Statistical file systems and archive statistics –

Experiences from Statistics Sweden

Bo Sundgren
2010-07-25

Introduction

When Svein Nordbotten in papers like Nordbotten (1960, 1966) challenged statistical agencies to make better and more systematic use of existing data in the production of official statistics, rather than relying mainly on data collected specifically for statistical purposes through censuses and sample surveys, it represented the beginning of a shift of paradigm that has turned out to become equally important as the shift of paradigm that occurred when probability-based sample surveys were introduced and replaced some total enumerations and more book-keeping like production methods.

Although Nordbotten's launching of his ideas of statistical file systems and archive statistics coincided with the introduction of computers in statistics production, the ideas are not as technology-dependent as one may think. This is illustrated by the fact that most of his ideas could very well have been successfully implemented on a large scale already when mainframe computers were less powerful than the smallest PC is today. Unfortunately this did not happen, even though Statistics Sweden was very close to doing it in 1974 – see Fastbom (1974) – but the failure was not due to technical limitations – or any shortcomings in Nordbotten's ideas, for that matter – but it was rather caused by normal, but regrettable, human and organisational inertia.

Now, 50 years later, many of Nordbotten's ideas have become widely accepted and implemented – not only in the world of official statistics. I am thinking of concepts like databases and data warehouses, integration of data and metadata, use of standardised software, etc.

Reuse of already collected data for the production of statistics is often regarded as equivalent with register-based statistics – see Wallgren&Wallgren (2007) – but the former concept is actually much broader than the latter. The basic role of a register is, in general, to provide a complete and up-to-date list of all objects belonging to a certain population, e.g. all persons living in a certain country at a certain time. A basic role of a register in statistics production is to serve as a frame for surveys. This means that a register, in addition to identities and names of the objects belonging to a certain population at a certain time, should also contain contact information like addresses and telephone numbers. Moreover, it is usually found to be practical among statistics producers to include in statistical registers stratification variables and other basic variables for statistics production. Administrative registers, which are usually the source of statistical registers, will typically contain variables that are necessary or useful for the administrative operations supported by the administrative register.

When planning for statistics production based on already collected data, one is not limited to considering only registers. There are many other kinds of data in society that could be reused for statistical purposes, e.g. data generated by all kinds of human and commercial activities in

society, both activities in the traditional world and in cyberspace. Also data collected by statistical surveys may be considered for reuse by other surveys or statistics production systems.

Like Nordbotten saw already 50 years ago, most statisticians are now able to see the potential of organising a lot of different kinds of data, from different sources, into statistical systems, where they can be used, reused, and combined in almost infinite ways; Sundgren (2010b).

Which are the main reasons for doing this, for changing the traditional survey paradigm for production of official statistics into a more systems-oriented and holistic paradigm? There are several good reasons:

- efficiency and costs
- quality, e.g. timeliness, coherence, precision

As for costs, Wallgren&Wallgren (2007, pp 62-63) refers to a report, Statistics Netherlands (2004), about the Dutch so-called "virtual census", another term for a census based on administrative registers and other already collected data. The Dutch virtual census of 2001 was completely based on already existing sources. The cost for a traditional census would have been about 300 M€, whereas the cost for the virtual census was only about 3 M€ – an impressive 99% cost saving. There were also other advantages of the virtual census. The willingness to participate in a traditional census had fallen, but the virtual census was easier for the general public to accept. A traditional census would have resulted in severe non-response problems, difficult to adjust for.

It is an interesting exercise to make a rough translation of the Dutch cost savings to the situation in Scandinavian countries, where almost all data used for statistics production nowadays come from administrative registers and other already existing data sources – in Sweden as much as 98-99% according to estimates made several times over the years. Let us say 99% for the sake of simplicity. If each one of these 99% of the observations used for statistics production costs only 1% to collect and prepare in comparision with each one of the 1% of the observations made by traditional surveys and censuses, it means that 50% of the total production costs is accounted for by the 1% of the data that are collected by traditional survey methods.

Nordbotten-inspired developments at Statistics Sweden

When I started to work for Statistics Sweden in 1968, Svein Nordbotten was already a highly appreciated advisor to Statistics Sweden and its top management, notably Dr. Ingvar Ohlsson, director general, and Dr. Lennart Fastbom, head of the planning department, and deputy director general.

Ingvar Ohlsson and Lennart Fastbom started a large number of projects with tasks of investigating the potentials in various directions of Svein Nordbotten's ideas concerning statistics production based on a so-called archive-statistical system (ARKSY). The projects were coordinated under a common umbrella called UKAS – in Swedish: "Utredningar kring det arkivstatistiska systemet"; in English: "Investigations of the archive-statistical system".

I became personally involved in this work. Among other things I became the secretary of a Working Group (WG 5) under the Scientific Council, dealing with the information processing aspects of ARKSY. The group was chaired by professor Börje Langefors, one of the designers

of the first Swedish computers, and the first professor of Information Processing as an academic discipline, shared by the Faculty of Social Sciences of Stockholm University and the Royal Institute of Technology. Other members of WG 5 were Christer Arvas, head of the new unit of data processing methods at Statistics Sweden, and Birger Jansson, expert of random number generators at the Swedish Defense Research Agency (FOA). We cooperated closely with Svein Nordbotten on the ARKSY developments over a number of years.

Data models

The model proposed by Nordbotten for structuring the data in a statistical file system has three basic dimensions: statistical units (objects in today's terminology, e.g. persons, enterprises), variables, and time; see, for example, Nordbotten (1967c), figure 1. Interestingly, this model is very similar to Langefors' model, based on e-messages as the basic building block, or atom, of information, developed and presented at about the same time. An e-message consists of (references to) an object, an attribute, and time; Langefors (1966). In comparison, the relational data model, as presented by Codd (1970) much later, is based on two of these three dimensions: rows (objects) and columns (variables, attributes). Time is missing in the relational data model. This was recognised by Codd in a conversation I had with him on a bus trip on Corsica in connection with the first international conference on databases; Klimbie&Koffeman (1974). I had just presented my doctoral thesis on an infological approach to databases, Sundgren (1973), based on the ideas of Nordbotten and Langefors.

By standardising the definitions and identifications of statistical units and variables, it would be possible to combine data from different files in a statistical file system. This was also clearly understood and explained in the early papers by Nordbotten and Langefors. See, for example, Nordbotten (1966, 1967a) and Langefors (1961a,b, 1963, 1966).

The basic concepts and modelling methods for structuring and describing data in general, and statistical data in particular, which were established by Langefors and Nordbotten were further nuanced and developed in the ARKSY development work at Statistics Sweden during the 1970's and up to today. The models are since long internationally known as data models and conceptual models (or information models). The latter focus on the information contents of the data, whereas the former also deal with data representations and more technically oriented aspects. See, for example, Sundgren (1973, 1974, 1995, 1999a, 2001b, 2004a, 2005b, 2005c).

Statistical file systems and databases

In today's terminology Nordbotten' s statistical file systems, or statistical archives, would be called databases and data warehouses.

The term "database" became popular around 1970. Börje Langefors was one of the promotors of the term. (He also invented the Swedish term "dator" for "computer" – in analogy with "motor", engine.) Langefors' preferred term was actually "data bank"; instead of making deposits and withdrawals of money, as in the case of a data bank, one would make deposits and withdrawals of data. A nice thing with data, in contrast with money, is that you can make a withdrawal of data without decreasing the data capital. You will also get good interest on your data capital in a data bank by combining the data you deposit with other data in the data bank.

Nevertheless, "database" became the prevailing term in Sweden and internationally. In the Nordbotten-inspired developments at Statistics Sweden, we made a distinction between

- microdatabases, containing microdata, data about individual objects (persons, enterprises, etc); cf "collection products" in Nordbotten (1966)
- macrodatabases, containing macrodata, aggregated data, "statistics", estimated values of statistical characteristics; cf "processing products" in Nordbotten (1966)
- metadatabases, containing metadata, "data about data": descriptions, definitions, explanations, quality declarations; cf data definitions and data descriptions in Nordbotten (1967a); the terms "metadata", "metainformation", etc, were established in Sundgren (1973), where the corresponding concepts were also defined and analysed in some detail

In another dimension, Nordbotten (1966) made a distinction between active files, historic files, and statistical registers.

Database management systems

We often had difficulties during these early years of database developments to make people understand the distinction between the database as such and the software managing databases, the so-called database management system (DBMS). The same acronym was often used for both, e.g. ARKDABA, TSDB, RSDB, etc.

In the early 1970's there were very few database systems available on the commercial market. IBM marketed its IMS system, but it was based on hierarchical data structures that were suitable for certain categories of business application, but not flexible enough for a statistical data archive intended for *ad hoc* retrievals and tabulations. Statistics Canada developed early the RAPID system, which was inspired by some commercial software products (TOTAL and System 2000), which were based on the inverted or transposed file technique, where data were stored by columns (variables) rather than by rows (objects), using the relational data model terminology.

After Statistics Sweden stopped the development of the ARKDABA system, because of the privacy and confidentiality debate (see below), we started to experiment with the Canadian RAPID system for managing microdata for internal production purposes.

Because of the privacy and confidentiality problems associated with statistical microdata, Statistics Sweden focused for a very long time, two decades, on databases and database management systems for macrodata. This resulted in the development of the AXIS software system for managing and making aggregated statistical data available to users of statistics and the public at large. The system was based on the multidimensional alfa-beta-gamma-tau model first developed in Sundgren (1973); see also Sundgren (2001b).

The AXIS system was probably one of the first metadata-driven systems in the world, that is, a system, where the data processing operations do not work unless the proper metadata have been loaded into the system before the data, and where the metadata can be updated independently of the data, thus modifying the processing of the data without requiring any reprogramming of the software. It was also one of the first systems for flexible management of large volumes of multidimensional statistical data.

The AXIS system was developed for IBM mainframes, and it was launched in 1976. It was an on-line system, accessible via dumb terminals attached via local and remote networks to the mainframe computer.

When PC:s entered the scene around 1980, a user-friendly front-end tool called PC-AXIS was developed for easy and flexible downloading and manipulation of mainframe-stored statistics to the PC environment.

In 1996, when Internet had conquerred the world in general, and the statistical world in particular, an Internet-based PC version of the AXIS system was launched – not to be mixed up with the front-end software tool PC-AXIS. The new, PC- and Internet-based AXIS system was still based essentially on the same logical principles and data/metadata model as the original mainframe AXIS system.

For more information about AXIS and Sweden's Statistical Databases (SSD) based on the AXIS software system, see Sundgren (1997).

Metadata

Through the early works by Nordbotten and Langefors, it also became obvious that the data collected, stored, and processed in a statistical file system, in a more or less continuous manner, must be carefully defined and documented, in order to make them useful for repeated uses and reuses. The chain of processes from the original data capture to the final use and interpretation by decision-makers may be quite long (also in time), and may involve complex combinations and analyses for purposes that have not been anticipated in detail, when the primary data were originally collected.

Nordbotten (1967a) outlines the concepts of a data definition language and a table definition language.

The first system for systematic data documentation that was developed at Statistics Sweden was the so-called variable catalogue, which was designed around 1970, with Lennart Fastbom, the deputy director general, himself deeply involved in the discussions, together with, among others, Svante Öberg, a later director general of Statistics Sweden. The variable catalogue focused on the three main data dimensions suggested by Nordbotten and Langefors: objects, variables, and time – and, of course, references or relations between data. It was also recognised that a full understanding of the meaning of data stored in a statistical archive will sometimes require rather detailed information about the survey processes behind the data, especially the questionnaire design and data editing. Hence the variable catalogue was designed to contain such information as well.

The implementation of the variable catalogue was not successful. Despite the support from top management, it was extremely difficult to convince the people responsible for the operations of the statistical surveys about the necessity of data documentation. They found the documentation work tedious and time-consuming and not very useful for themselves – after all they felt they knew everything they needed to know about the data they had collected, and if someone else wanted to know about this, e.g. the users of statistics, they were welcome to ask. On the other hand, from the point of view of the designers of the variable catalogue, there was a tendency to include too many details in the data documentation, just because these details "may be useful to have". Another problem was that the variable catalogue would not become really useful for practical use before it was reasonably complete, and, on the other hand, as long as there was not practical use of the documentation, it was hard to motivate people to prioritise the work necessary to make it complete – a Catch 22 situation.

The bad news about the variable catalogue is that it was never completed, and never became operational; after several years of struggling, the difficult decision was finally taken to abandon the project. The good news is that we took the time to analyse our failure thoroughly, and we learnt a lot from these analyses that we have used successfully in later metadata projects. See for example Sundgren (1993b, 1995, 1999, 2004a). Sundgren (2004a) formulates a set of Golden Rules for designers, project leaders, and top managers involved in the design of statistical metadata systems.

The lessons learnt from the early failures in the development of metadata systems have contributed to more successful achievements during the following decades, such as the metadata model of the AXIS and PC-AXIS systems, the SCBDOK methodology and templates for documentation of statistical microdata and statistical surveys, a standard for quality declarations of statistics, and, more recently, the MetaPlus system. For more information, see Sundgren (1993b, 1995, 1999b, 2004a, 2004b, 2005a), Rosén&Sundgren (1991), Lindblom&Sundgren (2004), Blomqvist&Lundell&Karling&Svensson (2007).

Privacy and confidentiality issues

During the 1970 population census, the first privacy debate exploded in Swedish media. The management of Statistics Sweden was taken by complete surprise and shock. The integrity and immunity to political and administrative pressure to release data for other purposes than statistical analysis and research had never been questioned before, and now Statistics Sweden was suddenly associated with the concept of a Big Brother society.

I will not go into the details of the privacy debate here, but it had the unfortunate effect that the management of Statistics Sweden became overly cautious about microdata, and in particular the flexible use of statistical microdata implied by the archive-statistical approach. The experimentation and development of archive-statistical microdatabases was stopped. The continued development was focused on making aggregated macrodata more available, by means of the AXIS system already mentioned. This development was very interesting and promising in itself, but the impact of this development would have been even more important, if it had been combined with an underlying microdatabase engine, ARKDABA, as was originally planned.

A positive side-effect of the privacy debate was that Statistics Sweden was given extremely generous appropriations for studying privacy and confidentiality problems of statistics production. We got the opportunity to carry out advanced research projects about how to protect the confidentiality of both person data and business data. The two problem areas turned out to be quite different, although they were both essential for regaining the confidence of people and enterprises.

Major challenges as regards the issue of privacy and statistical confidentiality are to strike the right balances between

- the right to privacy vs the need to know
- legal, methodological, and technical measures for protecting privacy and confidentiality

A major methodological problem is how to avoid inadvertent disclosures in statistical databases and publications, especially how to protect against reidentifications of persons and enterprises behind the figures in statistical tables and anonymised files of microdata. Depending on the background information in the possession of an intruder, as well as the

availability of public information about people and businesses, e.g. in registers, it may be relatively easy to reidentify anonymised microdata and data behind the cells in statistical tables.

It was only during the beginning of the 1990's that a new director general of Statistics Sweden, Professor Sten Johansson, had the courage to let methodologists like myself reopen the case of how to make statistical microdata more available to researchers and analysts. A number of projects were started to cope with the legal, administrative, methodological, and technical issues. New technology and software such as SuperCross of the Australian company Space Time Research helped in these endeavours, and both Denmark and Sweden developed secure systems for facilitating remote access to microdata. The Swedish system is called MONA. A new law was introduced in Sweden, criminalising all attempts to reidentify anonymised microdata.

For more information about privacy issues and issues of statistical confidentiality, see Fellegi (1972), Sundgren (1972, 1993a, 1999c, 2001a), Olsson (1973), Barabba (1974), Rapaport&Sundgren (1975), Block&Olsson (1976), Dalenius (1977, 1988), Flaherty (1989), Westergaard-Nielsen&Mathiesen (2003), Thygesen&Andersen (2003), Trewin (2006).

Svein Nordbotten also participated actively in the work on privacy and confidentiality. Among other things he was a member of the commission appointed by the Norwegian government for the preparation of Norwegian legislation in this field. See also Nordbotten (1968, 1971).

Generalised software

Around 1970, when the archive-statistical developments were taking off with full strength at Statistics Sweden, involving some 20 academically educated experts and doctoral students, the concept of generalised software was known and practised only in limited circles. Mathematicians developed and used standardised procedures and subroutines for mathematical and statistical computations. Computer-oriented, technical programmers (systems programmers) developed and used operating systems, assemblers, compilers, utilities, and other systems software for managing computers and application programs, tailor-made for specific applications by application programmers in machine-code, "almost machine code" (assembler languages), or so-called high-level languages, or 3rd generation languages (3GL), like COBOL, FORTRAN, and PL/1. At Statistics Sweden all statistical products, or surveys, had a group of application programmers allocated to them, who developed tailor-made programs for the operations of a specific survey, like the labour force survey or the population census, e.g. sending out questionnaires, entering of data from filled-in questionnaires into the computer, coding and editing of data, tranformations of received data into derived variables, aggregation of data into statistical estimates, tabulation and presentation of statistics. All in all, the development of tailor-made application programs occupied about 150 specialised application programmers at Statistics Sweden at this time. Developing tailor-made application programs was time-consuming, error-prone, and expensive, and it often resulted in inflexible, ill documented programs that were difficult to maintain for reuse, for example when more or less the same survey was to be repeated another month, quarter, or year.

In this respect, the situation at Statistics Sweden was very similar to the situation in other organisations that had now begun to use computers for administrative purposes like order processing, book-keeping, etc. The applications were not very advanced from a mathematical point of view, but they could involve quite complex data management operations. Computers

were more and more used as data processors rather than for mathematical computations. At first, the data management operations did not seem as obvious candidates for generalisations as mathematical computations. Although some of the operations of a statistical office are mathematical in the traditional sense, most of them are actually not, but resemble the data management operations of administrative applications.

Universities and commercial software developers had started to develop and market certain types of generalised software, mainly packages of mathematical and statistical procedures and subroutines, but generalised software intended for administrative applications and operations were seldom on a higher level than that of procedural languages like COBOL. Some simple, non-procedural report generators were emerging, though. A non-procedural programming language is a language where you only have to specify the input and the output, whereas the procedures necessary to tranform the input to the output are generated by a generalised software product, e.g. a compiler.

In this environment, Statistics Sweden started the development of four suites of generalised software products:

1. The non-procedural table generator TAB68 and a family of related software products for other typical operations in a statistical production process: data editing, estimation, etc.
2. The so-called DBC programs, intended for simple but frequently occurring data transformations.
3. The metadata-driven AXIS system for managing multidimensional, aggregated statistical data (including time series data), with their associated metadata, making aggregated statistics available on-line for all kinds of users of statistics, in a timely, flexible, and user-friendly way.
4. The Base Operator System, corresponding to a relational algebra for statistical operations, inspired by experiences from using the relational database management system RAPID, developed by Statistics Canada.

The TAB68 software family was developed during the 1970's. TAB68 itself was ready around 1973, and it became very popular – and much more so among the ordinary statistical staff at Statistics Sweden than among the programmers, who maybe felt a bit threatened. The TAB68 development was very costly, but also very profitable when seen in retrospect. The whole investment was paid back in about one year.

The Base Operator System was the result of a successful international cooperation during the early 1980's within the framework of UN/ECE. The product became popular and frequently used in some countries, but not in Sweden, maybe because commercial relational software had now become an attractive alternative to file processing systems. For a short description of the Base Operator System, see Sundgren (1999a).

Standardised interfaces

In the mid 1970's the software legacy at Statistics Sweden by and large consisted of traditional, tailor-made application systems, developed by application programmers. As was just said, TAB68 had become very popular among non-programmers,especially for responding to *ad hoc* demands from external users. This was clearly in line with Nordbotten's early vision of flexible statistical processing "on demand" of data already collected. However, the vast majority of "heavy" statistics production systems were still tailor-made, rigid, and difficult to maintain.

One reason for the complexity of the tailor-made application systems was that the data structures used for those systems were often tailor-made, too, optimised for machine-efficient processing. Thus many files consisted of complex hiearchies of variable-length records and subrecords, with little or no standardisation.

The first idea to cope with this situation was to continue the development of TAB68 and other emerging, generalised software tools, so as to be able to cope with different types of data structures, rather than only the "punch-card"-inspired flat files that we had used so far. However, we soon realised that such a development would be very complex and costly, and there would always remain data structures that we would not be able to cope with.

The second idea was the inverse of the first idea: why adapt the software to the existing data structures, why not instead adapt the data structures to the software, by transforming all kinds of data structures into the simple and uniform flat file structure? As a matter of fact this was exactly what the emerging relational data model proposed, using another, set-theoretically influenced terminology; cf Codd (1970).

It should be remembered that at this time, in the middle of the 1970's, the relational data model was regarded as a very theoretical model that would never work efficiently enough in a real, practical environment outside the academic world and its toy applications. We received the same criticism for our flat file approach. However, since I had just become the head of the unit with the 150 centrally placed application programmers of Statistics Sweden, I could actually order them to follow the flat file policy in combination with maximal use of generalised software. Since I could not be 100% sure that the new policy would work, I left an opening for exceptions. If the system developers had tried the new approach, and the resulting system turned out to become too inefficient because of this, I was ready to approve of alternative solutions, but only in those parts of the system, where the inefficiencies became intolerable. I never had to approve of any exceptions, and of course, not so long afterwards the relational data model became the industry standard for data management, and all talk about toys for academics disappeared rather quickly. "There is nothing more practical than a good theory." (Kurt Lewin).

New organisation

Nordbotten (1966) proposes also a new way of organising the production of official statistics. The author suggests that the observation data should be collected from respondents in a more continuous way, thus spreading these costly activities more evenly over time, and making more optimal use of the resources needed. Furthermore, he suggests that all data needed by different surveys from the same respondent, e.g. the same company, should be collected at the same time, instead of letting every survey make their own data collection, thus disturbing the same respondent, and reiterating the same costly data collection procedure, as many time as there are surveys requiring data from that respondent. As soon as incoming observation data have been properly prepared, they should be stored in the active files part of the statistical data archive and made easily available for further processing in combination with other data in the data archive, as needed. A number of predefined regular publications should be produced and published, but, even more importantly, a large and growing number of statistical outputs, tailored to the special needs of different users and usages, should be promptly, flexibly, and inexpensively produced on an *ad hoc* basis, as needs arise.

The top management of Statistics Sweden realised that in order to realise the full potential of the archive-statistical principles suggested by Nordbotten, a rather drastic reorganisation was needed – a reorganisation of both the processes and the staff managing and operating the processes. In the early 1970's the deputy director general, Lennart Fastbom, and his collaborators (among them Christer Arvas and myself) worked out the details of such an organisation. A proposal was ready for being presented and discussed in 1974; Fastbom (1974). However, even before the proposal had been printed and published, rumours about it were spreading around Statistics Sweden like a bushfire. And the reactions to the proposals were devastating. The proposal was criticised apart and together from the unions as well as from middle management. The archive-statistical principles were labeled as "inhuman", treating the statistical production process as if it were a comparable with a "mechanical" process in the manufacturing industry – which it actually is to a large extent, in my mind.

Fastbom (1974) was never published, but an almost complete manuscript is available.

What was then the contents of this proposal that aroused so much sentiments and aggression? Some key points in the proposal were:

- The traditional stovepipes, based on traditional statistical surveys, organised by topics and containing all production steps from data collection to publishing, would be broken up and reconsolidated into three major parts:
    - the input operations, responsible for the collection and preparation of data from different respondents and other data sources, e.g. administrative systems
    - the statistical data archive (data warehouse, as we would call it today) and associated thruput operations, taking care of input data, after it had been collected and prepared, transforming and organising the data in a well-structured way
    - the output operations, responsible for serving different users and usages by retrieving and combining data from the statistical data archive.
- The three major parts of the organisation would be specialised on quite different tasks, requiring quite different competences; these competences would be further developed by growing experience and methodological knowledge, stimulated by very focused research and development projects, supported by the best experts in the respective fields; for example, special methodological efforts would be devoted to the problems and opportunities of using administrative data and registers for the production of official statistics
- The output-oriented part of the organisation would be particularly customer-oriented, focusing on being very attentive to the special needs of different users of statistics and further developing these talents by taking active part in analytical and other tasks performed by users of official statistics

These visions and plans had to be buried. Instead the unions started a campaign for organisational changes in the opposite direction, labeled as "defunctionalisation", aiming at decentralising as many resources as possible from central, functional units to the survey-centred departments and units, organised by topics – and not by respondents or users.

As has already been described in earlier parts of this paper, some important parts of the archive-statistical vision were actually realised during the following years, in particular the developments of database-oriented statistics production, metadata systems, and generalised software. However, the impact of these developments would no doubt have become much more important and productive for all parties concerned, both users and producers of official

statistics, as well as for the respondents and for the quality of statistics, if the organisational changes proposed in Fastbom (1974) had been implemented.

It would take more than 30 years until a new attempt was made to change the organisation and production principles in a more radical way towards the archive-statistical vision. The initiative was taken by Kjell Jansson, director general, who focused on customer-orientation and standardised processes according to the input-thruput-output scheme. This new attempt started in a very promising way – see Sundgren (2007b) – but unfortunately lost momentum after about a year for reasons that I will not discuss here. This reengineering process is still going on and has become even more urgent because of some serious quality problems that have occurred at Statistics Sweden during the last few years. One of the main ideas with a process-oriented organisation lined out by Kjell Jansson, is exactly to come to grips with such quality problems. However, all organisations are dependent on people, and having the right people in the right places. Even if standardised processes should decrease the dependence on individual persons, the organisation as a whole cannot be made independent of competent and imaginative managers and collaborators – fortunately.

For a somewhat deeper discussion about how to best organise a statistical office for producing official statistics, see Sundgren (2004a,b).

Expected future developments

Statistical file systems, archive statistics, and a number of related concepts, like those described in this paper, have now become implemented, at least partially, at Statistics Sweden and in many other statistical offices. The well established international cooperation between statistical agencies, both bilaterally and multilaterally, within the frameworks of international organisations, have been important for spreading the ideas, exchanging experiences, and – at least to some extent – cooperating on the development of architectures, methods, and software. The development is still on-going and is being combined in a fruitful way with new trends like customer orientation, process standardisation, service-oriented architectures, and, most recently, cloud computing. Actually some these trends are not as new as they may seem to be. Like Nordbotten's original vision of an archive-statistical system, they may be seen as different aspects of a holistic systems approach to official statistics; Sundgren (2010b).

Towards an updated and extended vision

Several statistical agencies have recently formulated visions that may be seen as updated and extended versions of Nordbotten's original vision of a model of a production system for official statistics based on archive-statistical principles. See for example Statistics New Zealand (2004), Sundgren (2007a, 2007b). These visions often contains several parts and aspects, e.g. a business model, an architecture, and an organisation, including principles for governance and quality control.

Data editing in an archive-statistical system

In parallel with his early engagement in the development of an archive-statistical system, Svein Nordbotten was very active in developing methods for automating the statistical data editing and imputation process, which has traditionally been very labour-intensive and costly, and still is. Nordbotten published the seminal papers Nordbotten (1963) and Nordbotten (1965) more than 10 years before Fellegi & Holt published their famous paper within the same field; Fellegi&Holt (1976).

I am not sure whether Svein Nordbotten originally saw his contributions to a modernised and computerised data editing methodology as an integral part of his vision of an archive-statistical system, but in the updated and extended visions that I referred to above, data editing is certainly a process that is in focus, not least for economical reasons. Several studies have shown that data editing accounts for about 40% of the total costs of the production of official statistics, and this fact alone motivates focus on standardised, rationalised, and more "intelligent" data editing methods in any vision of a future integrated system for statistics production. Let me briefly mention some important progress that has taken place during recent years, and where Svein Nordbotten has played an important role.

Together with Leopold Granquist and others, Svein Nordbotten participated in the development of so-called macro-editing, later renamed "significance editing" or "selective editing" not to be mixed up with the kind of editing or "plausibility checks" that is often made just before final statistics, or macrodata, are going to be published. Selective editing aims at prioritising checks of those suspicious microdata, which, if really erroneous, would most significantly affect the estimates that are going to be computed by aggregating the microdata into statistics. If the estimates to be made are known in advance, as they are in traditional statistics production, selective editing is relatively straightforward, and may easily save 30-50% of the resources needed for data editing.

In an archive-statistical system, where many of the estimates to be made in the future are not known at the time of data collection, selective editing is associated with more difficult methodological problems, some of which still remain to be approached and solved.

Furthermore, Nordbotten has made innovative use of neural networks to manage editing and imputation problems.

For more information about modern editing methods, see Nordbotten (1995, 1996a, 1996b, 1997a, 1997b, 1998, 2000a, 2000b, 2000c), Granquist (1997, 2005), Charlton & Chambers & Nordbotten (2001), Norberg & Jäder (2005), Granquist & Kovar & Nordbotten (2006), Gåsemyr & Nordbotten & Andersen (2008). Further research is needed to investigate could best be used in a modern version of an archive-statistical system for statistics production.

New sources of statistical raw data

Yet another research area, where Svein Nordbotten has since long expressed interesting ideas, is how to exploit, for statistical and analytical purposes, the new and extremely rich sources of raw data made available through the computerisation of almost all processes in society, and not least through the rapidly growing penetration of Internet-based systems and activities. Capturing, transforming, and reusing these new sources of data for statistical purposes could turn out to become a fascinating extension of the original archive-statistical idea of reusing data from administrative systems and registers. However, there are considerable methodological problems that need to be tackled.

Participative design of statistical systems: reconciling conflicting goals

Designing a survey is a complex decision process, and it becomes even more complex when whole systems of surveys and registers should be considered, like in an archive-statistical system. In order to see similarities with other complex decision processes, one may describe the design of a statistical survey in the following way, translated from Sundgren (2010a):

- There may be many stakeholders in the process, e.g.

    - different kinds of users: ministries, researchers, analysts, business people, teachers and students, journalists, interested citizens
    - sponsors: taxpayers via the parliament and ministries, companies, organisations
    - respondents and other providers of data

- Different stakeholders will have different, and partly conflicting demands concerning the contents, costs, and qualities (relevance, timeliness, comparability, etc) of the statistical products and services

- Even every single stakeholder by herself may have contradictory demands

- There are a very large number of decision alternatives (possible designs), which may be difficult to overview in a systematic way – and even more difficult to evaluate and compare as regards the outcome in relation to the different (desirable) goals, especially since some factors may be quantifiable, whereas others are not

This type of decision situations occur in many different contexts in society. There is research on decision analysis, producing methods and tools focusing on

- getting a constructive cooperation between different stakeholders during the whole decision process
- structuring the decision alternatives in a manageable way, providing a good overview
- structuring the wishes and preferences of different stakeholders by means of weights
- making the decision analysis transparent, making sensitivity analyses
- finding "reasonably satisfactory" and Pareto-optimal solutions, which are acceptable to the stakeholders, rather than "optimal" solutions, which hardly exist
- building methods and tools (tool-boxes) supporting complex decision processes, as described above

A research group at Stockhom University has produced a decision process model and a tool-box for this kind of problems, "the DSV-DECIDE model for participative decision analysis and decision support". It has been tested in practice in a number of Swedish municipalities. The applications typically concern problems like the localisation of unattractive businesses, or projects with multi-faceted and complex environmental consequences. See Sundgren & Larsson (2009).

There are relatively few articles and books in the traditional statistical literature addressing the reconcilation between conflicting goals in the design of statistical surveys. Lars Lyberg has given a few simple examples of typical goal conflicts of this kind, and how they may be treated, in Biemer&Lyberg (2003), chapter 10.

Svein Nordbotten has expressed his interest to participate actively in this research.

Conclusion

I have been inspired and influenced by Svein Nordbotten during my whole career – at Statistics Sweden, from the day I started in April 1968, and in my role as an academic researcher. Svein Nordbotten was the faculty opponent when I presented my doctoral thesis in 1973 at Stockholm University. We have also done a lot of more practical work together, always very interesting and pleasant, e.g. the evaluation of Statistics Denmark; Sundgren &

Nordbotten (2003). One of the major achievements of Svein Nordbotten is his early vision of an archive-statistical system – and all the consequences of this vision in the statistical world, not least in Sweden. This has been the main topic of this paper. Fortunately Svein is still working at full strength, full of creativity as always, and I am looking forward to continued close contacts with Svein, both as a professional and as a friend.

With his open mind, broad and deep research interests, and holistic approach, Svein Nordbotten is a true champion of a systems approach to official statistics; Sundgren (2010b).

References

Barabba, V. P. (1974). *The Right of Privacy and the Need to Know*. Proceedings of the Social Statistics Section, American Statistical Association 33.

Biemer, P.P. & Lyberg, L.E. (2003). *Introduction to Survey Quality*, Wiley.

Block, H. & Olsson, L. (1976). *Backwards Identification of Person Information*. Statistical Review, 14, 135-144.

Blomqvist, K. & Lundell, L-G & Karling, C. & Svensson, E. (2007). *MetaPlus – a new metadata system for documenting microdata at Statistics Sweden*. Statistics Sweden.

Charlton, J., Chambers, R. and Nordbotten, S.(2001). *New developments in edit and imputation practices – needs and research*. 53rd Session of the International Statistical institute. Seoul 2001.

Codd, E.F. (1970). *A Relational Model of Data for Large Shared Data Banks*. *Communications of the ACM* **13** (6): 377–387.

Dalenius, T. (1977). *Towards a Methodology for Statistical Disclosure Control*. Statistical Review, 15, 429-444.

Dalenius, T. (1988). *Controlling Invasion of Privacy in Surveys*. Statistics Sweden, R&D Reports.

Fastbom, L. (1974). *ARKSY-utredningen*. Statistics Sweden, internal paper, in Swedish. Available for free downloading (in six pieces) from http://sites.google.com/site/bosundgren/ .

Fellegi, I.P. (1972). *On the Question of Statistical Confidentiality*. Journal of the American Statistical Association, 67, 7-18.

Fellegi, I.P. & Holt, D. (1976). *A Systematic Approach to Automatic Edit and Imputation*. Journal of the American Statistical Association, 71, 17-35.

Flaherty, D.H. (1989). *Protecting Privacy in Surveillance Societies: The Federal Republic of Germany, Sweden, France, Canada, and the United States*. University of North Carolina Press, Chapel Hill.

Granquist, L. (1997). *The new view on editing*. International Statistical Review 65 (3), 381-387.

Granquist, L. (2005). *Data editing - improving surveys?* In 55th Session of the International Statistical Institute, Sydney, Australia.

Granquist, L., Kovar, J. and Nordbotten, S. (2006). *Improving Surveyes - Where Does Editing Fit In?*, Chapter 4 in Statistical Data Editing, Vol. 3: Impact on Data Quality. UNECE Conference of European Statisticians. Geneva 2006.

Gåsemyr, S. & Nordbotten, S. & Andersen, M.,Q. (2008). *Role of Editing and Imputation of Sources for Structural Business Statistics"*, Proceedings from UN/ECE Workshop Session on Statistical Editing in Vienna. Geneva 2008.

Klimbie, J.W. & K. L. Koffeman, K.L. (1974) eds. *Data Base Management*, Proceeding of the IFIP Working Conference on Data Base Management, Cargèse, Corsica, France, 1-5 April, 1974. North-Holland.

Langefors, B. (1961a). *Information Retrieval in File Processing 1*, BIT, Vol. 1, No.l, Copenhagen, 1961, pp. 54-63.

Langefors, B. (1961b). *Information Retrieval in File Processing 2*, BIT, Vol. 1, No.2, Copenhagen, 1961, pp. 103-111.

Langefors, B. (1963). *Some Approaches to the Theory of Information Systems*, BIT, Vol. 3, No.4, Copenhagen, 1963, pp.229-254.

Langefors, B. (1966). *Theoretical Analysis of Information Systems.* Studentlitteratur, Lund.

Lindblom, H. & Sundgren, B. (2004) *The Metadata System At Statistics Sweden In An International Perspective* Invited paper for the conference "Statistics – investment in the future", Prague, Czech Republic. Available for free downloading from http://sites.google.com/site/bosundgren/.

Lundell, L.-G. (2009). *Strukturerade datalager för effektivare produktion*. (Structured data warehouses for more effecient production.). Statistics Sweden. In Swedish. Available for downloading from http://sites.google.com/site/bosundgren/my-library?pli=1.

Norberg, A. & Jäder, A. (2005). *A selective editing method considering both suspicion and potential impact*, UNECE Work Session on Statistical Data Editing, Ottawa, Canada.

Nordbotten, S. (1960). *Elektronmaskinene og statistikkens utforming i årene framover*, De Nordiske Statistikermøter i Helsingfors 1960, Helsinki 1961, pp.135-141. Available for free downloading from www.nordbotten.com.

Nordbotten, S. (1963). *Automatic Editing of Individual Statistical Observations*, Statistical Standards and Studies, No. 3, United Nations. Available for free downloading from www.nordbotten.com.

Nordbotten, S.(1965). *The Efficiency of Automatic Detection and Correction of Errors in Individual Observations as Compared with other Means of Improving the Quality of Statistics,* Proceedings from the 35th Session of the International Statistical Institute. Beograde 1965. Available for free downloading from www.nordbotten.com.

Nordbotten, S. (1966). *A Statistical File system.* Statistisk Tidskrift, Stockholm. Available for free downloading from www.nordbotten.com.

Nordbotten, S. (1967a). *On Statistical File System II.* Statistisk Tidskrift. Stockholm. Available for free downloading from www.nordbotten.com.

Nordbotten, S. (1967b). *Automatic Files in Statistical Systems.* Statistical Standards and Studies. Handbook No. 9. United Nations. N.Y. Available for free downloading from www.nordbotten.com.

Nordbotten, S. (1967c). *Purposes, Problems and Ideas Related to Statistical File Systems.* Proceedings from the 36th Session of the International Statistical Institue. Invited paper. Sydney. Available for free downloading from www.nordbotten.com.

Nordbotten, S. (1968). *Konfidensiell behandling av data, informasjonsnytte og klassifisering av data*, Statistisk Tidskrift, Nr. 5, Stockholm 1968. In Norwegian.

Nordbotten, S. (1971). *Lovgivning, og statistisk informasjon*, i Larsen, J. (red.): Lovgivning, administrasnon og samfunn, Universitetsforlaget, Oslo 1971. In Norwegian.

Nordbotten, S.(1995). *Editing Statistical Records by Neural Networks*, Journal of Official Statistics, Vol. 11, No. 4, 1995. Available for free downloading from www.nordbotten.com.

Nordbotten, S.(1996a). *Neural Network Imputation Applied to the Norwegian 1990 Population Census Data.* Journal of Official Statistics, Vol. 12, No. 4. Available for free downloading from www.nordbotten.com.

Nordbotten, S.(1996b). *Editing and Imputation by Means of Neural Networks.* Statistical Journal of the UN/ECE, 13. Available for free downloading from www.nordbotten.com.

Nordbotten, S.(1997a). *Models of Complex Human Screening and Correction of Social Data*, Computers in Human Behaviour, Vol. 13, No. 4, 1997. Available for free downloading from www.nordbotten.com.

Nordbotten, S.(1997b). *Metrics for the Quality of Editing, Imputation and Prediction*, UN/ECE Workshop on Statistical data editing, Prague. Available for free downloading from www.nordbotten.com.

Nordbotten,S.(1998). *New Methods of Editing and Imputation*, International Conference on Agriculture Statistics, Washington D.C. 1998. Available for free downloading from www.nordbotten.com.

Nordbotten, S.(2000a). *Statistics Sweden's Editing Process Data Project*. International Conference on Establishment Surveys II, June 17-21, 2000, Buffalo, N.Y. Available for free downloading from www.nordbotten.com.

Nordbotten, S.(2000b). *Meta-data about Editing and Accuracy for End Users*, UNECE Workshop on Statistical Data Editing, Cardiff, 2000. Available for free downloading from www.nordbotten.com.

Nordbotten, S.(2000c). *Evaluating Efficiency of Statistical Data Editing: General Framework* UN/ECE Conference of European Statisticians - Methodological Material. Geneva. 2000. Available for free downloading from www.nordbotten.com.

Nordbotten, S. (2010a). *Experiments with Image Content Analysis*, University of Bergen. Bergen. Available for free downloading from www.nordbotten.com.

Nordbotten, S. (2010b). *The Use of Administrative Data in Official Statistics – Past, Present, and Future – With Special Reference to the Nordic Countries,* Chapter 17 in Official Statistics in Honour of Daniel Thorburn, pp. 205–223. Available at officialstatistics.wordpress.com.

Olsson, L. (1973). *Measures to Protect Privacy in a Statistical Data Base*. Statistics Sweden, R&D Reports. Available for free downloading from www.nordbotten.com.

Rapaport, E. & Sundgren, B. (1975). *Output Protection in Statistical Data Bases* Conference of the International Statistical Institute (ISI), Warsaw, Poland.

Rosén, B. & Sundgren, B. (1991). *Documentation for reuse of microdata from the surveys carried out by Statistics Sweden.* Statistics Sweden. Original report in Swedish. English translation available. Available for free downloading from http://sites.google.com/site/bosundgren/.

Statistics Netherlands (2004). *The Dutch Virtual Census of 2001 – Analysis and Methodology*. Statistics Netherlands.

Statistics New Zealand (2004). *End-to-End Business Model – Business Model Transformation Strategy*. Statistics New Zealand. Available for downloading from http://sites.google.com/site/bosundgren/my-library?pli=1.

Sundgren, B. (1972). *Security and Privacy of Statistical Data Bases"* Statistical Review, 10, 4, 299-312.

Sundgren, B. (1973). *An infological approach to data bases*. Stockholm University and Statistics Sweden. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (1974). *Conceptual Foundation of the Infological Approach to Data Base Management* In Klimbie & Koffeeman (eds), "Data Base Management", North-Holland, Amsterdam.

Sundgren, B. (1993a). *Discussion – Computer Security* Journal of Official Statistics, Vol.9, No.2, pp. 511-517. Available for free downloading from www.jos.nu.

Sundgren, B. (1993b). *Guidelines on the Design and Implementation of Statistical Metainformation Systems.* Report for the UN/ECE METIS Group, established within the programme of work of the Conference of European Statisticians. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (1995). *Guidelines for the Modelling of Statistical Data and Metadata.* Published as Guidelines from the United Nations Statistical Division, New York 1995. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (1997). *Sweden's Statistical Databases - An Infrastructure for Flexible Dissemination of Statistics* Conference of European Statisticians.

Sundgren, B. (1999a). *Business Modelling in a Historical Perspective – Experiences from Statistics Sweden.* In " Perspectives on business modelling: understanding and changing organisations", edited by Anders G Nilsson et. al.

Sundgren, B. (1999b). *Information systems architecture for national and international statistical offices. Guidelines and recommendations*. Conference of European Statisticians Statistical Standards and Studies No. 51, United Nations 1999. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (1999c). *Increasing the availability of Sweden's official statistics* (Available only in Swedish: "Ökad tillgänglighet till Sveriges officiella statistik") Appendix to the evaluation of the reform of Sweden's official statistics, SOU 1999:96. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2001a). *Statistical Microdata – Confidentiality Protection vs. Freedom of Information*. Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. Skopje, FYROM. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2001b). *The αβγτ-model: A theory of multidimensional structures of statistics*. The MetaNet conference in Voorburg, the Netherlands, April 2001. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2004a). *Statistical Systems – Some Fundamentals*, Statistics Sweden 2004. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2004b). *Designing and Managing Infrastructures*, Statistics Sweden 2004. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2005a). *Modelling Statistical Systems* 55th Session of the International Statistical Institute, Sydney, Australia, 2005. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2005b). *Modelling the Contents of Official Statistics*, Meeting of the SDMX Group at the OECD in Paris 2005. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2005c). *A Conceptual Model of Society as Reflected by Official Statistics* Joint Statistical Meetings of the American Statistical Society, Minneapolis, Minnesota, USA, 2005. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2007a). *Lotta: konkretiserad målbild – scenario*. Statistics Sweden. In Swedish – English translation available: "Lotta - a scenario of the target state". Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2007b). *Process reengineering at Statistics Sweden* Meeting of the Management of Statistical Information Systems (MSIS), organised by the United Nations Statistical Commission in cooperation with the European Commission and the Organisation for Economic Cooperation and Development, May 2007. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2010a). *Design av statistiska undersökningar och statistiska system – komplexa beslutsprocesser*. (Designing statistical surveys and statistical systems – complex decision processes.) Paper submitted to the Scientific Council of Statistics Sweden. In Swedish. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2010b). *The Systems Approach to Official Statistics*, Chapter 18 in Official Statistics in Honour of Daniel Thorburn, pp. 225–260. Available at officialstatistics.wordpress.com.

Sundgren, B. & Larsson, A. (2009) *The DSV-DECIDE model for participative decision analysis and decision support*. Stockholm University. Appendix to Sundgren (2009). Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. & Nordbotten, S. (2003). *Review of Statistics Denmark.* Økonomi- og Erhvervsministeriet, Danmarks Statistik og Finansministeriet. Copenhagen. Available for free downloading from www.nordbotten.com. Also available for free downloading from http://sites.google.com/site/bosundgren/.

Thygesen, L. & Andersen, O. (2003). *The Danish System for Access to Microdata – From on-site to remote access*. Workshop on Microdata, Statistics Sweden, Stockholm. Available for downloading from http://sites.google.com/site/bosundgren/my-library?pli=1.

Trewin, D. & Task Force (2006). *Guidelines and core principles for managing statistical confidentiality and microdata access.* UNECE, Conference of European Statisticians. Available for downloading from http://sites.google.com/site/bosundgren/my-library?pli=1.

Wallgren, A. & Wallgren, B. (2007). *Register-based statstics – Administrative data for statistical purposes.* Wiley.

Westergaard-Nielsen, N. & Mathiesen, H. (2003). Research Data Experiences in Denmark. Aarhus School of Business and International Statistical Institute. Available for downloading from http://sites.google.com/site/bosundgren/my-library?pli=1.

Zhang, L.-C. and Nordbotten, S, (2008). *Prediction and Imputation in ISEE: Tools for More Efficient Use of Combined Data Sources*. Proceedings from UN/ECE Workshop Session on Statistical Editing in Vienna. Geneva. Available for free downloading from www.nordbotten.com.