Journal of Official Statistics, Vol. 11. No. 1, 1995, pp. 115–127

Changes in Statistical Technology[1]

Wouter J. Keller[2]

We discuss the changes in the organization of official statistics as a result of the changes in information technology (IT). It is argued that IT will have a great effect on our external relations, as well as on the means to improve these relationships. On the input side, problems with response burden, in particular for establishments, will require other ways of data capture than those used today. On the output side, our customers will demand better access to aggregated and unit record statistical data, as well as more timely and better coordinated statistics. This will require better use of technology, but also, it will place more emphasis on the concepts we use.

Key words: Information technology; Official statistics; EDI; electronic dissemination.

1. Introduction

The organization of the production of official statistics is changing rapidly as a result of new developments in information technology (IT). The rise of inexpensive and fast microprocessors, the presence of large storage facilities like hard disks and digital tape, and the realization of the electronic highway as a means of communication with nearly unlimited bandwidth will not only change society as a whole but also the technology used by statistical agencies.

From the three aspects mentioned, i.e., data processing, data storage and data transmission, it is the third aspect which, in my opinion, will have the greatest effect on the organization of our work. Fast electronic communication allows information to become available independent of time and place. Time emerges as an independent factor, since electronic messages can be sent at any moment and received at nearly the same moment, while a store-and-forward mechanism allows the receiver to pick up the message at his or her own convenience. Place also emerges as an independent factor, because in the future information will be transported at such speeds that it will be available worldwide at roughly the same point in time. The result is information that can be shared by anyone, at any moment, and at any place.

In the future the huge processing and storage capabilities that will be available to individuals and the facilities for concurrent data sharing have a dramatic effect on both the kind of work people will do as well as on the organization of their work

[1] This paper is a slightly revised version of the one presented at the ISI-conference on "The future of Official Statistics," September 1994 in Voorburg, The Netherlands.

[2] Statistics Netherlands, Research and Development Division, Prinses Beatrixlaan 428, P.O. Box 959, 2270 AZ Voorburg, The Netherlands.

in relation to other workers. Compared to the industrial revolution, the informational revolution has its own characteristics.

Industrial	*Informational*
Manufacturing	Servicing
Materials	Information
Pollution	Privacy
Centralized	Distributed
Hierarchy	Networking
Functional orientation	Process orientation
Batch-wise	Interactive
Macro-cycles	Micro-cycles
Sequential	Parallel
Specialization	Generalization

Even in manufacturing, the greatest assets will be related to information. Information about the products, the markets, the customers, the suppliers, etc., will be more strategic for an enterprise than the manufacturing itself. Repetitive, batch-wise production of large quantities of a single product will be replaced by production on demand, where the needs of an individual customer will drive the process. This process, in turn, is mainly oriented towards information, not production. In this world, repetitive tasks will be done by machines, computers, and the like. People will focus on the information, the knowledge itself: finding it, interpreting it, and combining it, and producing new information and new knowledge. And because any repetitive, rule-based and deterministic decision can be done better by computers, people will focus on the unstructured work: the "fuzzy" combination of information tempered by creative reasoning and decision making, resulting in new information, or more precisely, knowledge.

To enable machines to do the repetitive work, the exact meaning of information should be known. Often, we need well-structured information, that is information that obeys strict definitions and standards, both syntactically (in its appearance or form) and semantically (in its meaning). When we get information from others, as we do in statistics, the translation of information from different sources to make this information comparable and processable commands a great deal of effort. But before turning to this aspect (in relation to Electronic Data Interchange, or EDI), we should first look at present changes in statistical processing, which in turn suggest future changes in the organization of our work.

The intensive and more sophisticated use of IT increases efficiency and data quality, but it also creates new problems. Due to the decentralization of data processing activities there is a need for data sharing and standardization of tools. Furthermore, since most data processing can now be carried out by statisticians in subject-matter departments, there is a need for an integrated set of user-friendly data processing tools. This requires a reorientation of the role of the Electronic Data Processing (EDP) department. But it also requires changes in the organization of official statistics.

This paper focuses on the opportunities these changes provide, both for the

statistician as well as for the EDP specialist. In particular, we focus on the use and effect of several new ways of survey processing such as Computer Assisted Interviewing and EDI. Since in the world of technology predicting the future even three years down the road is a difficult and uncertain endeavour, I focus mainly on the next decade.

2. The Traditional Approach to Survey Processing

In the past, survey processing was organized around a single big mainframe computer. As a consequence, there was a central EDP department, a central processing unit, central data, central EDP management, and above all, central system development. The organization of survey processing reflected this central approach. Data collection and data entry, editing, imputation and tabulation were separate processes, often done in a batch-wise manner, with several batch cycles per process. These individual processes were often handled by separate groups or departments, with different types of specialization. As an example, let us look at an important step in the processing of survey data: editing the data.

Although the data editing processes differ from survey to survey, some general characteristics can be observed which hold for nearly all surveys. The traditional editing process is as follows. After collection of the forms, subject-matter specialists checked the forms for completeness. If necessary and possible, skipped questions were answered and obvious errors were corrected on the forms. Sometimes, the forms were manually copied to a new form to allow for the subsequent step of fast data entry. Next, the forms were transferred to the data entry department. Data typists entered the data in a dedicated data entry computer at high speed without much error checking. After data entry, the files were transferred to the mainframe computer system. On the mainframe an error detection program was run. Except for errors that were automatically corrected, detected errors were printed on a list. The lists with errors were sent to the subject-matter department. Specialists investigated the error messages, consulted corresponding forms, and corrected errors on the lists. Lists with corrections were sent to the data entry department, and data typists entered the corrections in the data entry computer. A file with corrections was transferred to the mainframe computer. Corrected records and already correct records were merged. The cycle of batch error detection and manual correction was repeated until the number of detected errors was considered to be sufficiently small. After the final step of the editing process the result was a "clean" data set, which could be used for tabulation and analysis. Detailed investigation of this process led to a number of conclusions. These conclusions are summarized below.

First, various people from different departments were involved. Many people dealt with the information: respondents filled in forms, subject-matter specialists checked forms and corrected errors, data typists entered data in the computer, and programmers from the computer department constructed editing programs. Transfer of material from one person or department to another could be a source of error, misunderstanding and delay.

Second, different computer systems were often involved. Most data entry was

carried out on minicomputer systems, and data editing programs often ran on mainframes. Transfer of files from one system to another caused delay, and incorrect specification and documentation could produce errors.

Third, not all activities were aimed at quality improvement. A lot of time was spent preparing forms for data entry, but not correcting errors. Subject-matter specialists cleaned up forms to avoid problems during data entry. The most striking example was assignment of a code for "unknown" to unanswered questions.

Another characteristic of the process is that it entailed "macro-cycles." An entire batch of data was run through cycles: from one department to another, and from one computer system to another. A cycle of data entry, automatic checking, and manual correction was repeated several times. Due to these macro-cycles, data processing was highly time-consuming.

Finally, the structure of the data had to be specified in nearly every step of the data editing process. Although essentially the same, the "language" of specification could be completely different for every department or computer system involved. The questionnaire itself was the first specification. The next dealt with data entry. Then, an automatic checking program required another specification of the data. For tabulation and analysis, again another specification was needed. All specifications dealt with descriptions of variables, valid answers, routing and possibly valid relations.

3. Reengineering the Survey Process

Before looking at new ways of processing surveys, we need to consider the main force behind the new methods. During the last decade, dramatic changes in the IT environment occurred. Of the triplet data processing, data storage and data transport, the changes in the first, data processing, are the most appealing. These changes are mainly the result of the introduction of the single-chip computer in the late 1970s, replacing mini- and mainframe computers. Using the performance of the old VAX780 minicomputer as the standard (1 MIPS), the present processor chips run 100 times faster. During the last few years, the increase in performance per dollar was approximately 60% a year. By the end of 1995, desktop computers with 200 MIPS processor chips are expected.

Given the dramatic change in IT, it is little wonder that organizations must also change. IT enables us to work more in teams than in departments. This is in contrast to the traditional Taylor-like organization, which relies on functional specialization, batch-work, and sequential processing. The team approach today, as seen, e.g., in the automotive and aerospace industries, is an integrated, interactive, and parallel approach where several of the traditional stages are done nearly simultaneously by the same group of people (concurrent engineering). This is made possible by more powerful computers which allow for greater degrees of interactive work. At the same time, computer networks allow for data sharing among team members.

While mainframes and large, serial, batch cycles were often the only means of processing large surveys in the past, the recent changes in IT also have an enormous effect on the organization of official statistics. Each person is surrounded by different

computers, shifting the focal point to the human-being instead of the computer. Survey processing will increasingly be interactive instead of batch-wise, allowing for much smaller cycles of development and the integration of several processes. Often processing will be done by record instead of by batch, making completely decentralized processing, even encompassing interviewing in the field, possible.

Decentralized, record-based and interactive processing is typical of future methods of data collection and data editing, as well as the other steps in survey processing such as coding, imputation, weighting, and tabulation. In the new approach, error checking and correction often are intelligent and interactive processes, carried out by a subject-matter specialist on his or her desktop computer. The traditional batch process in which the data set is processed as a whole, is replaced by a record-oriented process in which records are dealt with one at a time. So, macro-cycles are replaced by micro-cycles: the subject-matter specialist enters a record, and keeps working on it until no further error messages are produced. When the specialist is finished with a record, it should be error-free and ready for tabulation and analysis.

The collection of these new methods is often referred to as computer assisted data processing. Preferably, all software required for survey data processing is part of an integrated system. An integrated system for survey processing should be based on a powerful language for the specification of questionnaires. This specification is the "knowledge base," containing all knowledge (often called the metadata) about the questionnaire and the data themselves. The system should be able to exploit this knowledge, i.e., it must be able to automatically generate all required data processing applications. On the one hand this means an automatic generation of software for data collection, data entry, and data editing, and on the other hand an automatic generation of interfaces for other data processing software, e.g., for tabulation and analysis. In this way repeated data specification is no longer necessary, and consistency is enforced in all data processing steps. The most important element here is that both the data and the metadata are machine readable. This allows the computers to process automatically throughout all steps of the survey.

4. Computer Assisted Interviewing

While traditional data collection was very much a paper and pencil exercise, the modern mode is Computer Assisted Interviewing (CAI). This allows for the integration of various traditional steps such as data collection, data entry and data editing into one interactive cycle. This is most pronounced in Computer Assisted Personal Interviewing (CAPI) where the interviewing process, including routing and checking, is guided by the program in the interviewer's laptop computer. But also in Computer Assisted Telephone Interviewing (CATI) and Computer Assisted Self-Interviewing (CASI), the integration of what traditionally were separate steps results in a clean, machine readable record directly after the completion of the interview. If for some reason mail out/mail back is preferred (e.g., when the respondent needs access to records, etc.) the returned paper questionnaire can be entered and edited in one step with a Computer Assisted Data Input program (CADI). This program provides an intelligent and interactive environment for both entry and editing of data collected via paper and

pencil questionnaires. Whatever the form of data collection, the result will be a "clean" data file, i.e., a file in which errors cannot be detected.

In addition to changes in the interviewing process, the survey design will also change as new technologies are introduced. CAI exemplifies some of these changes. Typically, the design and testing of the interviewing program, including all the rules and edits, are more time-consuming than is the case for paper-and-pencil interview. On the other hand, post-collection processing goes substantially faster, which boils down to the bulk of survey work occurring early in the survey process.

The design of computer assisted interviews looks very much like the design of a traditional computer program. The tools used are what we call authoring programs, which allow the designer to control all aspects of the interview. In some ways this resembles the design of an expert system: each questionnaire is defined by specifying "facts" like questions, answer categories, etc., and "rules" to preserve consistency and a correct routing (skipping pattern).

Although CATI, CASI and CADI are conducted from a single central location, as opposed to itinerantly as in case of CAPI, it is preferable to use similar software and hardware for all computer assisted applications. This allows for multimode surveys where, for example, part is done by CATI and part by CAPI. The requirements of portability, cost effectiveness and interactive use are best solved by portable and desktop computers, preferably connected (wireless?) to the office in a Local Area Network for easy data sharing.

The most advanced method is CASI. In its simplest form, the respondent interacts with a computer program similar to the CAPI program for the interviewer. CASI and Electronic Data Interchange (EDI) are in many ways similar and we will turn to EDI after considering CASI in more detail.

5. CASI: Electronic Questionnaires

In contrast to CAPI, CATI, and CADI, CASI uses an electronic form which is completed by the respondent without engaging an interviewer. Since this necessitates that the respondent has access to a computer, CASI will often be used for establishment surveys, like the monthly foreign trade survey. To lower the response burden, a lot of the filling-in can be done automatically, by using available data (e.g., of previous months). The CASI program is linked to electronic administrations on the same or on different computers. This link could be done dynamically, for example, if the CASI program had direct access to the electronic administration. Networks (LAN/WAN) or other data communication means (like modems on telephone lines) could be used. Because of the technical and conceptual problems outlined below, one can also opt for a static link, using file exchange (export/import) between electronic administrations and the CASI program. To communicate with statistical bureaus, data communication over phone lines, electronic networks (e.g., by Internet, X400) and even floppy disk mail exchange can be used.

When supplying electronic questionnaires like CBS-IRIS (the form used for the Dutch monthly foreign trade survey), the statistical bureau can be accused of unfair competition by commercial software producers. The bureau could retort that they

would welcome statistical modules in commercial administrative packages, and even would withdraw its own CASI product from the market if commercial vendors could provide the same functionality. To support this claim, one might supply all software vendors with development kits to build CASI-like modules into their packages so that respondents can produce statistical messages directly from their electronic administrations (there is a problem of certification of messages here). Or, software vendors could augment or add to the bureau's CASI program. However, we think that "teasing" the software suppliers by first supplying the CASI program free of charge to respondents will also work well.

6. Data Capture in the Year 2000: EDI

The aim of reengineering the survey processes is, among others, to improve relations with the external world. Both respondents who provide data and users of that information comprise our customer base. On both sides we need a coordinated approach in order to circumvent the present day problems.

On the input (respondent) side, I focus on data capture at establishments (enterprises and institutions), since household/personal surveys are conducted effectively using CAPI and CATI techniques. It is of vital importance that we improve our relations with establishments. In my opinion it is absolutely unthinkable that we continue sending out numerous paper forms to establishments and just wait for a response. Here is a potential role for EDI. But before we look at this challenging new technology, we should look more carefully at the present process and its shortcomings.

Today, a medium size establishment (e.g., an enterprise with, say, 50 employees) will receive often more than 40 paper questionnaires (envelopes) a year from Statistics Netherlands (SN). Most of these questionnaires come from different statistical departments within SN.

As said, an establishment receives numerous paper questionnaires during the course of a year, completes them by retrieving information out of its own (electronic) administration, and mails them back. Such collections could use CADI to arrive at machine readable, clean records transferred directly to our computers. While the later steps are done more efficiently than in the past, the long data collection period (often several months for a yearly mail survey) and the cyclic nature of the process (where often the workload is spread till the next survey period, which comes over a year) often lead to a total calendar time of over a year. At the same time, the respondent experiences a high level of response burden. To make an already bad situation even worse, respondents must do a great deal of "translating" as survey concepts and definitions are different from, sometimes even foreign to, the concepts and definitions that respondents use in their own bookkeeping and administration.

While CASI can solve some of the problems experienced today, I do not think CASI or CADI is the ultimate solution for establishment surveys. In both CASI and CADI, data capture is done by a separate instrument, the electronic questionnaire, which still forces the respondent to make a translation between its own administration and the statistical questionnaire. Even if we add technologies like

optical character recognition or voice recognition to ease the data capture, we are still faced with a translation phase. Actual conditions must be reformulated so that they fit into the framework of the questionnaire, be it by the interviewer in the case of CAPI, or by the respondent in the case of CASI. We argue that in the year 2000 the burden of translation will probably rest on the statistical bureau. Information will be captured directly from electronic administrations and registers, without enlisting the time or effort of the respondent.

A medium size firm receives from SN dozens of questionnaires a year, often with similar questions but not always with the same definition for a particular concept (e.g., turnover or number of employees). In other words, the statistical bureau must improve the internal coordination of the statistical definitions used by different departments. But even within one department, there might be different definitions of the same concept if we look at the questionnaires, the data bases, and the various publications based on one or more surveys.

Furthermore, entrepreneurs often find our concepts to be different from their own or from those used by other public authorities like the IRS or the social security services. At the same time, there might be different administrations within the establishment (e.g., for personnel, finance, logistics, etc. but also for the corporation and for its divisions) each using slightly different concepts.

If we look at the logistics of the data collection process for establishments, we often see mail out/mail back processes, necessary to allow the respondent to check his or her records and make the necessary translations between administrative and statistical concepts. This step is heavily laden with response burden, and further, a nontrivial burden which totals several hundred hours a year for a medium size enterprise in the Netherlands. This translation not only involves different concepts, it also involves different media: the machine readable electronic record keeping is transferred to paper questionnaires. These paper questionnaires are sent back to SN where the reverse process takes place (from paper to machine readable) using a CADI-like system. What would be more natural than to skip the paper form entirely and do the entire data collection process electronically, by connecting the respondents' computers with our own?

Obviously this is easier said than done. Surprisingly the main difficulty is not technological, but conceptual (nevertheless, we do not mean to suggest that the actual electronic data interchange using data communication facilities is trivial). To connect Statistics Netherlands' computers to an establishment's electronic administration, we both must use the same concepts. If every establishment has its own concepts, we might do the translation at the bureau, but that involves knowing the concepts and the translation rules for hundreds of thousands of establishments. Clearly this would require an immense data base or expert system and huge maintenance efforts to keep rules current. Or it would require enforcement of strict standardization of administrative (and statistical) concepts at the national level.

Statistical offices are looking for alternatives to circumvent the problems described above. We believe there are two possible solutions:

1. Use of electronic CASI forms which make the data collection process easier for

establishments and still leave large parts of the (automatic) translation process up to the respondent. We looked at this solution in the last section.

2. The translation process needs to be shifted to what are called intermediaries (or focal points) between the respondents and the statistical bureau.

Examples of such intermediaries are software packages, service/administration offices, public and other registers, etc. For example, if we could work with software suppliers to incorporate statistical modules into commercially available accounting packages most conceptual translations could be done within the software itself. Any establishment using such a package for its administration could produce the desired statistical concepts in an electronic format, without any manual translation of concepts or media.

Additionally, if we could work with administrative service bureaus, which take care of the administration of small to medium size enterprises, again by using a single translation procedure (per bureau), we could handle hundreds of enterprises. The same holds for other service organizations like those which administrate personnel and wages for other enterprises; as with the other service bureaus, we could collect the necessary information from these bureaus instead of sampling individual enterprises. At the same time, savings would result from needing only one translation scheme per bureau.

Finally, as is already taking place on a large scale these days, we should look to other public institutions like IRS and social security, where a great deal of standardized information is available about persons and individual establishments, and again only one translation to statistical concepts is needed. Some of these intermediary sources are incomplete, making it difficult to translate to the exact statistical concepts. But instead of putting the burden on our suppliers, the statistical bureaus should examine other statistical concepts or other ways (by estimation, for example) to match statistical concepts with those readily available from suppliers.

If a statistical bureau wants to thrive in the year 2000, it must be flexible about concepts and not force the respondent to work with concepts that seem unnatural or fail to capture the respondent's experience. Unfortunately, the world is not administered according to concepts of the National Accounts! At the same time, all of the different surveys (40 envelopes a year) going to the same establishment should be coordinated: How necessary is it to have each statistical department send out its own questionnaire? Is it possible to combine these forms, thus lowering the response burden? Even when data collection and survey response are mandated by law, we should be careful of draining our capital: the respondents. Furthermore, no law mandates that the answers be of high quality. One result of the heavy burden we place on establishments is that questionnaires are returned to SN with increasingly dubious content! We should look for new ways to improve the quality and the speed of our data collection, and ways to lower the response burden, in particular, for establishments. Coordinated and electronic questionnaires, one for each electronic administration together with EDI-more efficient translation processes might be an answer for the future.

By the year 2000, the era where everyone was "willing" to do the translations and

fill in paper forms will be over. I think two problems will have to be resolved: from where will we procure the necessary information (preferably in electronic form) and how can we minimize the burden of translation for ourselves. I would not be surprised if statistical agencies end up paying other institutions (including public agencies like IRS and social security institutions, and private companies like large distributors and others with both economic and other data bases) for the use of their data. In the meantime, it might prove profitable to invest in the development of statistical modules in commercial software, data capture technologies like remote sensing (following, e.g., trucks), satellite observations, and other modes of data collection which do not rely on a respondent for their success.

7. Further Processing

After the interviewing and editing stage, several other steps are necessary. While the first stage can be viewed as the (often parallel) processing of individual records (e.g., one per CAPI session), the analysis stage in survey processing often involves the entire batch of records. Something between these two worlds is the coding operation, where codes are assigned to descriptions, e.g., for type of occupation, branch of industry or education. Although automated coding is preferred, in practice a substantial number of codes cannot be assigned by coding algorithms and are therefore done interactively.

Most coding algorithms focus on lexical or logical analysis, using already classified cases as a training set. Others focus on simpler matching rules (i.e., nearest neighbor) but industry and occupational coding can also be done using larger dictionaries (with descriptions of employer, industry, occupation, and duties) that require massive parallel computers.

Other steps in the analysis stage involve batch editing, imputation, post-stratification/weighting, and tabulation. Although batch editing is less and less necessary for CAPI/CASI interviews (since all checks can often be done during the interview), most mail surveys benefit from editing. An alternative to interactive editing is offered by fully automated edit and imputation programs. Some systems deal only with quantitative (mainly economic) data, while others focus on qualitative (often demographic) data.

There is no consensus on the question of batch vs. interactive editing. While there is some evidence that batch is more cost effective, in particular, for large surveys like censuses, experience at SN shows that interactive systems for economic surveys, like industrial production and foreign trade perform very well indeed. Batch programs no longer require a mainframe, and even large samples can be processed on super micros or workstations. Since so much of the editing and imputation involves rules and reasoning, I expect rule based systems to be used more frequently in the future.

Poststratification will gain importance in surveys of individuals, in view of increasing nonresponse. This holds also for mail surveys of establishments. A good sample frame will become essential in both cases. Joint endeavours between statistical organizations and public and private institutions (like IRS and insurance companies) will be essential in ensuring an up-to-date frame. EDI will also enable us to connect to external data bases, using data from these registers in many cases where we now use sample surveys.

Finally after post-stratification, the survey results are tabulated and published. As with many of the above steps, the programs involved should connect seamlessly. This, in particular, holds for the metadata (information describing the characteristics of the data itself, like type, format, labeling, definition, etc.). One of the greatest advantages of most CAPI/CASI programs is that after the survey design stage, it is seldom necessary to respecify the data or the results of the data processing. Often the output of these programs can be used directly, without respecification, as the input for other analysis programs like SPSS or SAS.

Some systems even provide their own integrated suite of programs for nearly all the steps in survey processing. The main part in such a system is the dictionary or knowledge base, specified by the subject-matter specialist, which describes the questionnaire: questions, possible answers, routing, range checks and consistency checks. With the specified questionnaire as input, the system automatically generates the necessary set-ups for all further data processing.

8. Dissemination

Most bureaus provides hundreds of statistical publications from several hundreds of surveys. This amounts to millions of figures, thousands of tabulations, and many, many different sources of information. Except for some special publications (like the National Accounts), each publication deals with a particular topic only, and users are confronted with "a gold mine" of data, which, nevertheless, can prove inaccessible, or at least daunting to access. Someone interested in, say, automobiles, has to look in more than a dozen publications to get a full picture, which would encompass the production of cars, exports and imports, use (time and mileage), energy consumption, environmental effects, etc.

At the same time, we sell only a small number of copies of each individual publication, often without recovering full dissemination costs, let alone collection costs. And finally, while users appreciate our impartiality and accuracy, they complain about the lack of timeliness.

Statistical agencies need to focus more on our external relationships, and make our activities fully customer-driven (with the respondents as our input customers and our users as our output customers). With respect to both the input and the output, this means more integration and coordination of what we do, both with questionnaires and publications.

While EDI and CAPI focus on more efficient and coordinated ways to collect data (the input side), other newly developed tools focus on data dissemination (the output side). Currently, our publications take many different forms: printed paper, floppy disks and CD-ROMs, automatic and human voice response, press release, videotex, etc. At the source of these different media, there is aggregated statistical information, often in machine readable form, e.g., as the output of survey processing systems. We need some sort of "one-stop" dissemination data base situated between the internal processing systems and the outside world, capable of producing many different media from one source, in a consistent, timely and efficient way. Such a system could provide the platform for better coordinated statistical data and easier access to the wealth of information at statistical bureaus.

As in CAPI systems, we can distinguish between the data themselves and their descriptions, i.e., the metadata. Just as CAPI systems focus on the individual data and metadata processed in the data collection and editing stages, the dissemination data base focuses on the aggregated data and their metadata. These metadata comprise both the definitions of the published items and their properties, as a description of the survey itself and how the items are derived.

Each item (e.g., number of employees) is often available for different domains, defined by crossings of discrete, categorical variables, like sector, region, or time. An important mechanism to coordinate the dissemination of statistical data is the standardization of these categorical variables, leading to classifications, e.g., for branches of industries, commodities, regions, etc. The basic representation of data used in such a data base is therefore the multi-dimensional matrix (sometimes called cubicles) where one dimension reflects the different variables (e.g., number of employees, profit, prices), a second one the (discrete) time axis (e.g., years and months), while other dimensions correspond to various classifications (industries, commodities, regions, etc.). The items inside the matrix reflect the measurements (number of) on a certain variable (employees) in the domain defined by the crossing of the categories on the other axis (in industry x in region y at time t). Often, categories are classified into different systems of detail (e.g., an n-digit industry classification, with $n = 1..9$) which are often (but not always) hierarchical, resulting in levels of classification.

Metadata (descriptions) in this data base of cubicles can refer to the total matrix, to the axes and their variables and categories, and to the individual items inside the matrix. Particular problems of meta data arise when the definitions of certain categories (such as regions, industries, commodities) change over domains and, in particular, over time. For example, take a region like a municipality. Not only is the number of inhabitants in Amsterdam in 1980 different from that in 1990, but the definition of Amsterdam itself also differs between the two years (e.g., because of border corrections). Similar problems arise when certain items are available only for certain categories or classification levels, making the comparison of different items in various domains difficult or impossible.

In such a data base of objects, most important is the high-dimensional object, or cubicle, explained above. The data base will contain many, many different cubicles, which might all share similar classifications along some of their axes. Besides these high dimensional objects, also simple ("flat") two-dimensional cross tabulations, as shown in most traditional statistical publications, have to be stored and presented in the data base, as well as (one dimensional) text objects like press releases. All this information is documented (metadata) inside the data base on various levels (from the total object down to the individual items or cells). A classification of objects into the well-known statistical domains (such as economic, social, and demographic statistics), and derivative classifications (production, environment, labour-market, well-being, etc.) make navigating through this immense data base feasible. A strong tool to locate desired data is a data base-wide keyword (thesaurus) system, which allows the user to find quickly the right object.

Such a dissemination data base should be based on the client/server concept, where the front end (running on a PC, outside the bureau) is separated from the back end

(the data base server, located at the bureau). Front end and back end are connected through data communication facilities (like Internet), making the wealth of statistical information available to anyone with an electronic connection.

All this holds for aggregated data. CD ROMs with microdata (at the individual respondent level) will increasingly become a separate product from official statistics. In particular, the demand for microdata from research and marketing organisations will increase greatly in the years to come. It is in this area where "the need to know" will have to be balanced against "the need for privacy." Possibly, official statistics will hold the microdata and provide semi-aggregated data on demand, meeting the demand to know while maintaining sufficient confidentiality.

9. Conclusions

Despite the display of buzzwords, I believe there is a place for multimedia, on-line transaction-processing, decision support systems, client/server, object-orientation, and parallel computing (to name a few) in statistics. But in my opinion, technology will be more important externally than internally to the statistical office. Because of the rapid changes in technology, respondents will stop filling in our paper forms and our customers will demand more than our fragmented statistical paper publications. Technology, coupled with the way we handle these pressures, will determine the future of statistics. To recapitulate:

1. We have to connect to external administrations and other data bases to collect data.
2. Customers must have better access to our huge collection of aggregated data.
3. We have to focus, possibly by partnering with other public and private institutions, on standardization of statistical and administrative concepts.
4. We have to lower response burden, in particular for establishments.
5. We must strive to better disseminate statistical data, in order to provide timely and well-coordinated statistics.

If we are unsuccessful in handling these external relationships, the private sector will assume our responsibilities. Most of the changes in our external relationships will be caused by changes in IT, while the changes in IT will provide the means to solve our external problems.

Will official statistics survive in this rapidly changing world? There is no single or sure answer. It will depend on how well we are able to strike the balance between pushing technology and being pushed by technology.

Printed in August.