UNITED NATIONS



Economic and Social Council

Distr. GENERAL

ECE/CES/GE.42/2007/SP/1 11 April 2007

ENGLISH ONLY

ECONOMIC COMMISSION FOR EUROPE

STATISTICAL COMMISSION

CONFERENCE OF EUROPEAN STATISTICIANS

Meeting of Experts on Business Registers

Tenth meeting Geneva, 18-19 June 2007 Item 6 of the provisional agenda

QUALITY IMPROVEMENTS IN BUSINESS REGISTERS AND IMPLICATIONS OF REVISIONS OF NACE (NOMENCLATURE GÉNÉRALE DES ACTIVITÉS ÉCONOMIQUES DANS LES COMMUNAUTÉS EUROPÉENNES) AND INTERNATIONAL STANDARD INDUSTRIAL CLASSIFICATION (ISIC)

A SYSTEM OF BUSINESS REGISTER'S QUALITY INDICATORS AS HANDY TOOL FOR USERS

Submitted by Italy

The meeting is organised jointly with the Commission of the European Communities (Eurostat) and the Organisation for Economic Co-operation and Development (OECD)

I. INTRODUCTION

1. The demand for statistical information is continuously on the increase. Political decision-makers, media, customers, require evermore detailed and timely information. Sometimes, though, the requests for statistics are rather contrasting, such as when having to produce both national economic statistics that are coherent at European and international level, and that satisfy, at the same time, each territorial community. While the first statistics must adhere to the specific standards – definitions, classifications, estimation methods, etc – that international organizations (UN, Eurostat, OECD, etc.) have identified, the latter type requires definite criteria for analysing the specific territorial analysis context.

- 2. Similarly, the offer of statistical information has also increased. Though often supported by public or association organisations, new private producers have engaged in the statistical data market, also partly because of the lacks in the official statistics. Nonetheless, the higher quantity of data available does not always make it easier to take decisions. On the contrary, they can even become source of confusion when contradictory messages are sent.
- 3. Users of statistical information mainly need to access easily objective information. While objectivity is not always an abstract concept, it is closely linked with both the users' needs (that is, to acquire information in relation to their own subjective needs) and with the purposes of who produces the information. Moreover, objectivity is also closely linked with transparency: an information is correct when it makes clear the objectives, concepts, definitions, methods for data gathering and treatment, result aggregation criteria. It is also correct when it provides an indication on the quality of the aggregate diffused. Transparency is what allows a user to easily comprehend the data, among those available, which can be used for his own needs.
- 4. Thus, if we take as product the Business Register (BR), its quality is closely related to how it is used and to whether it satisfies the users' needs. However, numerous users resort to the BR, and each of them have their own need. The BR's reference universe and updating period will be different if used for the Short Term Surveys rather than for the Structural Business Survey. If the Value Added is estimated based on the BR's reference universe, clearly the quality (e.g. activity code and size) of large units will be fundamental given the weight they have in the estimation of such variable. On the other hand, if the indicators of the Business Demography take the BR as reference, the quality of the smaller units will be very important given their higher involvement in demographic events.
- 5. Thus, if the BR is a complex product because of its numerous and different users, the criterion for evaluating its quality will also be more complex and, even more, the synthesis for disseminating such evaluation to the users.
- 6. To measure the quality of a BR is a difficult task. Consequently, it is necessary to use the statistical surveys that take the BR as reference basis, to establish the over-coverage indicators (but not under-coverage indicators) of the register or accuracy indicators for some of its variables. This is not enough, though, and can even be misleading: Italian enterprises are indeed known for not indicating any modification or, even more often, errors in their economic activity code usually pre-printed in the questionnaires.
- 7. In order to have an overall overview of the quality of a BR, it is necessary to develop a system of indicators that analyse and measure the different aspects of quality, identified by some criteria. For each criterion, the quality indicators are constructed, using pieces of information drawn from each phase of the process or from certain parts of the register (e.g. subsets of units, variables, etc.) and evaluating their trend overtime.

II. THE QUALITY COMPONENTS IN RELATION TO THE BUSINESS REGISTER

- 8. Eurostat has identified seven quality components to evaluate a statistical product. The quality components are: relevancy, accuracy, completeness, coherence or consistency, comparability, timeliness and punctuality, clarity and accessibility.
- A Business Register (BR) must be pertinent to the needs of the users. In other words, it must contain relevant units and variables for constructing survey samples and/or lists for implementing these. It must be accurate in such a way as to reflect correctly the reality. A BR can be considered complete when it includes all units of the target population and all the variables requested. Consistency includes both internal coherence and coherence with other registers. While internal coherence involves a consistent treatment of the register data, consistency with other registers is obtained using and archiving reference data. For the BR to be comparable, it is necessary to evaluate two aspects: space and time. In relation to space, since its comparability is ensured at a European level through a regulation, this component could be measured evaluating its level of adherence to such regulation. Comparability over time means to be able to compare both the units' data and the aggregates' data at different temporal periods. As regards timeliness and punctuality, the timeliness with which the BR is updated in order to reflect the events that occur in reality can be an important quality criterion even though it could be in conflict with the coherence and comparability criteria. However, the problem of updating the variables could be solved by keeping two versions of some variables, one version with the last variables updated and one with frozen values at a certain date (1 year). Access to the data could involve the possibility for internal users to obtain single data by directly connecting to the database, or for external users to obtain synthesis tables. Generally, the facility with which aggregated data can be accessed must be considered as an important quality component. Another aspect of accessibility is the easiness with which the BR information can be interpreted. Quality measurements, thus, involve the availability of documents necessary to correctly understand the information.

III. THE SYSTEM OF QUALITY INDICATORS FOR THE BUSINESS REGISTER

10. It is necessary to understand the way in which a product has been obtained in order to measure its quality. Istat updates the BR every year through a process that, by conceptually and physically integrating several administrative sources, develops appropriate methodologies (often probabilistic) to estimate the variables of the register units and uses, for larger units, information drawn from statistical surveys and from the profiling procedures of the more complex units. The data from the administrative archives are essential to update the archive of small-medium enterprises (over 4 millions), for which it is impossible to obtain direct statistical information since Istat is unable to sustain such costs. Thanks to the administrative sources that regularly provide low-cost information, the statistical burden on enterprises is rather limited. Table 1 shows the administrative sources that feed the BR and their frequency.

Table 1
Business Register's administrative input sources

Sources	Timing	N° of records
Tax Register	Annual – September	10,7 mil
Ch. of Commerce	Annual – September	7,4 mil
Social Security register	Annual – September	1,8 mil
Yellow pages	Annual – July	2,3 mil

- 11. The quality of the register can be ensured only when verifying the quality of the register's input (sources of origin) and the quality of the processes used for treating and integrating such inputs. A system of quality indicators must take into account three aspects:
 - (a) Quality of the INPUT;
 - (b) Quality of the process (matching, merging, editing, updating);
 - (c) Quality of the OUTPUT.

A. Quality of the Input

- 12. The quality of the administrative source is already a very useful indicator of the BR's quality. Nevertheless, it is impossible to control ex-ante the quality of each administrative archive from the BR's point of view. Rather, the quality of the source can be evaluated only ex-post by means of suitable analyses to identify any error in the supply of data and to adopt adequate corrections through integration processes.
- 13. Thus, simple indicators of the quality of the source are:
 - (a) Time lag: difference between the date on which the data are supplied and their reference year;
 - (b) Indicators of completeness of the variables: for each variable indicated (company name, address, etc.), it is possible to calculate the ratio: $1-N_{missing}/N_{total}$.

Most of the problems that arise when using administrative sources regard changes adopted in the source itself that are not known: changes in the classification criteria, in the registration and cancellation rules, in the administrative control processes used. Substantially, "big" changes that occur during a certain period (year) are considered as not very plausible. Therefore, the main objective is to verify the "stability" of the sources and avoid that merely administrative changes produce "non-real" structural changes in the BR. Simple indicators are based on the comparison of the values provided in two different years:

(c) Weight % of the variations (per character): once a synthetic indicator (median) of the number of variations has been selected, further analytical verifications are carried out when the weight of the variations in a year exceeds the average level.

The fact of comparing the different supplies of data from one same source is also fundamental for analysing the completeness of the enterprises' creation and cessation dates. Such analysis is

essential to avoid problems of under-coverage or over-coverage of the BR. A complex indicator is used to evaluate the informative loss or gain and identify the data concerning the enterprises' cessation dates:

(d) The gain of information is obtained by comparing the t+1/t yearly supplies and the loss of information by comparing the t+2/t yearly supplies.

Table 2 provides, as an example, an estimation of this indicator in relation to a lag in the registration of the cessation dates with the Chambers of Commerce. As regards the cessations, clearly, if we choose a shorter time for the acquisition of administrative data (the tth supply=reference time), an important share of information can be missed. At the same time, it seems inefficient to wait for a longer period (t+2) in the hope of some improvements in the registration of the dates.

<u>Table 2</u>
<u>Lag in the registration of the Cessation dates of businesses with the Chamber of Commerce</u>

Cessation date acquired by the Chamber of Commerce	2002' supply for 2001 BR realisation	2003' supply for 2002 BR realisation	2004' supply for 2003 BR realisation	2005' supply for 2004 BR realisation	Gain respect to the supply's year	Lose respect to the theoretical optimum
year 2000 year 2001	245,138 241,383	285,119				18.1
year 2002 year 2003 year 2004	123,493	283,957 92,974	289,418 271,974 94,541	275,505 266,305	129.9 192.5 181.7	1.9 1.3
year 2005				57,110		

B. Quality of the Process

- 14. A complex system like the one producing a Business Register requires the process be maintained under control through the use of some quality indicators calculated ad hoc for that phase of the process.
- 15. The Italian BR production process presents, as an example, some critical points. Especially those determined by the complex system of logical and physical integration of records of input more than 20 millions -, the procedures for estimating units' characters, the editing and imputation plan. The process can be logically divided into three macro-phases. Each of them can be described with associated indicators controlling the quality.

Macro-phase 1: The integration of records of input from administrative sources

16. The purpose of the first macro-phase of the BR production process if the integration of administrative archives and the creation of clusters referring to the same entity (the enterprise). On the inside, two different sub phases can be distinguished and then described by quality indicators.

<u>First sub-phase</u>: Link intra-archive – Inside each input source, records that pertain to the same legal entity are integrated (where the legal entity is represented by the Fiscal Code). Synthetic quality indicators are:

- (A) Weight % of fiscal codes duplicates (by source) /total number of supplied records, temporal consistency. A decrease of this indicator over time indicates an increase of quality;
- (B) Number of new records (with respect to the previous t-1 supply) by source in the year t. It indicates a coverage measure with regards to unit creations. In particular, by comparing the weight % of new records using time series, a graph is useful to compare trends at source level and among different sources.

Second sub-phase: Link inter-archives – Integration of records coming from different sources and related to the same unit in order to build-up a cluster of record for the same enterprise. The administrative benchmark that is the base used to integrate all the other sources if the Fiscal source. This phase if very peculiar having as aim the determination of the set of administrative information available for each legal unit. Mistakes that could be done in this phase – missing or wrong links – can affect considerably results coming from the following step. Quality indicators are:

- (C) Number of clusters of records in year (t) with the presence of the Fiscal register (reference for the BR year t) (1)
- (D) Number of clusters of records in year (t) without records from Fiscal register (2)
- (E) Number of clusters of records in year (t-1) without records from Fiscal register (3)
- (F) Under-coverage indicator: $[(3) \cap (1)]/(1)$

Macro-phase 2: The estimation of characters

- 17. The aim of the second macro-phase of the BR production process consists in the imputation of main attributes to each unit and in the identification of active units in year t. Each attribute's estimation procedure can be evaluated using ad hoc quality indicators. These indicators make use of outputs produced step-by-step by the implementation of the procedure. In particular the setting up of the frame of active units is of high priority in the process: other characters are checked only for active units that determine the reference universe for sampling and for the economic structure.
- 18. The developed methodology for the determination of the state of activity to each enterprise of the BR is structured into two parts: a) a probabilistic approach where estimates of the state of activity are based on a logic model where explanatory variables are represented by some administrative variables (signals of activity) b) then auxiliary information coming from partial administrative and statistical sources are used to cover sub sets of units and deterministic rules are applied to adjust the probabilistic estimates. Compliance rates can be calculated with respect to the main sources that hold information strictly related to the active/not active status of unit.

(G) Percentage of concordance/discordance rates calculated between the BR state by source indicating active/not active units (SME surveys, structural changes database), active units (Foreign trades survey) and not-active units (bankruptcy database).

Macro-phase 3: Editing and imputation procedures

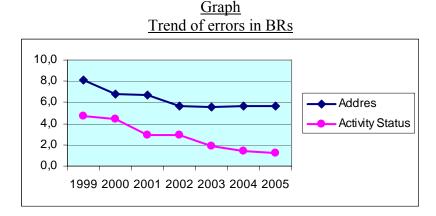
- 19. The third macro-phase of the BR production process relates to the editing and imputation process and aims to obtain the final identification of the universe of active units for the reference period year t. In order to measure the quality of the adopted edits and procedures quality indicators can be calculated. The check plan has been developed as a group of projects working as separate modules to be executed in order; they can be also changed in number and composition. Each project is characterized by a set of common rules having a similar structure and affecting data in the same direction. The main projects are:
 - (a) Cleaning rules determining the exclusion of some units from further checks;
 - (b) Deterministic if/then clauses that if verified determine the automatic change of some characters:
 - (c) Errors assessment and errors rules that if verified determine the warning for a follow-up.
- 20. Some remarks need to be done. Since the number of edits could be very high, deterministic rules focus on peculiar subset of information and on specific characters (for example some that combine Nace code and size). Editing process allows to produce warnings of supposed errors. Even in that case the need to contain the number of warnings that is possible to verify, reduces the number of edits that can be checked. Trained staff concentrates controls only on meaningful larger units instead of less economic significant ones. The calculation of quality indicators based on the number of warnings or errors edited and imputed by each project is misleading. Appropriate indicators can be built looking at trends over time that measure the increase/reduction of units involved by each type of edit. A synthesis can be done as follows:
 - (H) Variation (%) between t and t+1 of units involved by type of error;
 - (I) Variation (%) between t and t+1 of units automatically changed by deterministic rules;
 - (J) Variation (%) between t and t+1 of units having a warning and manually verified.

C. Quality of the output

- 21. There are several approaches to measure the BR quality at the time it is available for users. In the Italian case some typical instruments, as the quality satisfaction survey or any other ad hoc surveys aimed to measure quality, are not (yet) used. A different low cost method is to integrate some indicators built up upon:
 - (a) Results coming from surveys based on the BR frame;
 - (b) Results from external sources:
 - (c) Information available inside the BR.

1. <u>Indicators based on data coming from statistical surveys</u>

- 22. One of the main tasks of a BR is to represent the mailing list and the sampling universe for statistical surveys. Those surveys are both a fundamental source for the BR updating and their results allow to verify the BR quality. In particular data coming from surveys allows to comply with the accuracy component where it is assumed that responses be true. The basic hypothesis is that the survey results are unbiased and that respondents describe correctly reality. As these assumptions are not proved accuracy never will be perfectly measured. A way to take under control the quality of a survey is by paying attention at the processing of data: by doing twice the processing of a subset of data and by controlling consistency.
- 23. Derived indicators from statistical survey results are mainly used to measure the compliance with some variables in the BR like the economic activity code, size (in terms of employment and turnover), state of activity and localization variables. At last, absolute indicators (as the number of wrong economic activity code) do not provide a sufficient information, it is more effective to build up and disseminate information showing the trend of errors. Graph below shows trend of errors with regards to addresses and state of activity of the BR; they are computed by using the SBS sample survey for enterprises sized below 100 number of persons employed.



2. Indicators based on the use of external sources

24. It becomes common the availability of external sources – pubblic and private – that register information on enterprises and local units, at least to cover specific and limited set of data. For example data supplied by Ministry, list as the yellow pages, data provided by trade associations. On the basis of such sources it is possible to build up some coverage indicators. Large discrepancies, unless because of the peculiarity of that source, highlight some under or over coverage in the BR. Even with reference to some characters like the employment, the use of external sources can support the BR quality evaluation. Estimations of employment – better the yearly variations- can be get from the National Accounts or from the Labour force survey.

3. Indicators that use information inside the BR

25. Information present inside the BR can be furthermore exploited to build up indirect indicators on the quality of the same BR. These indicators are more complex to build up but can produce relevant information.

- A measure of under/over coverage of enterprises, with reference to a given year t, can be computed by counting the number of reactivations i.e. those units that restart activity after a period that have been not active. Generally a reactivation can be identified by comparing three consecutive years: t with t-1 and t-2. The total number (or the average, if one wants to differently weigh sub sets of units) of reactivations, within a yearly updating procedure, is divided by the total number of units (%). If the result increases over time there is a signal of any anomaly.
- 27. With reference to <u>timeliness</u>, an indicators is given by considering that the higher the degree of newness of information and its updating frequency the greater is the quality. The average number of updating and the source from which the updating come from provide a good information on the quality of each variable. In the Italian case the presence of some auxiliary variables into the BR allow to make quality evaluations. Those variables describe:
 - (a) The modality of attribution of that character (attributed from administrative archive, estimated, corrected by check, by profiling activity);
 - (b) The process through which that character was assigned (administrative sources integration or online updating);
 - (c) The date at which the updating process has been done.
- 28. <u>Indicators of completeness</u> can be calculated on the basis of the coverage of a specific stratification variable (economic activity), contact variable (telephone, e-mail) and identification variable (name, address, etc.).
- 29. Not always the completeness of variables indicates high accuracy. The possibility to evaluate accuracy depends on some factors like the amount of available resources. In fact only some subset of units can be checked: the economically relevant units. The number of ON-line updating units or variables controlled by expert can be interpreted as an accuracy indicator.
- 30. Other more complex indicators of accuracy can be computed by using time series of some variables recorded into the BR. Those units showing coherence over time become more reliable for the selection of sample; while those more subject to variations (because of errors in the sources of updating, errors in the characters estimation process) have an impact not only on surveys but also on some analysis such as the structure and the business demography. In particular, with reference to the state of activity the following indicator compares the time series of the state of activity over two years t-1 and t-2 with the state of activity attributed at time t. Starting to the calculation of the variable sst-2_t-1_t (represents the sequence of 0=enterprise not active/missing and 1=active), a measure of error (0 / 1.5) is associated for each subpopulation of the register identified by the variable (see Table 3).

Table 3
Time series comparisons of the state of activity

$Ss_{t-2_t-1_t}$	Frequency	Population	Error
001	408,922	Entries in year t	0
000	2,145,242	Out never active	0
111	3,559,637	Active	0
110	346,218	Exits in year t	0
100	217,721	Exits in t-1 and not active in t	0
011	361,247	Entries in t-1 e and active in t	0
010	54,618	Dis-activation	1.5
101	36,873	Reactivations	1.5

31. Each numerousness is weighted with the error evaluation. The following indicator may assume as maximum the value of 100 in absence of error:

$$I = 100 - (x*val)/x*100 = 98.1\%$$

32. The indicator can be computed for specific subset of units (by size class, sectors of activity, geographical area) and for different time lag.

IV. THE QUALITY DECLARATION

- 33. Once the BR is realized an important issue is let users have tools to interpret BR contents behind the data thanks to the dissemination of quality indicators. A quality declaration (QD) on a complex product like the BR is based on the concept of transparency; this means to supply all the meaningful and useful tools to measure different quality components in relation to each stage of the process. A QD consists of a rich documentation made up of a set of important direct and indirect indicators, having a time dimension for data, sources and variables. It contains:
 - (a) A set of meta-data taken from the database;
 - (b) Some indicators are represented by tables, diagrams, graphs;
 - (c) Calculated synthetic indicators;
 - (d) The high priority indicators;
 - (e) Indicators that can easily interpret and chosen in accordance with the main users needs.
- 34. The following schedule organizes and describes the list of indicators identified in the QD and disseminated to "internal" users. They have been chosen on the base of relevance criteria both in terms of sources and variables. Moreover they cover different quality criteria combined with the different steps previously identified as "input", "process" and "output". Indicators are calculated for each reference year t on the basis of available microdata coming both from the administrative sources used as input in the production process and from the BR itself i.e. the realized output. Time comparisons (generally two consecutive years) and calculated variations add further indications on the BR quality.

Legend: t = reference year of the BR data, s = administrative source of input

A. Quality of input

Criterion: Timeliness

(1) Time lag, measured as the time difference expressed in months between the supply date (st) and the reference date (t)

Criterion: Coverage

- (2) Number of received data from s
- (3) Number of cessation dates by s i.e. the loss of information in t (in terms of cessation dates) when using the administrative supply of t: I=1-supply(t+1) / supply(t) [cessation year=t]
- Number of beginning the activity dates by s i.e. the gain of information in t using the t+1 supply: I=supply(t+1) / supply(t) [beginning activity year =t]

Criterion: Completeness

- (5) Enterprise name, by s: Number (%) of records with missing information
- (6) Legal status, by s: Number (%) of records with missing information
- (7) Address, by s: Number (%) of records with missing information
- (8) Economic activity code, by s: Number (%) of records with missing information
- (9) Municipality code, by s: Number (%) of records with not valid codes
- (10) Number of employees, by s: Number (%) of records having zero employees
- (11) Economic activity code not codified according to the last classification table, by s: Number (%) of records having the old classification

Criterion: Accuracy - Indirect indicators

(12) Number of changes by variables (s): % of the variations between supplies t and t-1 compared to the median (t-n, t)

B. Quality of process

Criterion: Coverage

Information coming from the macro-phase 1

- (13) Missed matching: Number (%) of records of s not matched with s=F (fiscal register) by fiscal code
- (14) Under-coverage due to time lag: Number of clusters of records without s=F (t-1) linked with clusters of records with s=F (t) as ratio over clusters of records with s=F (t-1)
- (15) Structure of clusters of records by number of linked sources: Number (%) of clusters of records made up of 3 and 4 sources

Criterion: Accuracy

Information coming from the macro-phase 1

(16) Nace code: number of records by s with Nace code –old classification – not codified in the new classification over the number of total records (with old Nace classification)

ECE/CES/GE.42/2007/SP/1 page 12

Information coming from the macro-phase 2

(17) State of activity: number records (in terms of units and employment) having a probabilistic status modified by deterministic rules and coverage sources

Information coming from the macro-phase 3

- (18) Number of units (%) with Nace code/state of activity/employees with errors and corrected by deterministic standard rules (automatic)
- (19) Number of units (%) with Nace code/state of activity/employees with errors
- (20) Number of units (% over total errors) with Nace code/state of activity/employees with errors and corrected by deterministic ah-hoc rules (automatic)
- (21) Number of units (% over total errors) with Nace code/state of activity/employees with errors and manually corrected
- (22) Number of units (%over total errors) with Nace code/state of activity/employees with errors but considered valid

C. Quality of output

Criterion: Coverage

- (23) Number of active enterprises in t and variations on t-1
- (24) Number of reactivations in t, as percentage of active enterprises
- (25) Number of units (in terms of enterprises and employment) by modality of attribution (from surveys, estimation, profiling) of a character (Nace code, employment, state of activity)

Criterion: Completeness

- (26) Incomplete Nace codes: number of units having Nace code at 3,2 and 1 digit
- (27) Enterprise name: number of units with names missing
- (28) Address: number of units with addresses missing
- (29) Postal code: number of units with postal code missing
- (30) Telephone: number of units with telephone (fax, email) missing

Criterion: Timeliness

- (31) Time lag, measured as the time difference expressed in months between the BR dissemination date and the reference date of data (t)
- (32) Information available into the BR database having reference time more updated with respect to the reference time t.

<u>Criterion: Accuracy</u> - Direct indicators (Compliance with Statistical Surveys):

- (33) Number of refused questionnaire (%) by typology of errors
- (34) Number of units (%) with wrong address (trend over the last x years)
- (35) Number of units (%) with wrong state of activity (trend over the last x years)
 - (35a) Number of units (%) with changed state of activity and corrected (after the manual check)
 - (35b) Number of units (%) with changed state of activity and with errors (after the manual check)

- (36) Number of units (%) with complying employees
 - (36a) Number of units (%)with complying employees and corrected (after the manual check)
 - (36b) Number of units (%) with complying employees and with errors (after the manual check)
- (37) Number of units (%) with economic activity code discordant
 - (37a) Number of units (%) with Nace code changed and corrected (after the manual check)
 - (37b) Number of units (%) with Nace code changed and with errors (after the manual check)

<u>Criterion: Accuracy</u> - Indirect indicators

(38) State of activity: analysis in time series of the BR (see § 3.3.3)

<u>Criterion: coverage – updating – transparency – dissemination</u>

(39) Updating of the BR_t information during the dissemination of the BRt+1: number (%) of false active units (not active) in t; number (%) of units with changed Nace code in t; number of units and employment (%) with changed employment in t.
