Dok nor 9, 28/07/10

Technological development and the vision of an archive statistical system in Statistics Norway

Rune Gløersen, Olav Bjerkholt, Svein Gåsemyr and Jenny Linnerud, Statistics Norway

1 Introduction

At the tri-annual meeting of Nordic statisticians in Helsinki 1960 Svein Nordbotten of Statistics Norway presented his ideas and vision for an archive statistical system, [1]. The presentation led to similar activities in all Nordic countries, where the respective national statistical institutes (NSIs) initiated early steps in a development towards an archive statistical system as outlined by Nordbotten, see e.g. [4]. The 1960 presentation thus became the starting point for the development of what has become known as the Nordic approach to official statistics. An English version of parts of the 1960 paper was published in 1966, [2], see also [3]. Nordbotten defined an *archive statistical system* as "a production of statistics with continuous compilation of data independent of the traditional statistical branches and census dates and where the processing of the filed material is undertaken as the user needs arise and completely independent of the compilation." In more recent terminology *register based statistics* has been used as an expression which, even if not perfectly coinciding with Nordbotten's ideas as originally stated, covers the core content of these.

The technological development in Statistics Norway since the 1960s discussed in this paper is viewed to a large extent as to whether it has been conducive and supportive of promoting and fulfilling Nordbotten's vision. It is a history of ups and downs, coloured by optimism and pessimism. However, it should be noted that important elements of the archive statistical system are not dependent on specific technological developments. The fascinating emergence of computer technology, still at an embryonic stage in the late 1950s, may have served as a catalyst for Nordbotten to form his vision, and the ensuing technological development as an enabler to reach important objectives, but the main obstacle for an overall implementation of the vision and the reason for its lack of fulfillment has not been technological.

2 Early developments

2.1 Towards the universal machine

Before 1940, compilation of official statistics was mainly based on manual grouping and counting of paper forms for statistical surveys, statistical censuses and administrative sources. The US Bureau of Census started to use punch card machines for the 1890 Population Census. Statistics Norway acquired Hollerith punch card machines in 1894 (several keypunches, a sorter and a tabulator), in preparation for the Norwegian 1900 Population Census. On a large scale manual methods were not replaced by punch card machines until the 1940s and 1950s. By and large, we can say that the 1950s represent the culmination of the era of the punch card machines. At the same time, from the very beginning of the decade, the interest in emerging universal electronic computers started, in the wake of the pioneering development in the USA

in the immediate postwar period. Statistics Norway became one of the front runners among national statistical institutes in Europe in taking first-generation computers in use for statistical purposes, [5], [6a] and [6b].

A punch card machine was specialized to work out a standard process. The statistical production was organized as a sequential chain of processes using different kinds of punch card equipment. The processing sequence could typically be:

> The **keypunch** transferred information from paper documents to cards
> The **verifier** accepted punched cards for re-keying. Erroneous cards were re-punched
> The **sorter** sequenced cards into any of 80 columns. The cards were routed into the correct pocket
> The **collator** sequence-checked, merged, matched, and selected cards
> The **reproducer** reproduced duplicate cards, punching copied detail cards from a single master card, performed mark sensing, and summary punching
> The c**alculator** was capable of addition, subtraction, multiplication and division
> The **tabulating machine** was capable of adding, subtracting, summarizing totals and printing tabulated reports

Other machines like the Summary punch and the Interpreter were also used for specific purposes. In 1950, Statistics Norway procured the famous IBM 101 (Electronic Statistical Machine) that could count and summarize and in parallel, sort the cards and print tables on paper.

A universal machine that could work out all or several of these functions did not exist, but became a vision for the expectations of future possibilities, see [6b].

Preparatory investigations of the possibilities for getting access to a universal electronic computer started in Statistics Norway soon after Nordbotten started working there, in 1952. Many had read popular and speculative articles about ENIAC and other pioneering efforts. Ragnar Frisch at the Institute of Economics at the University of Oslo, who had close relations with Statistics Norway, had established contact with prominent US researchers in the field for his own interest in numerical analysis for economic modelling. Petter Jakob Bjerve, the Director General of Statistics Norway from 1949, had spent 1947-49 at economic research centres in the USA. It became quickly known that the most important American statistical institution, *Bureau of Census*, in 1951 had acquired the first commercially available computer, Univac 1, for its statistical work. Nordbotten became involved in the discussions of computers at an early stage and after only one year in Statistics Norway, he was sent to the USA financed by a program for exchange of civil servants under the Marshall Plan, see [7]. Other staff members were also sent on similar missions and US experts involved in input-output analysis visited Norway.

The briefings and information gathered through these study visits about computers was important at a time when computers were not yet available commercially. The deliberations within Statistics Norway about computers went clearly in the direction that it would be of great interest to look further into the possibilities for taking computers into use both for statistical purposes and for research. The acquisition of a computer would, however, require an investment of substantial funds and would not be possible without approval and funding decided at a high political level. From 1956 Nordbotten was put in charge of planning and

preparation in connection with the possible future acquisition and use of computers in Statistics Norway. This comprised the making of a comparative survey of available computers. The preparatory process ended successfully in permission by the government and funds secured for procuring the computer DEUCE produced by *English Electric*. The ultimate decision was made by the Storting (Parliament) in May 1958. The huge computer was put in place in Statistics Norway's quarters and was in operation from February 1959. Nordbotten was in charge of supervising the installation, and later the programming work and the operation of DEUCE and also other computer equipment.

DEUCE was at the time a fast and powerful computer using a magnetic drum for storage and needless to say, with rather limited memory capacity. Input and output was only by punched cards, no printers attached. Already in 1961 Statistics Norway purchased an IBM 1401 computer, which introduced high speed input/output data transfer and high storage capacity using 4 magnetic tape drives. In addition the IBM 1401 operated a 600 line pr. minute printer.

The two computers were used to complement one another, and together they formed at least the chimera of the functions of the universal machine. DEUCE was scrapped in 1965 and replaced by an IBM 360/40 in 1966; the era of the mainframe had begun.

Nevertheless, from 1961 the technology was present in terms of functionality and capacity to be able to start to implement the ideas of the archive statistical system on a larger scale.

2.2 Background for the idea of the archive statistical system

In a presentation for the Director Generals of Nordic NSIs, 1966 [8], Svein Nordbotten stated that the ideas of the 1960 presentation on archive statistical system were not new. He referred to an ISI report of 1872: "Registre de Population" and commented: "The great difference, however, is that today we may be able to implement these ideas". This also illustrates how technological developments influence non-technological challenges.

Of more interest should be that in 1959, the American Economic Association (AEA) considered the need for preservation and use of micro data files for economic research. The AEA recommended that the Social Science Research Council set up a Committee to study this problem and undertake a program of action. The report of this initiative was published in 1965, see chapter 5.2 below.

When a statistical micro file was stored on punch cards, the cost of reuse of data was limited by the cost of using punch card machines to access the data and to link to other files. Compared with the cost of reusing information directly from paper questionnaires, which could be of the same order as the original cost of processing the survey, the use of punch card was a major achievement, also with respect to the potential reuse of information. In practice these possibilities were not utilized. For example: to execute one run of the micro file of the Norwegian 1950 Population Census, 8000 kilograms of punch cards had to be moved from the archive – in and out of the several punch card machines – and finally back to the archive. To update the statistical information about the population on a regular basis, which is basically the idea behind the archive statistical system, would require frequent operations of this kind.

The logistics in dividing a dataset for a population census into smaller sets which were more practical to handle, and at the same time exploiting the capacity of the different machines and

keeping track of the status of the different subsets and their whereabouts, were discouraging.

When magnetic tapes were the main tool for sorting and linking stored statistical micro files, *reuse* of these files became a much more attractive option. To be able to establish and maintain a data archive with the purpose of further and future utilization of the collected information required the following:

- A permanent and stable identification of the basic statistical units.
- Standardisation of the definitions and characteristics of the unit attributes
- Time specification of the observed value for these attributes

The idea of an archive statistical was elaborated in different papers towards a descriptive model named The Statistical File System, abbreviated SFS hereafter, was described as a "data box" with the above three dimensions. The SFS in a way introduced another main objective for a statistical institution; in addition to producing statistics, keeping an archive of the collected micro data for secondary use became a clearly defined task.

These ideas were not new, but since they now could be realized by the use of the emerging technology, the need to describe and structure these needs into a consistent model of an archive system became evident. The papers on SFS, [9], form a rather complete information model that links directly to our current challenges and developments in the society of statistical institutes. See figure 1, as copied [9], giving an example of one "record" in the data box which holds the observed value at a specific time for the statistical characteristic "income" of the (statistical) unit "Person A".

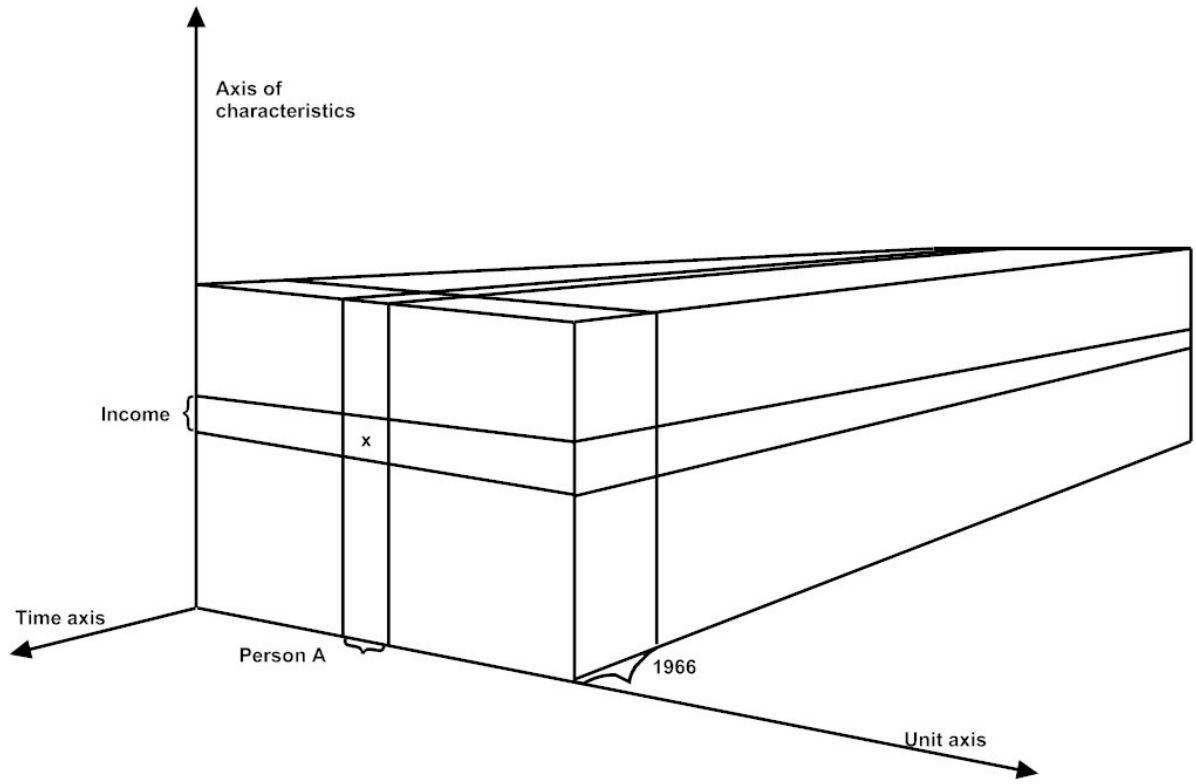


FIGURE 1. The Data Box and the Storage of Income X for person *A* in 1966

2.3 The implication and role of IT

As mentioned, the SFS was presented by Nordbotten around 1960, when the first computers were installed in Statistics Norway. One question striking anyone who reads the papers about the SFS in retrospect, is how it was possible to foresee the possibilities of the new technologies at such an early stage. This question can only be answered by one person, but one reason is perhaps that the new technology did not come completely out of the blue. The computers did actually carry out the same tasks as the various punch card machines did before, just with an impressive higher speed and with the possibility to carry out more than one task at a time, and with the ability to store and retrieve data with fewer manual operations. The processing capacity increased exponentially, giving the opportunity to carry out calculations that hardly were possible beforehand, not because they were not thought of, but they took more time than was considered reasonable for the outcome of the task. Furthermore, the storage capacity was immense compared to the old punch cards. It could be noted though that the technical jargon was to a fair extent a continuation from that of the punch card machines.

Would it have been possible to establish and maintain a complete, coherent archive of all statistical units, their characteristic attributes and the observed attribute values at any specific point of time, at the time when these ideas were launched?

From a storage point of view, the answer is yes. The SFS is a system, a principle of data archiving, where the archive could be broken down into more domain specific cubes or boxes. We should also consider the fact that from the beginning, IT delivered services to store at any time all data retrieved and to process all statistical information requested. However, within service levels not always found acceptable to the end users. The SFS system was challenging in many aspects, but it would also have led to less duplication of stored data, and to an early possibility of concentrating efforts and resources around more common, cross sectional production systems.

In a way, the feasibility had already been shown by the fact that Statistics Norway had kept a statistical Business Register since the business census in 1953. The register was maintained on punch cards until 1965, with regular updates from various sources, leaving no doubt that technology was not an obstacle in itself.

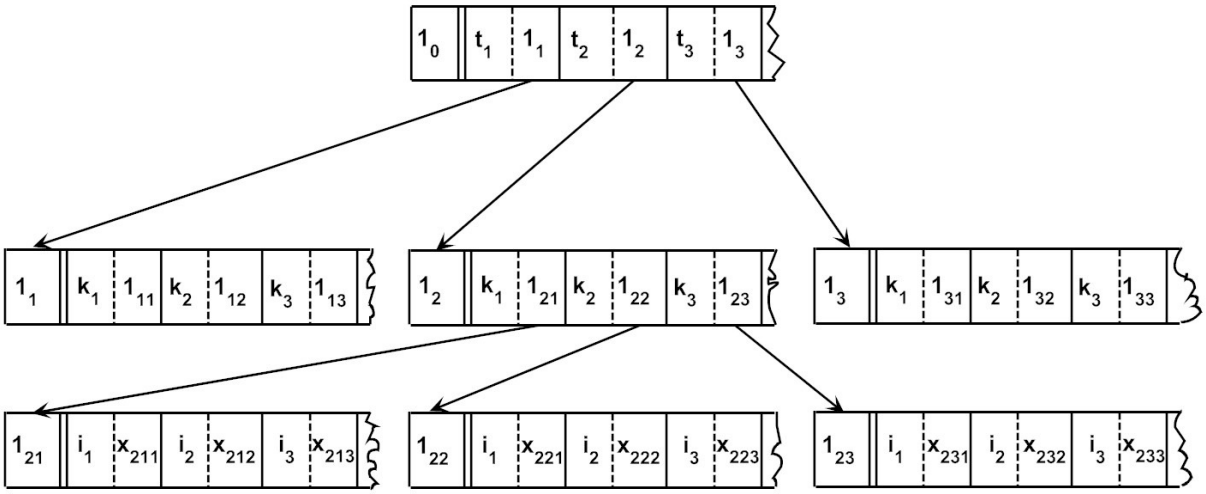
From 1964 the Central Population Register (CPR) was in operation. The task of the CPR was to assign official and unique PIN, (Person Identification Number). The PIN was needed by businesses to reduce response burden and by government agencies to establish electronic administrative data system. The CPR was the first register based on continuous updates, established as a system of chronological files, following the principles of the SFS. This was first implemented on the IBM 1401, on a machine which we today can hardly believe was able to cope with such a task. Once again, technology showed the possibilities.

However, technology also represented a bumpy road. The update of the chronological file was time consuming, the secondary usage perhaps not very comprehensive. This is reflected in Nordbotten's papers, where the plans and developments are always accompanied by a roadmap on expected technological developments. It is obvious that the developments were needed in order to enhance the speed and functionality during update and retrieval of data in the archive. Therefore, already in the late 1960s, Statistics Norway discussed plans for

moving the CPR to a database platform, or at least to move it onto the on-line, random access disc systems, [6].

Nordbotten outlined the SFS as a system of hierarchical files, [9] as shown below in figure 2.

FIGURE 2. File Hierarchy



Statistics Norway implemented a naming convention standard for file labelling in the 1960s based on this. Most likely, the SFS principles were just implemented on the surface. The logical relations in the hierarchy, based upon precise, content based references between data elements and files were most likely not followed. This would have required a supportive documentation system, which to our knowledge did not exist, other than as paper templates for file descriptions. An implementation of such a system would have resulted in an elaboration of what Nordbotten described in 1967 as a “Data Description Language”, later on renamed by Bo Sundgren as “Metadata” which has become international terminology.

Throughout the 1960s, under the supervision of Svein Nordbotten and with a great management support by Director General Petter Jakcob Bjerve, Statistics Norway continued to establish and enhance the base registers. The need for registers in addition to base registers on person and business in areas like land properties and buildings, education, employment etc was outlined. These registers are now well established and managed and maintained outside Statistics Norway, and form the base register infrastructure that is typical for the Nordic countries, extremely profitable for the statistics production, [12] and [15].

At this stage, in the beginning of the 1970s, Statistics Norway was more or less forced into outsourcing of the IT services to the newly established Government Data Centre (GDC). The GDC decided, based on a tender, to procure a Honeywell Bull mainframe. This was almost disastrous for Statistics Norway. All available IT resources were put into a long term work on converting systems from our IBM environment, and more or less all innovations were put on hold. Core competence was lost, and a dark period for IT development in Statistics Norway started, which lasted for 10-20 years. This also coincides with Svein Nordbotten leaving Statistics Norway.

An example of delay in development was that the plans in the late 1960s of moving the CPR onto a database platform did not become a reality until 1985. The CPR was then developed on

a Bull network database platform. This introduces a recurrent topic that always causes controversy among technologists; which technology is the best for a specific purpose?

If we look at technological developments, in the end the brilliant technologies do not always penetrate the market. Concerning network technology, Token Ring was sophisticated, but the more primitive Ethernet won the market, TCP/IP was regarded a migration strategy towards X.400, but the ideal ISO standard more or less vanished. Oracle was not the preferred choice of database experts nevertheless, it became the leading choice. No one could imagine the bandwidth we now can achieve using the simple copper telephone cables. A lesson that could be learnt from Nordbotten is to look beyond the limitations of existing technology is valuable at all times.

The question about which type of database technology that has been regarded as best suited for statistical institutes has also caused a lot of debate. Network databases used physical pointers, which are pointers understood by the database management system only, to locate and relate data. They were well suited for transaction based systems, but not equally good for traversing all data at one time. The relational database model constituted logical relations, but within a rather simple model, at the beginning suffering from bad performance compared with existing hierarchical and network based systems. However, the principles were easier to understand, and brought technology closer to the user.

The Neuchâtel model for managing statistical classifications was first implemented in an object oriented database system. When XML-oriented databases occurred, they were found very promising for part of the statistical data. This reflects for instance that up until now, the technical database system has been discussed and regarded as closely related to the complexity of the data models. The SFS in the end will be managed and structured by the description of the content, which are the metadata and the structures of metadata. Nowadays, we see that this can be achieved by using rather simple structures of the actual data, but maintaining and reflecting the logic, relations and the complexity of the information in the interface that provides access to the metadata and data. The database itself is no longer the core issue in the design of the systems.

There is no doubt that the full realization of all visions following the early thoughts of the SFS would have required uptake of new technologies and would have been dependent on the general development in terms of performance and storage capacity etc at all times. The limitations of existing capacity in the IT systems should have led to tactical choices, but should never have influenced the overall objectives and strategies. An early decision to follow the principles, would have led to a stepwise development where a lot of the plurality, overlaps and stove pipes (traditional stovepipes, based on traditional statistical surveys, are organized by topics and contain all production steps from data collection to publishing, [16]), that we face today, could have been avoided. However, there is also a tendency in all attempts observed to develop systems and the organization needed to fulfil the ideas, that they were too ambitious and comprehensive. They became regarded more as revolutions than evolutions, the latter actually being envisaged from the beginning.

3 Metadata strategy and the uptake of the old visions

While the 1970s were spent on converting systems from IBM to Honeywell Bull, the 1980s were actually spent on moving back to IBM. Statistics Norway purchased an IBM mainframe

for the 1980 Census. From 1982, systems started to be moved back to the IBM environment. By the end of the 1980s, Statistics Norway had two IBM mainframes, one in Oslo and one in Kongsvinger. Systems were actually not converted back from Honeywell Bull, but they were to a large extent rewritten using the old system as the specification for the new one. However, database technology was introduced.

The main target for the developments in the 1980s was to combine data registration and data editing in one operation. Looking back, this was not a particularly visionary decision. First of all, it led to a very slow uptake of using new and more effective statistical methods in data editing. Editing while punching data is slow and very resource demanding, and you spend equal amounts of resources on every unit irrespective of the impact on the total quality. The result was basically that every survey or statistical product got their own, proprietary system for data editing. This was definitely not in line with the ideas of Nordbotten.

3.1 Out of the dark

From the mid 1990s, Statistics Norway started developments in three basic areas which are related to the old set of ideas:

- Dissemination systems and uptake of the Internet possibilities
- Establishment of master metadata systems
- Developments in data collection and reduction of the response burden

We leave aside the issue of dissemination systems here, the other initiatives were both directly related to the ideas put forth by Nordbotten. Statistics Norway had ended up being very stove piped, while the pressure on better availability to statistical information, more flexibility in combining information, and especially the pressure on collecting one piece of information once and later on reusing it for many purposes was increasing. A coherent statistical production system was needed. The foundation needed to begin this work was to establish the principles and elements of a complete metadata system.

3.2 Metadata developments

The basic idea behind the statistical file system is to store statistical micro files in such a way that they can easily be reused. Statistical micro files must be linked at unit level to the actual base register, to micro files of other statistical domains and to the same file for another reference period. A prerequisite to do this is standard definition of units and variables and an efficient metadata information system for Statistics Norway. Statistics Norway planned to develop a common documentation system including a catalog of variables in the 1960s, but the project never took off.

The documentation of the database for the 1980 population Census was based on a virtual storage access method file system, (IBM VSAM). The experience of the 1980 Census was used to develop a database system for population statistics. The new systems were closely linked to the development of the first attempt to create a common metadata system in Statistics Norway – BIMS. The idea of BIMS was to have a system that could not only store and reuse the metadata, but also actively use them in production. New micro files could be created by selecting variables from different files, and new variables could be derived.
However, this attempt was not a success, either.

Statistics Norway has in the course of time developed many different metadata systems. This led to the same information being stored several times in several places making the

availability of updated and consistent information difficult. In recent years, there has been a strong focus on the need to link existing systems and a requirement that new metadata systems should not be built in isolation. To facilitate this, Statistics Norway developed a metadata strategy, which was approved early in 2005. The strategy focuses on establishing a conceptual framework, clear roles and responsibilities, and a stepwise development involving integration and linkage of systems.

The current objective of Statistics Norway's work on metadata is to develop an integrated metadata system that will contribute to effective and coherent statistics production and dissemination, in addition to improved quality of statistics. Different metadata systems are being linked together making the metadata more easily accessible for all users. Metadata should be updated only in one place.

Current metadata systems:

- *Datadok* – file descripions
- *Vardok* – variables documentation system
- *Stabas* – standard classification database
- *Metadb* – metadata-base for event history data
- *Metadata Service library* for the master systems
- *Metadata portal* (implemented on Internet, under further development on intranet)

3.3 Improved data collection and reduction of the response burden

The other major idea of the SFS, that data collection could be regarded as a common process, establishing the micro data separate from the statistical analysis, reflecting a view on the data archive as a growing data capital of the NSI, was not followed up from the beginning. In general, the uptake of the possibilities caused by the established registers in the statistical production was rather modest. The registers served as frames for sampling, but sampling and sample administration were not particularly standardised. Most opportunities were not fully explored.

New technologies for data collection emerged with the Internet. Statistics Norway established from 2000 browser based solutions for data collection from electronic questionnaires in the data reporting from municipalities and county administrations (Kostra system). From 2004, driven by a Government directive, all questionnaires within business statistics could be filled in electronically. This caused a centralisation and standardisation of the data collection systems. Direct reporting from business internal systems was introduced. This has led to improved use of the business register as the source for sample administration, and a reduction of the number of these types of specialized systems throughout the organization. Equal improvements have been implemented with respect to the statistical CPR. The role of managing the base Statistical Population registers have been significantly improved, now playing an important role in the coordinated use and reuse of the register based information.

The pressure to reduce the response burden has led to coordinated efforts and developments across government agencies. A common portal for data reporting to the government sector was established in 2003 (Altinn). This portal has been further developed, acting as a two way channel for communication between businesses/citizens and the government sector. The improved usage has forced the public administrations not only to use a common channel, but to actually harmonise their needs, and share the information provided by the users.

For Statistics Norway this implies that more and more data will be collected and reused independently of the subsequent statistics production. Again, the visions presented by Nordbotten are about to become a reality, driven by the Internet technology, government directives and the end user needs and preferences.

4 Organisational aspects

4.1 Petter Jakob Bjerve on adapting the organization of national statistical institutes to technological and methodological development

Petter Jakob Bjerve became Director General of Statistics Norway in 1949 and took over an institute that suffered from fairly limited organizational development since it was founded in the 19th century. Bjerve, whose period as Director General lasted 31 years, worked very hard, particularly in the early part of his reign, to modernize the stove piped organization he had inherited. A joint programmatic paper by Bjerve and Nordbotten on automation of the production of statistics in 1956, [5], initiated the discussion about a more functional organization. After having led Statistics Norway during its first steps into the computer era Bjerve presented to the Nordic Chief Statisticians in 1963, [10], his views on the impact the technological development ought to have on the organizational structure and the division of labour within the national statistical institute.

Bjerve was well aware that there was a lot more to come than had been seen so far of computer development of relevance for the production of statistics. By 1963 the computers had not yet made much impact on the organization of the statistical work, although they opened up, in Bjerve's view, for progress in the production of statistics that was nothing less than revolutionary. He reviewed briefly the impact of half a century of punched card technology on the ways of working in Statistics Norway before moving on to considering the impact of foreshadowed technological progress, noting in passing that the methodological development currently lagged behind the technological progress. A guiding light in Bjerve's synoptic overview was his conviction that the future demand for statistical information would emphasize integration, richness of details and actuality. The role of general statistical publications would decline.

Bjerve ended his paper by stating seven conclusions on the impact of the technological development on the organizational structure of national statistical institutes. We render his conclusions in English in a short version below and in full in annex B. They may be read as principles on how to organize the statistical institutes in the future as a consequence of technological and methodological developments:

His introduction to the presentation of the principles indicates that he felt these changes would become challenging. The principles are as follows

1. The organizational structure and working relations of the national statistical institute must adjust to technology and methods in such a way that the needs of statistical production can be supported, at all times, in the best possible way.
2. Developments in technology will replace manual operations by computer and increasingly in such a way that the computer operations become integrated into larger, automatic processes.

3. This will lead to the release of resources in statistical and analytic units, the expansion of process oriented units, and increased challenges for the top management in communication and coordination.
4. A further implication is that the working relations within the institute will have to be increasingly formalized and precise, with work schedules prepared well in advance.
5. Statistical methodology will become increasingly refined in interaction with the technological development, and methodology specialists will need to be placed in central servicing units.
6. As more primary statistical data become continuously registered in electronic archives, the overall preparedness for accommodating demands from users is increased and the border line between the production and the analysis of statistics becomes less distinct.
7. The trends in development works towards breaking asunder the traditional subject matter oriented organization as functional units find their place. Different solutions are possible, larger statistical institutes may be more attracted than smaller ones to a more complete functional superstructure.

That the organizational reform of Statistics Norway in 1991 ran counter to these principles (at least partly) was paradoxical to say the least. In the 1991 reorganization central services were decentralised. The register divisions were included in the Department for Industry Statistics, (the statistical Business Register) and in the Department for Social Statistics (the statistical CPR). The decentralisation was a result of several factors and was viewed at the time as inevitable. The statistical departments had lost trust in the central IT services. Other comparable organizations had already decentralised, and the principle of decentralisation was riding high as organizational hype. As often is the case, the reorganization also reflected the need to change people in central positions.

The lack of continuation of the developments based upon the ideas from Nordbotten was caused by IT wandering in the wilderness of the Government Data Centre in the 70s, and thereafter the lack of leadership in modernising Statistics Norway in the 80s. The decentralisation was definitely not an action to improve common solutions and standardisation. Even though IT was coordinated by a specific committee, the decision power was placed in the respective departments. The development of stove piped solutions continued.

In 2003 IT and Data Collection was merged into one department and placed directly under the top management of Statistics Norway. However, the statistical departments still held onto their own IT divisions. From 2009 the Department for IT and Data Collection was split. Partly based on an external analysis, IT was centralised into one department together with the Division for Statistical Methods. Currently, IT and Data Collection respectively have become centralised and part of top management, along the lines of the principles outlined in 1963.

While Statistics Norway has throughout its history had highly proficient leaders whose management to a large extent was founded on a strong interest in the statistical products or in the analysis and economic research related to statistics, the strong focus of Nordbotten and Bjerve on the need for developments in the production processes has later on to some extent lost attraction as a fundamental part of leadership responsibility.

5 Some reflections on upcoming challenges

5.1 Further pressure on cooperation within public administration

The demand for digital public services has pointed out the need to describe work processes, systems and stored information for external usage. The pressure to reduce response burden has led to an increased need for collaboration and interaction. When Nordbotten first described his data box as in figure 1, he also described what the IT industry some 25 years later presented as a data warehouse. It took the IT industry another 10 years to understand that the metadata needed in such systems should comprise much more than technical descriptions. The awareness of the role of sufficient metadata when turning services digital is increasing. Still it is not evident for everyone that the demand for cooperation and coordination within the government sector depends on systems that are completely metadata driven. That means that they are completely controlled by the description of the rules and information the system handles, which in turn means that the system is controlled and altered by the user and not by the IT expert. The systems interact by sharing common metadata.

A project has been started in Norway to collect all relevant information on jobs and income from employers in one operation. An equal project has been carried out in Denmark. This is an example where the collection and primary data editing should be carried out in an operation separated from the further statistical analysis. The first reflex on how to handle this in Statistics Norway is to split the data once they are retrieved by Statistics Norway in data needed for the separate statistical domains, and carry out business as usual. However, this is probably not the last example where we will be forced to retrieve and maintain data across statistical domains. It is obvious that we should take the opportunity to keep the data integrated, and plan for unexpected future usage according to the old ideas.

One of the principles of Norwegian e-government policy is that the respondents shall never have to report the same information more than once, [11]. As a consequence a common metadata system for the government sector has been under development for several years, first to monitor collected variables, secondly to hold the metadata describing information held by public administrations. The latter initiative is still under development – the Semantic Register for Electronic Interchange (SERES), which contains metadata that describe the semantics and the information structure of data to be exchanged with and within the government sector. Metadata systems in Statistics Norway will interact with SERES.

Researchers frequently use data collections from Statistics Norway for their research. However, the process from finding out what you need, to actually getting the data, may be long and troublesome, especially for inexperienced researchers. Statistics Norway has therefore (with support from the Research Council of Norway) developed a website to make information about this process more easily available. Among other things, this page provides the users with documentation of several data collections. Each data collection has a general description e.g. of data quality, and it also contains a list of relevant variables, including variable documentation from Vardok. A new system is being scoped, hopefully with even more automatic solutions, [17].

As a consequence of the above mentioned issues, Statistics Norway should ensure that all data are stored according to master specifications and that the data stored are completely controlled and made available according to the defined rules and definitions. In other terms, we need to realize the SFS as of today.

5.2 Privacy issues

Discussions on privacy issues started already in the 60s, with the report from the Ruggles committee published in 1965, [13], see also [14]. In 1967, Nordbotten included privacy mechanisms in the SFS represented with the idea on security sets, [9]. The Norwegian Act relating to official statistics and Statistics Norway (Act No. 54 of June 16 1989) has given Statistics Norway the right to collect all kinds of micro data and under specific conditions to archive them for future use. However, there is an increasing pressure on the need to remove identification elements from the archived data. The new legislation on health registers clearly state that identification elements shall be removed from the data sets, in such a way that only internal identifiers within the specific registers are allowed.

Before the introduction of the person identification number and the legal unit identification number, Statistics Norway had to operate with internal identification numbers. For the purpose of the reuse of already collected information across all domains in the statistical office, the public unique identifiers are not needed. It is only when statistical data need to be matched with other external data that the need for public or official identifiers occurs. The reason for storing data with public identifiers is related to the need for contacting respondents during data editing etc. If the pressure on ensuring privacy increases, one solution could be that Statistics Norway returns to keeping data with stable and unique internal identifiers, establishing specific procedures when the need to link to external (public) identifiers occur. In such a case, another reason for separating data collection from the statistical analysis emerges.

5.3 The emerging semantic web

While working on this document, a workshop on how statistical institutes can contribute to and benefit from the emerging semantic web took place. Both in the US and in the UK work is underway in order to be able to describe, retrieve and link available data by their content, an example being www.data.gov. The technology is already in place, and the technological developments take place under the umbrella named Linked Data or Linked Open Data (http://en.wikipedia.org/wiki/Linked_Data).

Statistical institutes have a specific knowledge of classifying information by their content. This knowledge should be exploited in the ongoing work in this area. On the other hand, there is a need for a common, available and also agreed high level model of statistical information. In fact, it more or less exists through the ideas spread from Nordbotten, and the further work undertaken by Bo Sundgren. However, in reality,, we do not have a master model to point at. Furthermore, it is a fact that every time new technologies like these occur, people tend to start all over again, believing that old know-how is not applicable for new inventions. One model that is used for modeling statistics as linked data is called the Statistical Core Vocabulary. To some extent the model outlined by Nordbotten in the 1960s is more comprehensive than the Statistical Core Vocabulary model.

The work on linking data today reflects exactly the same visions that were described by Nordbotten. Data should be collected and kept as a capital, to be prepared for further utilization that could not be envisaged at the time of their capture.

6 Conclusions

The need to standardise work processes and systems has become evident to all statistical institutes. A seminar entitled *The Modernisation of Statistics Production* in Stockholm in 2009 pointed out the needs for new solutions which correspond remarkably well to the visionary ideas presented 50 years ago [18], see also [19].

The cost of developing systems with overlapping functionality can no longer be ignored by the statistical institutes. Each statistical institute is also obliged to produce statistics based upon international demands and standards, facing huge costs in adjusting to these demands. The possibilities for increased international collaboration in order to standardise and reuse systems across institutes have been intensified. Because of this, high level models of the statistical production models have been elaborated. The Generic Statistical Business Process Model (GSBPM) describes the way statistics are produced. A high level model describing statistical information is equally needed (a Generic Statistical Information Model – GSIM). Such high level models describe the common foundation of the statistical industry, and serve as the basis for better cooperation in development and use of best practise methods and systems, at the end as a basis for enhanced interoperability when needed. The first model in this area was drafted by Svein Nordbotten.

Nordbotten stated that providing statistics should be regarded along the lines of common production theory. Following those lines, today we work with the aim of making statistics production an industry across nations, jointly stocktaking the potential of future technological developments.

References

[1] Nordbotten, S. (1960). *Elektronmaskinene og statistikkens utforming i årene framover*, (Computers and the future form of statistics). De Nordiske Statistikermøter i Helsingfors 1960, Helsinki 1961, pp.135-141. Available for free downloading from www.nordbotten.com.

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

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

[4] Luther Georg. Statistikens historia i Finland till 1970, Helsingfors 1993

[5] Bjerve Petter Jakob og Nordbotten Svein. Automasjon i statistikkproduksjonen (Automation of the production of statistics), Statistiske Meldinger nr 6, 1956. Reissued with an additional section on the developments 1956-1969 as Artikler no. 28, 1969, Statistics Norway.

[6a] Aurbakken Erik. Den elektriske metoden kommer til Norge. (The electric method arrive to Norway). Oslo 1998

[6b] Aurbakken Erik . Fra hullkort til PC – Glimt fra databehandlingens historie i SSB 1959 – 1990, (From punched cards to PC - the IT history of Statistics Norway 1959 – 1990), Oslo, 1999.

[7] Bjerkholt Olav and Gåsemyr Svein. Svein Nordbotten at eighty – Computers in the service of statistics. Draft not published. Oslo 2010.

[8] Nordbotten, S. (1967): *Om arkivstatistiske systemer*. Annex 4, in: Statistical Reports of the Nordic Countries, Vol. 14. Copenhagen 1967.

[9] 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.

[10] Bjerve Petter Jakob. Kva konsekvensar vil utviklinga i teknikk og metodar ha for organisasjons- og arbeidsformene i det statistiske sentralbyrå? (The consequences of technological and methodological developments for the organization of the national statistical institution), Proceedings from the Nordic Chief Statisticians' Meeting, Reykjavik 1963, Annex 14.

[11] Regjeringen.no (Government web site): Reuse and coordination of information

[12] 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.

[13] Ruggles, r., et al. J. (1965): *Report of the Committee on Preservation and Use of Economic Data*. Available online at: http://www.archive.org/details/ReportOfTheCommitteeOnThePreservationAndUseOfEconomicData1965

[14] Miller, A. (1967): The National Data Center. The Atlantic. Reproduced at http://blog.modernmechanix.com/2008/03/31/the-national-data-center-and-personal-privacy/

[15] UNECE (2007): *Register-based statistics in the Nordic Countries*. Review of best practices with focus on population and social statistics. United Nations Economic Commission for Europe. Also available online at http://www.unece.org/stats/publications/Register_based_statistics_in_Nordic_countries.pdf

[16] Sundgren, B. Statistical file systems and archive statistics. Nordisk statistikermøte København 2010

[17] Hægeland, T et al. Infrastructure in social science - Access to register data for research purposes. Report for The Research Council of Norway - by an inquiry team. Oslo 2003

[18] Seminar Stockholm 2009, The Modernisation of Statistics Production, (MSP). http://www.scb.se

[19] Gløersen, R and Sæbø, H. V. Standardisation for improvements in Statistics Norway, MSP seminar Stockholm 2009, http://www.scb.se

Annex A. UNECE meetings on use of computers in statistics and on utilizing base registers and others administrative records as a source for official statistics

The first UNECE ad hoc meeting/seminar on use of Data Processing Electronic Machines in statistics was arranged in 1957. DG of Statistics Norway, Petter Jakob Bjerve was elected as chairman. Morris Hansen leader of department of methods, Bureau of Census, US, was the main actor in the meeting or more correctly the seminar. He presented experiences with using computers for processes in production of statistics. A Norwegian paper on the planning for providing a computer was presented, a) collect technical data about existing machines, b) specify problems and get experience in programming, c) training staff – a new way of thinking. It was decided to organize a permanent Working Group on use of computers in statistics.

The first WG meeting was arranged in 1961. DG of Statistics Sweden Ingvar Ohlson was elected chairman. Svein Nordbotten was the Norwegian member of the WG. The WG decided to concentrate on two topics, **i)** automatic editing and **ii)** statistical file system, (i.e. the archive statistical system). The second and later WG were chaired by Knut Medin, Statistics Sweden. Svein Nordbotten was consultant for the second WG meeting and responsible for the preparation of documents for the meeting and for the report on the results of the discussions. The report is published as a UN Handbook: Automatic Editing of Individual Statistical Observations, 1963.

Svein Nordbotten was consultant for the third WG meeting. Also in this case his report of the meeting was published as a UN Handbook: Automatic Files in Statistical Systems, 1967.

The UNECE is still organizing Work Shops in the field of editing (statistical data editing and imputation) and statistical file system (use of administrative records as sources for official statistics).

In 2007 UNECE published a report: Register-based statistics in the Nordic Countries. Review of best practices with focus on population and social statistics.

Annex B

Translation of the concluding part of Petter Jakob Bjerve's presentation to the Nordic Chief Statisticians in 1963, see [10]

1. The organizational structure and working relations of the national statistical institute must adjust to technology and methods in such a way that the needs of statistical production can be supported, at all times, in the best possible way.

2. Developments in technology will make it possible to eliminate some working processes in the production of statistics, such as punching, typographical setting and proof-reading, but the most important consequence will be that production steps that currently are manual in time can be taken over by computer operations and more and more become integrated into larger, automatic processes.

3. In organizational terms this will lead to the release of resources in statistical and analytic units, and, furthermore, that process oriented units will be expanded and restructured, and that top management must be strengthened by, for example, staff units for administration, planning and control, because the challenges associated with communication and coordination will move upwards in the organizational pyramid.

4. The consequences of this development for the working relations within the institute will be that internal communication must, to a greater extent than earlier, be precise and as a rule in writing, that all working routines must be explicit and carefully specified, and that the work must follow plans prepared well in advance.

5. The development of statistical methods, which partly is influenced by and also in turn may have an impact on the technology, will have similar consequences for the organization and the working relations. In particular, methodological developments will require relatively more specialists for some work operations that – at least in small national statistical institutes – will need to be placed in central servicing units for the specialist capacity to be fully utilized.

6. The technical and methodological development will after a while enable the statistical institute to provide at short notice and to an increasing degree statistics for specific analytical purposes. The necessary requirement for such a development, as indeed the demands from users of statistics strongly encourage, is that more and more primary statistical data are continuously registered in electronic archives, so that the overall preparedness for accommodating demands is increased. At the same time, the border line between the production and the analysis of statistics will become less and less distinct, because analysis will be closely integrated with the production of statistics in a more or less automatic process. This may necessitate for some national statistical institutes to put relatively more emphasis on analysis than earlier.

7. It may seem as if the trends in development as set out above are bound to break asunder the strongly subject matter oriented organizational structure currently characterizing most national statistical institutes. The most radical alternative is a strongly function-oriented organization with e.g. collection, processing, and perhaps analysis as well, as the main units. However, if the centralization of work operations really went so far that it seemed rational to

replace the current subject matter oriented organizational superstructure with a functional organization, it could only be in a statistical institute so large that the functional entities would be subdivided by subject matter units.

Annex C. Milestones in the development of an archive statistical system in Statistics Norway

The list covers use and developments of punch card machines, electronic computers, base registers, metadata systems and organization of a National Statistical Institutes

1893: Anders N. Kiær, Director General, Statistics Norway, met Hollerith, during an ISI meeting

1894: Statistics Norway rented a Hollerith punch card system to be used for an income survey

1900: Hollerith punch card machines (several keypunches, a sorter and a tabulator) were procured for the 1900 Population Census and used with grate success to process census data

1910: Improved versions of the 1900 machines were used for the 1910 Census

1910: From 1910 and into the 1920s automation of production of statistics stagnated

1920s: Statistics Norway cooperated with the Norwegian engineer Fredric Rosing Bull and introduced his Norwegian produced punch card machines, with Powers Tabulating machines, which were used for processing foreign trade statistics up to the beginning of the 1950s

1930s: A new break through for the punch card method in the Norwegian production of statistics came when Statistics Norway acquired the equipment necessary to realise Kiærs plans for expanding/developing foreign trade statistics. The greatest improvement was the acquisition of a punch card tabulator.

1946: Statistics Norway obtains responsibility for the Central Office for the municipal population registration

1947: A central unit is established in Statistics Norway, responsible for operating punch card machines

1950: The IBM 101 Electronic Statistical Machine was procured for the 1950 Census

1953: Svein Nordbotten was sent on a study tour to USA – the program included study of electronic computers

1953: Business Census used as the initial micro file for the statistical Business Register

1956: Petter Jakob Bjerve and Svein Nordbotten published a paper on automation in the production of statistics

1957: UNECE arranged the first meeting on the use of computers in production of official statistics

1958: The first computer obtained in Statistics Norway – Deuce from English Electric with punch card as data carrier, 402 word (32-bit) memory and a magnetic drum with storage capacity of 8192 words

1960: Svein Nordbotten presented his ideas for the archive statistical system

1961: IBM 1401 procured for the 1960 Census – 2 (from 1962 another 2) magnetic tape drives, 4 Kbyte core memory

1963: Petter Jakob Bjerve presented 7 principles to organize a National Statistical Institute. The principles were his conclusions of discussion about functional organization that started with paper [5] in 1956

1964: Central Population Register, (CPR) - the initial micro file was based on i) 1960 Census micro file, ii) reconciliation with the municipal population register and iii) updating of the 1960 initial micro file for the period 1 November 1960 – 1 January 1964

1964: A modern statistical Business Register was established – historical data registered from 1959

1965: A department for statistical production was organized - IT, base registers, data collection and statistical methods

1966: IBM 360/40 was rented, 64 Kbyte RAM, 4 tape drives and 2 disk drives with 29 Mbyte storage

1967: A new system for documentation of statistical micro files was in operation – this documentation was based on Svein Nordbottens box, (unit, variable and time) + a fourth dimension the *version.* The fourth version is related to the sequential chain of processes in statistical production

1967: A model group for analysis of the unit of person was established – later a group for Socio-economic Research Group was established – the research was based on reuse of statistical micro files and development of new methods to utilize statistical micro files

1968: An administrative Central Employer Register operated by Social Security Norway was in operation

1970: The 1970 Census - name, address and PIN for all residents of CPR were pre-printed on Census forms sent to the households

1970: An administrative Central Register for VAT units operated by Tax Norway was established

1971: Svein Nordbotten left Statistics Norway. He was appointed as Professor in informational science at the University of Bergen from January 1972

1973: The mainframe computing was outsourced to the Government Data Centre (Statens datasentral) running Honeywell/Bull – all existing IBM software had to be converted

1975: The final and complete CPR micro file per 1 November 1970 was created - based on the 1970 Census and reconciliation with the municipal population register

1975: The 1970 census was the initial file for the start of registration when a person started and fulfilled an educational program. Educational attainment of a person is based on these two data sources

1976: The Business Register was established on a Bull IDS 1 database platform, being the first database application at Statistics Norway using commercial database management systems (network database principles)

1978 An administrative central register of employee jobs was established

1980: An IBM 4341 was obtained for the 1980 Census

1980: The use of Honeywell Bull mainframe was reduced, but continued to the beginning of the 1990s.

1982: Systems started to be moved back to the IBM environment

1984: The Central Population Register was established on a Bull IDS 2 database platform (network database management system)

1984: Statistics Norway was responsible for the central office of the administrative base register GAB, (Ground property, Address and Building). The Map Agency became responsible for the central office in 1987. A new and advanced register was in operation by the Map Agency in 2003 – the Cadastre

1995: The administrative Legal Unit Register (LUR) was in operation in 1995. The LUR is an integration of 3 administrative registers, Employer, VAT, limited Companies and the statistical Business Register

1998: In house development of a metadata system for documentation of file descriptions

1999: Statistics Norway last mainframe computer was turned off before the year 2000. Statistical production continued on a Unix/Windows platform

2000: In house development of a metadata system for documentation of definitions for variables

2001: Housing Census 2001 – creation of the initial micro file of the unit of Dwelling. Dwelling is the forth unit of the Cadastre

2002: In house development of a metadata system for documentation of classifications in cooperation with Statistics Denmark started, operational in 2004.

2005: Metadata Strategy 2005- approved for Statistics Norway

2005: In house development of a metadata services library started, completed in 2008

2005: In house development of a metadata portal started, available on the Internet in 2006

2007: The FOSS project started – a parallel to the Swedish LOTTA

2008: Business process model for Statistics Norway approved

2009: Portfolio management of development projects

2009: IT architecture/model for Statistics Norway approved