DATA-CAPTURE AND DATA-BASE SYSTEMS TECHNOLOGIES FOR STATISTICAL DATA-PROCESSING[1]

by W. T. Torres[2]

I. Introduction

This presentation is mainly concerned with two things:

(1) to make the interested computer user aware of the availability of certain computer hardware and software technologies that may be used for more effective and efficient data-gathering and data storage and retrieval; and

(2) to indicate a number of ways by which such technologies can be organized in combination with human processes in order to improve the systems performance of data-capture and data storage, maintenance, and retrieval.

It is generally accepted that the use of computers has greatly increased the productivity of most statistical activities. This is true in number-crunching applications that deal with studies about populations, e.g., data manipulations and calculations on data samples of those populations. Perhaps this is also true in data analysis where statisticians are concerned with "massaging" the masses of data (which are thought of as representing an entire population) and attempting to understand what the data say about the population. However, in both cases, the situation is that the data had already been captured and stored in computer-readable format. The preceding steps of data-gathering and organizing the data in the computer remain difficult and are quite costly. These are our areas of concern in this paper.

---

[1]Presented at the Philippine Statistical Association's Annual Conference on October 16, 1981 in response to PSA's request to the Philippine Computer Society.

[2]Senior Vice-President, Development Academy of the Philippines

Classical Data-Processing

In the 1950's, at the beginning of data processing by computers, the computational procedures and the data description(s) were of necessity integrated in the same user program. (See *Figure1*). The programmer had to know exactly how the data, typically key punched on cards, were physically arranged on each card as well as in the entire card deck. Then he codes his computational procedures – writes his program – always keeping in mind this organization of the data set. If some cards were missing or somehow the cards got interchanged, the program does not work!

Then in the early 1960's, when monitors were first installed, user programs could be run with several data sets. The data could be separated from the program but the physical data description remained in the program. (See *Figure 2.*) To a limited extent, data sets could be created and maintained separately from the programs, however, the programmer may have to rework his data descriptions within the computer programs whenever any of the data set is changed.

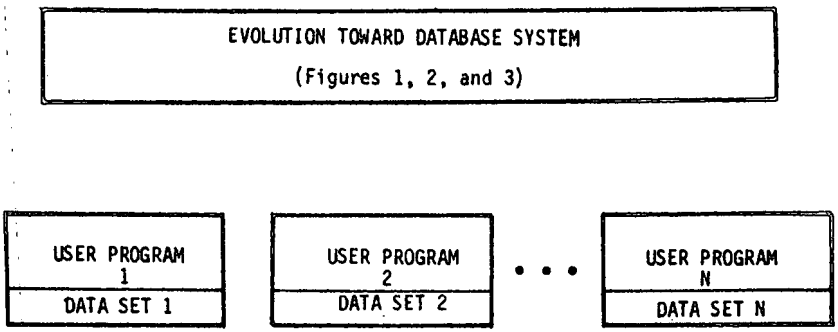

Fig. 1. DATA IS PART OF PROGRAM

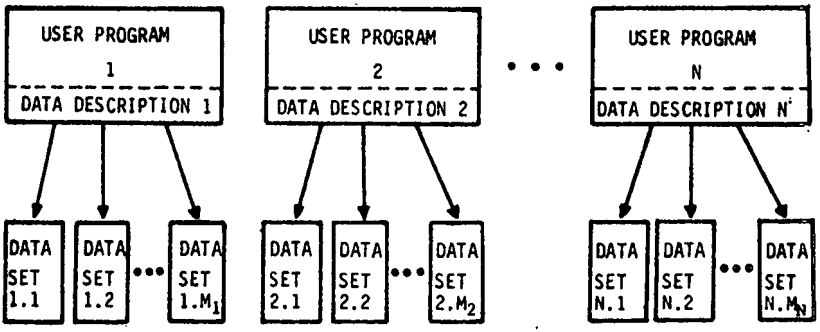

Fig. 2. DATA SEPARATE FROM PROGRAM

File Management Systems

As more and more people used the computer and as exponentially more data were stored in it, it became uneconomical for one person to use the data. Data were (then and still are) better used simultaneously by several users. A more efficient use of computer storage, which was relatively costly at that time, was possible if there were less redundancy of data. To be able to efficiently handle the growing volume of data, *file management systems* were introduced. (See *Figure 3.*) With FMS, several users could share different data files. Note that the physical data record description would still be needed in the user programs. Data redundancy was still present but the more serious and still unsolved problem is the possible inconsistency of the data. Updating the data in the different files was (and still is) a very difficult task in regard to maintaining their consistency. A new software technology was needed and eventually this was developed and made available by computer manufacturers and software houses. Thus database management system (DBMS) software came into the data processing scene. We shall return to this subject later.

Batch Data Entry

Figure 4 depicts the manner in which a data processing shop is organized to be capable of processing large amounts of data using the

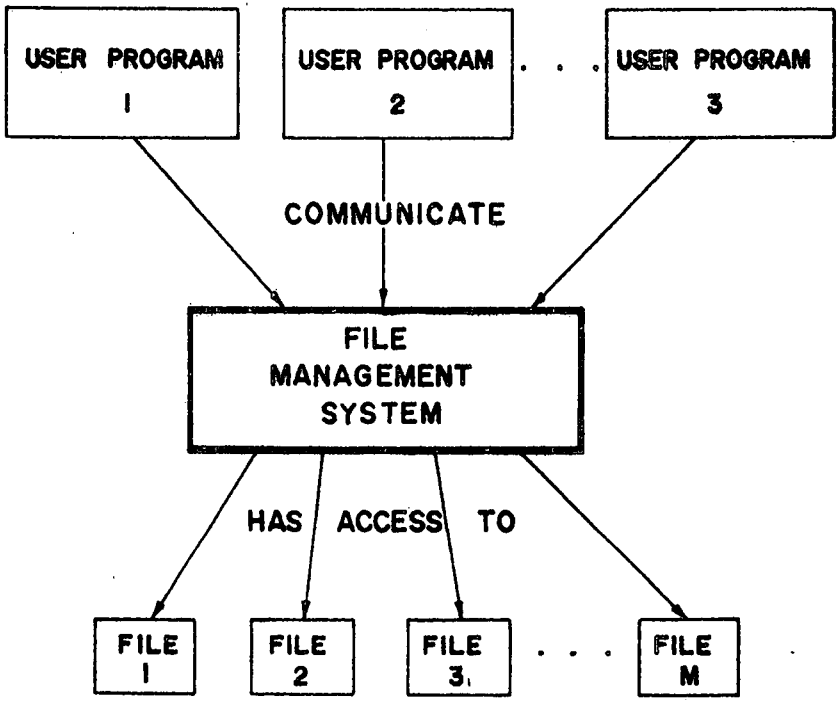

Fig. 3. FILE MANAGEMENT SYSTEM

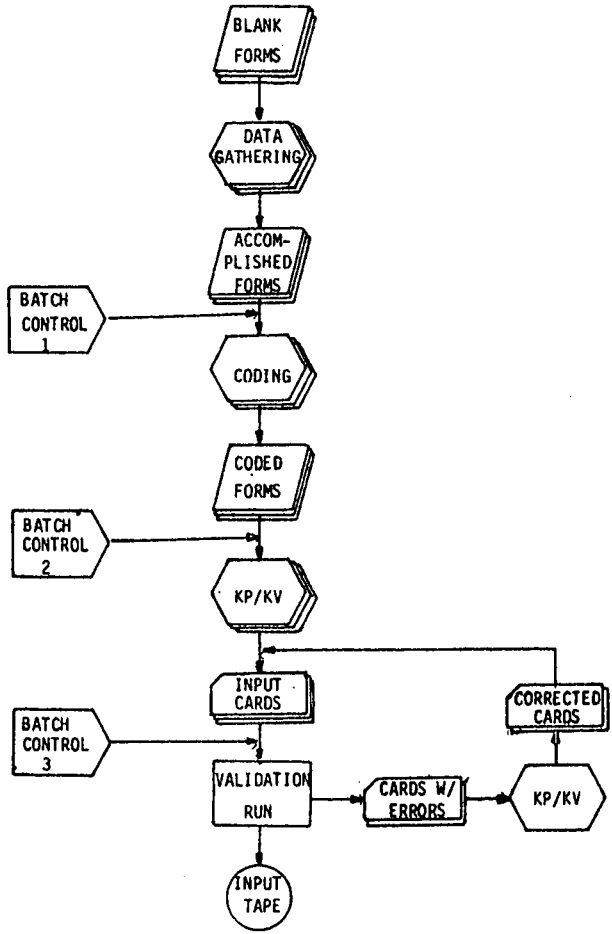

Fig. 4. TRADITIONAL BATCH DATA PROCESSING OPERATION

traditional batch data-entry process. After data-gathering, the accomplished forms are coded, key-punched and key-verified, and then computer-validated before the data are stored in a computer-readable medium, say, magnetic tape. On top of these processes, manual control procedures are instituted to minimize human error, such as, for example, not having key-punched a form or not having corrected an error. Because of various types of human intervention by different people at various steps in the entire data-entry process, the probability that the data stored in the computer is different from the data gathered is uncomfortably high.

To reduce this variance to a manageable minimal level is time-consuming and quite costly. (In the U.S. a five-year old cost-estimate for data-entry is fifty cents per thousand characters; this may be close to a dollar at present because of increased labor costs. Locally, this cost at present is more than two pesos per thousand characters just for key-punching and key-verifying records.) What is obviously needed in order to reduce, if not totally eleminate, the possible sources of human error is a mechanism for enabling the person who gathers the data to directly input the data into computer-readable format.

II. Data-Capture Technology

Centralized On-Line Data Entry

Recent advances in computer hardware and software technology have made possible various levels of integration of the basic manual data-entry processes, namely, "*keying*" (or "entering into the computer") the data, *verifying* the keyed-in data, visually and/or with the help of the computer, and even *checking* the data *against* relevant *reference data* that were previously stored. (See *Figure 5.*) In any of these arrangements, statistical workers may still pass the data, encoded into forms, on to the data-entry clerks. The data-gathering activity benefits from the skill and speed of the latter (however, the former may still have to go over computer-printed lists for a final check before the computer files are created/updated). On-line data entry has been shown to be a great improvement over the

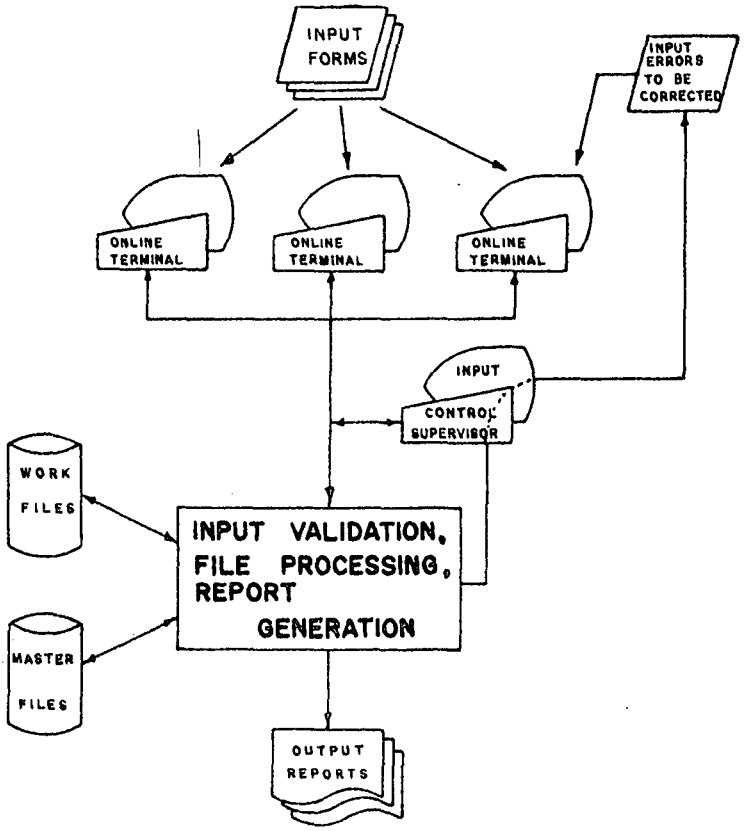

Fig. 5. ON-LINE DATA-PROCESSING OPERATION (CENTRALIZED)

batch-oriented off-line procedure. However, this organization of the "data-capture" activity is still not ideal in cases where the statistical workers are located at a considerable distance from the data-entry shop. The encoded forms have to be transported to the data-entry shop and may get lost or misplaced in the process; the computer printouts also have to be transported back to the statistical workers for their visual checking.

Of course, the statistical workers may themselves key-in the data at the terminals. While this is technologically feasible, in general, it is quite costly because of many other factors that must be taken into consideration (e.g., statisticians are paid higher than data-entry personnel, continuous use of expensive data-entry equipment is not possible, etc.)

Data-Capture at Source

In many industrial applications, errors due to human intervention during data-gathering have been eleminated through *source-data automation.* (Source-data automation refers to the creation or preparation of input data for the computer, automatically). Equipment that directly sense operating conditions (e.g., temperature, pressure, fluid flow, position, number of items that pass through a point in the plant or factory, etc.) are connected to micro-processor-based control systems and can therefore directly transmit data to the data-computer for further statistical analyses.

This concept can be adapted for commercial and financial application (e.g., inventory control, bank tellering) in business and government enterprises as well as in other social research applications (e.g., surveys, interviews, etc.). The key departure from industrial source-data automation is only that the statistical worker (or information/system analyst) is the one who enters the data into computer readable medium at the time (or close to the time) during which he/she gathers the data. There are quite a number of alternatives for this scheme; we described a few typical ones below.

A *"dumb" on-line data-terminal* may be connected to mini-computer or main frame via telephone line and data entered at the terminal is immediately transmitted to the central computer. On the other hand, *"intelligent" on-line terminals* can store and process a limited amount of data, and then at the same time (*on-line mode*) or at later time (*off-line mode*) hooked-up to transmit data to as well as receive responses or reports from the data center. This type of terminal is quite expensive and, as indicated, requires the support of costly communication hardware and software facilities.

Worth looking carefully into is *going off-line completely.* There are inexpensive data-entry devices that can accept data through key boards and optionally, do some basic processing, and store data in cassettes or floppy disks which later can be sent to the data center via cheaper modes of transportation. Such devices can, of course, also read from the cassettes or diskettes and can display the date for viewing. With such devices, verification of the data can be done in the field by the data-gatherer and thus, possibly, eliminate most computer validation/verification procedures at the data center. A notable variation of such a device is one that allows *hand-printed input.* For example, a device called MICRO-PAD allows alphabetic and numeric hand-printing of standard characters which are simultaneously displayed as they are hand-printed.

For capture of voluminous data (e.g., in national surveys) we are just beginning to explore the potential uses of optical pattern/mark recognition machines. For instance, the same OMR (optical mark reader) machine used (in a very simple mode) in the NCEE when combined with a computer-printed form will enable us to economically handle certain applications which require repeated responses overtime from the same sources. (here, we pre-print on the data-gathering form previously gathered information together with an *identification code* that uniquely identifies the form. After the form is accomplished, only the identification code and the additional data need be read by the OMR since the other data are already in the computer and were, in fact, pre-printed on the form only for the convenience of the data-gatherer/respondent.

There are countless other devices that are available and various schemes of combining such devices with the computer still awaiting invention. What is needed is man's imagination on how such hardware and software technologies can best be configured for the particular set of applications at hand. *Figure 6* shows the systems architecture of a microcomputer-based system that may be considered for data-entry operation at or near the data source. The "data-capture unit" in this configuration could be a very simple and inexpensive device that, incidentally, I think, can easily be manufactured in the Philippines right now!

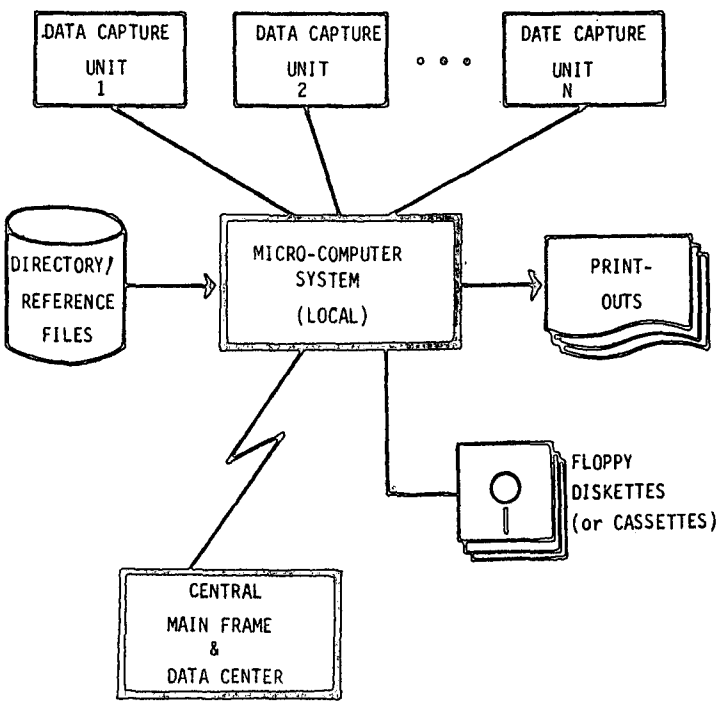

Fig. 6. INEXPENSIVE LOCAL DATA-CAPTURE SYSTEM WITH SOME PROCESSING CAPABILITY

The systems architecture of a DBMS is shown in *Figure 7.* Conceptually, it is rather simple, however, in any of its implementations DBMS requires no less than professional attention.

Without promising a thorough discussion, shown in *Figure 8* are the basic components of a database system and its "object application." Please note right away that the lower portion of the diagram attempts to depict the computer programs *as they are being executed.*

Apart from the necessity of understanding the usual components of "input data transaction," "output report," and "user process," we note that the DBMS user has to contend with many other components including those indicated in the diagram as "system data." The diagram also indicates that "user data transaction" and "output report" processes may be run separately from the computational and data manipulation processes.

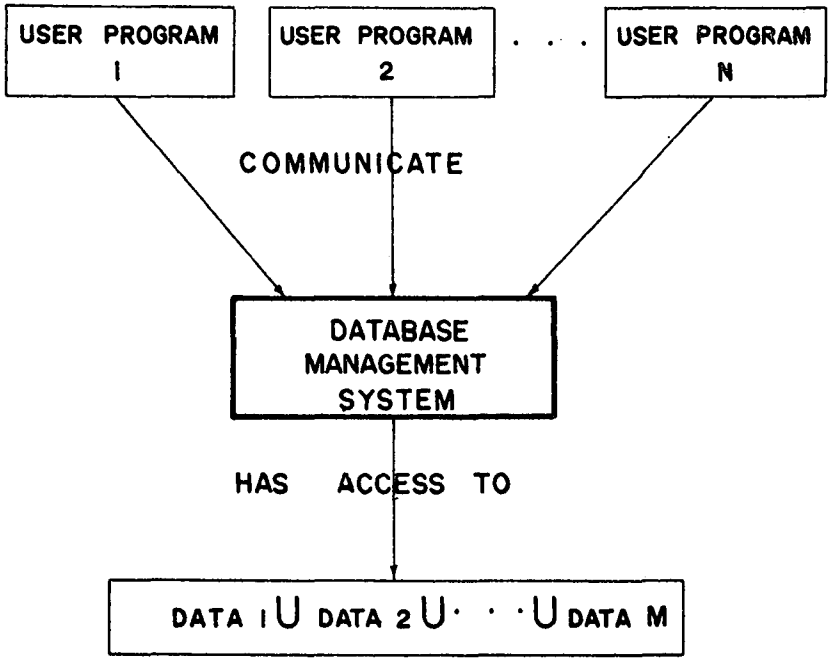

Fig. 7. DATABASE MANAGEMENT SYSTEM

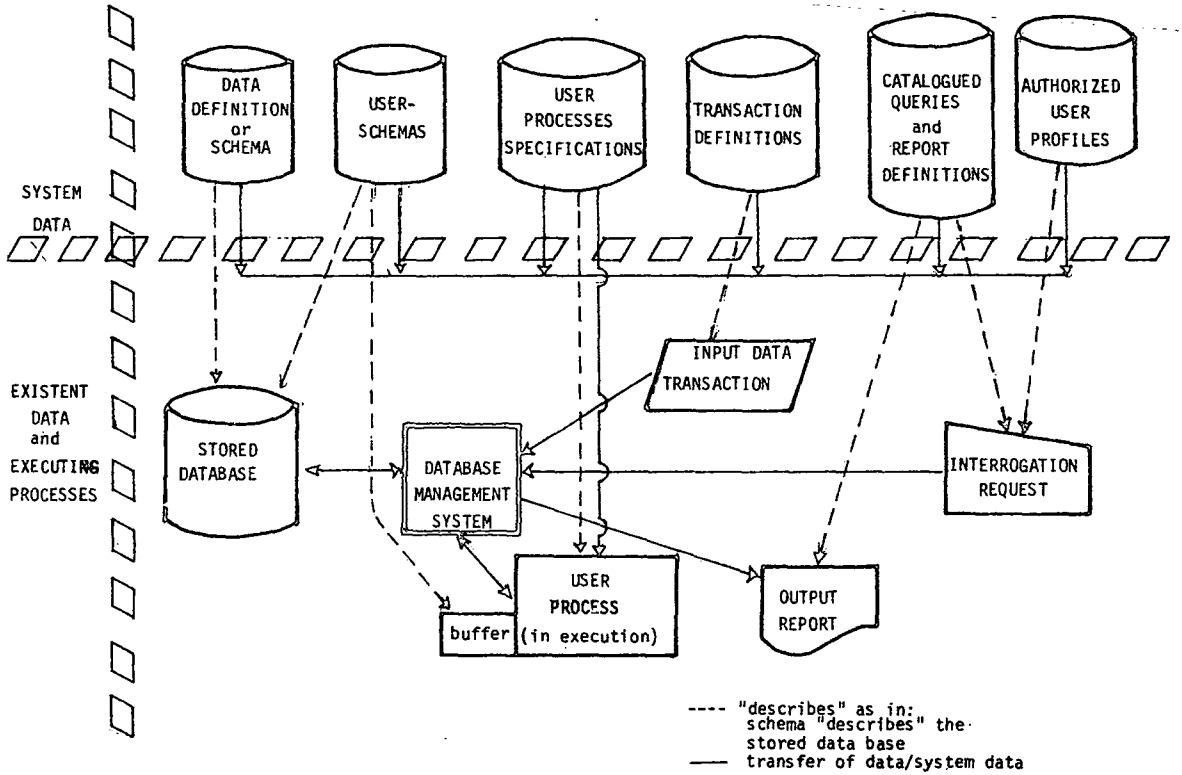

Fig. 8. COMPONENTS OF THE OBJECT APPLICATION/DATABASE SYSTEM

The complete *logical* and *physical* definition of the stored database is called the database SCHEMA. A USERSCHEMA is the view of a particular user or application process has in a portion of that database. Using the DBMS software, the systems programmers and database designers, who are concerned with system performance, can effectively specify how the data are positioned in the hardware, how data are indexed or located, and what data compaction techniques to use; the database administrator and those systems analysts who are concerned with all of the data and their relationships can easily specify such data elements and relationships in the database (logical definition); and the application programmers, who are usually concerned only with portions of the database, can also very readily specify the data they want for their respective programs. In other words, DBMS provides many features not only for effective storage and maintenance of data in the computer but also for ease of retrieval for multiple applications processing.

There are other concerns than what we have indicated above. We hope that by stating the objectives of database technology we can have an overall view of these concerns. These objectives are:

1. to make an *integrated collection of data available to a wide variety of users;*
2. to provide for *quality* and *integrity* of the data;
3. to insure retention of *privacy* through *security measures* within the system; and,
4. to allow *centralized control* of the database, which is necessary for efficient administration.

The terms in italic in this enumeration of objectives have quite complex connotations, too technical to elaborate here. May I refer the reader, who may be interested to pursue this further, to an excellent technical tutorial on the subject, ACM Computing Survey's special issue on Data-Base Management Systems, March 1976. For the purpose of acquiring basic database knowledge, James Martin's book, Principle of Data-base Management (published in a low-cost edition by Prentice-Hall of India in 1977) is as good as any book on the subject.

The development of DBMS software was motivated mainly by two parallel development:

(1) *Direct access storage devices* – the storage capacity and the cost/ performance of these equipment have improved tremendously (the cost per byte had decreased by a factor of 200 over the last 15 years and by a factor of 100 during the last 3 years);

(2) *Application requirements* in all fields, especially in business and industry, have become more complex and more sophisticated.

In conclusion, in statistical work, where massive volumes of data are to be processed, it seems that many possible statistical and analytical studies may be undertaken by different data analysts, if only the data were more readily available to them. Database technology can be a major tool in this direction. The technology has matured enough and it is now time to make the decision to seriously try it out.

11 October 1981
Mandaluyong, Metro-Manila