basically **CROSS-SECTIONAL SURVEYS and TIME SERIES**. We should locate these at **LEVEL ONE** as absolutely **BASIC MATERIAL**, on which we will have to work **first**, in order to come to an effective integration of dispersed data which anyway are not produced in a manner to facilitate comparative use.

II

TOWARDS SCIENTIFIC ENTERPRISES

But we will have **to go further** - and in fact some of us are already putting their resources together to move further, **by producing more complex data and trying to produce them in ways that should make them comparable across the involved countries.**

Let us call them **LEVEL TWO** enterprises, and in our personal case these objects are **LONGITUDINAL PANEL STUDIES AND COMPARATIVE PANELS**. The reasons I am putting them forward in this orientation paper, are twofold:

The **first** reason is clearly egocentric: at our own center and in our networks we are devoting large amounts of resources to run our national longitudinal studies, to develop comparability between already existing country panels, and to insure, for nascent panels, the greatest possible comparability right at the start. The **second** reason: these complex endeavors make visible in a most appropriate way some of the basic issues we are confronted with at this seminar - at least insofar as we are dealing with micro-social, -economic and -demographic data, the ones CIESIN is possibly interested in for some conceivable integrated databases.

A brief incursion into the field of longitudinal micro-social and micro-economic panel studies will be necessary. We should have in mind that they may deal with households and individuals or with businesses and firms.

Why <u>longitudinal</u> survey studies?

The rationale for conducting longitudinal studies lies in the fact that **the time dimension is necessarily to be taken into account**

- whenever we want to know about a phenomenon what is permanent and/or periodic, temporary or episodic, random or an isolated instance, or

- whenever we want to ascertain whether the units under observation (persons, households, firms) react in a similar or dissimilar way to **changes** in their environment or in the sets of conditions wherein they function,

- in short, whenever we want to **observe and analyze dynamics**.

Why <u>comparative longitudinal</u> studies?

The coordinated establishment of several, comparable, panel studies, each in its own political and economic context, is a worthwhile undertaking for a number of reasons:

- scientific reasons: to enhance our understanding by means of a cumulative body of knowledge produced under jointly defined or even controlled conditions,

- practical reasons: to enlarge the frame of reference for our observations, analyses and conclusions,

- political reasons: to stimulate innovative processes in social and economic policy by exchanging comparable experiences, gained under identifiable conditions, regarding successes and failures.

Our basic concern in this endeavor is **to secure simultaneously**

(a) the **variability** of the factors as well as of their environmental conditions, and

(b) the **comparability** of the data which document this variability.

*** Let us focus now on the very stringent exigencies for <u>comparability</u> in panel research.**

- In any single panel study observations must be **comparable over time**, which means from one wave of data collection over to the next waves. If we fail to ensure this basic type of comparability, we prevent ourselves from **determining intra-individual, inter-temporal variance**.

- In a set of different panel studies, the data must also be **comparable in space** (from one regional or country panel to another). Otherwise we will lose the possibility to observe what inter-individual variance may exist at the inter-regional or the inter-national level.

- Comparability should whenever possible be taken into account **at all levels by all partners**: the shared goal of achieving comparability requires common work in

    - initial conceptualization of issues and goals,
    - operationalization of hypotheses, variables,
    - defining sampling and corresponding statistical characteristics, ...

If any of these tasks are disregarded, comparability may be seriously compromised.

* At this point, I should insist on a very important requirement which is implicit in the exigency of comparability: that of **quality control** of data, data processing and whenever possible of data production. Quality control related to the data as such and related to data comparability.
Those who are professionally involved in establishing comparability across datasets are well aware of the differences between

- **consumption** of information and data, and

- **production** of information and data.

We are consumers of information when our studies have to rely on data we have not produced ourselves. We often do not know the conditions under which the data have been produced or how they were gathered. We are not in a position to assess their quality. This is not a value judgment but a mere statement.

Furthermore, we frequently work with data that were gathered for administrative purposes and, in their original form, can only with difficulty be put to scientific use.

In the actual state of research, it is of high priority to achieve comparability between microdata sets that were not produced with comparability in mind. And this needs innovative thinking and action. In one particular case, I think we succeeded in a somewhat conspicuous way at CEPS/INSTEAD, where **since 1983 our LUXEMBOURG INCOME STUDY Division processes cross-sectional micro-datasets from more than twenty industrialized countries and works towards achieving the best obtainable degree of comparability.**

In the case of **already existing** household panels as well - where structure is considerably more complex than in the cross-sectional datasets used by LIS - we are currently in the process of setting up a network of experts and centers to work towards comparability among these panels. **For the last three years this endeavor has been strongly supported by the European Science Foundation in the form of an ESF Scientific Network of Household Panel Studies**.

For practically all of these longitudinal studies on living conditions of households, the model has been the United States' Panel Study of Income Dynamics (Ann Arbor, Michigan 1968). While those who have set up these studies are indeed the **producers** of their own data (and attach great importance to the **quality** of those data), **only few** have seen it as their task to ensure **comparability**. In fact a formal concern for the strictest possible comparability right at the time of inception was noted only in the case of the Luxembourg Panel (1985) and the Lorraine Panel (1985), which were conceptualized as twin-panels, by strongly interrelated teams. - The support offered over the last three years by the European Science Foundation to the originally informal panel network has been very instrumental in helping the producer teams of existing panels to focus more intensely on comparability across countries and in helping the teams of nascent panels (in the EC and in the Eastern countries) to join the network already at the planning stage.

Maybe it is useful to mention at this point, in regard to panels and related comparability problems, that at our center and in our networks we deal not only with longitudinal studies on households, but also on firms. In fact we started on firms a year earlier than on households. And once we had research partners in other countries, **we tried from the onset to solve our comparability problems** according to the following rules defined and controlled at the round table at CEPS/INSTEAD:

- panel data are produced according to rules and conditions defined in common at the start and checked over time,
- comparability has to be planned in advance,
- quality controls and comparability controls are to be carried out during the production process,
- to ensure this we necessarily have to work interactively and within a network.

Here we come closest to a situation where as a network of centers we have common control over the conditions of data production for ensuring data quality and data comparability. Of course our panels on firms are not of the same size as the panels of households and, with the exception of Luxembourg, they do not cover countries but only regions. A differential analysis would be interesting but would lead beyond the scope of the present introduction.

What is important to be stressed here, is a characteristic these studies have in common:

LONGITUDINAL STUDIES AND COMPARATIVE LONGITUDINAL STUDIES ARE SCIENTIFIC ENTERPRISES

Such extensive and costly projects go beyond the confines of the academic traditions of social sciences, as well as they do not fit into the framework of an individual academic career. **They are scientific ENTERPRISES, which as such place new demands on the humanities, the social sciences and economics, at all levels: scientific, organizational, financial and political**.

Economics, the social and human sciences have not yet reached a stage similar to that of **physics, natural sciences or technology**, which have

already a long history in generating not only **cumulative knowledge transmissible** across cultural, linguistic and national boundaries but in generating also **far-reaching cooperation and communication networks** as well as **large-scale facilities for knowledge production**.

III

TOWARDS LARGE-SCALE NETWORK FACILITIES FOR ECONOMIC, SOCIAL AND HUMAN SCIENCES

Having given some thought to this issue over the last decade, while developing our own center, I would like in the present context to insist on the necessity of creating in the social and economic sciences large- scale network facilities. Please note that while the physicists do have large-scale **facilities** (for object related reasons), we should rather consider to have **network facilities** (for corresponding task oriented reasons).

Let me give you a possible profile for such a large-scale network facility. (You will understand that any resemblance between this profile and our own center and its networks is purely intentional).

The configuration may comprise

* a **research institution** or a limited **consortium of institutions** organized for carrying out micro-economic and micro-social **studies** and for creating the corresponding **data bases**, and committed to

* - producing **innovative information,**
  - adding value to conventional data by creating **compatibility** and **comparability**
  - developing **innovative methodology** and
  - new **information instruments** useful either for monitoring policies or transferring technology

* - consolidating the **networks** through the **joint execution of trans-national projects under contract**.

* organizing **training programs** and offering on site(s) **training opportunities** to young economic and social sciences researchers so they may learn early in their intellectual career about data production, comparability production, and about turning data into transnationally comparable, scientifically grounded and cumulative knowledge, usable for public policy analysis.

... INTEGRATING THE ENVIRONMENTAL DIMENSION ? ...

At the present stage, this means for us and the partner institutions involved:

* developing the conceptual framework for bringing together in a multi-layer mapping system for **comparative** information

- selected demographic, social, economic **and** environmental variables
- presented at different levels of dis-aggregation, including for each country/region involved the lowest administrative level given,

* knowing that in the test area we have chosen