Data processing at the Central Statistical Office of Zimbabwe

Lessons from recent history

M. Jambwa

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1990:4

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 1990:4. Data processing at the Central Statistical Office of Zimbabwe. Lessons from recent history / M. Jambwa.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1990-X101OP9004

Data processing at the Central Statistical Office of Zimbabwe

Lessons from recent history

M. Jambwa

R&D Report
Statistics Sweden
Research - Methods - Development
1990:4

Från trycket Maj 1990
Producent Statistiska centralbyrån, utvecklingsavdelningen
Ansvarig utgivare Åke Lönnqvist
Förfrågningar Birgitta Lagerlöf, tel. 08-783 40 36

ISSN 0283-8680
Printed in Sweden
Garnisonstryckeriet, Stockholm 1990

DATA PROCESSING AT THE CENTRAL STATISTICAL OFFICE- LESSONS FROM RECENT HISTORY

M Jambwa, Central Statistical Office, Harare.

1. Introduction

Data processing has been a major problem area for the Central Statistical Office (CSO) for the last six years. This period also coincides with the initiation of three huge electronic data processing (EDP) projects by the Department. The projects include:

- the processing of the 1982 population which as well known is a huge undertaking on its own.
- the processing of the modules of the National Household Survey Capability Program (NHSCP) which is also a huge task given the multi-surveys undertaken.
- the transfer of the CSO systems that existed prior to independence from one arm of the Government's Central Computing Services (CCS) to another wing of the same Department. This category includes some large systems like Foreign trade, the Census of Industrial production, Commercial Agricultural Censuses, etc. In total the systems are around thirty and they entail a mixture of monthly, quarterly and annual runs. In 1982 it was decided that all these systems were to be implemented on the newly acquired Perkin-Elmer mainframe belonging to the CCS.

In retrospect, it would appear that the decision to implement huge projects such as these at more or less the same time was not appropiate especially given that for most of the period and historically, CSO has had little control over computing resources despite their criticality to its production processes. This, as will be explained later, has proved to be the important reason leading to the computing bottleneck.

Apart from the Perkin- Elmer, the hardware that has been available to CSO includes over twenty CPM Jet 80 micros (including two MOLIM network machines) and around 15 IBM PCs. Most of these machines were acquired to provide stop-gap solutions to the problems encountered with the centralized system but they could not provide a long-term solution due to the huge volumes of datasets handled by the CSO. Also their optimal use was not aided by the fact that CSO was not accorded EDP personnel due to the policy of centralizing all personnel at CCS though quite a large propotion of staff at CSO has now been acquainted with computers.

The above situation has pertained for most of the last six years. However, the situation has radically changed due to the acquisition of a Micro Vax II computer by the CSO at the beginning of 1988. At the time there was also a slight reversal in Government policy vis-a-vis EDP personnel and CSO was allowed to set-up its own EDP unit persons. The unit was set-up mainly to solve EDP problems for NHSCP modules but the traditionally large systems like the population census, foreign trade, etc. were to remain at CCS.

2. Bottlenecks

(i) Excessive centralism in Government's EDP policy has adversely affected the processing of CSO work for a variety of reasons. These include:

- The CCS has controlled the hardware, software and software development resources for statistical work. These are key resources for any national statistical office. In many cases, for example during work for the voter's roll of 1985, the CCS had to give priority to projects other than Statistical ones. This is inherent in the centralized structure but no doubt a clear disadvantage for CSO. Time scheduling for CSO projects has been almost impossible since CSO-CCS plans have in most cases not converged.

- The lack of expertise in systems design for statistical applications has been another major problem. The strength of CCS has been in programming rather than systems design. This is also understandable because to a lot of extent systems design is much more subject matter oriented than programming. This is especially true in the area of statistical EDP work where the design of the EDP system in many cases is closely tied to the statistically oriented design of the survey. At CCS - where the skill to design systems in general has been a scarce resource - it has not been possible to develop the special skill for the design of statistical EDP systems.

- Another consequence of the centralized structure is that CSO has not had much say in the selection of hardware with the result that the hardware available has not enabled the adoption of more versatile statistical software such as SAS, for example. As a consequence of this most systems have had to be implemented in Cobol and Fortran and have consumed much more resources than they would have otherwise.

(ii) High turnover of computer staff has been a continuous plague for all EDP activities within Government. As a result,through no fault of its own, the CCS has not been able to allocate the requisite number of programmers to CSO work. Costantly the number allocated has not only been insufficient but very inexperienced for the timely implementation of CSO systems.

(iii) The bottlenecks encountered by CSO with regard to EDP have not been entirely exogenous but also endogenous.

- Up until recently, the bulk of the staff have not been acquainted with computers. As a result of this specification of the expected output from the systems has presented a lot of difficulties. The specification of user-requirements has tended to be not optimal because CSO has either not been able to provide detailed requirements in time or the requirements made have been too ambitious and hence required more EDP resources than normal. In essence the CSO till recently has lacked this expertise with the result that many communication problems have ocurred between the statistians at CSO and the programmers at CCS.

- Especially for the NHSCP project, there has been a persistent imbalance between plans and existing resources. It can easily be argued that vis-a-vis the project the lack of EDP resources was relatively less of a problem especially after micros had later been acquired specifically for it. The EDP problems encountered

do not emanate from the lack of EDP resources per se but are only components in a larger complex of problems where the effective remedy is to have a lower level of ambition and better planning for the whole NHSCP.

- Perhaps the most important endogenous variable, which may even encompass the above, has been the lack of a rational and effective scheme for processing CSO systems. The lack of such a scheme has led to a rather unco-ordinated development of the various systems and has made links between them very difficult. Also due to this there has not been a standard reference for the statisticians to follow when they are specifying their requirements. Problems faced with the NHSCP modules and the Integrated Agricultural System (IAS) typically manifest the result of lacking such a scheme. The structure of the surveys/censuses is quite similar from one module to the other with the household/farm as the basic object that information is collected on together with one or two other objects related to it. What this implies is that general solutions can be worked out to cater for most of the modules of each of these systems. However, this virtue of design has not been exploited and to a lot of extent EDP resources have not been optimally utilized. This issue will recur a lot in this paper because if it is tackled I see it as a long-term solution to the EDP problems ecountered in the NHSCP and the other CSO projects.

In summary the data processing bottlenecks encountered by CSO vis-a-vis her EDP projects can be viewed from a historical perspective as well as being due to both variables exogenous and endogenous to CSO. The lack of control of her main EDP work has been one of the main limiting factors to CSO's performance. However, this has become less important recently because CSO has acquired significant equipment of her own and has now set up an EDP unit. The main endogenous variable, as already pointed out, has been the lack of a general scheme to be followed for CSO systems. As should be acknowledged, the availability of an adequate hardware system is a necessary, but not a sufficient, pre-requisite for successful EDP development in a statistical office. Software methodology, training and recruitment policies are equally important preconditions.

3. Towards an EDP model for CSO

3.1 A general framework for the development of CSO systems

Section 2 was an attempt to describe the various EDP problems in general within the last six years. Perhaps the most important point made relates to the lack of a general scheme for the design of systems. The result of this has been the adoption of ad hoc solutions for particular systems and these have mainly used "smart coding" with the result that either the output has come out very untimely or it has never come out at all. Hardware and the lack of requisite skills could have also been major bottlenecks but it can easily be countered that lack of a comprehensive framework has led to most of the significant problems. Also as a result of this, the plans adopted for eg. the NHSCP have been over ambitions because there was no scheme to check, for example, the feasibility of the table requirements versus the resources that would need to be enlisted.

The point being made here is that a common model is required because there is need for common concepts and language between subject matter specialists and EDP experts. All people involved in systems development at CSO, whether subject matter or EDP personnel should be able to apply the model in the conceptualisation of their work. Thus the systems development model would become a common frame of reference and communication tool in all EDP related work. In addition to this, emphasis would be put in providing broad

training in the use of chosen packages eg SAS for solving typical statistical problems within the context of the uniform systems development methodology. I do not make any points regarding the issue of policy towards training and hardware in this paper. The latter I feel is currently less of a constraint on CSO though caution needs to be exercised vis-a-vis the compatibilty between any additional hardware which may be acquired and that already existing. Regarding the former it is important that a deliberate policy of training within the context of the systems design model is formulated.

The pre - occupation of this section is to describe the framework currently being pursued by CSO to rectify the above. This partcular sub - section explores the general approach for the NHSCP and the subsequent one goes into the actual model being pursued. Most of the CSO systems can be fitted into the general approach but the details must be developed for each particular system. In this instance the NHSCP is used for illustration.

First lets look at some general concepts:

(i) Structure of data

- Basic object - the basic object in most modules of the NHSCP is the HOUSEHOLD
- Associated objects - there could be several associated objects to the basic object but these have to be determined for each survey. In the ICDS, as shown later, we have the objects PERSON and DECEASED associated to HOUSEHOLD.
- Variables - for every object there will be several variables of interest. A variable being a property that is attached to the object eg. age and sex for PERSON.
- Identification - every object has a unique identification. The identification of an associated object indicates which basic object it belongs eg. for PERSON we have the household id and the person serial number combined.

(ii) Input to the data processing system

The input is based on values observed and measured by enumerators according to a questionnaire and the enumeration is household based.

(iii) Output from the data processing system

This consists mainly of tables and these can differ with respect to the type of object, type of variable and characteristic shown. The variables in the tables could be the original or also have to be derived from the original variables eg. the size of household (SOH).

(iv) File organisation

- Data is best organized in flat files with two different file structures, one for the input stage and one for the output stage.
- The input file structure should be designed with the overall structure from the questionnaire and there should be a specific set of record types for each type of object.
- The file structure should be designed according to the requested tables. If there are tables for several types of objects there should be a set of output record types for each type of object.
- If the structure of data is simple the output record structure might be the same as the input record structure.

(v) System flow chart

A reasonably detailed flow chart is important for each survey to an over view to help when making the time schedule and the estimation of manpower needed to complete the survey. Typically in the processing of the NHSCP modules, three different stages can be distinguished and these should always be reflected in the systems flow chart. They include:

- Data entry and editing stage including preliminary tabulations for editing purposes.
- Transformation of the data structure used at the input stage to a data structure that is suitable for tabulations.
- Tabulations

The basic aim should be to use generalized software and packages for the first and third stages while the second stage typically requires utility programs (sort/merge) and sometimes small tailor made programs.

In essence the above point to the considerations which CSO has taken towards the development or adoption of an already existing framework for the developments of its EDP systems. Without having to reinvent the wheel, it has been found that the infological model already developed by Statistics Sweden goes a long way to cover the requirements. It was therefore decided to adopt this scheme and adapt it to the CSO's own circumstances. The ICDS was one of the first systems to be tackled using the model and hence its use, to be described in the next section, is to be done using the ICDS as an example.

3.2 The infological and datalogical models

The methodology which CSO is adopting consists of two major phases: one contents-oriented (infological) phase and the other technique-oriented (datalogical) phase. During the former, the structure and contents of the planned information system is specified in terms of objects, relations and variables. This phase is user-oriented and the design work requires very close co-operation between subject matter statisticians and EDP specialists. It is mainly concerned with the contents and purposes of the system , i.e. WHAT? and WHY?, and not with the technical aspects of the system. As already indicated in the previous section, in the past a lot of the EDP communication problems at CSO occured mainly due to the lack of this kind of analysis at the inception of systems. During the datalogical phase the resulting infological model is systematically transformed into a model of the data files and processes. This is done in such a way that all files become flat files/relational tables to facilitate the optimal use of generalized software. The production system is modelled and the main consideration is HOW? best to satisfy end-user requirements.

Each of the phases is subdivided into two parts, viz,

• Infological phase	Datalogical phase
- reality analysis	- datalogical analysis
- infological analysis	- physical realization

Reality analysis - this provides a verbal and general description of the aim and function of the system; its subsystems; its relations to other systems; the rules and confidentiality governing it; the timetable for production; etc. It aims to point out some of the major problems in the subject matter area and indicate how the system could help solve these problems. For a example, the function of the ICDS in Zimbabwe has been taken as to gather demographic characteristics of the population, and its aims have been specified as:

- to update the sampling frame and design of the NHSCP
- to provide a basis for updating data on variables collected in the 1982 Census.
- to serve as a pilot study for the 1992 Census.

The ICDS is related to the 1982 census which provides its sampling frame while it also provides

- an update to the 1982 census
- the basis for updating the sampling frame for the other NHSCP modules
- a pilot study for the 1992 census.

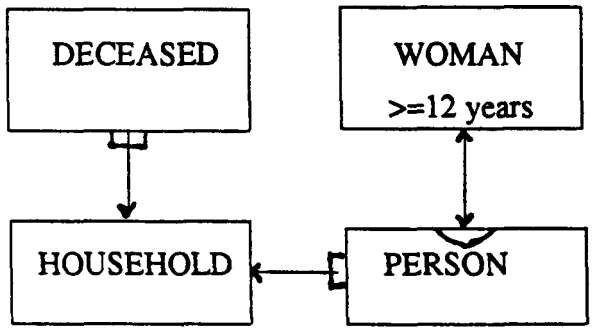
All these provide its relations to other systems.

With regards to confidentiality and security, it is specifically stated that the information can only be published in an aggregated form. No information on individuals/specific households is to be divulged, and the schedules are not to be kept in a way accessible to unauthorized persons.

Infological analysis-this entails three basic concepts, viz:

- objects
- properties
- relations

An Object is the entity for which the information system is aimed to have information about. In the ICDS the objects are identified as:



HOUSEHOLD - group of persons who normally live and eat together

PERSON - usual member of household and visitor last night

DECEASED - deceased who was usual member of household in the last 12 months

WOMAN - woman>=12 who is a usual member of household or visitor

From the schedule these four objects could be identified and encompased all items covered by the questionaire. Objects have properties.For example, the object PERSON has age, sex, etc. These properties are referred to as variables in the model. Variable description for each object indentified therefore forms one of the main steps of the model and this has details encompassing

- name of variable
- description/role
- values/codes that it may take.

All properties of interest should be defined and described per object. To a great extent this step coincides with the record descriptions but the latter is still stated in a more formal way in the datalogical analysis.

Objects are usually related in one or more ways and the types of relations are depicted in object graphs within the model. The object graph for the ICDS could, for example, be charted as in the above graph.

Persons belong to households and you can have more than one person to a household ie a "many-to-one" relationship. Similarly for the deceased-household relationship. We have a "one-to-one" relationship between the object WOMAN and PERSON but the former is a sub-group of the latter and this is indicated by the semi-circle under the arrow in the box for object PERSON. It should be noted that relations can also be formulated as "many-to-many" eg. persons can be employed by more than one firm each.

Another very important element of infological analysis is the very careful statement of the output required by the user. This entails looking at all the tabulation requirements. In CSO, the approach we are adopting is to solicit dummy tables from the subject matter sections and then transform these into the simple methodology of the infological model called alpha-beta-gamma analysis.The alpha refers to the subsetting variables for the particular table, the beta to the analysis variable/statistics, and the gamma to the classification variables.

The structure is defined as follows:

- alpha: for <objecttype> with <property/combination>
- beta: give variables or statistical measures
- gamma : by variables

Examples of this analysis vis-a-vis the ICDS are as follows:

Table on Total population by size of household, age and sex

For PERSON with USUAL MEMBERCODE = 1 or 2

give number

by AGE (group) *

SIZE of HOUSEHOLD (group) *

SEX

Note: (group) means that the classification variable is grouped or translated eg. a numeric code may be translated into text.

Table on Total population 10 years and over by main kind of work, age and sex

for PERSON with AGE >= 10 and WORK =00 - 98

give number

by WORK *

AGE (group) *

SEX

Table on Total number of children ever born and children surviving by sex amd main kind of work of mother

For WOMAN with AGE >= 12 years

give sum BOYS TOT

sum GIRLS TOT

sum (BOYS TOT - BOYS D)

sum (GIRLS TOT - GIRLS D)

by WORK

By adopting this approach we have found that we eliminate the source of most conflicts between EDP and subject matter because this is done jointly and also all important information needs are covered well in advance. Also this analysis gives EDP specialists a basis for judging some of the complexities of the datalogical phase.

Datalogical analysis - the infological specification to a lot of extent implies

- the file structure and record layouts
- the updating or editing operations
- the retrieval and tabulation processes

In this phase efforts are made to transform the infological model to the datalogical model in such a way that

(i) all files are flat files or relations according to relational theory ie. the aim is to arrive at a simple file structure with fixed record lengths and layouts so that in the end what is basically required are programs for sorting, matching, selection and aggregation, also called base operatiors.

ii) the processes can be covered by generalized software to the maximum extent. With such a set of standard procedures and a simple file structure based upon the idea that each homogenous object group (as per our object graph) should consist of one file, retrievals that logically can be performed should also actually be possible to execute using only these standard procedures with a limited number of batch news. Available software may place limits on this ideal and also sometimes it may be necessary to have somewhat more complex file structures in order to perform the immediately desired tabulations within a reasonable time. However what we are trying to do within the NHSCP in Zimbabwe is to find a set of standard procedures which reasonably fulfills the above basic functions and this will determine how close we get to the ideal simple file structure. The above describes the basic tenets of datalogical analysis. Specific aspects include the following:

(a) Datalogical overview.

- an overviewing of the system encompassing the specification of the various flat files, production steps starting from the basis of the sampling frame itself through to data collection, manual processing, data entry, editing, tabulation, etc
- systemsflow showing the processes applied to the flat files and the various operators used in the processes.
- archiving - showing which files should be archived and the rules applied.

(b) Data descriptions - this is simply the detailed description of record layouts or metadata based on each of the object types in our original object and it differs from the variable description in that it also includes the positions, lengths and types of the various fields.

(c) Process descriptions - specifying the EDP processes to be applied to each of the flat files. This includes stating the input and output of each process and the editing rules to be applied.

At this point it should just suffice to point out that all these specific aspects have been followed in the context of the ICDS. All the aspects are very well documented.

Physical realization - this is a subphase of the datalogical phase and it basically includes all the instructions for the actual production process.

In summary the approach currently implies, inter alia, that:

- systems design should be done in very close co-operation between subject matter people and EDP specialists
- the infological model, containing an objects systems graph, definitions of objects, relations, variables, and a formalized description of known information needs (tabulations) should always be worked out before the technical construction of the data processing system starts
- the data should be structured into a set of flat files (or relations in the sense of relational database theory) and the design of the data structures should be based on the object graph in the infological model
- the construction of the data processing system should be based on a systems flow that has been broken down to the level of the typical and fundamental file operations in a statistical data processing system like: selections, projections, sorting and matching of files, derivations of new variables, aggregations, tabulations, and graphical presentations.
- the processes thus identified in the systems flow should be implemented by generalized software used in a well structured way and "home woven " complex constructions such as those often made in tailor-made programs should be avoided.
- technically oriented optimization and "smart-coding" should only be utilized when it has been proven to be absolutely necessary for the sake of machine-efficiency but this should be minimized as far as possible.

4. An evaluation of the new approach

At this juncture, it is still premature to give a conclusive evaluation of the model since it has only recently been adopted. The main applications on which it is currently being applied are the ICDS and the Integrated Agriculural System(IAS) made up of nine systems which had existed independently at CCS.

However, what can safely be stated is that these are the only systems to date which have been thoroughly tackled, at least in accordance with the requirements of infological analysis are these two. Specific points that can be made here include:

- The numerous discussions between the EDP and subject matter units regarding the questionnaires, coding and interviewer manuals, editing rules and other logistics.
- The very detailed systems documentation formulated as a joint effort between the subject matter and EDP units and these are to be used as examples for other systems to be developed by CSO.
- The speed with which the results especially for the ICDS round 1 have been produced in relation to the performance on the other surveys. In the case of the IAS we have now managed to implement the two biggest subsystems ie. the Crop and Livestock censuses for large scale commercial farms. The CCS made two attempts to implement the IAS and both failed.
- Adoption of generalized software (SAS) for most of the processing of the systems. This can be viewed as one of the main reasons for the speed achieved.
- The less communication problems encountered between subject matter and EDP as a result of clarity facilitated by the model.

5. Conclusion

The status of data processing at CSO has undergone radical changes in the last six years. The changes encompass both the availability of more powerful hardware as well a shift towards a more systematic approach in the design of systems. The lessons from the experience of this period include:

(i) In the planning of any enquiry by the CSO, it is imperative that emphasis is taken on the need to achieve "vertical coherence" vis-a-vis all the steps entailed in the statistical production process ie. in the planning and implementation of data collection, processing, analysis and dissemination. Disequilibrium between any of these steps, entails the risk of not achieving the ultimate objectives of the particular enquiry. For example a mismatch between too much detail collected without considering the data processing capabilities entails the risk of either getting the results very untimely or never getting the output at all. Similarly, unless adequate provision is made for the dissemination of the results, little is gained by collecting or even processing data.

(ii) Equally important is to ensure the achievement of "horizontal coherence". It is important to ensure that the various surveys and censuses undertaken are complementary and mutually reinforcing. Given the scarcity of resources it is important that any given enquiry benefits as far as possible from the other enquiries undertaken. This approach is especially beneficial in the context of a generalised model for the various NHSCP modules as already cited. However, it has also been found very appropiate to the IAS especially in veiw of the need to integrate the results of the various sub-sectors of Zimbabwe's agricultural sector.

(iii) The design of systems need to take cognisance of the above two factors. In addition it is foolhardy to try and short circuit the process of systematic design,especially for ongoing systems, because in the end one is likely to be caught up with some difficulty. A case in point is the repeated failure by CCS to implement the IAS. To me the deep-seated reason for this was the tendency of those involved to start straight on to coding the actual production systems without bothering to take time thinking and charting the various production steps. Very often because of lack of infological analysis there were clashes between subject-matter and EDP staff.

R & D Reports är en för U/ADB och U/STM gemensam publikationsserie som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna. I serien ingår även **Abstracts** (sammanfattning av metodrapporter från SCB).

R & D Reports Statistics Sweden, are published by the Department of Research & Development within Statistics Sweden. Reports dealing with statistical methods have green (grön) covers. Reports dealing with EDP methods have yellow (gul) covers. In addition, abstracts are published three times a year (light brown /beige/ covers).

Reports published earlier during 1990:

1990:1 (grön) Calibration Estimators and Generalized Raking Techniques in Survey Sampling (Jean-Claude Deville, Carl-Erik Särndal)

1990:2 (grön) Sampling, Nonresponse and Measurement Issues in the 1984/85 Swedish Time Budget Survey (Ingrid Lyberg)

1990:3 (grön) Om justering för undertäckning vid undersökningar med urval i "rum och tid" (Bengt Rosén)

Kvarvarande beige och gröna exemplar av ovanstående promemorior kan rekvireras från Elisabet Klingberg, U/STM, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 41 78.

Kvarvarande gula exemplar kan rekvireras från Ingvar Andersson, U/ADB, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 41 47.