2. Modelling the Information and Processes of a Statistical Organization (Statistics Austria)

back to case study home page

2.1 Statistical information model

2.2 Adoption of GSIM

2.3 Statistical business process model

Within the framework of the STAT+ project, a model of the statistical life cycle (called the "4-layer-model" because of the four data systems it defines) was elaborated at Statistics Austria in 2002. The model distinguishes between the following types of statistical projects: surveys, registers and analytical systems.

Surveys are the most "typical" and most commonly occurring form of statistical projects at Statistics Austria. One can differentiate between primary surveys (in which the collection of raw data is one of the steps of the process) and secondary surveys (which process data which already exist and often were collected for non-statistical purposes).

There are also mixed types, e.g. surveys in which data from secondary sources are used to augment the data collected by questionnaire. Some surveys are only undertaken once, others are repeated at regular or irregular intervals - although the sets of variables collected in each repetition do not have to be identical. It is therefore useful to further subdivide the survey structure: each survey consists of one or more survey versions, and each survey version consists of one or more survey instances, i.e. concrete executions. E.g., in the case of a survey with monthly periodicity the data collection of each year might be seen as a new survey version with twelve survey instances.

In contrast to the data of a survey instance, which pertain to a certain reference date or period, **registers** are usually updated continuously. Maintaining a register is thus a core process that is typical for a statistical project of type "register" but unknown for projects of type "survey". Another fundamental difference is that register data are used as resources for workflows in other statistical projects, e.g. when drawing samples, for addressing, registration of incoming questionnaires and administration of reminders. Commonly, specific (database) applications are developed to carry out these functions.

Analytical projects and systems (as, for instance, national accounts) characteristically do not collect raw data on specific observation units, but use data from other statistical projects and evaluate them or combine them into a coherent, integrated model.

Data which form the input to a statistical project (or which are collected in an early phase of processing, in the case of primary surveys) often pertain to individual observation units, e.g. individual persons, households, enterprises, events, etc., and are termed "microdata". However, the input may also consist of macrodata, i.e. data pertaining to collectives. In addition to these, metadata also enter into a statistical project and form an important resource for the steps carried out in its processing.

The output of a statistical project consists predominantly of macrodata - cross-classified tables, multidimensional data cubes and time series being the most important categories - and metadata. More rarely, (anonymized) sets of microdata may be produced. Macrodata and certain related metadata are often combined into an "information object" (e.g., a press release consisting of a table and descriptive text). Such information objects may also be composites of smaller information objects (as with a printed publication containing several tables and descriptive metadata - e.g., analytical texts).

The Statistics Austria model of the statistical life cycle distinguishes between the following phases in the production of statistical information (this description applies to statistical projects of the type "survey". Registers exhibit different core processes - creation, maintenance, and use of the register -, although the contents of a register may also form the basis for production of statistics and information, which can be identified with the relevant phases of a survey. The line between surveys and analytical projects is not always clearly defined - in fact one could certainly argue that analytical projects are a special type of statistical survey creating statistical information from input data, the difference being mostly that methods are applied which differ from those used in "typical" surveys.):

Phase 1: Planning, design and system development

The survey is set up in this phase. Given specific requirements (e.g. EU regulations) and the information needs of internal and external parties (e.g., statistics users in government and the economy), the survey must be prepared to satisfy these as best possible while simultaneously

minimizing the effort of the statistics producers and the burden on the data providers.

Output of this phase are metadata of various kinds - e.g., description of the survey's goals; description of the characteristics to be surveyed; definitions; value domains; classifications; questionnaires and comments explaining them; list of validation rules, etc. The metadata created in a survey may form the input to other phases or other surveys and be reused there.

The development of tools for actually conducting of the survey (e.g., electronic questionnaires, editing software, programs for checking consistency and plausibility) is also a component of the first phase.

Phase 2: Data production

Whereas the decisions taken in the design phase and the metadata and tools which are created there apply to the whole survey or at least a survey version, the focus of the activities undertaken in the following phases lies mostly on the current survey instance (excepting activities in which data from more than one survey instance are processed, as in the creation of time series).

Data production can be subdivided into three sub-phases:

- In pre-production activities such as drawing the sample, printing and posting paper questionnaires, loading Web questionnaires with respondent-specific initial data, etc. are undertaken.
- The actual survey/measurement/observation of the statistical raw data is termed core production. This includes conducting interviews, filling in paper or electronic questionnaires, registering and roughly checking questionnaires which arrive, mailing reminders, data entry from paper questionnaires, etc. In secondary surveys, this sub-phase includes acquisition of the secondary data and, if necessary, reformatting or recoding it. The collected data are stored in the so-called Raw Data System (RDS).
- Post-production includes all activities necessary to improve the quality of the raw data. Among these activities are validation and consistency checks, examination and correction of dubious information, and imputation of missing values. The results of this phase are the "authentic" data (ADS: Authentic Data System); of these several versions may exist, especially in complex and voluminous surveys (e.g., preliminary version at a certain cut-off date and final version at a later date).

A large part of the metadata created during the design phase enters the second phase as input. As also in later phases, however, new metadata also are produced (e.g., the answer rate, which is an attribute of the survey instance).

Phase 3: Statistics production

In this phase, the contents of the Authentic Data System (consisting mainly of microdata) are processed further. To do this, data from other surveys may occasionally be accessed. Some of the processing steps undertaken are aggregation into macrodata, calculation of statistical measures and indices, diverse methods for increasing quality and comparability of statistical information (e.g., seasonal adjustments), and creation of time series. The results are data sets which are stored in the Statistical Data System (SDS) and are at the disposal of internal, often also external users. The SDS primarily contains multi-dimensional data cubes, although anonymised sets of microdata may occasionally also be created in this phase.

In part, the transformations which are carried out here have already been planned in the design phase and are applied to the data of each survey instance. Partly, however, ad hoc analyses may also take place, which use the existing data material in ways not foreseen whenthe survey was planned. This underlines the importance of comprehensive and easily accessible documentation of all a survey's design decisions, of the data sets and of the transformation processes (in whatever phase of the statistical life cycle they may be created or executed).

Phase 4: Information production

In this last phase, "information objects" such as tables, charts, articles, press releases etc. are created from the data stored in the ADS and the SDS and their metadata and disseminated via various media (internet, print publications, etc.).

The following figure presents the phases described above, the data systems, the registers and the meta-information systems. On the one hand, the latter provide input and various services for the activities carried out during the statistical life cycle, on the other they also accept the metadata created as output from each phase. Thus metadata systems and registers form an infrastructure layer accompanying the whole production process. The data systems are drawn as broader than the "process arrows" in order to point out that they contain data from various surveys and that an individual phase of a survey may accept input data from more than one statistical project.

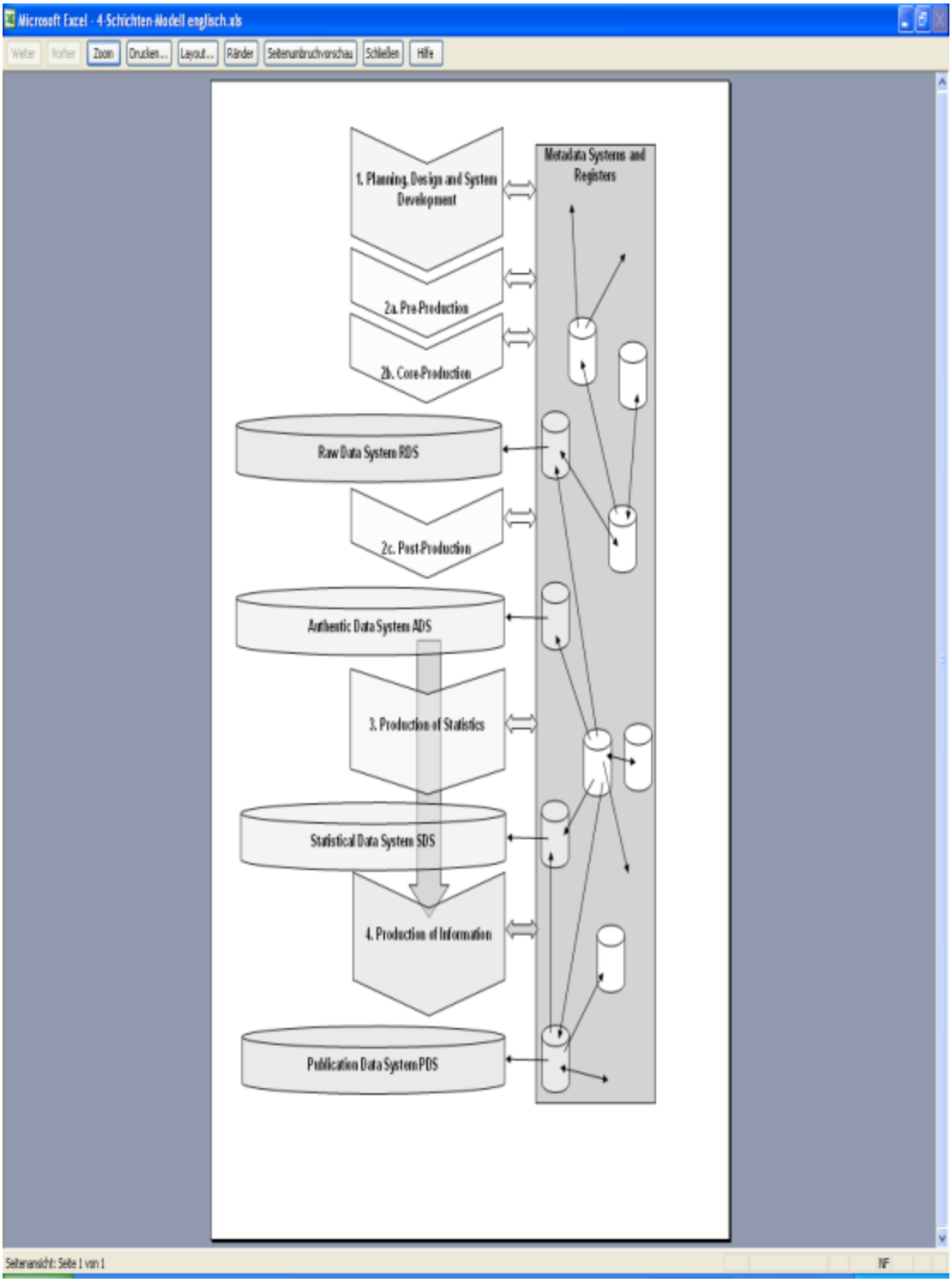

Figure 4: "4-layer-model"

In actual fact the workflows are of course not quite as linear as the figure might suggest; on the contrary, complex control flows (branches, loops) often occur. Moreover, events in later phases may have retroactive effects on the survey's design and lead to adjustment of the current or future survey instances (e.g., changes to the validation rules).

Phases 1 to 7 of the Generic Statistical Business Process Model ("Specify Needs", "Design", "Build", "Collect", "Process", "Analyse" and "Disseminate") can be mapped to the "4-layer-model". The phase "Evaluation" has been considered as an ongoing process, but is not explicitly mentioned in the model. The over-arching process of "Metadata Management" is represented by the "Metadata systems and registers infrastructure layer". "Archiving" and "Quality Management" are not part of the model.

The Generic Statistical Business Process Model incorporates many more details than the 4-layer-model. Therefore the GSBPM (with the addition of the four data systems) would appear to be appropriate for use in future metadata projects.

2.4 Relation to other models