Journal of Official Statistics, Vol. 15, No. 1, 1999, pp. 103–114

Applied Systems Analysis in Official Statistics

Klaus Neumann[1]

Statistical Information Systems (SIS) are a dominant part of official statistics and are classified among large-scale systems. According to systems theory, this class has distinguishing features which not only determine its behaviour but also its handling. Therefore applied systems analysis is an ideal tool both for streamlining the design of SIS and for gaining insights into their efficiency, evolution and management. This article offers a framework for such an approach, which is based on the application of some ideas from systems theory and systems analysis to the context of official statistics. This type of framework can be used to examine a specific SIS both in its entirety and in its modules. Such an approach reveals the dynamics, its life cycle (the entity) and sublife cycles (the different modules). Such an approach, furthermore, is a prerequisite for the generation of new SIS, their subsystems and the regeneration of existing systems.

Key words: Statistical Information Systems; large-scale systems properties; system design; metadata; life cycle phases; new technologies.

1. Introduction

During recent decades, official statistics has developed in terms of quantity, quality and timeliness not only with respect to its input and output but also (or even more) in its ''production method'' (in a wide sense: the design, development, implementation, real production of statistics and their dissemination) and in its position in a changing environment (social and technological).

As this evolution proceeded, it became increasingly obvious that official statistics are a special type of system,

- with characteristic properties and behaviour,
- divided into different components (subject matter and process oriented, methodological and organisational),
- which constitute a network of subsystems.

The ''systems aspects'' are not always obvious to the statistician in her or his daily work. Nevertheless, a few examples show the frequency with which we are confronted by such aspects.

* Each term which is used to describe a special statistical object is only useful in its relation to other terms, being a part (an ''element'') of a universe (a ''system'').

[1] Barnimstrasse 14, D-10249 Berlin, Germany. Freelance consultant.
Acknowledgments: This article is based on Neumann (1993).The views expressed in the article are influenced by the extraordinary work of international teams, above all the EDP Working Group of the Commission of European Statisticians and its ISIS-Seminars which I have had the honour and the pleasure of attending for the last twenty years.

* Tools like classification and nomenclature are a classic type of ''hierarchical systems'' with more or less related ''subsystems.''
* Modern technologies entail very complex systems which provide a permanent interaction between human beings and the powerful trinity of: hardware, software and – often neglected – orgware (i.e., organisation specific hard or software).

As the theory and practice of official statistics have developed, there have been different attempts at discussing statistical systems and a system view (see e.g., Langefors (1966); Sundgren (1973); Fellegi (1977); Klas (1978); Westlake (1987)); but these attempts have been more or less sporadic. Nor has the parallel development of systems theory been steady.

This is especially true with regard to Applied Systems Analysis which deals with the properties and behaviours of different classes of systems (as understood by cybernetics) and which aims to optimise systems design, operation, maintenance, and control.

2. Applied Systems Analysis – Selected Aspects

From the vast area of systems theory, we select topics of value to the special context of Statistical Information Systems, for instance:

- those aspects which are important for applied systems
- those features that belong to information technology, since it is mainly that part of Statistical Information Systems which interests us.

The cybernetic systems theory postulates the following definition:

A system is a set of elements and relations between them (see Klaus (1968, p. 634) or more generally, as shown by Steinmüller (1993, figure 1).

But Steinmüller maintains that this definition does not answer four questions, which are important for applied informatics:

1. Who selects who or what belongs to the system? Without him/her the system does not exist obviously.
2. Who determines its spatial and temporal limits? (→ system definition).
3. Why does he/she select exactly these elements and relations and why does he/she set the limits just there? (→ system function).
4. Who is responsible for the behaviour of the system?

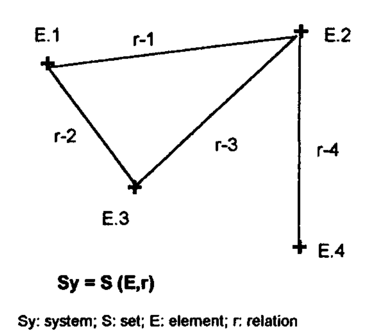

Fig. 1. A system according to Steinmüller (1993), p. 163

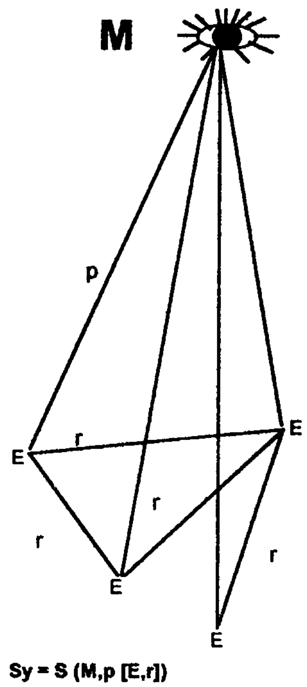

Fig. 2. A system according to Steinmüller (1993), p. 165

This consideration of the system's genesis requires a completion of the given definition:

- A system comes into existence only when a subject observes a ''reality,'' reducing the image according to special purposes and enlarging it by the observer's experience.
- Without the observer and his/her purposes, we do not have a specific system. Without the observer it is not clear what belongs to that system and where its boundaries lie.

For these reasons, Steinmüller (1993) introduces the term ''system's master'' (to emphasise the dominant role of a system subject, which belongs to the system *after* its creation). This extends the definition to:

A system (Sy) is a set (S) of elements (E) and relations (e), selected by a subject/master (M) for a purpose (p).

Taking into account the extraordinary role of the human being as the subject, or master, in all the phases in a system's life cycle such as exploration, development, operation and management, evaluation and dismantling (Sundgren, 1993), it is very important to apply the rules of a system-oriented approach especially in the context of information systems.

Two aspects of the system-oriented approach are helpful:

- the ''systematising'' (top down) approach,
- the ''systemic'' inclusion of the appropriate holistic system.

Systematising or the systematic approach mainly has the task of coping with the variety of details. Primarily, this is an analytical process that divides an existing or imagined system into its functional parts. For Statistical Information Systems, this is a main characteristic, because it is impossible to observe a statistical population as such. What we observe are its constituent elements. Without a systematic view, without their grouping and their classification, these elements are nothing more than a chaotic set of numbers.

On the other hand, overemphasising the implications of the elements can lead to a one-sided view and miss the global functionality.

The **systematic approach** is the indispensable pendant to the systematic view, resulting from systems thinking. It aims to recognise the interactions of the elements and to discover the systems structure, mainly by:

- tracing the holistic property of the system by heuristical testing;
- disclosure of the effect of the holistic system, using step-by-step theoretical abstraction, which is not deducible from the partial elements;
- taking into account the influence of the past, the present and the (expected) future and the environment (in a broad sense) on systems development;
- analytical testing of the results, thus closing the loop in the systematic approach.

(For further details, see Steinmüller, 1993, pp. 72–77.)

Another aspect of the systems theory is the differentiation between different classes of systems, different not only in size and function but also with regard to more general characteristics, especially their behaviour and handling.

For our subject – in particular Statistical Information Systems – we are mainly interested in the class of **large-scale systems**. These are generally characterized by:

- *human beings* actively handle both the process executing and controlling
- *transformation* of matter, energy, and information in the sense of conjunctive combination to achieve the process results
- *complexity and complicity*, i.e., not only large numbers of elements but a multiplicity of couplings and different operators for transformations
- *process stochastic*, based on a diversity of external and internal operational factors
- *inference control* by multistability
- *systems evolution* including the ability to adapt behaviour (Golenko et al., 1977).

To apply systems analysis methods to Statistical Information Systems we should examine both their features and their special properties.

3. Features and Special Properties of Statistical Information Systems (SIS)

Some of the general attributes of large-scale systems are of great importance for SIS.

- The *active role of human beings* is not only characteristic of the SIS itself (its design, functionality, maintenance, and control) but of its environment too, which consists of supervisors, data suppliers and a multiplicity of clients. This is what makes the SIS a system of the type

  Human Being → **M**achine → **H**uman Being,

  which is in some ways different from the more technical type of system

  Machine → **H**uman Being → **M**achine.

While the M → H → M-type drives a (fully) *automated* system, the SIS is (like all social information systems) limited in the degree of automation, as it is always based on an interactive man-machine communication. This is one of the main reasons that statistical production and consumption are such prosperous fields for new technologies (see for instance GMD, 1992 and SMI, 1993).

- The *transformation aspect* in the SIS is highly (but not only) significant for the transformation of data into information and then for the transformation of that information into different types of information. This process starts with the acquisition of data from different sources, is followed by different cycles of processing, for instance data and metadata processing, transport and storage, and finally evaluation, presentation and dissemination.

It is very interesting that we find different but, nevertheless, similar distinctions between the sequences of transformation processing. These sequences are manifested in different phases:

Steinmüller (1993) differentiates between special computer oriented steps (a–g) and the general phases of an ''ideal'' problem-solving process (1–10).

These ten phases are indeed identical to the usual procedure of daily statistical work. Even if this is not fully recognised (due to the more or less routine character of such transformations) it becomes increasingly important in the context of modern communication facilities. Without an exact description of the applied transformation methods and the current status of the data, the worldwide use of statistical data (which is at present a reality) can become very dangerous. There could be very serious consequences from further use. It is crucial that information on all transformative actions is available not only for those who produce the statistics, but for those who use the statistics at some point in the future. This is one of the reasons that metadata descriptions and meta information are available as understandable documents of the applied transaction

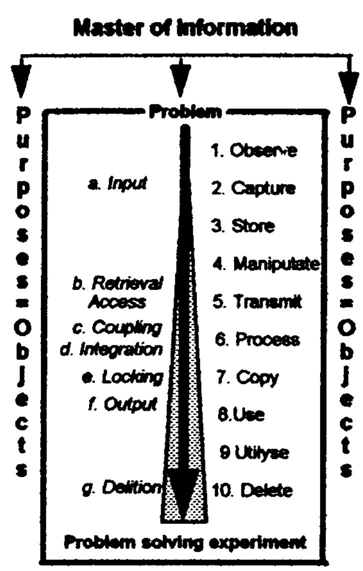

Fig. 3. Sequences of transformation processing according to Steinmüller (1993), p. 229

procedures. It becomes increasingly obvious that process oriented – or more precisely process based metainformation is indispensable for a holistic perspective on information processing both in the statistical offices and by the end users. Codd (1994)[2] formulates the conceptual basis for a statistical information system focusing on the external end user. Such a system, called the On-Line Analytical Processing (OLAP) Mandate, should ensure:

1. Easy access to all kinds of data irrespective of their position inside or outside the user's processing system.
2. Clear definitions of all processed data and their dimensions.
3. Freedom for the user to choose both the way and the context in which he/she wants to view, manipulate and animate the data model.
4. An interface that makes all these functions accessible and which is familiar to the end user.

Codd expressed his mandate in the following twelve rules:

OLAP Product Evaluation Rules

1. Multi-Dimensional Conceptual View
2. Transparency
3. Accessibility
4. Consistent Reporting Performance
5. Client-Server Architecture
6. Generic Dimensionality
7. Dynamic Sparse Matrix Handling
8. Multi-user Support
9. Unrestricted Cross-dimensional Operations
10. Intuitive Data Manipulation
11. Flexible Reporting
12. Unlimited Dimensions and Aggregation Levels

The rules' main emphasis is on the end user – in our context all those who are using statistics in their own profession as politicians, managers, journalists, scientists, etc., or last but not least as responsible citizens.

Figure 4 shows another series of sequences than those used by Steinmüller (1993). Sundgren (1993b) differentiates between five major phases in the life cycle of a Statistical Information System[3]. This approach is extremely clear for the design of statistical information systems and much more directed to an active use of metainformation even for such a design. It also suits the design of a variety of subsystems, e.g., surveys, subject matter areas, regional statistical systems, etc., for holistic information systems on different levels.

This figure also clearly depicts the use of system analysis procedures with respect to the application of graphical design methods in the design process. Unfortunately, at present such effective instruments as case tools are rarely used for these purposes. The

[2] Parallel to the reworking of this paper a contribution about Codd's publication was written: see Neumann (1995).

[3] The source of this figure, the ''Guidelines on the Design and Implementation of Statistical Metainformation Systems'' by Bo Sundgren, is important not only in this context but also in general as an example for Applied System Analysis.

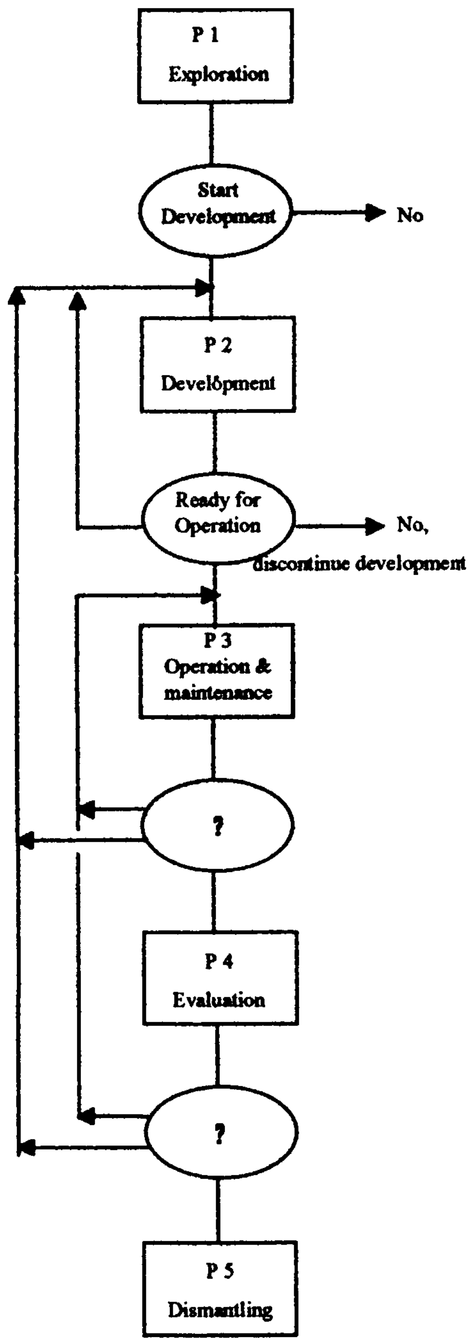

Fig. 4. The life cycle of Statistical Information System according to Sundgren (1993b), p. 13

Canadian experience is the best proof of their effectiveness, (see Statistics Canada 1992a, b, c; Statistics Canada 1993.)

- The terms ''*complexity*'' and ''*complicity*'' express properties of diversity. The first reflects the systems structure (relations between the elements) and the second the functions (transformations). In an SIS both complexity and complicity are caused by the multiplicity of subfunctions (Fellegi 1977. See Fig. 5), which results in their modularity (Neumann 1992; Appel 1993).

Just these properties are the potential basis for the standardisation of modules and the interfaces between them and between the SIS and its environment. One of the most exciting examples of the repeated attempts to tackle the problems of complexity and complicity is the ''long march'' of Statistics Canada, marked by the creation of the powerful Data Base Management Systems. This project started in the 1960s with CANSIM, generalised in the 1970s with the functional schema of an ideal statistical system (Fellegi 1977). In the early 1990s it developed into an Information Technology Framework (ITF) (Statistics Canada

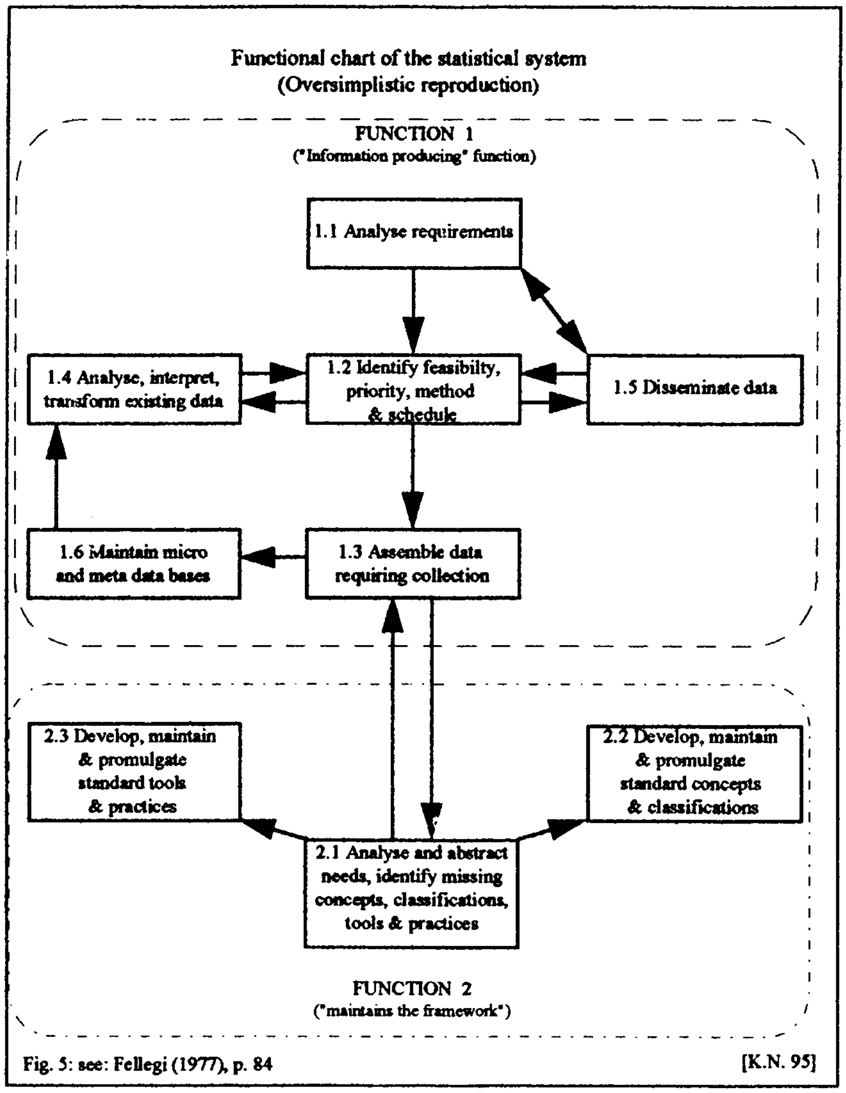

Fig. 5. Functional chart of the Statistical System (simplistic reproduction) after Fellegi (1977), p. 84

1992a; Statistics Canada 1992b; Statistics Canada 1992c; Statistics Canada 1993). ITF was an excellent example of the advantages of new technologies. ITF used CASE TOOLS in a consistent fashion both for the design of such a large, complex and dynamic system, and for its continuous updating. In this way the work of the systems designer was made easier, and subject matter specialists and managers were also able to use the framework for their tasks.

Systems analysis had had an uneven effect on the development of statistics systems because of the conflicts inherent in the following two goals:

1. Providing the most complete description of the functions and properties of the Statistical Information System, and
2. The availability (or lack of availability) of tools both to create a sufficient model and to modify (changes, updates, etc.) and document the status of the system and its subsystem.

An example of this conflict is the different descriptions used by various groups (subject matter specialists and computer experts; system designers, methodologists, and operators; producers and users, etc.). Only after the metadata/metainformation approach gained increasing acceptance in the statistical community worldwide, did adequate new technological tools, e.g., for computer aided design procedures, become available. That is one of the reasons that system analysis aspects are often found in the context of metadata/ metainformation (see e.g., the above-mentioned Canadian papers; or for the pioneer approach of the Australian Bureau of Statistics: ABS (1979), and more recently, Sundgren (1991).) This approach creates a new view of the statistical work by using the metadata/ metainformation for more than a passive, mainly identifying description of data and procedures (as was previously common when databases and database management were introduced in the information processing of the statistical offices – the Data Dictionary/ Data Directory Area).

For an interactive handling of metadata it is necessary to:

- uncouple masses of statistical data from the descriptive metadata, which in turn will preserve the essentials of all the elements, characteristics and relations;
- abstract from the overwhelming level of detail that is part of the statistical system and to focus the users on the essentials, presenting them with an X-ray picture of the holistic system or parts that are of interest to them;
- manipulate, simulate and test elements and relations without the burden of unnecessary masses of data, thereby saving time and avoiding confusion.

Sundgren (1993b) states that the metainformation infrastructure of a statistical office must support many complex relationships, for example production and retrieval systems, local and global metadata, and the different phases in a system's life cycle.

These relationships are often complex in the sense indicated in the upper part of Figure 6, which shows two sets of entities of some kind, which are related to each other in what is called a ''*many-to-many*'' pattern. For example, one of the two sets could be a set of production systems, and the other could be a set of retrieval systems.

This gives rise to the development of a unique communication between each related pair of entities to the $n*n$ interfaces and the adoption of a standardised solution to maximal $m + n$ interfaces.

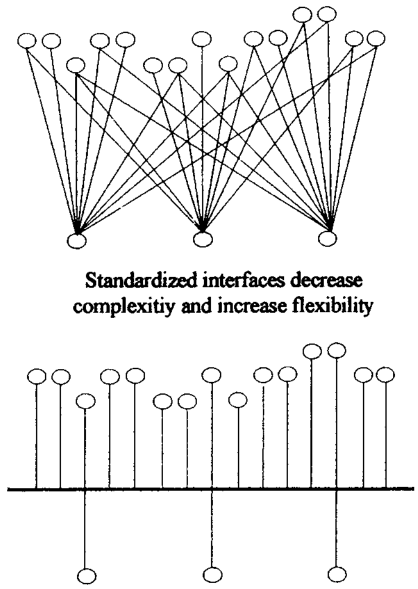

Fig. 6. See Sundgren (1993b, p. 45)

An additional benefit is a highly flexible architecture. Additions are easily absorbed into the system as communication with all other entities is established through the standardised intermediary interface.

Large systems are also characterised by *systems evolution* as statistical information systems normally have long life-spans and must continuously adapt to changing environments (Neumann 1989). For this reason it is a paramount task for managers to ensure that the system can adapt to and meet with the regeneration of the systems architecture if necessary.

From the viewpoint of systems regeneration, the handling of metadata is crucial. For example, the entire statistical process is affected by the interaction between the creation and use of different kinds of metadata and metatexts, as illustrated by recent development by the Statistical Office of Berlin (Appel 1993a).

Even if these selected examples illustrate only some aspects of applied systems analysis and their performance, it seems obvious that its implementation may assist official statistics in the continous improvement of their statistical information systems.

4. Some Conclusions and Potential Fields for Research and Development

The systems analysis methods can now be applied by means of new technologies such as case tools. That allows them to be used for different tasks which requires that the gap between theory and practice will be bridged by cooperation. Now it is mainly management's job to better use their knowledge of the functionality of information systems in general (because serving them is their daily work) and to refine the handling and managing of the statistical system in a manner adapted to large-scale systems.

Progress in this direction is also dependent on further research in fields like:

- *Dynamic Design of Statistical Systems*, using the experiences not only of the different groups of statisticians but also of those who supply the official statistics with data. There are a growing number of (real or potential) clients whose business can only be captured by exploiting the expected new technologies.
- *Data and metadata modelling*, which needs not only tools, but also a sound background of ''brainware'' in the form of a mixture of knowledge and futurology.
- *Information management*, which above all uses the richness of the information.
- *Controlling procedures* and information management are integrated instruments for the permanent regeneration of both the different subsystems and the holistic system.
- *Exchange of SIS modules according to their life cycle*, as part of a permanent regeneration, taking into account that farsighted acting is much more effective and smooth than rash handling in emergency.
- *Multimedia aspects of the dissemination of statistics*, which include again an increasing flexibility and adaptability of the Statistical Information System to keep pace with the expected changes in the social environment.
- *Harmonisation of the communication inside and outside the Statistical Information System*, which should become a driving force for the acceptance of official statistics, because of its decisive impact on quality and quantity in a client/server environment.

Of course this all requires not only the availability of modern technologies but, much more, a fruitful cooperation within the statistical offices, between statistical offices and between statistical offices and different scientists – both on the national and the international level.

Two projects – expected to be realised in the frame of the ESPRIT programme of the European Community (see NTTS 1995) – may be a further step in the direction mentioned;

1. The project ''IMIM'': Integrated Meta Information Management.
2. The project ''IDARESA:'' An Integrated Documentation and Retrieval Environment for Statistical Aggregates.

5. References

ABS (1979). Statistical Computing Environments: A Survey. Australian Bureau of Statistics, June 1978; revised February 1979, Canberra.

Appel, G. (1993b). Zum Entwurf eines metadatenbasierten statistischen Informationssystem. Allgemeines Statistisches Archiv, 77, 68–91.

Appel, G. (1993b). Outline of a Rule-Based Statistical Metainformation System. ECE – Work Session on Statistical Metadata (METIS). Conference Room Paper No. 7. Geneva, 15–18 November.

Codd, E.F. (1994). Adding Value to Relational and Legacy DBMS: The OLAP Mandate. Business Intelligence, October.

Fellegi, I.P. (1977). Functional Analysis of an ''Ideal'' Statistical System. ECE Seminar: Statistical Services in Ten Years Time. Washington, DC: Pergamon Press.

GMD (1992). New Techniques and Technologies for Statistics. Bonn: Gustav-Stresemann-Institute.

Golenko, D.I. et al. (1977). Kybernetik zur Steuerung ökonomischer Prozesse, Berlin: Akademieverlag.

Klas, A. (1978). Structuring of the Computerised Information Systems from the Aspect of Integration. CES/SEM. 10/22.

Klaus, G. (1968). Wörterbuch der Kybernetik. Berlin: Dietz Verlag.

Langefors, B. (1966). Theoretical Analysis of Information Systems. Studentlitteratur Sweden and Auerbach, 1966.

Neumann, K. (1989). New Metainformation Aspects in a Changing Environment. Bulletin of the International Statistical Institute, Contributed Papers, 47th Session, 193/205.

Neumann, K. (1993). System Analysis Aspects in Statistical Production and Consumption. Bulletin of the International Statistical Institute, Contributed Papers, 49th Session, 213/214.

Neumann, K. (1995). New Standards for Data Analysis Tools? The possible impact of Codd's OLAP mandate. Submitted and accepted contribution to NTTS-95, Bonn, Nov. 20–22.

NTTS (1995). 1995 Seminar on ''New Techniques and Technologies for Statistics.'' Bonn, November 20–22, 1995.

SMI (1993). Statistical Meta Information Systems Workshop. Eurostat/World Systems. Luxembourg.

Statistics Canada (1992a). Information Management and Metadata in Statistics Canada. February.

Statistics Canada (1992b). Information Technology Framework. July.

Statistics Canada (1992c). ITF REFERENCE LAYER Proposal. September.

Statistics Canada (1993). Information Holdings within Statistics Canada: A Framework. November.

Steinmüller (1993). Informationstechnologie und Gesellschaft – Einführung in die angewandte Informatik. Darmstadt.

Sundgren, B. (1973). An Infological Approach to Data Bases. Stockholm 1973.

Sundgren, B. (1991). Towards a Unified Data and Metadata System at the Australian Bureau of Statistics. Final Report, December.

Sundgren, B. (1992). Organizing the Metainformation Systems of a Statistical Office. R&D Report, Statistics Sweden, no. 10.

Sundgren, B. (1993a). Modelling Meta-Information Systems. SMI, 59–79.

Sundgren, B. (1993b). Guidelines on the Design and Implementation of Statistical Meta-information Systems. ECE – Work Session on Statistical Metadata (METIS), Working Paper No. 8. Geneva, 15–18 November.

Westlake, A. (1987). The Impact of ''New Technology'' on Statistical Offices: Hardware, Software and Systems for Survey Processing. International Statistical Institute Research Centre.

Received July 1994
Revised February 1997