**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Topic (ii): Editing administrative data and combined sources

ROLE OF EDITING AND IMPUTATION IN INTEGRATION OF SOURCES FOR STRUCTURAL BUSINESS STATISTICS

Invited Paper

Prepared by Svein Gåsemyr, Svein Nordbotten, and Morten Q. Andersen (Statistics Norway)

Abstract

      This paper focuses on use of administrative data in the statistical value chain processes of business surveys. Most processes of a traditional business survey can be improved by use of administrative data as auxiliary or supplementary sources. More and more administrative data is available for business statistics. The statistics presented is annual structural statistics of manufacturing industry. The methods to utilize linked files of business surveys and administrative sources need to be improved. In general, the efficiency of editing is increasing with the information available for each unit. By linking two or more files from different sources, the number of variables in each resulting integrated records increase, and so does the potential for effective control and correction. The strategy of Statistics Norway is to develop standardized modules for statistical processes. The modules for editing and imputation within this system will have an important function in integration of sources and in improving the quality of statistics.

I.   INTRODUCTION

1.   Through nearly half a century, the *Nordic* countries have been working on an infrastructure for electronic reporting from enterprises and households to government agencies for administrative purposes. Common government policy in these countries is also to promote reuse of existing administrative data. The result is an increasing number of administrative data systems available as sources for official statistics. The *Nordic* approach of register-based statistics is presented in a report published by **UNECE**, 2007, [1].

2.   Register-based statistics are so far most extensively prepared in the field of social statistics. For some years we have, however, also seen a fast increase in use of administrative data for business statistics. Annual enterprise accounts and annual general trading statements are now for example available in an electronic database for nearly all Norwegian enterprises of the private sector. Administrative data for enterprise and establishment units are also used both as a source for register based economic statistics and as supplementary or auxiliary sources for business surveys. Administrative data is defined as data collected by government agencies for administrative purposes.

3.   This paper focuses on use of administrative data in the statistical value chain processes of business surveys. Most processes of a traditional business survey can be improved by use of administrative data as auxiliary or supplementary sources.

4.   The paper presents the interaction between:
- the statistical *Business Register*
- the database to coordinate the sample design for business surveys
- the micro file of the statistical survey
- the database of enterprise accounts and general trading statements, and
- the menu of data editing, imputation and estimation

5.   The paper concentrates on use of administrative data in annual structural statistics of manufacture industry.

II.   DATA SOURCES FOR STRUCTURAL BUSINESS STATISTICS OF THE MANUFACTURING INDUSTRY

A.   Statistical units and variables registered in administrative base registers and data systems

6.   The statistical *Business Register* (**BR**) is more or less an integrated part of the administrative *Legal Unit Register*, (**LUR**). *Statistics Norway* has succeeded to register the statistical unit of establishment in **LUR**, and is responsible for profiling an enterprise in two or more establishments when suitable [2]. The profiling of an enterprise is common for **BR** and **LUR**. The statistical establishment unit is also registered in the *Social Security* data system of employee jobs. The unit of employee job is linked to the statistical unit of establishment. This is an important advantage for register-based employment and production statistics. As there is no need for the statistical unit of establishment in administrative procedures, *Statistics Norway* has to allocate resources for controlling the reports from complex enterprises on employee jobs. The main statistical classifications such as **NACE** (industry), **ISCO COM** (occupation) are used in administrative data systems.

Administrative Legal Unit Register
7.   Registers that are integrated to **LUR**:
Administrative register of employers
Administrative register of limited companies
Administrative register of foundations
Administrative register of value add tax units
Statistical business register
Administrative tax-register of companies and self-employed

Because of the strong attachment of **LUR** to several administrative business registers and the statistical BR, the coverage of different populations should be of very good quality.

Statistical Business Register
8.   The statistical *Business Register* (**BR**) is the sampling frame for business surveys and is also an important tool in using administrative sources for business statistics. All administrative micro files of enterprise and establishment units are linked to the **BR** by the coordination function of **LUR**. Several data sources are used by the **BR** to determine if an establishment is active or not at a given month or date.

Database to coordinate the sampling of business surveys
9.   A common database to coordinate sampling for business surveys have been in operation for some years. The data collection strategy of *Statistics Norway* aims to reduce the response burden for enterprises. Measures to achieve this are:

- Increased use of administrative sources in business statistics
- Small enterprises are to be excluded from statistical business surveys
- Rotation of sample to distribute the response burden to many medium size enterprises
- Promote direct electronic reporting from enterprise data systems

- Use of electronic tracks of third part as sources for business statistics

B.   The annual structural business survey of manufacturing industry

10.   The composition and development of the Norwegian enterprise and establishment populations is shown in *Table 1*. A complex enterprise is profiled in two or more establishments in LUR.

Table 1. Number of enterprises and establishments, 2001 and   2005

	Number of Enterprises		Number of establishments		Percentage of production value	
Year	*2001*	*2005*	*2001*	*2005*	*2001*	*2005*
Complex enterprises	1 056	1 086	1905	1 768	55.5	59.4
Single establishment Enterprises	20 052	19 416	20 052	19 416	44.5	40.6
Total	21 108	20 492	21 957	21 184	100.0	100.0

Census and sampling design

11.   All complex enterprises participate in the survey. Small complex enterprises respond to a simplified questionnaire, 285 establishments in 2001 and 319 in 2005. Other establishments of complex enterprises and large single establishment enterprises are surveyed by a census. In 2001 there were no uses of sampling. In 2005 medium-size establishments are surveyed by sampling. Small single enterprises are excluded from the survey. The excluding threshold is 10 employed persons. During the 4 years, 2001 – 2005, the sampling rate for medium size enterprises has been reduced. The reduction is based on a controlled process to be informed on how the reduced sample affects the quality of the statistics. *Table 2* presents enterprises that participate in the survey, census and sampling and units where administrative data is the only sources.

Table 2. Establishments by sources, the survey, administrative data and combination of sources

	Number of establishments (Business register)		Employment in 1000 (Register based job file)		Production in per cent. Diff. sources	
Year	*2001*	*2005*	*2001*	*2005*	*2001*	*2005*
A. Census survey	4 691	2 096	240.9	178.7	90.9	80.7
B. In sample and selected	0	1 251	0.0	21.0	0.0	7.5
C. In sample and not selected	0	1 092	0.0	18.0	0.0	4.2
D. Small complex enterprises using simple form	285	319	1.2	1.9	0.3	0.7
E. Small establishments excluded	16 981	16 426	45.7	34.7	8.8	6.9
Total	21 957	21 184	287.8	254.3	100.0	100.0

Variables and sources of the statistical micro file

12.   The statistical micro file for structural business statistics includes all units. The main data source is the general trading statement reported for the unit of enterprise to the *Tax Agency*, (**TA**). A supplementary statistical survey collects information for the unit of establishment for group **A**, **B** and **D** in Table 2. The reporting is done by mail or electronic medium.

13.   Variables on the unit of establishment: compensation of employee, gross value production, cost of goods and services consumed, value added, and value added at factor prices.

14.   Variables on the unit of enterprise: operating income, operating cost, operating profit/loss, of the year.

15.   The number of establishments is based on the statistical *Business Register*. Other variables that are based on the *Business Register* are industry and institutional sector. Statistical surveys are used to update the administrative **LUR** and the **BR**. Employment, measured by employed persons and hours worked, is collected from the register-based job file.

16.   The editing of the statistical survey is done manually with the following prioritizing: i) large single enterprises and large complex enterprises, ii) new enterprises and enterprises with no reported activities for the reference year iii) other complex enterprises iv) other single enterprises.

17.   Information about accounts variables and turnover for single enterprises that not are selected for the sample survey or excluded are based on: electronic reported trading statements, register of company accounts and the database of **VAT** units.

18.   Automatic imputation is applied for variables that not are specified in the administrative sources. Simple models of rate estimates are used.

Updating on industry code for small single establishment enterprises

19.   One problem related to the fact that small single establishment enterprises do not participate in statistical surveys is that there is no source for updating of **NACE** code. Information of the quality for variables such as address and **NACE** code for small single establishment enterprises are needed.

20.   The *Tax Agency*, (**TA**), is now planning to use the **NACE** code of the **LUR** for controlling enterprise accounts and general trading statements. The **TA** has to collect information to ensure that the **NACE** code of **LUR** (and **BR**) is correct. We hope this practice would result in a source for annual updating of the NACE code of **LUR** (and **BR**) for enterprises that do not report to statistical surveys.

C.   Database for all available sources of the units of enterprise and establishment

21.   The plan is to develop the existing database of enterprise accounts and trading statements to include all available data for the establishment and enterprise units. For both unit types, the data have to be organized as a longitudinal database. The variables of a unit can then be followed through time and the time series can be compared at unit level through time and across available sources such as **BR**, records of business surveys, enterprise accounts, **VAT** data (turnover), and register-based job files (employed person, hours worked, compensation of employee). The starting point for the edit procedure will be a comparison of the data collected in the survey with all available administrative sources to check if the compared time series of several variables seem to be reasonable.

Enterprise accounts and general trading statements

22.   Enterprise accounts refer to the unit of enterprise. For a single establishment enterprise the enterprise accounts refer also to the establishment. For complex enterprises it is of importance for business statistics that the sum of a variable for all establishments within the complex enterprise adds up to the figure for the enterprise unit. In principle, the opposite is not possible, i.e. it is impossible to distribute figures from an enterprise accounts to its profiled establishments. In practice, however, an estimated distribution has to be done for some complex enterprises.

Job files

23.   Job files have an important functioning in integrating sources at unit level. The job unit identification is the combination of the **PIN** of the employed person, the **BIN** of the work place and the period in which the job is active. The variables of a job file are therefore related to 3 different units: person (age, sex and educational attainment), establishment (**NACE** code, size group) and the unit of job (occupation, hours worked, compensation of employee). All available variables that are related to an employed person can be aggregated to the level of establishment, (structure of staff for age, sex and education).

D.   Combined use of business surveys, base registers and other administrative data

24.   Examples of variables not registered in the database of administrative data, but needed in business statistics, are input and output variables of the production. These variables are calculated for establishments by imputation.

25.   Unit and item non-response are also imputed at unit level, and so are units not selected for the survey sample and variables not registered in administrative sources. An important property of a statistical micro file completed by imputation is that it can be linked to other sources at unit level for the total population.

III.   METHODS FOR INTEGRATION AND EDITING OF DATA SOURCES

26.   Use of a statistical standard across data sources is the traditional tool for developing integrated statistics. In the *Nordic* system, linkage of records from different sources at unit level is the approach applied. Integrated micro records from such linkages represent a far more reliable and flexible approach to preparing consistent and integrated statistics. The use of base registers and the common official **ID** numbers for the administrative and statistical units provide the efficient infrastructure permitting this approach.

Reliable record linkage of enterprises and establishments

27.   Even if there is an efficient infrastructure for record linking some errors arise. An example: an employee job is registered as active in the *Social Security*, (**SS**), data system but without compensation of employee in the *Tax Agency* (**TA**), data system. Usually the reason for this inconsistency is use of different **ID** number for the work place in reporting to the **SS** and **TA**. A rather complex procedure to correct for this type of inconsistency has been developed. [3].

Editing of a data from a specific source

28.   Usually, the owner of an administrative source executes some editing followed by some additional controls when the data file is received by *Statistics Norway*. Methods for editing of a single administrative source are about the same as of statistical surveys. Usually administrative data files are very large and the editing has to be based on automatic editing.

29.   Traditionally the editing of structural business surveys was done manually. Recently, selective editing strategies have been used according to which the units are divided into categories by their importance for selected population estimates, and editing efforts and resources have been allocated to make these estimates as accurate as possible.

Editing of linkage of data sources

30.   When two sources are linked at unit level always some cases of inconsistency arise. The Norwegian experiences in editing and imputation of base register micro files and other administrative data are presented in [4].

31.   In general, the efficiency of editing is increasing with the information available for each unit. By linking two or more files from different sources, the number of variables in each resulting integrated records increase, and so does the potential for effective control and correction. The human editors will get a wider insight into the individual enterprise and establishment, and a better possibility to do their jobs

efficiently. By the same token, the increased number of variables in the integrated record will significantly enlarge the number of possible edits and imputation functions.

32.   To utilize these new possibilities to perform a more effective editing, the human editors must have extensive knowledge and long experience. One problem is, however, that the number of available specialists with such background is limited. The linkage of data from multiple data sources has reinforced the justification for developing computer-based methods and support for the editing.

33.   When working with linked data, a frequently appearing situation is that some variable values are missing for a part of the population. The cause may be non-response, that one of the data sources is a sample survey, or that some observation have been lost in one or another process.

34.   Twenty years ago, it was strongly believed that by interviewing the experts it should be possible to construct computerized experts. Ten ears ago, the hope was that by recording and analyzing the decisions made by experts as a response to different tasks it should be possible to simulate the expert. So far most editing methods are created as logical rules and functions.

35.   Most edits used at present are control of linear relations and ratios, and most automatic corrections use linear imputation functions frequently based on regressions, and hot deck imputations. With the number of background variables in integrated records from linked data sets, we anticipate that non-linear imputation in some cases will provide higher imputation quality than those presently used. Some of our ideas with respect to imputation are presented in a separate paper [5].

IV.   STANDARDIZED MODULES FOR STATISTICAL PROCESSES – INTEGRATED SYSTEM OF EDITING AND ESTIMATION

36.   The *Integrated System of Editing and Estimation* (**ISEE**) is a project aiming to develop standardized modules for statistical processes. The project was initially started for developing a system for 20 new price statistics. Instead of developing 20 tailor-made systems for editing and estimation, it was decided to develop one common editing system for price statistics.

37.   The standard system for price statistics contains 2 interactive applications, **DYNAREV** for editing and **PRICE** for estimation. The statistician responsible for processing can switch between the 2 applications and instantaneously inspect the effect of any changed values on the price estimates.

38.   The project became a success and **ISEE** is now being developed for production of other statistics. In addition to **PRICE**, another application, **STRUCTURE**, is designed to estimate population totals, averages, etc. All activities within ISEE are logged.

39.   The objectives of **ISEE** are to:

- Improve the quality of statistics
- Develop efficient IT solutions and processes
- Increase flexibility so that the need for contribution from experts on IT, methods and editing experts for a given statistics can be reduced.
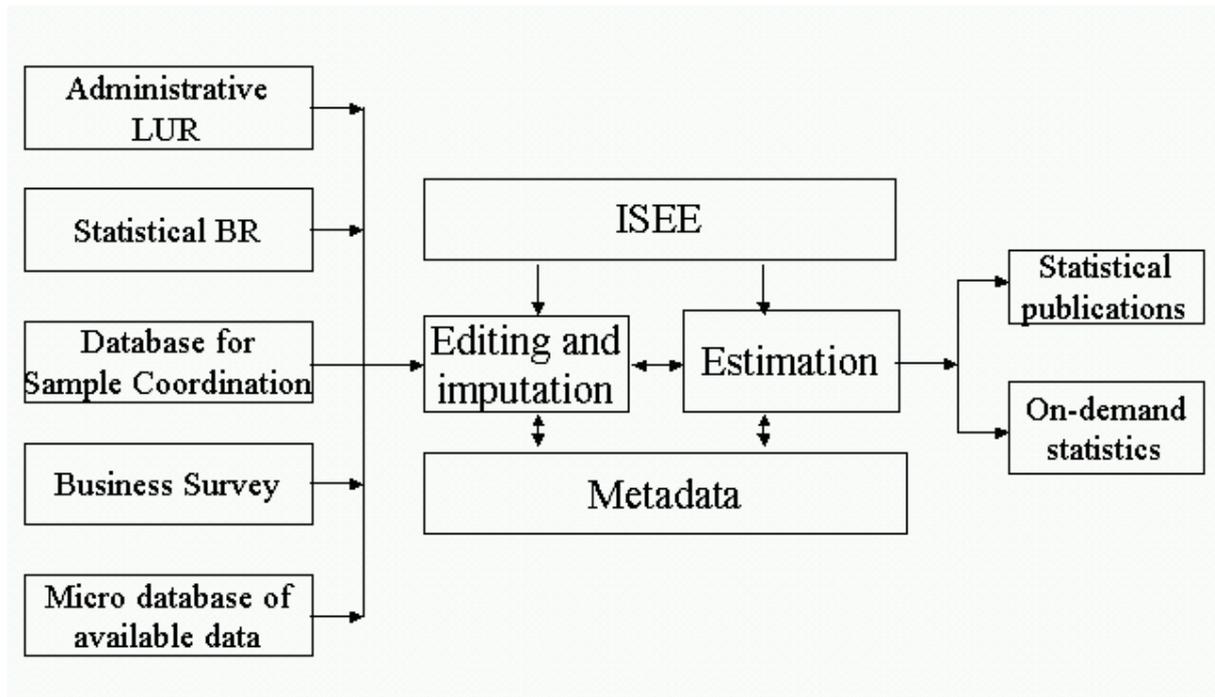
Figure 1. Data sources to be integrated for business statistics by ISEE

V. THE NEED TO MEASURE THE QUALITY OF MICRO FILES OF COMBINED SOURCES

A. Need for quality information

40. The integrated records from combining statistical and administrative data will in general contain a large number of variables, and be the basis for an even larger number of potential statistical products some responding to anticipated request, but most will be ad hoc products from future user requests.

41. Both the users and the producer of the statistics will want information about the quality of the different statistical products. The concept of statistical quality has in recent years been widely discussed [6]. It is generally agreed that it comprises a number of dimensions such as relevance, accuracy, timeliness, availability, etc. In the present context, accuracy is the focal point, but also other quality dimensions can be very important for producing statistical products from combined data sources.

42. The users wish quality information to decide if the statistics serve their needs while the statistical producer wants quality information to determine if more or less resources in the future should be allocated to the production processes of these statistics, or if there are other possibilities for improving the qualities of the products.

43. Producing statistics from integrated records from combined data sources differs from the ordinary production of statistics in several respects. The quality of statistics from a single data source may be adequately indicated with one quality indicator for a main population parameter estimate. Because the statistics of a set of integrated records are, however, based on 2 or more different sources that can vary with respect to quality, the quality of a single aggregate may not be representative for the products that can be derived from a set of integrated records. The combination process, the linking of records from different sources, can also be an error source affecting the quality of the resulting integrated record.

44. A more basic quality measuring approach is required. We advocate the use of quality metrics that reflect the set of integrated record quality, and that can be used for deriving the quality indicators for population estimates when needed.

**B.	Process errors in integrated records**

45.	Statistical quality, more specifically accuracy, is determined by the errors made in processes as data collection and processing. Several types of errors can be made in preparing integrated records of which we shall concentrate on the following:

- Collection errors due to incomplete or erroneous registers,
- Linking errors occur when identifiers in one or more record sets to be matched are erroneous or incomplete,
- Observation errors can be made because of inadequate questionnaires or interviews, object repulsion,
- Detection errors are made when during editing not all errors are detected, or correct values are erroneously considered incorrect,
- Imputation errors are introduced when assigning a new value to a record.

46.	Working with administrative data sources in *Norway*, where common identifier systems are used by the different administrative authorities from which *Statistics Norway* can get administrative data. This situation provides the most powerful tool for detecting coverage errors when comparing data from 2 or more sources. Using data from a different source can also be an important way to prevent observation errors during the collection process.

47.	Linking records from 2 or more sources is the core of administrative register based production of statistics. Its success depends completely on the correctness and continuous updating of the identifier system applied. *Table 3* illustrates several types of matching results can appear:

Table 3: Units identified in 2 registers

Units in register 1: Units in register 2:	Not existing	Incorrectly existing	Existing with incorrect ID	Existing with correct ID	Sum
Not existing in register	00	01	02	03	0.
Incorrectly existing	10	11	12	13	1.
Existing with incorrect ID	20	21	22	23	2.
Existing correct	30	31	32	33	3.
Sum	.0	.1	.2	.3	..

48.	Each unit can be classified and counted in one of 4 categories for its position in either register, i.e. in one of 16 categories. Category 00 shows the number of units neither included in set 1 nor 2, and these errors may be considered the most difficult to identify and correct. Category 11 contains all units incorrectly included in both registers and can be difficult to detect. The number of units correctly identified in both register, category 33, is hopefully in majority. Units in the remaining categories are all units identified erroneously.

49.	Different strategies are needed for attacking the different categories of errors, and probably we shall never be able to make the registers perfect. The table can, however, serve as a framework for evaluating the quality with respect to errors in unit registration and linkage.

50.	Observation errors have been studied by questionnaire designers and survey experts for many years. They are also a major aim of editing processes. Integrated records from 2 or more data sources represent a corresponding greater challenge to the editing. Fortunately, the more information we have about a unit, the more power full we can make the edits. In designing the edits of integrated record sets, one important aspect is to evaluate the relative reliability of the data sources that have been integrated.

51.   The editing process frequently introduces errors that are important to evaluate to avoid to extensive editing. The effect of the editing process can be summarized as *Table 4*:

Table 4: Edited records by result and by raw record status

Edited records: Raw records:	Records rejected	Records accepted	Raw records
Records with errors	00	01	0.
Records without errors	10	11	1.
Edited records	.0	.0	..

Here we meet again the requirement about information about the correct records that is unavailable, but can, as we shall discuss below, be estimated. The table summarizes the result of an editing process by assigning each edited record to one of 4 categories, Category 00 represents the results of successful edit controls, while categories 10 and 01 represent the errors made by the control process. Category 01 corresponds to the Type 1 errors, i.e. accepting as correct a n erroneous record, and category 10 corresponds to Type 2 errors, i.e. rejecting a correct record. The table is a reasonable frame for evaluating the errors introduced by the editing process.

52.   A similar table can be used to denote the errors introduced in the imputation process for correcting errors detected by the editing, the records in categories 00 and 10 in table 2. As for editing, imputation can be made much better for records integrated from several data sources than if the imputation process is carried out for each data set separately because of the much larger set of variables available.

C.   Statistical quality indicators

53.   In the previous section we outlined possible errors that can be made during collecting and processing of integrated records, and stressed that measures of quality are needed for the micro data to guide both users and producer in their decisions. The users are interested in quality measures for the processed set of records, and in particular measures that can easily be used for creating measures for the special estimates they want. The producer on the other hand, is interested in measures reflecting the quality of the individual processes in order to consider changes in processes.

54.   All error and accuracy measurements have in common that some correct value set exists. The problem of accuracy occurs because the costs of accessing the correct values are prohibiting. In practical work, the only possibility is to estimate the deviation between the processed data values and the correct values of the unit variables from a small evaluation sample of units. This sample must be processed as effectively as possible through the sequence of processes expecting approximately correct unit values. In parallel the raw sample unit values must also be processed according to regular procedures. The 2 sets of values for the sample units can then be compared and statistical accuracy measures computed.

55.   These value sets make it possible to compute estimated of numbers as those in *Tables 3* and *4* which will give useful error frequency information for the producer when compared with resources used on different processes.

56.   We recommend computing mean square errors for the individual unit values and organize the computations so that measures for subgroups can easily be extracted. To provide the producer with further information about the processes, we need to compute estimates of the mean square errors of selected variables for each process. For the users, estimates of mean square errors of the variables for all processing of the individual records are required to make them ready for estimating the wanted population parameters. The mean square error estimates can easily be used, or supplemented with new computations from the sample value sets, for computing corresponding quality measures for the wanted statistics.

VI.   WORK TO BE DONE

57.   Declared aims for *Statistics Norway* strategy are to make the statistical production processes more effective and to improve the service to the users [7,8]. Two important aspects are continuous work on better utilization of available administrative data sources, and preparing useful quality indicators associated with the statistical products.

58.   In the context of use of administrative data sources in statistical production, we want to emphasize the need for:

1. Continuing work to improve and promote the system of unit identifications.
2. Adjusting the standard processing system **ISEE** to the extended requirements to and possibilities for editing and imputation capabilities of integrated records from combined administrative data.
3. Testing and evaluating the system in controlled experiments with empirical data from several administrative and statistical business data sources.
4. Developing a system for designing and implementing suitable evaluation samples, procedures to be followed, saving results from process evaluations, resources needed and their practical organization.
5. Deciding how to use the process evaluation data for future process designs and for statistical product quality declarations.

References

1. UNECE, (2007): *Register-based statistics in the Nordic countries.* Geneva.
2. Gåsemyr, S. Børke S. and Andersen M. (2007): *A strategy to increase the use of administrative data within an integrated system of business statistics.* Register seminar. Helsinki.
3 Gåsemyr, S. (2005): *Editing and imputation for the creation of a linked file from base registers and other administrative data.* WS on statistical data editing. Ottava.
4. Børke, S. and S. Gåsemyr, S. (2006): *Editing and imputation of linked micro files to be used in statistics, research projects and administrative procedures.* WS on statistical data editing. Bonn.
5. Nordbotten, S. and Zhang, L-C. (2008): *Mass prediction and imputation in ISEE.* WS on statistical data editing. Vienna.
6. Eurostat (2000): *Assessment of Quality in Statistics.* Luxembourg.
7. Statistics Norway, (2007): *Strategy for data collection.* Oslo
8. Statistics Norway, (2007): *IT strategy 2000.* Oslo