Statistiska centralbyrån Statistics Sweden

Quality Aspects of a Modern Database Service

Pat Dean and Bo Sundgren

R&D Report
Research - Methods - Development
1996:4

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 1996:4. Quality aspects of a modern database service / Pat Dean; Bo Sundgren
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1996-X101OP9604

Statistiska centralbyrån Statistics Sweden

Quality Aspects of a Modern Database Service

Pat Dean and Bo Sundgren

R&D Report
Research - Methods - Development
1996:4

Från trycket Januari 1997
Producent Statistiska centralbyrån, utvecklingsavdelningen
Ansvarig utgivare Lars Lyberg

Förfrågningar Bo Sundgren
telefon 08-783 41 48
telefax 08-783 45 99

Pat Dean
telefon 08-783 47 27
telefax 08-783 45 99

ISSN 0283-8680

Printed January, 1997
Producer Statistics Sweden, Department of Research and Development
Publisher Lars Lyberg

Inquiries Bo Sundgren
telephone +46 8 783 41 48
telefax +46 8 783 45 99

Pat Dean
Telephone +46 8 783 47 27
telefax +46 8 783 45 99

Paper presented at International Working Conference on Scientific and Statistical Database Management, Stockholm, Sweden, June 1996

ISSN 0283-8680

Quality Aspects of a Modern Database Service

Pat Dean and Bo Sundgren

Statistics Sweden

Research & Development

S-115 81 STOCKHOLM

Sweden

Abstract: This paper briefly reviews Statistics Sweden's statistical databases with an emphasis on quality issues, the role that metadata play, the documentation system and its ancillary templates. Statistics Sweden's agency-wide policy for quality in surveys is outlined and the main characteristics of the agency-wide quality declarations are cited. The paper concludes with some general statements about metadata: their importance, their uses, and their requirements.

Key words: Data provider; survey documentation and documentation templates; quality measures; metadata.

1. Introduction

Statistical databases are becoming one of the most important dissemination media for data and statistics produced by government agencies. Electronic publishing in general has replaced many of the uses and functions previously fulfilled by paper reports. Furthermore, electronic publishing promises to meet new needs and requirements, some of which have emanated from the availability of the databases themselves.

Electronic databases per se are nothing new, and have been around in various forms since 1970s, but their content, flexibility to be used in diverse and novel ways, and their general use by government, academia, researchers, and the general public has grow exponentially since 1990. For example, daily users of Statistics Sweden's databases are: Ministry of Finance, scientists, public sector analysts, private sector analysts, financiers, international organizations and organs, journalists, politicians and campaign managers, teachers and pupils, interested citizens. These uses are as diverse in their needs as they are in their technical and statistical sophistication. Analysts and governmental organs require data at a level of reliability and precision that far surpasses what pupils or journalists require or are even interested in. And yet, databases must be constructed in such a way that all levels of needs and requirements are met simultaneously. This is the challenge of building and operating a database service: fulfilling a variety of needs, needs that can even be conflicting, with a single product.

Statistics Sweden has followed three guiding principles or philosophies when navigating through the quagmire of users' needs. These are: The legal obligations to provide public data as mandated by the Swedish government, Statistics Sweden's own policy on quality matters, and the legal restraints imposed by the Data Inspection on the protection of confidentiality.

2. Statistics Sweden's Obligations as a Data Provider

The Swedish law on official statistics, promulgated by the Swedish Parliament, states, on a general level, the rules and obligations of official statistics. Official statistics should be objective and publicly available. Furthermore, they should be collected and disseminated in full accordance with the laws which guarantee respondents and others in registers full confidentiality. Also specified are the instances where companies and other legal entities are obliged to provide data for official statistics.

There is a law of slightly less authority, called a decree or even an ordinance which states further guidelines, again on a very general level. For instance, it states that official statistics shall be documented, with a statement of quality, and made available in a format stipulated by Statistics Sweden. This empowers Statistics Sweden to dictate how these data should be documented. These regulations are binding. Statistics Sweden has issued a directive entitled Directions and Recommendations on the Publishing of Official Statistics (MIS 1994:3). An updated version is currently underway which is more detailed and specific regarding metadata. This document mandates that all official statistics should be accompanied by a quality declaration. The quality declaration should be published in Sweden's Official Statistics.

3. Statistic Sweden's Quality Policy

In MIS (1994:3), Statistic Sweden's quality policy is described as both a formulation of a notion of quality for official statistics and guidelines for quality declarations of official statistics. The guidelines are intended to serve both the producers and the users of official statistics. Quality is a fluid concept which changes over time. Recently, quality has come to include a holistic aspect, a focus on "total" quality. Fundamental precepts in this new quality thinking encompass (1) that the user and his/her needs are central. The user's perspective defines the quality of a product (goods and services) and its utility. Product development should also assume the user's perspective. To operationalize this notion of quality, (2) that quality expresses itself in all aspects of a product and should strive to satisfy all aspects of the user's needs and expectations. The user's "needs and expectations" should be interpreted as both the product's performance and the service required to procure and use the product. For the user, quality is relative to intended use. A single product could have very high quality for one use and very poor quality for another use.

3.1. Quality measures: quality declaration and documentation template

Statistics Sweden employs two main tools to evaluate or make statements about data quality. In practice, these are two separate entities, but they can both be seen as expressions of the same descriptive model or quality philosophy. On the other hand, there is no single "quality metric" that reflects the general level of quality like an index number can reflect the general level of inflation. Quality is too multifaceted to lend itself to a single measure. Instead, we use a number of different measures that each reflect different characteristics of the data. (Or, these measures reflect the same characteristics, but from different perspectives.) From this composite picture of characteristics, the level of quality is inferred.

3.1.1 Quality declarations

The quality declarations result from describing or evaluating the survey on the basis of the following four main components: content, timeliness, reliability, and availability. Content reflects the statistics' characteristics

(objects, population, variables, statistical measures and presentation groups), and even comparability with other statistics. Timeliness suggests all time-related aspects, including comparability over time. Reliability describes different types of error sources and their effects. Availability entails that the statistics should be both intellectually comprehensible and physically available. As stressed earlier, the notion of quality should be user-oriented and formulated, to the extent possible, in terms that capture the user's perspective. All Swedish official statistics should include a quality declaration. This holds regardless of the dissemination form (paper publications, statistical abstracts, disks, electronic media, etc.). The requirement is that enough information always be provided so that the user can determine whether the data are suitable for his/her uses and that the data are used in a statistically defensible way. The table below presents Statistics Sweden's quality concept in main and sub-components.

Contents

Statistical quantities
- objects and populations
- variables
- statistical measures
- presentation groups

Comparability with other statistics
Reliability

Total reliability
- error sources
- coverage
- sample
- measurement

nonresponse
processing
model assumptions
Presentation of error measures

Timeliness

- reference point
- production time
- punctuality
- periodicity
- comparability over time

Availability

- dissemination
- presentation
- documentation
- primary material
- directions and information

3.1.2. Documentation template

The documentation template is supported by an EDP tool that both documents the survey process, step by step, and provides the basis for the metadata. More specifically, the metadata are extracted from the information entered into the documentation template. Figure 1 contains an example of a documentation template. It should be emphasized that the template contains sub-templates that give the documentation the required structure so that the metadata can be derived from the documentation. The entire documentation system is called SCBDOK and it is supported by PCDOK which operates in a windows environment

4. Statistics Sweden's Database Services

Statistics Sweden's databases are at present in a transitional phase. The new and the old systems will be operating in tandem for a couple of years until all underlying production systems have been transferred from the old mainframe to the new, entirely PC-based client/server platform. The mission of

the new system is to provide a database service that disseminates data that are of interest to many sectors of society. It should:

* be user friendly, offering system solutions adapted to users' needs;

* allow access to microdata for users to conduct exploratory data analysis and process the data themselves;

* contain documentation (metadata) that allows and enables re-use of statistical micro and macromaterial.

DOCUMENTATION TEMPLATE FOR A STATISTICAL SURVEY

0 Administrative information 0.1 Survey name 0.2 Branch of statistics 0.3 Responsible organization, person, etc. 0.4 Approximate cost of the survey 0.5 Purpose and history of the survey 0.6 Users and usages 0.7 Voluntary/mandatory response 0.8 Confidentiality and data annihilation 0.9 EU regulations and requirements	1 Summary 1.1 The survey plan 1.2 Contents: statistical quantities 1.3 Output: statistics and microdata 1.4 Time frame 1.5 Documentation
2 Data collection 2.1 Frame and frame procedures 2.2 Sampling procedures 2.3 Measurement instrument 2.4 Data collection procedures 2.5 Data processing	3 Observation register 3.1 Target and observation objects 3.2 Variable lists 3.3 Physical organization 3.4 Experience from last survey cycle
4 Statistical processing and presentation 4.1 Estimates: assumptions and computation formulas 4.2 Presentation and dissemination procedures	5 Data processing system 5.1 System summary and system flow 5.2 Processing 5.3 Database models 5.4 Database tables 5.5 Database accessories 5.6 Reports 5.7 Other data sets
6 The log book	

Figure 1. *Documentation template for a statistical survey and its production system.*

Database services will consist of basic service and additional services. There are seven basic services which are for the most part publicly funded.

- Use of Statistics Sweden's information service for assistance in using the databases. No fee is attached to this service.
- Computer access to all nonconfidential material on disks, CD-ROM at university level libraries and county libraries (no

fee). Copying of disks or CD-ROM at marginal costs.
- Material available at Statistics Sweden's library, but not at university or county level libraries can be ordered (a regular interlibrary loan) at no cost. Copying at marginal costs.
- Access to de-identified microdata at marginal costs.
- Macrodata available via Internet or equivalent at no cost.
- Statistics published on CD-ROM or disk are sold at marginal costs.

All other services are considered additional database services and entail fees.

Database customers will not notice such a great difference in the new system other than the new system will contain a much improved documentation system, the metadatabase, and that it will operation in a windows environment.

5. Metadata

A (potential) user of statistical data will need metadata for three major purposes:

* searching for potentially relevant and useful statistical data;

* evaluating the adequacy of available data and the cost/benefit of using them;

* retrieving, interpreting, and analyzing statistical data.

First, statistical metadata are needed as a basis for search operations. The (potential) user is looking for statistical data that could be relevant and useful in describing, analyzing, or solving a certain problem. The traditional approach is for the user to turn to a statistical office. Staff members of statistical offices are often very helpful, but today this approach is not sufficient. There are far too many potential users for a statistical office to cope with face-to-face. In addition, many users need to combine statistical data (and other data) from several sources, and no particular staff member, or even organizational unit, of a statistical office will have the necessary overview. Moreover, manual help-functions are relatively expensive and slow, even if they are computer-assisted. Today a user will expect the metadata needed for search tasks to be organized and disseminated in such ways that he himself can search for relevant data on the basis of widely available, computerized metadata. The process may start from a relatively vaguely expressed information need. The computerized, metadata-supported process should help the user to better understand his own needs, and it should result in explicit references to available statistical data, which are likely to be relevant for the user's problem.

Second, once the user has identified some statistical data of potential relevance for his problem, he will have to determine, if the data are really adequate for the intended purpose. This means that the user has to evaluate the quality of the data, and to consider whether it is really worth the effort and cost to retrieve, interpret, and analyze the data.

Third, if and when the user has come to the conclusion that certain available data are of sufficient quality to justify the efforts and costs to use them, he will need metadata in order to actually retrieve, interpret, and analyze the data. Retrieval may be accomplished by downloading data and accompanying metadata to the user's own PC or by obtaining a disk or CD-ROM copy. Interpretation and analysis will require the same kind of metadata as were needed for making the preliminary judgment of the quality of the data. However, at this stage it may be necessary to obtain deeper and more precise information about how the data

were collected and processed, before they resulted in the available statistics.

The documentation template in Figure 1 identifies metadata items that are desirable or even necessary as a basis for responsible usage of statistical data emanating from a particular statistical survey. If appropriately compiled with the corresponding metadata for other surveys they may also serve as a basis for search operations. See also Sundgren (1991a, 1991b, 1992, 1993a, 1993b).

It is an equally important task for a statistical office to produce metadata concerning its surveys as to produce the survey data themselves. In order to be able to accomplish this task in an efficient way, the statistical office must carefully design its metadata flows. Metadata should be captured when they naturally arise for the first time, e.g. as the result of a design decision. Then they should be automatically transferred and transformed whenever the underlying survey data are transferred or transformed. Furthermore, they should be automatically updated, when the survey processes are changed, e.g. as the result of new design decisions.

The metadata describing a statistical survey and its data outputs are a combination of formalized metadata, e.g. code lists and record descriptions, and free-text metadata like verbal descriptions of variables and processes. Thus software systems for handling statistical metadata may require different types of software components to be combined, e.g. relational database management systems and software for managing and searching large amounts of text data. Hypertext software (like in advanced help functions and high-level Internet-tools) will also have a great potential for enabling the users to navigate and associate in available statistical data and metadata and to process them in efficient and intelligent ways.

6. References

Ad hoc Committee on Databases (1994). Statistics Sweden's Future Database Service - Final Report. Stockholm: Statistics Sweden.

Malmborg, E. and Sundgren, B. (1994). Integration of Statistical Information Systems - Theory and Practice. Proceedings of the Seventh International Conference on Scientific and Statistical Database Management, Charlottesville, VA, USA.

Shoshani, A. (1982). Statistical Databases: Characteristics, Problems, and Some Solutions. Proceedings of the 8th International Conference on Very Large Databases.

Statistics Sweden (1994). Quality Definitions and Recommendations for Quality Declarations of Official Statistics. Stockholm: Author.

Sundgren, B. (1993). Guidelines on the Design and Implementation of Statistical Metainformation Systems. Statistics Sweden R&D Report 1993:4. ECE Work Session on Statistical Metadata, November 1993, revised versions 1994 and 1995.

Sundgren, B. (1995). Making Statistical Data More Available. Proceedings of the 50th Session of the International Statistical Institute, Beijing.

Förteckning över utkomna R&D Reports

R&D Reports är en för U/ADB och U/STM gemensam publikationsserie, som 1988-01-01 ersatte de tidigare "gula" och "gröna" serierna. I serien ingick fram till årsskiftet 1992-93 även **Abstracts** (sammanfattning av metodrapporter från SCB).

Reports published during 1996:

1996:1 (grön) On Sampling with Probability Proportional to Size *(Bengt Rosén)*

1996:2 (grön) Bortfallsbarometern nr 11 *(Antti Ahtiainen, Stefan Berg, Margareta Eriksson, Åsa Greijer, Dan Hedlin, Monica Rennermalm, Anita Ullberg)*

1996:3 (grön) Regression Estimators in Theory and in Practice *(Tomas Garås)*

Tidigare utgivna *R & D Reports* kan beställas genom Ingvar Andersson, SCB, U/SIB, Box 24300, 115 81 STOCKHOLM (telefon 08-783 41 47, fax 08-783 45 99, e-post ingvar andersson@scb.se).

R & D Reports listed above as well as issues from 1988-1994 can - in case they are still in stock - be ordered from Statistics Sweden, att. Ingvar Andersson U/SIB, Box 24300, S-115 81 STOCKHOLM (telephone +46 8 783 41 47, fax +46 8 783 45 99, e-mail ingvar andersson@scb.se).