Role of the information technology in official statistics

Juraj Riecan
Director, Statistics Division, Economic and Social Commission for Western Asia

ABSTRACT

The paper highlights the importance of information technologies for official statistics in the 21st century. Statistics is a profession based on collecting, managing, processing and disseminating information. Therefore, it has important demand for IT support.

The paper further presents the Generic Statistical Business Process Model[1], and discusses what IT support is needed at each stage. The paper also provides information on software tools that may be of interest of statistical offices. Some of this software was developed by national statistical offices in other regions. The business process model is an interesting tool, when considering the architecture of statistical information systems from the process viewpoint.

At the end the paper makes also suggestions related to the role of the IT department in the statistical offices and the necessity to own and control the core IT tools by the statistical offices. The paper also makes a suggestion for the future exchange of experience and expertise on issues related to statistical IT and statistical infrastructure.

A. INTRODUCTION

Statistics is a profession related to information. Therefore, the success of statistics depends on the way in which we are able to collect, process, safeguard and disseminate this information. Having entered the 21st century, the management of statistical information is unimaginable without a support of adequate tools of modern information technologies.

The last two decades witnessed an important move towards statistical information systems that integrate the tasks throughout the statistical activity, whether it is a census or survey. Leading statistical offices integrate the data repositories and processing systems, harmonise the tools used, with a view of having efficient, transparent and easy to maintain systems. This is the general trend in statistical offices at present.

However, the integration of statistical information systems requires taking into account various specific needs of individual phases of data processing, and in various subject matter contexts. Therefore, creators of the integrated statistical information systems need, as a first step, to create a model describing the functions of the statistical offices.

Let's take a look at a model that may be useful in describing the statistical operations, and think about the role of the information technology at each phase of the process. And let's base our journey on experiences of statistical offices that looked in-depth into these issues.

B. GENERIC STATISTICAL BUSINESS PROCESS MODEL

1. Specifying needs

If we take an example of any statistical survey, it responds to a concrete demand for statistical information. Therefore, at the beginning of the survey cycle, we have to specify the needs. At this stage that we will mainly communicate with the potential users of the data, usually those who initiated the task.

[1] For more information about the Generic Statistical Business Process Model see: http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model

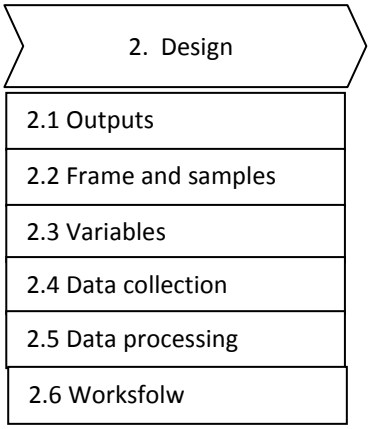
Initially, we the standard business and communication software would be sufficient in this phase, in particular, in determining the needs for information, consultations with the users and defining the output. However, we will need an access to the existing database(s), in order to check data availability. We can again rely on the standard office software in preparing the business case.

2. Design

This is the phase, when statisticians begin their own work on the future survey, census or other activity. Based on the results of the first phase, it means the specification of the needs, it is necessary to specify outputs, possibly using software for generating statistical tables.

At this stage specific statistical tools like Survey Processor (Croatia), Questionnaire designer (Australia), Blaise (Netherlands), etc. are used to perform specific design tasks including sampling design, metadata for variables, data collection (questionnaire, registers, etc.) design, choice of data processing methodology and workflow design.

The design phase is the first step towards formalising the requirements. Usually, it is possible to achieve the formalization sufficient from the statistical and IT viewpoint rather easily. In case of more complex design it is possible to use a modelling approach lithe UML (Unified Modelling Language) that is supported by commercial software like Rational Rose, etc.



3. Build

The extensive use of the information technology in the statistical business process cycle requires an attention to the information systems and their components. This should be completed well in an early stage of the cycle, and a due attention should be paid to the testing. This will ensure that the actual statistical activity will be smooth, and there will be not delays due to technological issues.

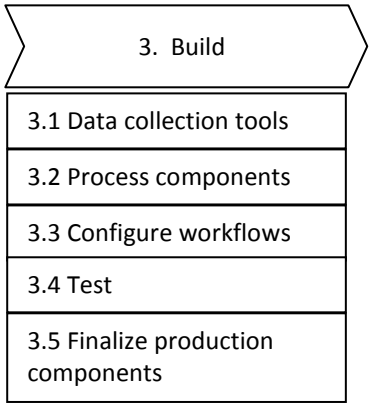
A particular focus is on building and tailoring the tools that are used in the data collection and the data processing phases. The configuration of the tools will have to take into account specifics of the statistical subject matter treated. Therefore, it is important to have a clear data model. The data model comprises

details of the multidimensional structure of variables and indicators, metadata, code lists, relationships between these components, etc.

Another aspect, when building the information system components, is to look not only in the current survey, but consider, at the same time, the reusability of components for other surveys. Vice versa, we may not need to build all components from the beginning, because they may already exist. In simple words, it is preferable to have in mind consistent information architecture of the statistical office rather than taking ad-hoc solutions. While the latter may give faster results from the immediate perspective, consistent information architecture will pay back in the long term.

So the components that will be built and/or configured in this phase include the data collection instruments and process components. These will be used immediately at the beginning of the survey. Further, we will configure the workflows (between components, units, etc.), and test all components and workflows so as to avoid unpleasant surprises during the survey itself. After the production systems are finalised the core statistical operation (survey, census, etc.) may begin.

Individual tasks in this stage may be supported by the software produced by some statistical offices, for example Blaise (Netherlands), SIV (Sweden), Quat (Netherlands), Questionnaire Development Tool (Australia) and VVIS (Estonia).



4. Collect

The data collection tools, built in the previous phase, will depend on the sources of data. In case of the classical survey and/or census, when we use traditional printed questionnaires, these may include software for a manual data entry or for an optical character recognition.

In case of data from administrative registers and records, we need either an off-line module for import and pre-processing of administrative data or a link to the administrative source itself. In either case, if we use administrative registers, it is advisable that the statistical office maintains a statistical register based on the administrative data. It means that the data collection consists primarily on the update of the statistical register, rather than on a full acquisition of all data. The two best known examples are statistical business registers (combining data from the commercial registers, licensing of small enterprises and individual entrepreneurs, etc.), or a statistical population register (updated from the population registers of the Ministry of Interior, but also from population censuses and municipal records), etc.

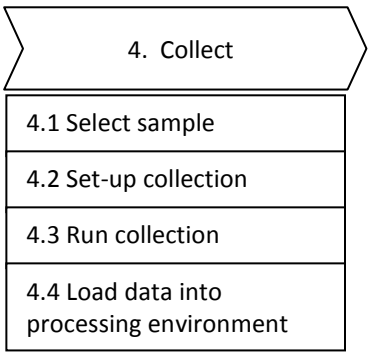
There are statistical fields that benefit from advanced technologies. For example the statistics on environment may use automatic registration of environmental variables at designated points of measurement (population, climate, etc.). The statistics on land use is a potential field of application of remote sensing tools – for example the satellite imagery.

So when we select a sample (when applicable), set-up a collection and run a collection, we ideally end up by having the data loaded into the production system. However, we should look into the quality of data and

perform at least basic data checks already at the data collection phase. The data checks may be built into the manual data entry software that would flag suspicious record. When we use scanning and OCR, we have to check first the quality of scanning and data recognition, and subsequently proceed with the basic statistical checks for outliers or otherwise suspicious data.

A specific case is the data collection via internet, when respondents themselves enter data into electronic on-line questionnaires. The questionnaire forms need to be equipped with the basic consistency checks that would flag the suspicious data and would prevent the respondent from proceeding further. In this respect let's stress that the built-in checks in on-line questionnaires should not appear abusive. We should focus only on the most important checks. If we would check for minor discrepancies, we may consider it legitimate from the data quality viewpoint, but the respondent may be discouraged, abandon the questionnaire, we would en-up with a non-response. In conclusion, the computer assisted checks in the data collection phase are important, but on the other had there is a trade off between the data quality and the risk of non-response.

The sampling is supported for example by the following tools developed by statistical offices EHE Sampling (Norway), Generalised sampling System (Canada), MAUSS-R (Italy) and Survey Processor (Croatia). The data collection setup and execution are supported by VVIS (Estonia), Blaise (Netherlands), Quat (Netherlands), SIV (Sweden) and xcola (Finland). The Dutch Blaise system also supports loading of data into processing environment.



5. Process

This phase combines a large number of tasks that are aimed at protecting the confidentiality, ensuring quality, integrating data with data already existing in databases and performing all necessary calculations and estimations. There, therefore, an important complexity of IT tools that are used in this phase. We may combine the existing statistical packages like SAS and SPSS with specific modules prepared by statistical offices (see examples bellow). In all cases it is important to use an integrated data repository and an integrated metadata system, and provide for linking and data sharing between all tools employed in this phase.

The tasks listed in the Generic Statistical Business Process model are all possible tasks. A particular survey or census may not need to perform all of them at this stage. Let's begin with standardising and anonymizing the data. This may be facilitated already in the previous phase, depending on the design and method of data collection. If data were collected exclusively for statistical purposes, we may not need to focus too much i=on standardising, unlike if the data are collected from administrative sources.

As far as anonymizing is concerned, a simple removal of direct identifiers may not be sufficient. In order to prevent potential intruders would from obtaining confidential private information, software tools for avoidance of statistical matching are available – usually produced by national statistical offices.

With respect to the coherence of statistical information system the calculations and estimations should not take place off-line on a desk top of expert statisticians. These should be supported by the system. It is also important to integrate all data into a database, so that we have a record of the raw data, all intermediary

aggregates as well as the final aggregates and estimators. The integration of data also implies the integration of underlying metadata, coding and classifying.

We have mentioned that basic checks already at the data collection phase. These allowed us to get back to the respondent in order to verify suspicious and/or missing data. At the data processing phase we may employ data editing and imputation tools that are more profound. These would allow us to go perform more in-depth evaluation of data quality. Unlike in the data collection phase, we would not go back to the respondent, but rather replace outliers and missing data with estimates. Two aspects should be kept in mind when using the editing and imputation tools. We should not use them abusively, the data should remain natural. Overdoing editing and imputation would lead to a risk of creating a synthetic data set away from reality. Another aspect is that the errors and edits should be fully recorded, and we should analyse them for systematic errors. The goal is to make adjustments t5o the design of data collection and avoid the same systematic errors in the next instance of the survey. We will come back to this at the evaluation stage of the cycle.

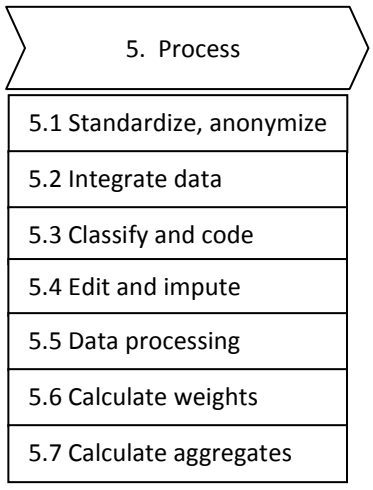
Finally we perform calculations and estimates needed for our outputs, that is deriving new variables, calculating weights and aggregates.

The standardisation and integration are supported by G-link (Canada) and RELAIS (Italy). The classification and coding process is supported by a range of software available at national statistical offices, like ABS Autocoder (Australia), Codificator automatic (Spain), G-Code (Canada), Postcode register (Netherlands), Sicore (France), VVIS (Estonia) and Blaise (Netherllands).

There is a whole variety of tools for statistical data editing. These are intended for specific editing and imputation approaches, some at the macroediting stage, some for editing microdata. More universal tools are Blaise (Netherlands) and Survey Processor (Croatia). More specific tools for data editing and imputation are Banff (Canada), CanCEIS (Canada), CONCORDJAVA (Italy), DIA V3 (Spain), ISEE (Norway), IST (Serbia), LogiPlus (Canada) and SELEKT (Sweden).

For calculating the weights and aggregates the Dutch VHM (Vullen Reference Database) may be considered along with the generalised Estimation System (Canada), Price Index Processor System (UN Economic Commission for Europe), Price System Implementation Project (Australia) and StatMx (Canada).

Other software tools that may be considered at this stage, and that were produced by national statistical offices are DIGROS (Netherlands) AND Re-GENESEES (Italy).



6. Analyse

This phase aims at creating desired outputs from the collected and processed data. Again we shall use statistical data analysis tools mentioned above. The set-up of the technology should take into account that this phase is performed by the statistical experts from individual subject matter domains. Not only that we

should have in mind user friendliness, as they are more skilled in statistics than in IT. We should also reflect their substantive knowledge into the design of the IT tools.

Again we main need a variety of standard and tailored statistical packages that we have to integrate into a coherent statistical information system. These will help us to perform statistical analysis and prepare draft outputs. We may need tools for verification of outputs (for example checking whether all assumptions were satisfied for employing the methods employed).

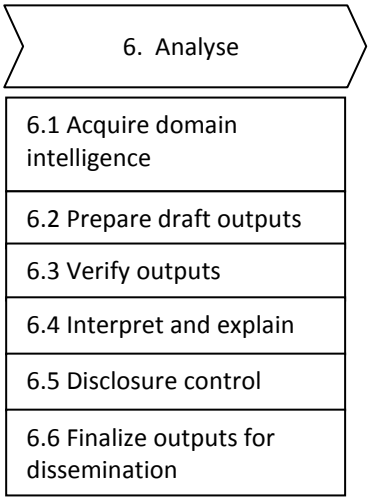
Interpreting and explaining the results is based on the personal expertise of statistical experts. However, they should have access to the metadata system and record their findings, so that these can be used when disseminating and communicating data to the users.

Before the final aggregated data and results of analysis may be presented to users, we should ensure that the private and confidential information is protected. Appropriate tool for data disclosure control help us to make sure that neither direct identification, neither statistical matching can be used to obtain confidential information.

Once the subject matter specialists finalize the outputs for dissemination, these can be passed on to specialists in dissemination and communication.

Statisticians may again use either the marketed IT tools like SPSS and SAS or tools developed by statistical offices: Demetra+ (Eurostat), G-Series, G-Tab and Price System Implementation System (Canada), PX-Edit (Sweden) and Re-GENESEES (Italy).

The quality control is supported by Stat Control (Netherlands) and EVER (Italy). The disclosure control is a specific task. Specialised software suitable for official statistics was developed by Statistics Netherlands under the name of Mu-Argus and Tau-Argus. Other disclosure control tools are Confid2 (Canada) and Rounding (Norway).



7. Disseminate

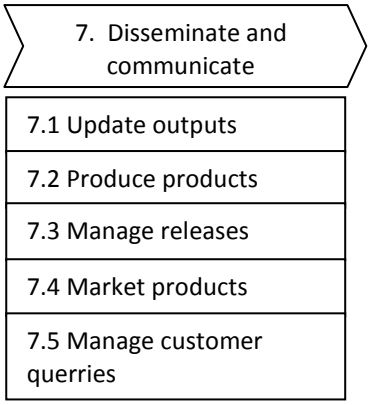
The statistics is meaningful only when it reaches the user. Official statisticians think usually about the statutory users, who already expect the data to be presented to them. However, we should not omit the public, media, researchers, etc. Therefore, this phase includes also an outreach to new users through marketing and maintaining public relations. This phase has a potential to use already existing IT tools for publishing – whether printed or electronic, and in particular over the internet. However, there is still a need for specific tools enabled with statistical capabilities. Some of these may be the ESCWA Statistical Information System (ESIS), PC-AXIS produced by a consortium of about 30 statistical organizations around the World led by Statistics Sweden, DevInfo offered by UNICEF, OECD.net that is available from the

Organizations for Economic Cooperation and development, etc. We may need to add to these GIS tools for cartographical interpretation of data, for example Mapresso, produced by the University of Zurich.

Other tools for data dissemination are SDMX tools (a consortium of international organizations), StatFlow and StatWeb (Netherlands), CoSSI (Finland), ISTAR (Italy), Business Tool Box (New Zealand), Jaxi (Spain) and REEM (Australia).

When we talk about electronic on-line dissemination, the data are normally continuously available on the internet. In that case we update the existing outputs rather than create new ones. However, we may also produce print or electronic products (CD-ROMs, etc.). Depending on the culture of the statistical system the office may have to respect a strict release calendar. Therefore, some statistical offices developed specific tools for managing the data releases.

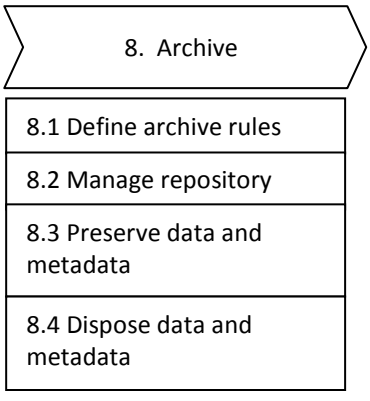
We may use standardised IT tools available on the market for marketing and for managing customer queries. The latter can be supported by tools usually used by call centres and helpdesk for managing requests.



8. Archive

Once the dissemination is completed, the core tasks seem to be over. However, taking into account the periodic character of statistical surveys, it is important to archive the information for the future round and for the institutional memory, and to feed the lessons learned to the future rounds of the survey. The archiving phase is typically supported by the traditional database management software. The records stored in the archives should be managed also by a set of rules, preferably automated. The rules primarily refer to the retention, safeguarding and access. It is important to highlight that the archived statistical information may still carry some confidential elements, and the access rules should preserve these similarly as for the "live" data.

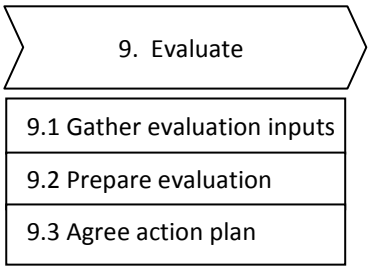
Once the rules are defined, the archiving module of the statistical information system should manage the repository, preserve data along with all associated metadata and manage also disposal of records that are beyond the specified retention period.



9. Evaluate

As mentioned above, the statistical surveys are generally periodic. However, it is important to ensure that the issues and problems tackled in the present round of the surveys do not repeat again. Therefore, in the processing phase we mentioned that all edited outliers and missing data should be recorded. These records should be then carefully analysed for a pattern of systematic errors. This is essential for ensuring data quality, because it is not meaningful to continue editing and imputation with respect to the same systematic errors in the future rounds of the survey..

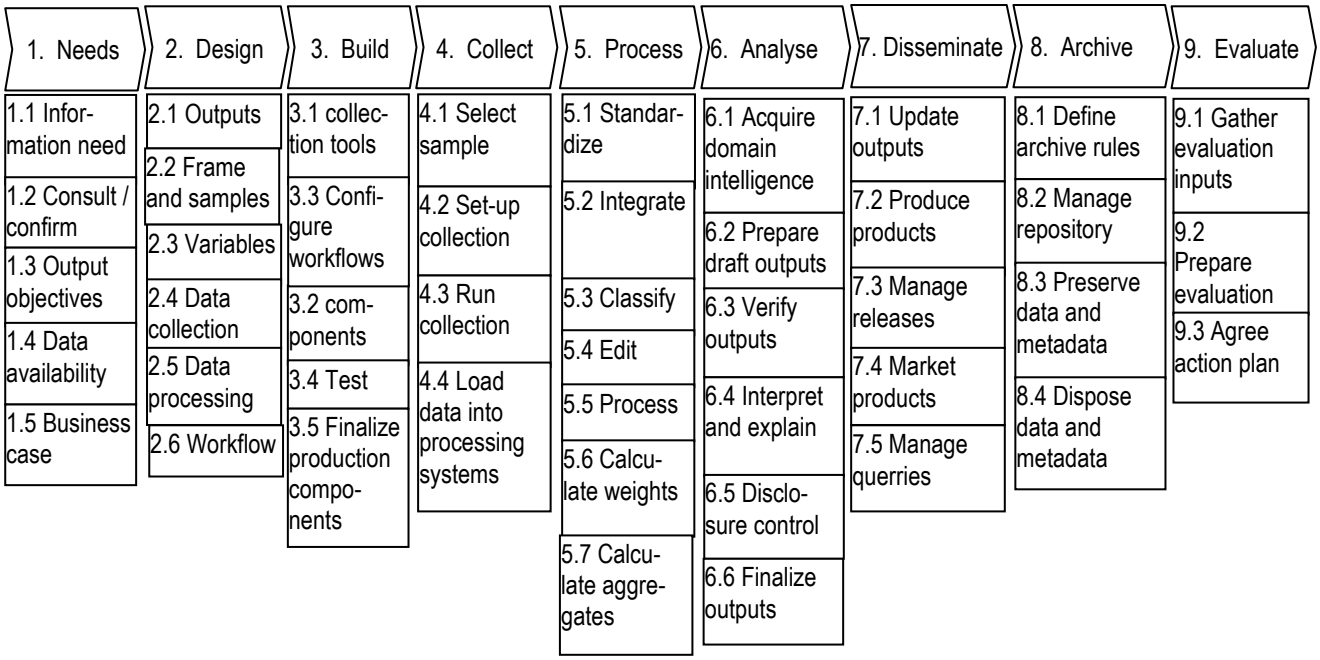
Similarly as for the initial phase, we will mostly use the standard office and communication software. The evaluation phase concludes our journey through the Generic Statistical Business Process Model.



C. SUMMARY SCHEMA OF THE STATISTICAL BUSINESS PROCESS MODEL

The following schema summarises the model presented in the previous section. It is important to keep in mind that this is a generic model of the statistical business process cycle. Each survey will comprise only those phases and tasks that are vital for that survey. It means that some of the phases and tasks may be skipped in concrete surveys.

The model is used by a number of statistical offices in order to create process oriented architecture of their statistical information systems. It was also adopted by the Sharing Advisory Board that facilitates sharing of software tools among mostly European statistical offices – this initiative is open to any interested statistical office[2].



[2] See: http://www1.unece.org/stat/platform/display/msis/Software+Sharing for more information about this software sharing initiative. The initiative is open to statistical offices, but restricted to tool developed in the framework of official statistics. Commercial software and producers are excluded.

D. ROLE OF THE IT DEPARTMENT IN THE STATISTICAL OFFICE

The basic question is whether a statistical office needs a department of the information technology or not. If we agreed at the beginning that statistics is tight with the information that is collected, processed, preserved and disseminated. There are diverse opinions about this issue. The author believes that statistical offices should be in control of the core information technology that they use.

The tools and components of the information systems should be considered according to their relationship to the core statistical business. We may summarise it in the following points.

(a) It may be possible to outsource the networking and hardware infrastructure, it is possible to implement out-off- the shelf office and communication tools for the basic tasks through an external supplier.
(b) On the other hand the statistical offices should be fully in control of the data collection, processing and analytical tools, its databases and metadata systems, and should fully control the archiving environment.
(c) It is also important that when the statistical office outsources some of the IT operation, it should be able to verify the quality of delivery by its own staff, and should not depend for the maintenance on a unique external supplier.
(d) If a software component needed for a core statistical process is supplied by an outsourced external supplier, the statistical office should be the full owner of the intellectual property rights to this component, and should have an ability to utilise, adjust and modify the component as needed.

Most of the core statistical tools are used through the phases 2. Design to 8. Archive, and the most critical phases are 4. Collect, 5. Process, 6. Analyse and 7. Disseminate. These should be fully in the hands of the statistical office.

E. A WAY FORWARD

It is important to facilitate the exchange of expertise and experience related to the role of IT in official statistics and other issues related to the statistical infrastructure. The statistical offices in the region have each some experience that may be important to other countries. The issues of interest for this cooperation may be the following:

(a) Statistical information systems and IT tools in support to official statistics;
(b) Sharing the development of IT tools and components of statistical information systems between statistical offices;
(c) Statistical metadata;
(d) Statistical quality control, including methods and techniques for editing and imputation;
(e) Use of administrative registers and records, linking of registers and maintenance of statistical registers, including statistical business registers, statistical population registers and others;
(f) Protection of confidentiality of statistical data and control of disclosure risks;
(g) Issues related to the use of geographical and cartographical tools in support of official statistics (GIS in data collection, processing and dissemination);
(h) Advanced methods for dissemination and communication of statistics and relationship with users.

Should the statistical offices in the region be interested in the exchange of experience and building of expertise in these areas, the UN-ESCWA Statistics Division is ready to support this in substance.

* * * * *