DUVA – Information Management System:

collecting - processing – presenting Information

Holger Pietschmann, Statistical Office of the Free State of Saxony

Modern societies are information societies. An increasing need for information and the enormous speed of technical development requires a new handling of information.

In every project, company and administration information is available. Providers of information are urged to process reliable data without delay, under consideration of legal regulations and for a vast variety of customers. On the other hand, the internet allows new ways of data distribution and data mining especially for persons without specific knowledge in data-analysis or access to analysis-programmes. A gap exists, however, between the need for information and its accessibility. Still, calls for information must be answered by expert staff-members. The need for information often can't be satisfied "just in time".

Thus, up-to-date information management requires conceptual and technical solutions, which allow both expert and non-expert users the access to reliable information.

The DUVA-Project

The aim of DUVA is the cooperative development of a cost-effective, reliable and easy to use information management system. The DUVA-project was founded by a nationwide group of representatives from communal statistical offices in 1989. It's first objective was to develop a software tool for a comprehensive and flexible evaluation of the data of the German 1987 census. Due to the successful completion of this project, DUVA became a continuous organization. The portfolio of DUVA meanwhile covers the field of data entry, a common database and means to aggregate and to analyse data. All data is listed with a complete set of metadata. Existing metadata is used to control the entire process of data entry, data mining and data analysis.

DUVA is part of the KOSIS-Verbund (KOSIS-network). The KOSIS-Verbund was founded in Germany in 1982 as a common platform for different self-help projects in the field of statistical information management. Currently, more than 150 communes and other public institutions are associated to the KOSIS-Verbund; 57 members such as the city of Berlin, the city of Frankfurt and the cities of Nuremberg and Dusseldorf as well as several federal institutions belong to the DUVA-section of the KOSIS-Verbund at the moment.

Processes of the statistical work

Statistics involves activities such as data collection, structuring and ordering, data analysis by applying descriptive and inferential techniques as well as tabular, graphical and textual data presentations. Although statistical data derives from different sources and belongs to different fields or specialties commensurate with the ordering method applied, there is one thing they have in common, i.e. they reproduce certain parts of the real world at fairly high levels of accuracy and abstraction. This applies to customer data of companies, to measured groundwater quality data as well as to data collected by residents' registration offices, by research laboratories and data contained in private address lists.

The measurement procedure or collection method applied determines the accuracy or objectivity of information transfer from the real world. Clearly defined physical measurements of distances or weights are characterized by a level of objectivity that is different from empirical social research based, for example, on non-structured interviews. In the first case, a standard scale (e.g. 1 m = 100 cm) and a calibrated measuring instrument determine the quality of measurement. In the second case there are no universal measuring scales or measuring instruments. They must be defined for each individual case in a specific data collection concept based on pertinent theories and assumptions. Measurement procedures or data collection concepts are indispensable prerequisites of any collection

of statistical data, providing as they do a verifiable and reliable reproduction of segments of the real world.

Metainformation and physical information

The statistical processes described above must be considered as a process chain rather than isolated, single steps. The objective of statistical investigations beyond the field of technical issues and statistical computations is to collect and maintain all process-related information and make this set of information available for comprehensive retrieval processes.

In this context two information qualities must be taken into consideration:
- metainformation (all descriptive information) and
- physical information (measurement or survey results in the form of codes, values, frequencies or textual data).

Metainformation is derived early on during initial reflections on research objectives. All subsequent steps lead to an increase in the amount of metainformation. As to the amount of physical data the situation is different.

Following data collection, physical information is available in the form of measurement or survey results. The information load comprises all attributes collected for each statistical unit (e.g. each respondent). Since this unit-level information load is unsuitable for conveying knowledge, the body of data is reduced by means of several processes of abstraction. For example, primary data are first compounded into aggregate data (e.g. frequency tables) or into statistical measures (e.g. means, measures of dispersion, contingency measures, etc.).

The DUVA approach

Statistical work is not a static process chain. It requires maximum flexibility to address content-related issues, data source issues (primary, secondary statistics) as well as logistic (e.g. data flow, data communication), formal (e.g. data formats, etc.) and technical issues (e.g. data processing systems, programs for further processing, etc.). A comprehensive conceptual approach to data management of statistical information systems is needed.

A possible minimum manual solution to data management must involve detailed and systematic written specifications of every single step (from defining a research objective to data communication in the Internet via dynamic tables of results). Databases, spreadsheet programs and statistical analysis programs are very efficient means for maintaining and processing physical information. An obvious disadvantage of such minimum solutions is the fact that metainformation and physical data are stored on different media (media splitting). Attempts to integrate both types of information require searches on the medium concerned and a time-consuming integration process at every single step.

The basic idea underlying DUVA system development was to provide a management system for meta- and physical information without splitting between different media and to provide metainformation interfaces for all program elements forming part of the statistical production process (data bases, aggregation and analysis programs, tabulation, graphics, plotting programs, etc.). The content-related flexibility of the system is extended to include technical flexibility. Furthermore, the system should offer exhaustive search functions (e.g. search for individual data, linked information, analytic results, etc.).

The referral system (Nachweissystem) is the core element and control center of the DUVA information management system. The referral system collects and maintains metainformation and uses this set of information for system control purposes. Although meta- and physical data are held by separate media, the NWS referral system provides access to physical information. System access and individual access rights are controlled by the user administration module.

Metainformation input follows the principle that 'metainformation is systematically collected at its source and is subsequently reused and augmented repeatedly throughout the statistical production process'.