June 2012

Before the Q2001 and after the Q2012[1]

by

Svein Nordbotten, Professor Emeritus

University of Bergen, Norway

Contents

1. Introduction 1

2. Before the Quality Conference in Stockholm 2

3. After the Quality Conference in Athens.....5

4. References 13

1. Introduction

This paper reflects the content of an invited keynote speech to the European Conference on Quality in Official Statistics in Athens May 28-June 1, 2012 referred to as Q2012. This was the most recent in a series of quality conferences organized by EUROSTAT and National Statistical Institutes (**NSI**s) of which the first was held in Stockholm in year 2001. The conferences have discussed a wide set of topics related to quality in official statistics, and the author's intention was twofold: to recall some personal memories of quality promoting activities carried out by the official statistical community before the intensive treatment of quality aspects in the mentioned quality conferences, and to present some visions about what may happened in the years to come.

[1] The title refers to the conferences on Quality in Official Statistics in Stockholm year 2001 and in Athens year 2012.

The quality concept as used by the official statisticians is composed by a number of dimensions and comprises the effects of most activities in the statistical system and its interactions. Figure 1 gives examples of dimensions frequently referred to.

- coverage errors
- sampling errors
- observation errors
- processing errors
- production cost
- relevance
- reliability
- timeliness
- accessibility
- response burden

Figure 1: Examples of quality dimensions if official statistics

Producers of statistics are focused on quality concept dimensions anchored in international frameworks and standards while users of statistics have concepts derived from their individual needs.

We do not know much about which concepts are used by whom and how the different actors are ranking the dimensions. A single, quality indicator representing all dimensions would be desirable, but seems, however, meaningless. When referring to statistical quality, we usually refer to a quality dimension determined by the context.

2. Before the Quality Conference in Stockholm

During the 20th century, the quality of official statistics was improved by introducing new methodology and technology in the production process. Figure 2 presents a list of the more important areas of improvement. Attention was first paid to the data collection. Since most of the data used was collected for administrative purposes, statisticians were not always satisfied with the data *relevance*. Complete surveys for statistical purposes only were, however, expensive. New, statistical collection methods based on representative samples were slowly adopted [Kiær 1897].

- **Collection methods incl. sample surveys**
- **Screening of collected data (editing, imputation)**
- **Registers and identification**
- **Re-organisation NSI organizations**
- **Integrating and storing statistical data**
- **Estimation and calibration**
- **Extending user services**
- **Standardization and cooperation**

Figure 2: NSI activities to promote the quality of official statistics

After World War I, the scientific development of sampling theory led to the impressive methodological development of sample survey methods within the US Bureau of the Census.

The introduction of punched card equipment at the end of the 19th century was an important factor for improvement of the *timeliness*. With manual equipment, the processing of American population censuses could not anymore be completed within the decade before the next census [Truesdell 1965]. The punched card equipment was in general use in many **NSI**s up to the middle of the 20th century. This technology was also used to control the consistency of codes used.

However, the revolution in **NSI**s and for quality of official statistics was initiated by the introduction of electronic processing equipment in 1952 in the US Census Bureau rapidly followed by other **NSI**s.

Data *editing*, i. e. control and correction of collected data have always been a major quality improving processes in the production of statistics. From the mid-1950s the effects of statistical editing has been extensively studied [de Wall 2011 a. o., Nordbotten 1955]. The programmable computers opened up great possibilities for automatic editing, and the ECE Conference of European Statisticians initiated

preparation of a handbook containing theoretical considerations as well as the state of the art in the early 1960s [Nordbotten 1963]. More than 35% of the total computer processing time was at that time used for controlling and correcting collected data.

The art of automatic editing made substantial progress in the last decades of the previous century. In USA, Canada and several European countries, large editing systems were developed and extensively used. Many of these were based on the famous paper by Fellegi and Holt in which they demonstrated how to locate a minimum set of attributes needed to be changed in a record is rejected by an edit. This principle preserved as much as possible of the collected values at the same time as each record was made acceptable [Fellegi and Holt 1976].

Early in the 1990s computers had become important tools for simulating the human brain by means of artificial neural networks. Since control and correction of data record were based on intellectual experience, application of editing of statistical data by means of artificial neural networks was proposed and several types of approaches were investigated in the EU project EUREDIT 2000-2004 [Nordbotten 1995, Charlton a. o. 2001].

The ECE/UN Working Group for Electronic Data Processing was substituted with several specialized groups. From the late 1980s, the Work Session on Editing has met regularly and made substantial contributions for improving the quality of official statistics [UNECE 2006].

The capacity and speed made it possible to build and maintain large registers effectively. The Nordic countries soon got central population register systems in which each inhabitant had his/her unique and permanent identification number. By offering service from the population registers to public authorities and major private organizations as banks, insurance companies, etc., significant public resources were saved. As a by-product, administrative data obtained by the **NSI**s were uniquely identified and could be integrated with data from other sources and data collected for statistical purposes forming a basis for more complete statistical pictures of the societies [Nordbotten 1966]. Already in 1981, Denmark Statistics as the first **NSI** produced their population census mainly on data collected from

administrative sources using the personal identification numbers as integration keys [Thygesen 2010].

The Conference of European Statisticians asked already in the 1960s for a second handbook from the Working Group of EDP dealing with the possibilities for identification, integration, organization of collected register-based data in automatic file collection for effective re-use [Nordbotten 1967]. Future use and potentials were discussed in a number of papers [Nordbotten 2010a, Nordbotten 2010b].

3. After the Quality Conference in Athens

From the first conference on quality in official statistics held in Stockholm in 2001 to the present conference, they have been a major meeting place and opportunity for exchanging experience and ideas about quality improvements in official statistics. It can therefore be of some interest to speculate about what will be central quality topics in the future.

Today, the situation for many NSIs can be summarized by two negative trends and two positive trends. Most NSIs are fighting against budgetary problems at the same time as the public are complaining about the response burden in connection with information requested from public authorities. The positive trends in today picture are a growing demand for official statistics from old and new users, and a continuing technical development.

My vision for the future is that the **NSI**s will focus their attention and development work on the following areas:

- Metadata systems
- Identification and integration
- Increased use of administrative data
- Use of electronic footprints (EOD)

- Conversion of non-numeric data
- Storage and re-use of microdata
- Online instant statistical service
- Global standardization and cooperation

The purpose of a *metadata system* associated with a collection of statistical files is to inform about the content and properties of the data the system to help the users to navigate, select and give processing instructions. In principle, the metadata system should represent the relations among the components of the statistical system as indicated in Figure 3. In an automatic system, it is essential that any set of new data collected or generated data and any processing of data are automatically recorded in the metadata system and accessible to the users.

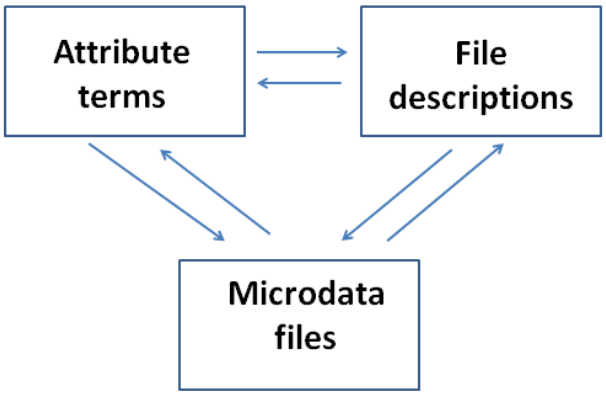

Figure 3: Relations among the components of a statistical file system

During the last 40 years since Bo Sundgren introduced the metadata concept in statistical systems, progress in development and implementation of metadata systems for official statistics has been considerable. However, there are still several steps to be taken before comprehensive metadata systems are satisfactory included. In particular, the fusion of the metadata system with the statistical data system must be improved. No data sets, new or generated, should be made available before its content has been included in the metadata system. As response to a search by a user, the metadata system should give a complete description of

the data requested with links to their locations. Ideally, the interface between the metadata system and the microdata system should be seamless.

If there are no identification references to the real life objects to which the collected data refer, the dataset is cannot be used as a brick in building a comprehensive statistical picture of the real system to be reflected. The situation was dramatically changed when unique and permanent object identifiers were introduced and used by different data collectors. Now data for the same real objects collected by different organizations and at different times could be *integrated* if the links to other objects were also observed and recorded. This does not mean that NSIs should store microdata with public identifiers. On the contrary, the public identifiers should be converted to internal identifiers used for stored microdata to reduce the risk for misuse of the data.

There are 2 different methods for integrating sets of identified records. Integration by *matching* can be carried out if the records of 2 sets of data refer to the same type of objects, for example personal identification numbers.

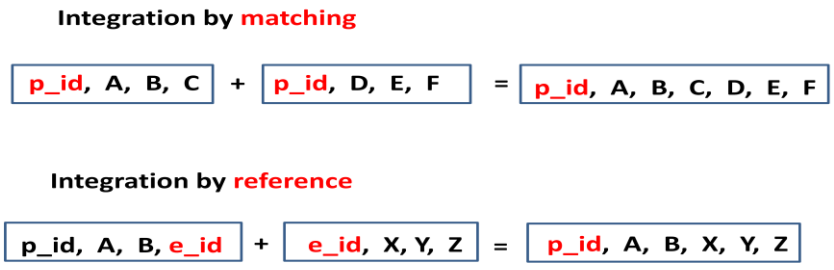

Figure 4: Two approaches to integrating records from different files.

In some cases, an attribute in a record set of persons is referring to an identifier of another object type, for example an enterprise identification number. These two sets of data can be integrated by *reference* so that the records for the persons are extended with attributes of enterprises to enterprises to which they are related, for example by employment.

In countries with operating central identification systems or registers for different real objects, the task will be to extend the set of national identification registers to new types of objects by arguing with the savings in operation and maintenance of

one single set of registers. Other countries do not have this lucky situation, and have to develop other methods for integrating data sets from different collectors.

The by-product for the **NSI**s from national identification systems is the pre-identified data available from administrative organizations. Since the cost of using administrative data are much lower than collecting the data for statistical use only, obviously the road to wider statistical service is through extended use of administrative data.

In recent years, the electronic revolution has made progress in a number of fields, some of which are referred to in Figure 5 [Nordbotten 2011].

1. Electronic recording (Credit cards, passports, tracking, sencors)
2. Radio frequency identification (RFID)
3. Mobile phone networks (GSM)
4. Global positioning system (GPS)
5. Communication satellites (COMSTAT)

*Figure 5: Technologies of interest for **NSI**s*

Wireless electronic recording is exploding. Some of these data can be of interests for **NSI**s. As an example, consider customers to a food supermarket equipped with commodities coded with small RFID chips, or alternatively with 'old-fashion' bar codes (Figure 6). When a customer passes the cashier, the RFIDs are automatically read, or in case of bar codes scanned manually by the cashier. The commodities are paid by credit cards which are read by a reader. The details of the customer's visit to the store are electronically recorded and can be sent immediately to the respective **NSI**.

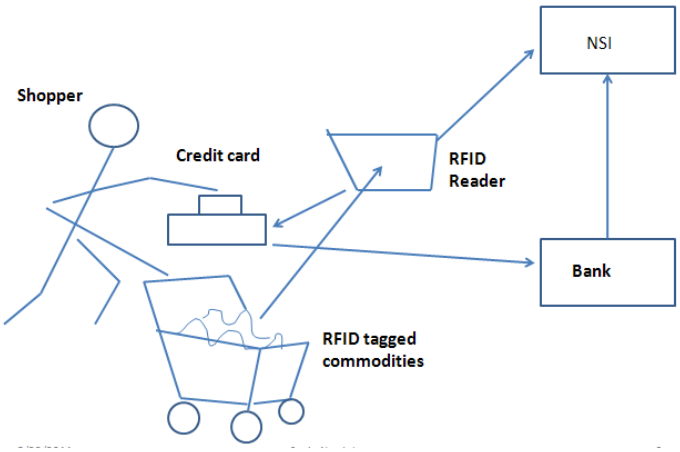

Figure 6: Shopping

Figure 7 indicates the type of data which can be transferred to the **NSI**. If the **NSI** has access to the relation between credit cards and owners, the **NSI** can by integration create a file of records as indicated to the right in the figure, By proper organization and setup, the time between recording and availability of the integrated files in **NSI** can be almost instantaneous. Official statistics indicators for retail trade broken down on commodity groups could in principle be available on a daily basis if wanted.

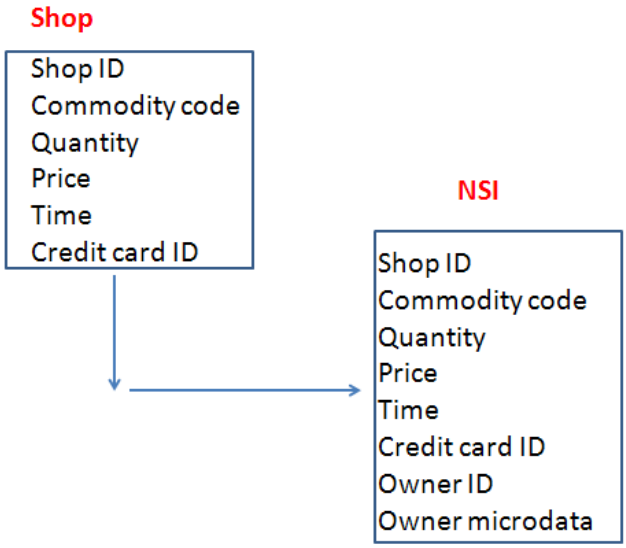

Figure 7: Records from the shopping

Another example can be selected from the transport/communication sector. Toll roads are becoming a usual way of collecting payment for the availability of roads.

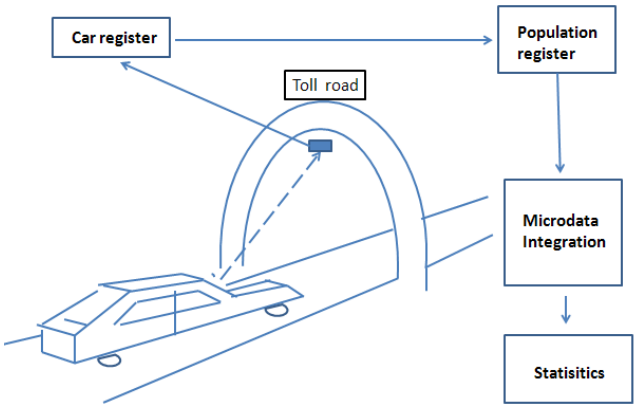

Figure 8: Car passes on toll roads

Automatic toll gates recording cars passing by means of RFID are becoming usual (Figure 8). The basic recordings are sent to the center responsible for the respective toll roads which build the records as exhibited in Figure 9 builds on the car register.

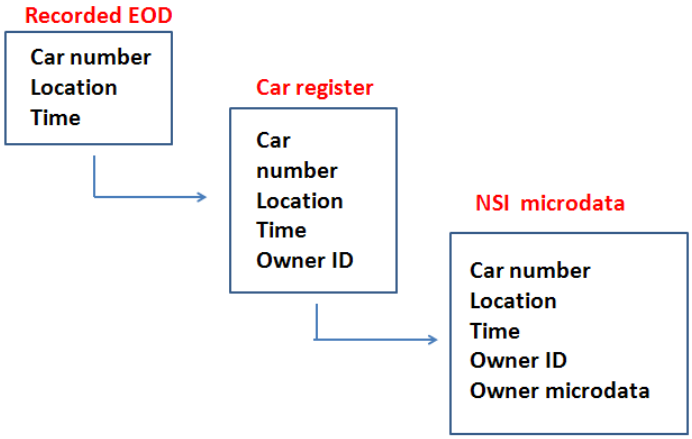

Figure 9: Car traffic characterized by owners' attributes

If these records are passed on to the **NSI**, the statistical organization can integrate the received records with owners' microdata, and produce hour-to-hour traffic statistics as well as traffic statistics based on owners' socio-economic properties.

There are a number of methodological challenges implied. Because of the huge number of events recorded in supermarkets and by toll road readers, a selection of records must probably be done. How many records and how to select will be new tasks for the sampling methodologists.

Another potential source of inexpensive data which can be very useful for official statistics is *textual data* on internet expressing opinions of different aspects. Software extracting words and expressions has for many years been used for tagging and retrieving documents, and can also be used in counting the frequency of opinions and attitudes in populations. Some estimates, which today are based on data from interview surveys, may be substituted by faster, possibly as accurate and less expensive data collection by scanning public blogs and other media on internet.

Relevance and *accessibility* are necessary conditions for successful dissemination of statistical information. Extending and improving user services for users with special needs will to a large extent require user access to and integration of stored microdata files.

Previously, official statistics were prepared for the governing authorities and the processing was planned well ahead of the production. Re-use of the microdata was expensive and only exceptionally done.

In the 21st century, official statistics are considered a public service, and the aim is to serve as many needs as possible. By organizing the microdata files in a statistical file system including an integrated metadata system and utilizing available technology, an online and instant statistical service could be offered without getting into conflict with confidentiality requirements.

Some **NSI**s have already been offering on-site and/or remote microdata file access to authorized researchers for some time [UNECE 2007]. Users, who do not qualify for research association, e. g. business analysts, individual researchers and students, can also have important and legitimate needs for data and statistical services.

A key to a general *online* and *instant* statistical service is a well organized microdata file storage system which can be accessible by a repertoire of pre-approved analytical programs operated interactively by the remote users. The user could by means of the approved software remotely access and process files with microdata without being able to 'see' the individual records. By using analytical programs, confidential problems are reduced, but each approved program must still have imbedded confidentiality control of all results for protection of confidentiality.

The use of mobile devices such as smart-phones and tablets has become part of daily life. In addition to serve their original purpose, smart-phones are used for emailing, gaming, surfing and connecting to social media, frequently by use of small software programs referred to as 'apps' which facilitate connection to different types of data sources. This technique can be used to bring official statistics out to the 'man in the street'. So far, only a few apps have been developed for easy access to statistical information. By either themselves or by establishing suitable interfaces for external developers, **NSI**s can by introducing apps for different statistics, make official statistics information available for the users independent of the time or the users' locations are as long as they bring their mobile phones. Statistics can be consulted when at work or reading a newspaper, in a debate or in discussions at the breakfast table.

In Figure 10, the UN Fundamental Principle on Confidentiality is displayed.

"Individual data collected by statistical agencies for statistical compilation, whether or not they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes ".

Figure 10: UN Fundamental Principle on Confidentiality

The fundamental principle is supplemented by national legislations to protect confidentiality and privacy. The ideas suggested above to further develop quality of official statistics may not meet the requirements of the existing legislation, and legislation may either have to be adjusted or the ideas rejected.

In my opinion, the value of collected organized microdata is increasing as knowledge bases for understanding and running complex, modern societies, and cannot be wasted. A legal basis for safe conservation of all microdata will be required to solve the conflict between the wish to protect privacy and the request for more efficient societies. Microdata collected by public resources .and considered of future statistical and research value should be conserved in one single organization with the responsibility for keeping the data "*strictly confidential and used exclusively for statistical purposes*". With the public trust enjoyed by the **NSI** in most countries, the **NSI**s should be charged with this responsibility.

Finally, in a world growing more and more global, international statistics will also grow in importance. It will require continuous international concept standardization, cooperation and exchange of methodological and technical experience among statistical agencies.

4. References

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.

de Wall, T., Pannekoek, J. and Scholtus, S. (2011); *Handbook of Statistical Editing and Imputation*. John Wiley & Sons. N.J.

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

Kiær, A.N.(1897): *The representative method of statistical surveys*. Reprinted by Statistics Norway. Samfunnsøkonomiske Studier nr. 27. Oslo 1976.

Nordbotten, S. (1955): ***Measuring the Error of Editing Questionnaires in a Census.***, American Statistical Association Journal. Volume 55. pp. 364-369.

Nordbotten, S. (1963): ***Automatic Editing of Individual Statistical Observations***, Statistical Standards and Studies, No. 3. United Nations. United Nations. New York and Geneva 1967

Nordbotten, S. (1966): ***A Statistical File system.*** Translation from Norwegian of an article in Statistisk Tidskrift in 1960.

Nordbotten, S. (1967): ***Automatic Files in Statistical Systems.*** Statistical Standards and Studies. Handbook No. 9. United Nations. New York and Geneva 1967.

Nordbotten, S. (2010a):***The Use of Administrative Data in Official Statistics – Past, Present, and Future – With Special Reference to the Nordic Countries***, Official Statistics – Methodology and Applications in Honour of Daniel Thorburn, pp. 205-225. Available at Officialstatistics.wordpress.com. .

Nordbotten, S. (2010b): ***The statistical archive system 1960-2010: A summary***. Nordisk Statistikermøde i København 11.-14. august 2010. København

Nordbotten, S. (2011): ***Use of Electronically Observed Data in Official Statistics***. Presentation at the 58th World Statistics Congress of the International Statistical Institute, Special Topic Session 25: Analyzing Internet Traffic Flows and digital footprints for statistical purposes, Dublin August 2011.

Thygesen, L. (2010): *The Importance of the Archive Statistical Idea for the Development of Social Statistics and Population Censuses in Denmark*. Nordisk Statistikermøde i København 11.-14. august 2010. København.

Truesdell, L. E. (1965): *The Development of Punch Card Tabulation in the Bureau of the Census 1890-1940*. Bureau of the Census. US Department of Commerce. Washington D.C.

UNECE (2006): *Statistical Data Editing*, Vol.1-3. United Nations. New York and Geneva 2006.

UNECE (2006): *Managing Statistical Confidentiality & Microdata Access. Principles and Guidelines for Good Practice*. United Nations. New York and Geneva 2007.

.