UNITED NATIONS - UNITED NATIONS POPULATION FUND

TSS/CST Workshop on Data Collection, Processing, Dissemination and Utilization

New York, 15-19 May 1995

INT/92/PH1/doc. 12/95
15 May 1995

Original: English

Why aren't database packages used?

by

Ari Silva
Consultant, Brazil

Why aren't database packages used?

by Ari N. Silva

Abstract

If the database packages are so powerful, why they are not used more widely in the statistical environment? After "sampling" experiences on some countries, the conclusion is that there are three specific faults that affect most of them. First and most important, is their lack of tools directed towards data documentation (metadata) and search facilities. Second, the end users prefer to ask a programmer to produce the information they need rather than to use the databases directly themselves. And third, any package that is to be installed in an institution needs to have a "champion" or a "godfather", that is, a person to push it and "force" its usage.

The document is focused on the Latin American statistical offices of Brazil, Bolivia and Peru, and their efforts on the implementation of "statistical databases" using "database packages", either commercial or in house developments. The concept of "database packages" adopted in this document is very broad and practical, being applied to "any piece of software that can be employed to store statistical data and to produce information in any format". It covers microdata databases such as REDATAM+ and Supercross, aggregated databases such as SIDRA II, BISG and INFORMIX, tabular form databases like PC-AXIS and SISCOP, and even packages for data display like Atlas GIS.

It starts with a brief description of the packages, the scenario on each country and how the packages are being used. Then it detects the similarities of the problems and suggests basic actions to be taken.

There is no intention to compare and/or judge either the packages or the procedures used by the countries; they were used only with the purpose of establishing the grounds to disclose the common deficiencies.

I. Introduction [1]

The purpose of this document is to discuss the problems related to the usage of statistical databases at the National Statistical Offices (NSO). The hypothesis we want to check is that there are common deficiencies that are independent of the specific NSO, the database software and/or the procedures involved.

In order to do that, we have selected some Latin American countries and some packages used (or tried) there, defining what we call "case studies". The examples are from the NSO of Brazil, Bolivia and Peru, and the software set is composed of a variety of microdata databases such as REDATAM+ and Supercross, aggregated databases such as SIDRA II, BISG and INFORMIX, tabular form databases like PC-AXIS and SISCOP, and even packages for data display like ATLAS-GIS. As it will be shown, this "sample" covers Specialized software, General purpose software and In-house developments.

There is no intention to compare and/or judge either the packages or the procedures used by the countries; they were used only with the object of establishing the grounds to disclose the common deficiencies. This document has no benchmark purposes and in no way is recommending or disapproving the usage of any software mentioned here.

It starts with a brief description of the scenario on each country and how the packages are being used. Then it describes the packages, their known strengths and weakness both at the database creation phase and later on the database usage and specific objective. Finally it detects the similarities of the problems and suggests basic actions to be taken.

II. Terminology and Database descriptions

II.1 Some definitions

It is difficult to address the subject of statistical databases without first having to undergo a series of formal definitions of the concepts and terminology used, for example, what is a "database", the difference to "data bank", is a tape library a database?, and things like that. Or the concepts of microdata and aggregated

[1] The author wishes to thank Mr. Arthur Conning for his comments, suggestions, ideas and criticism. Some of the issues presented here were generated during personal discussions with him.

information. Instead of trying to define precisely these concepts (and most probably failing in the attempt), we decided, for the purpose of this specific document, to accept the extremely broad, short and non-academic definitions that follow.

Statistical database and packages

A "database" is any form of data storage, from its lowest implementation (a file or a tape) up to the most sophisticated ones like multidisciplinary and hierarchical integrated databases. Therefore, a "database package" is any piece of software that can be employed to store statistical data and to produce information in any format. This broad definition is needed to allow us to consider software that could hardly be called "database packages", like PC-AXIS or ATLAS-GIS, but in doing that, it brings very useful illustrations to our assumptions. Later in this chapter we will give a small and rudimentary description of each one of them, listed in no specific order.

Microdata and aggregated data

By microdata we designate the data that are stored at the level it was collected, that is, information that is represented the same way they are shown in the questionnaires. For example, in a population census, there are usually a set of questions for the household (water, sewage, etc.) and another set for the persons in the household (sex, age, etc.). These are the microdata (that is, there is information for each variable of each household and person).

On the other hand, we call aggregated data the information that is consolidated (or totaled), often for the higher geographical levels of hierarchy. For example, the total number of males and females for each county.

Tabular form database

The information is organized in such a way in the database that instead of retrieving one "cell" (or one value) at a time, it produces a set of related information in a tabular form. For example, the population of each county by sex and age groups.

Metadata

The concept of metadata (sometimes called metainformation) is "data about data", or "information about information". That is, information that is used to define, describe, comment and document the data stored in a database, in order for the user to understand it. This metadata can have several forms like published documentation and data dictionaries in magnetic form.

II.2 Description of the database systems [2]

Atlas GIS

"Atlas GIS is a full-featured information mapping system that combines the extensive analytical and presentation capabilities of mainframe mapping with the ease and affordability of desktop software" (see Strategic Mapping). Even though Atlas GIS is not really a database system, it is one of the "building blocks" of the BISG system, used to display the numeric information in a map form.

BISG[3]

The BISG is also another approach to an aggregated database system developed internally at IBGE by its Department of Social Statistics (DEISO). It aims the integration of the different sources of information in a standard fashion, allowing the display of thematic maps (see Sousa e Silva).

Although the title is somewhat pretentious, the system is an integration of two commercial software, SAS and ATLAS-GIS. The information is stored in the mainframe using SAS files at the município (county) level. SAS is used to retrieve the variables, and transmit it to the PC environment, generating a **.dbf** file, which is processed by ATLAS-GIS to produce the thematic maps.

Bolivia's in house application

This is another example of a commercial software combination, like the BISG above, where INFORMIX and ARCINFO where put together to store and retrieve the aggregated population census data (see Zuñiga). The information is stored in INFORMIX using DB2, and by means of Standard Macro Language (SML) programming in ARCINFO, the user can display simple thematic maps on the screen.

PC-AXIS

PC-AXIS is a tabular form database software developed by Statistics Sweden that was designed to operate (make calculations within and between) tables, producing results also in diagram form, and providing conversion facilities to export data to other commercial software (see Statistical Databases).

[2] Listed in alphabetical order

[3] Banco de Indicadores Sociais Georreferenciados (Geocoded Social Indicators Base)

REDATAM-Plus[4]

"REDATAM-Plus is a user-friendly, interactive, microcomputer-based system that provides access to hierarchically arranged combinations of very large data files, including the microdata of national censuses, aggregate statistics and large survey files, ... organized in such a way that any tabulations or other statistics can be produced readily by the user for the smallest hierarchical area defined in the data, such as city blocks, or for any grouping of such hierarchical units" (see CELADE).

REDATAM-Plus is being used in various Latin American and Caribbean countries, mainly to store the population census microdata, make tabulations for small areas and do hierarchical processing. Some countries generated a truly multidisciplinary database, by adding other data to the same database, like Honduras, for instance.

SAS, ARCINFO and INFORMIX

These are well known commercial software that need not be described here.

SIDRA II[5]

The SIDRA is an on-line aggregated database system developed internally at IBGE that works in the mainframe environment. It can be consulted by local or remote terminals, and it is available also to users outside IBGE. It retrieves, displays and exports information about several subject areas stored in the database (see Cabral and CDDI).

SISCOP[6]

SISCOP is another tabular form database system, developed at the National Institute of Statistics and Informatics (INEI) of Peru, that stores in magnetic form the tabulations already published in their normal publications. It uses **.wks** files (that is, Lotus-like files) to store the information, and a menu-driven system to navigate through all the available tables. One of the main usage of SISCOP is the Statistical Yearbook, that is going to be offered to the general public in magnetic form.

Supercross

"Supercross is a crosstabulation package that provides rapid access to large databases. ... Instead of writing a program to perform a crosstabulation, fields can be quickly selected, criteria set and the data retrieved using the friendly interface" (see Space-Time Research).

[4] Retrieval of Data for Small Areas by Microcomputer

[5] Sistema IBGE de Recuperação Automática (IBGE's System for Automatic Retrieval)

[6] Sistema de Consulta a Publicaciones (Publications' Consult System)

Supercross is also a microdata database, being used in countries like Australia, New Zealand and India to store their population census data. There is also a pilot project to implement it at IBGE (Brazilian Institute of Geography and Statistics) for the 1991 population census that will be explained later in this document.

II.3 General notes

REDATAM+, PC-AXIS and SISCOP are microcomputer-based, working under DOS. REDATAM for Windows (winR+) is under development and is expected to be released around the end of 1995. PC-AXIS was said to have a Windows versions to be available also during 1995. Supercross and Atlas GIS are microcomputer-based, developed under Windows. SIDRA II works only on IBM mainframes, their developers already started to design a new version for PCs under Windows. There are SAS versions for several platforms, the ones being used by BISG are for the IBM mainframe and for the PCs. INFORMIX works under UNIX.

III. Countries' scenarios

This chapter intends to describe the interrelationship between the users and the packages, in the context of *real situations* found in the Latin American countries. It tries to cover both the database creation and usage phases, focusing on the major difficulties detected. Each situation is used to pinpoint certain problems (which will be called "diseases" in the document) that, far from being specific to the scenario (or the package), are very common to most of them, and will be commented later. The situations are listed in no specific order.

III.1 REDATAM+ at INEI (Bolivia)

The INEI, the NSO of Bolivia, is using REDATAM+ to store their population census data of 1980 and 1992 in separate databases. The creation phase of each one of them was done directly by their programmers, with little reported problems. However, the decision of using two separate databases instead of trying to combine both data in a single one shows one of the general "diseases" that affect the statistical databases: the natural evolution of the lower geographical divisions in each country (department, county, district, etc.) makes it difficult to produce a "time-independent" database.

Both databases are installed in a Novell network, and are used to produce *"ad hoc"* tabulations from the census data. Although REDATAM+ is known to have a user-friendly interface, the queries are written by programmers, centralized in the Informatics department at La Paz. A copy of the databases was installed at the Social Statistics department to be consulted directly by the end users, but with no avail: the requests keep being forwarded to Informatics.

III.2 Supercross at IBGE (Brazil)

Supercross was installed at IBGE, Brazil's NSO, on a trial basis: a small part of the 1991 population census data was sent to Space-Time Research for them to generate the database (there are no records about the complexity of the tasks involved in the process). The product is being evaluated and initial negotiations to acquire it are beginning to take place.

The first analysis were positive: the package is very fast and the interface (under Windows) is interactive and friendly. As there was no in depth evaluation of it at the moment, it would seem to be unfair, if not risky, to include it in this paper, since there was no real usage, that is, no real scenario, but it serves to clarify a specific point, which is also another common "disease": there is plenty of help for the package, but none at all about the data itself.

III.3 SIDRA II at IBGE (Brazil)

"The SIDRA II was developed by the Directory of Informatics of IBGE to store and access its huge collection of aggregated data, and that could be accessed by the IBGE user's community (both internal and external, government or not)" (free translation of Cabral, M.). It was designed to work in the environment of a big IBM mainframe, thus the set of IBM-like tools such as RDBMS, DB2, CICS, CSP, etc., none of them might be called user-friendly (it seemed that it was unavoidable to deal in that kind of environment, since the system was to be accessed by outside users). Nevertheless, they have done a fairly good work and the system is considered to be acceptable.

The system documentation claims to have a good metainformation (that is, information about information), and that all the variables are described and fully documented, available at query time. In order to update this information SIDRA has a very complex module maintained by the data base administrator, which is not a burden in itself as it is transparent to the final user.

What is the problem then? Navigation. Even if the designers affirm that the user is "guided by a series of windows up to the desired results", this is not really the case. The system does not offer a good user interface to be considered independent of further documentation. It seems that you either know what you want to get or you will never get it. In other words, the system is good if you are very familiar with it and its contents. Besides that, the export tools are very incipient, which presents a different kind of problem when you want to combine several queries, or if you want to "map" the results.

III.4 BISG at IBGE (Brazil)

Exactly because of the difficulties presented by SIDRA II for the end users, one of the IBGE's departments, Department of Social Statistics (DEISO), decided to develop a more user-friendly system called BISG, without all the complexities posed by SIDRA's environment: in this case the decision was to use SAS as the basic tool to store the information (aggregated), under TSO in the mainframe computer. An additional goal was to be able to display the information using a Geographical Information System (GIS), obtaining thematic maps. Results are migrated from SAS-mainframe to .dbf form on the microcomputers, and from then on it is up to Atlas GIS.

The procedures to load the database with new information are not well documented, but they work. Each subject-matter area specifies the data (absolute as well as relative) indicators they want to store in the BISG, describing shortly each variable, and the group of programmers working at DEISO takes care of the rest.

The system performs, but it is not as much used as it was intended to be, and the reason is that only the programmers that created the BISG seem to know how to access it, no matter the effort to push it through the non-programmers (documentation, demonstrations, etc.). In other words, the system was designed to be used directly by the end users at DEISO, but they kept asking the programmers to produce the results they need. This situation could be understandable if the argument was that they (the end users) needed to know other basic tools for the mainframe operation, like JCL, TSO, and some SAS (which is true), but what about the Atlas GIS part? Atlas GIS is very user-friendly and there is no difficulty at all to understand and use it (several times the programmers executed the first part of the system, exporting the data to .dbf, in the hope that it would be easy for the end users to produce their thematic maps, but it did not work that way, and the programmers had to finish the job).

III.5 INFORMIX and ARCINFO at INEI (Bolivia)

INEI at Bolivia had to cope with two kinds of environments: a) PCs with DOS used to process the 1992 Household and Population Census; and b) a SUN Sparc 10 System running under UNIX that was installed as a central processor for all the other surveys, and to host a database that would be developed.

Naturally, when they decided to design an aggregated database, the population census, which by that time was already processed and published, was chosen as the pilot data. By definition, the database was to be stored in the SUN workstation, and as they did not have enough personnel (nor time) to do an in-house development, their strategy was to use available software, combining them as needed. Thus, INFORMIX was chosen as the DBMS (Database Management System) to be used, and ARCINFO was defined as the GIS for the map display.

The first intent of using the database was through the SQL (System Query Language), but as they soon realized, it was not meant for the end users, and even the computer specialists had to be trained on this new language, to be able to access the database.

Then, the next step was to design a more user-friendly front-end interface for the users to specify their requirements, which solved the problem partially, as it became clear that more complex queries had still to be programmed using SQL. Besides that, as it was stated from the beginning, this front-end was just designed to show the feasibility of the process, not the real application, so, it was very limited on the map display, since it produced only one variable at a time.

The impression one has by using the system is that it lost the flexibility the host languages (INFORMIX and ARCINFO) have, while not accomplishing any actual gain in the user interface side of it, that is, the system is still usable only by the programmers who created it (and even they have to make frequent use of printed documentation to get through).

III.6 PC-AXIS at IBGE (Brazil)

The PC-AXIS system was recently introduced at IBGE. The pilot project was the creation of a database with information from the 1991 population census composed of around 40 tables at county level (4,491 counties in total). The second database was generated with information contained in an annual survey on medical care establishments, some 30 tables at different levels (department, metropolitan areas and other department capitals). Two more databases are being prepared, one also with population census data and the other with themost recent household survey data.

PC-AXIS has some internal tools to help on the database generation and maintenance, although they showed to be insufficient for a great number of related tables (a small set of support programs were designed for that matter).

The databases seemed to be well accepted by IBGE's internal end users, who worked directly with them with little help (sometimes none at all) from the programmers. The policy that is being discussed right now is about the commercialization of the package to the outside users.

Which are its main weaknesses? Poor metainformation (the system only allows for footnotes at the table, variable or category levels), poor tables organization system (only one level of subject areas can be defined), and no navigation or help available for the data itself.

III.7 PC-AXIS at INEI (Bolivia)

PC-AXIS presented also a very interesting situation at INEI Bolivia, which had nothing to do with PC-AXIS *"per se"*, but with the way people reacted about it. Somebody from INEI heard about PC-AXIS, and they asked for (and received) a database example from Statistics Sweden. An agreement was reached between INEI and Statistics Sweden for the usage of PC-AXIS at INEI, something was even discussed about the royalties that INEI should pay to Statistics Sweden for each copy of PC-AXIS distributed to outside users (US$30).

And what happened? So far, PC-AXIS was not used there. From our point of view, there are two probable causes: a) not one cent was paid for it; and b) there was nobody inside INEI to "push" PC-AXIS, someone that really believed in the package and would act as its "godfather", promoting its usage inside the institution.

III.8 SISCOP at INEI (Peru)

This system was the first trial at INEI of Peru towards the establishment of a magnetic data dissemination policy, avoiding the traditional paper publications. The idea was simple and easy to accomplish: a series of Lotus-like (.wks) files containing statistical tables managed by a menu-driven program that should guide the user in determining which table should be displayed. The actual display is executed through any spreadsheet commercial software that handles .wks files.

The database creation is an easy task, existing tables being fed to Lotus or Quattro-Pro, and a bit of organization to define the various levels of subject areas for the navigation. The first database was created using the Statistical Yearbook information.

The system presents the same general "diseases" mentioned before: poor metainformation and the navigation facilities to search for the desired information are restricted to a series of top-down lists of subjects. The purpose of mentioning SISCOP here is to provide grounds to the discussion of another problem: the database is too large to be commercialized in a diskette form, requiring a different technology (CD-ROM).

IV. Problem Analysis

This chapter is dedicated to the discussion and analysis of the situations and problems mentioned above, attempting to classify them into groups in such a way that it would become possible to generalize and later to propose alternatives for them.

IV.1 Database generation

By looking at the descriptions above, the first thing that becomes very clear is that the failure to implement a given database is not related with the database generation (or update) steps, as complex as they might be for a given software compared to others. This is very understandable since, in most cases, this task is executed just once, by skilled programmers, and within a controlled environment.

Assuming that the database system is already designed, database generation cannot be blamed, although there are people who think that this is precisely the bottleneck, mainly at the design stage, where "...the main obstacle to build a database is the lack of qualified human resources at the NSO. The institutions have very few people, which are always leaving for better salaries elsewhere" (see Zuñiga).

IV.2 Data-friendliness

On the other hand, even if it was not clearly mentioned in the description of every scenario, the common condition that affects all of them is produced by a combination of two factors: a) the deficit of metadata to support the databases; and b) the lack of good "navigation" systems to guide the users in finding the information they need. In other words, they are not **"data-friendly"**.

"Data-friendliness", in this context, is the capacity the systems must have to help the users in working with the data, looking for it, and understanding what they mean, from the view point of the data itself, not the system. It is important not to mistake it for the, now popular, concept of "user-friendliness": All the systems mentioned above are user-friendly (some better than others). They have help-keys, help-lines, on-line information on how to use the menus and options, context-sensitive information on the several facilities each system has, beautiful screens and windows, etc., but all of them, without exception, are directed on **"how to use the system"**, not on **"how to use the data"**!

Besides, by not having some clear definitions of the data in the database, it is very easy for the end users to reach the wrong conclusions because of a misunderstanding or a misinterpretation of the values. For example, when the database has a variable containing the population from the census, nothing is said about its domain, like if it includes the people in collective living quarters (generally it does), or for the number of households, if it contains places not intended for living, or vacant households. These questions might be irrelevant for the people who are very familiar with the census or survey that produced the information, but are severely important for the ones that consult the database less frequently. This corroborates, or reinforces, the end users' problem (see below), where the databases appear to be used only by their "creators" or by people which were already "initiated" or "introduced" in the database mechanisms.

All the systems do not have anything like a table of contents, or an index by keywords, or a list of the existing tables, graphics or charts, or a general "map" to show where the user is in a complex structure of menus, and finally, when the result is shown, there is no help to interpret it, and no information at all about the concepts used.

Helps are meant for people who do not know the system very well, and/or people that do not work with the system in a daily basis. That is, data-friendliness is more important for the outside users, that is, persons that were not involved with the production of the information in the database, or for persons that are not programmers themselves (see next item). Even for the everyday user (or for the programmers), data-friendliness is also important when the databases become more complex and involve several subject areas.

The ideal would be databases behaving as "expert" systems, or analysis oriented systems, with some embedded "artificial intelligence" guiding the user towards the solution to his queries, but this is much easier saying than doing, and, pragmatically speaking, too much to ask for.

IV.3 End users

Some of the scenarios above presented the condition that the databases were not operated directly by the end users, but by the computer programmers or analysts that were familiar with the system, the data, or both. That is, the end users kept asking the programmers to produce the tables or the results they needed for their analysis, instead of consulting the database they have available.

In most cases, this could be due to the lack of data-friendliness (see above), or the scarcity of microcomputers for the users (which is a fact), or the need for training (also true), or insufficient publicity (see Dekker), but there is another component that influences this kind of behavior, which is based on the laziness of the human nature: "it is much easier for me (the user) to ask him (the programmer) to do that". Or on the maintenance of the "status quo" to corroborate the first: "they (the programmers) have been doing that for ages, why should I (the user) start doing it?".

This symptom is important because it refutes the very basic reason or principle for having software packages, which is to approximate the data to the end user, thus removing the existing barrier (the programmers) between them.

IV.4 Package start-up

If one does an inventory of the existing software in any NSO of any country, the list is immense: different packages for the same purpose (word processing, spreadsheets, databases, editors), different versions for different operating systems, etc. The list size seems to grant a kind of "status" to the Institute and its personnel; the bigger the list the more important they appear to be. However, the list of the packages really used is much smaller.

One of the reasons for having unused software, besides the obsolescence, is that they were obtained for free, no effort or "energy" (or money), that is, any kind of "payment". was involved to acquire the software, which appears to degrade its value in the user's mind.

Another reason for that is dependent on the need for each software to have a "godfather" in the Institution, someone that believed in the package, acting as its "champion" or "sponsor", forcing its installation and promoting its usage. The chances the package will have to be accepted are directly proportional to the level of this person in the hierarchy.

Conning goes beyond and states that personal sponsors (or "true believers" as he calls them) are not enough, "because they can leave the institution, or tire". He says that "... in the adoption of a new technology (which he calls *institutionalization*) there are various phases. Initially there is a need for the personal sponsor, but on a second phase, an institutionalization is required, where the use of the system becomes *normal* for everyone."

Conning raises another issue, closely related to this, which is the "technical support", such as the ones provided by the commercial companies. "Without that, it is difficult to spread the use of any database system very widely". And, on this case, much of the support must be for the data (see item IV.2) not just the software.

By the way, these problems are not a "privilege" of the database packages; basically they affect all the software that are distributed and/or maintained by non-profit organizations, like IMPS, POPMAP, etc.

IV.5 Multidisciplinary databases

To establish databases that have several sources of information, and moreover, that allow linkages between these data, like population, education, health, vital statistics, etc., or "time-dependent" information like consecutive censuses, is no easy task. Two of the most common problems that affect these databases are: a) the natural evolution of the lower geographical divisions in each country (department, county, district, etc.); and b) the census is the only survey that can be used at the lower geographical levels.

The first one is difficult to address. For example, in 1991, when the last population census was taken, Brazil had 4,491 municípios (counties). From then on more than 500 were installed (which means that population projections will be more complex, and historical comparability will be hard to achieve), besides the problems of having to store data with different "geographies" in the same database.

The second is related to the singular characteristics of the other surveys that will be combined with the census information. Most of the times they are sample surveys that are not valid at the lower geographical levels, like the household or the income surveys, or even if they are theoretically sound at the lower levels, like the vital statistics, for example, the number of cases at these levels might be too small (and the percentage of errors too high) to be trusted.

Both problems mentioned above are peculiar to small-area databases, either microdata or aggregated data, that is, information stored at lower geography levels. Databases at the department level, for example, are not affected by them, but they are not as useful as the small-area ones.

IV.6 Emerging technologies

Statistical databases, specially the microdata ones, are voluminous. The concept of huge volumes is relative, varying from one country to the other (population databases have a different dimension for Brazil and Mexico, for example, compared to Montserrat or Saint Lucia), and depending upon the computer environment and the media (mainframes deal with Gigabytes, while microcomputers are still in the Megabytes [7]).

On the other hand, the worldwide tendencies of downsizing and decentralization are unquestionable. The NSOs are migrating to "smaller" machines, the users are becoming more and more familiar with the microcomputers, government and planning are being done at lower levels, and finally, the value of statistical information is being reassessed in the developing countries, bringing forward the concepts of commercialization and marketing inside the NSOs.

The combination of these events is determining that statistical databases be available for microcomputers, to be used inside the NSOs or sold to the general public. The former is easy to implement by means of a Local Area Network, where the database would be stored in the server's hard disk, but the latter can be achieved only by using the CD-ROM technology, and then thinking on new forms of presentation, containing graphics, maps, etc.

Of course this is not new to anybody, the market is full of Encyclopedias to prove this point. There are also statistical databases in CD-ROM already such as Australia and the Nordic countries, to name a few, but not as many as it should be.

For the NSOs, CD-ROMs can play another very important role, which is the ideal media for back up purposes, mainly for the census data. There are several examples of censuses of the last decade being lost because of inappropriate storage media, like magnetic tapes.

IV.7 Geographical Information Systems (GIS)

Everyone talks about GIS, and this paper will not be an exception. It looks like every NSO is creating (or already has) a working GIS. However, the great majority is only producing thematic maps displaying census variables, whereas a real GIS is much more difficult and complex than what has been done so far. This is not totally bad in the sense that, even if thematic mapping is not true-GIS, at least the direction is correct, and serves to generate a GIS mentality that will be needed for later works.

[7] Although there are already microcomputers with hard disks of 2 or more Gigabytes in the market, the great majority is still in the couple of hundred Megabytes range.

"The point is that GIS is mainly being used for display, while - so far as population is concerned - this technology offers a way to integrate population factors in decision making in a variety of fields relevant to social and economic development" (Conning, personal consultations). The significance of this combination (geography and population data) is based on the fact that both are "common denominators" for any social or economic project, providing a reliable foundation to incorporate data from other sources. From one hand, the geography of a country is unique, no matter all the differences between surveys, censuses or other type of information. On the other hand, population is the basic demand for the same projects - everything is done for the people (hospitals, schools, roads, etc.).

Historically, the lack of digitized information (cartographic data), mainly at the lower geographical levels, was the key limitation in the process - as soon as people have access to it, they immediately start producing ideas on how to use maps in their studies. Unfortunately the procedures to create these maps are very expensive and time-consuming, sometimes the original maps are not available or non-existent at all.

Another aspect to be considered in relation to cartographic data is the identification that must be applied to the objects (polygons, line or points), that is, a unique code that will be used to bind the cartographic and numeric information together. Past experiences have shown that the NSO responsible for census (and surveys) processing, with some exceptions, did not bother to check (and correct) the codes for the lower geographical areas (sector and blocks), causing several inconsistencies during the binding process. The rationale was that, until recently, census data was only used to produce tabulations at the higher geographical levels, such as department, province and sometimes district, so, it did not matter if some set of data had a wrong block code, provided it did not "migrate" to a different district. Today, with the arrival of systems like REDATAM-Plus, directed towards the study of small-area data, the NSOs are giving due importance to the lower level geographical codes.

So much for "dataware". On the software side, there are some development being done in the direction of facilities for using a combination of tools that allows the end-user to look at alternatives, to help them come to a more rational decision. These "spatial decision support systems (SDSS)", as their designers[8] call them, "do not decide the decision, but allow the consideration of different alternatives" (see Hall). The apllications are based on REDATAM for Windows and a specially designed GIS (hence the *winR+GIS* cognomen), where the REDATAM-Plus (R+) and GIS, as such, will be hidden from the user. The four applications of this scheme, as defined in Conning and Hall, and their pilot sites, are:

[8] The systems are being developed in a joint effort by CELADE and the University of Waterloo (Faculty of Environmental Studies), with a grant from the International Development Research Centre (IDRC) of Canada.

EDUPLAN: Planning of educational resources taking into account both demand for and supply of education (Santiago, Chile: the municipalities of Conchali, Huechuraba and Recoleta of the northern zone of Santiago with Programa Interdisciplinario de Investigación (PIIE), a chilean NGO which has a Ford Foundation grant to work with the three comunas to improve education, and within which EduPlan will be one component of the work).

ACCESS: Determination of access of women to family planning and primary health care and allocation of facility resources (San José, Costa Rica: the Caja Costaricense de Seguro Social and Central American Population Program of the University of Costa Rica, using data from the entire Valle Central).

TOURGIS: Identification of land parcels for development meeting specific criteria and formulation of land development scenarios that are directed toward future land use change (Grand Cayman Island: Department of Planning and the Statistical Office).

ZONPLAN: Creation of common indicators to help identify target group populations and to define planning and other zones. ZonPlan provides functionality to access Redatam-Plus census and multisectoral databases, compute and process socioeconomic indicators, link them with a cartographic base, and determine spatial areas of distribution, and then, from these areas, extract selected groups, and, for more advanced users, to return to *winR+* for further analysis of the population identified. To be tested in the Municipality of the Cantón de Escazu, Costa Rica.

V. Proposed solutions

Unfortunately there is no easy answer for the situations mentioned above. It is clear that the most frequent cause that applies to all "case studies" is the data-unfriendliness, which is the failure of the database packages to document, describe, search, and organize data in such a way to allow users to find what they want easily and fast. It would require a major redesign of the existing database software, whether they are specialized, general purpose or in-house development. The last are the only ones that we can manage directly (since we built them!), but it is easier to say than to implement such a change, considering that it involves a conceptual modification of the database structures to store the extra-information (metadata), plus the programming of the routines to access, display and search them. The other groups of software, commercial or public domain ones, are not under our control, therefore the only thing we can possibly do is try to influence the developers to implement these changes in the future versions. There is a procedure that one might undertake to ameliorate the data-unfriendliness, which is to execute a training and education program to induce the users in establishing a better relation with the package. Besides that, even if the benefits are marginal, the solution is to try to document as best as one can the variables and the information on the database, using the restricted tools each package offers up to the limit.

Training could help also to solve the distance between the end users and the database, although the real solution requires a mentality change. The users must be aware of the handicap they have by not being able to access the data directly, maybe with a psychological push from the management, by difficulting or prohibiting the interaction with the programmers. This strategy might not be as direct as that, but for example, they could give the programmers such a workload that would delay any further user's request.

As for the lack of a "sponsor" to implement the software, the solution would depend on the availability of such a person with the necessary profile in the organization. Most of the times, however, this is not sufficient, since the manager himself must "believe" in the system, be it an outside package or an in-house development. It is interesting to note that, for some reasons, this problem occurs even with their own tailor-made systems, one of them being the changes in management that accompany the political elections in the country. The "institutionalization" of the package is the key.

VI. Conclusion

Independently of the complexity of the database, its environment, scope, data kind (microdata or aggregated data), type of software (commercial, public domain or in-house development), all of them fail mainly because of **"data unfriendliness"**, which is the lack of the ability to describe, search for, and access data in an easy and human way. In order to solve it, developers must concentrate software design much more on the metadata approach rather than on software documentation which seems to be the tendency today.

Other reasons were mentioned and commented upon, like the historical way with which end users interact with the data (through programmers), and the need of a person to "push" the package inside the NSO, and last but not least, specific technical difficulties proper to statistical databases, such as multidisciplinarity, or the political boundary changes.

To solve them directly means changing the package design, which can be made only if it is an in-house development, or trying to cope with the problem indirectly by extensive training and documentation.

VII. Bibliography

Cabral, M. S. and Figueiredo, L. A., **"As Principais Características do Sistema SIDRA II"**, IBGE/DI/DEBAD, DEBAD Series, Year 4, Volume 8, February 1994.

CELADE, **REDATAM-Plus: User's Manual**, CELADE, Santiago, Chile, Series A-201. LC/DEM/G.90 December 1991.

Centro de Documentação e Disseminação de Informações (CDDI), **SIDRA, Manual de Consulta**, IBGE, December 1994, Rio de Janeiro, Brazil.

Conning, A., **"Description of the R+GIS Tourism Application Tools Project"**, CELADE, March 1995, Santiago, Chile.

Dekker, A., **"Building Population Databases"**, International Seminar on Optical Technology for Development of Population Databases, 30 September to 3 October, 1992, Budapest, Hungary.

Hall, G. B. et al, **"Implementation of Integrated Decision Support Tools for Education and Primary Health Care Planning in Latin America"**, University of Waterloo, 1994, Ontario, Canada.

Space-Time Research, **Supercross User's Guide**, Statistics New Zealand, December 1994.

Souza e Silva, D. et al, **"Uma Experiência na Implantação de um Banco de Indicadores Sociais Georreferenciados"**, IBGE/DPE/DEISO, October 1994, Rio de Janeiro, Brazil.

Statistical Databases, **PC-AXIS User's Guide**, Statistics Sweden, September 1993.

Strategic Mapping, **Atlas GIS Reference Manual**, 1994.

Zuñiga, E., **"Construcción de una Base de Datos y su Utilización en el Sistema de Información Geográfico"**, INEI, Subdirección de Informática, June 1994, La Paz, **Bolivia.**