Doc. Class/05/NL/03 - EN

2005-->Rev.1.1??--->---->2008-->Rev.2!!

Implementation guide

Contents:

Introduction

In January 2008 NACE (Nomenclature statistique des Activités economiques dans la Communauté Européenne) gets a new version Rev 2, to be able to reflect economic development. Most EU countries have national classifications from which NACE can be derived. So as NACE changes these national classifications will also change at the more detailed level to reflect national requirements. Because this may vary considerably from country to country, the amount of work may also vary.

From February 2005 onwards the Task Force on Implementation (TFI; for a list of members, see annex 1) has tried to give some guidance to Member States in implementing NACE Rev 2 in the business register and statistical outputs. This paper forms part of that guide and focuses on issues relating to statistical business registers. This paper tries to outline good practices so that readers can judge what might be most appropriate in their context.

The first action involved sending a questionnaire to all national statistical institutes, to be able to discuss items which were related to the implementation and which were important for all countries.

In the April meeting the results were discussed. Although there are many differences between Member States, there seems to be consensus about a lot of problems the Task Force has to deal with.

In the July meeting Hans van Hooff presented the first version of this paper. It was felt it had to be further developed. By adding information from Steve Vale (the Implementation Guide and confrontation with Canada's Experience with Naics 1997 Implementation and Backcasting it was enriched.

It is our impression that this enumeration remains incomplete. Yet it should boost discussion, define mutual problems clearly and lead to:

- Listing items to take into account during the implementation process
- Determining items where assistance may or should be given to the new Member States
- Setting a critical time-path for all countries

Acronyms

ACTR: Automatic Coding by Text Recognition

CATI: Computer Aided Telephonic Interviewing

CPA: Classification of Products by Activity

GBR : General Business Register

ICT: Information and Communication Technology

KAU: Kind of Activity Unit

NACE: Nomenclature statistique des Activités économiques dans la Communauté Européenne

SIC: Standard Industrial Classification (in this document: the national version of NACE)

TFI: Task Force on Implementation

A: Start of project

A1. Point of departure: tools from Eurostat

The structure of NACE rev. 2 will be finalised by the NACE/CPA working group meeting in September 2005. Then it will be submitted to the SPC, after that to the EU-council and the Parliament. In fact it will be 'final' after publication in the Official Journal, expected October, 2006.

According to the 'Operation 2007 calendar of activities' the explanatory notes will be final by January 2006 (a draft is already available within Ramon). Then they will be translated into French and German. It is still uncertain whether they will be translated into all the other 17 official languages, and if so, when. Translation by national statistical institutes may be necessary, which of course means additional work. Eurostat will find out which languages need priority. Countries which make a national version of the explanatory notes themselves might have less priority than those countries which have to wait for the translation of the original explanatory notes in their language. For the first countries later on the national versions can be used as base for the "official" translations.

For the correspondence tables (old-new and new-old) preliminary versions existed in February 2005 and they will be revised until September 2005. These are really important for all countries

B: Preparation

B1. First step: developing an implementation plan

The operation 2008 at the level of individual countries will be a wide-ranging project. The seven main elements are:

- Development of a national classification
- Implementation in the business register
- Implementation in the statistical collections and outputs, including social surveys
- Implementation of CPA/Prodcom (in some countries after the development of a national version)
- Communication with users
- Contacts with administrative sources providing the business register with activity codes
- Revision and approval of legal acts

B2. Development of a national version

Although NACE is the level for reporting to Eurostat, most countries may develop a SIC that is more specific than NACE. The reasons may include statistical demand in the country, legal conditions or complexity of activities etc. Developing an own national SIC requires some preconditions:

a. The national classification should map to NACE Rev 2.

b. Criteria should be developed to determine which 5^{th} digits (subclasses) will be accepted in the SIC. Each split is costly and requires allocation of resources. A decision should be taken if only statistically relevant or administrative criteria are applied. It is important to determine whether the classification will be built only for statistical use or also for administrative use. Other criteria may concern the size and number of units in the subclass, the homogeneity of activities classified in the subclass, the intensity of use of it outside NSI, user needs etc.

c. If the SIC is also meant for administrative use, an inventory should be made of the reasons for inclusion in the administrative systems.

d. The structure, indexes and explanatory notes should be developed and maintained.

e. Coding tools and other systems should be modified for changes in SIC.

B3. Correspondence tables and transition codes

Once the new SIC has been developed, correspondence tables should be made between old and new classifications. The following situations are feasible:

- 1 to 1: the old code can directly be transformed into the new one
- n to 1: the old codes that have to be transformed into one new code can also be directly transformed, though correspondence tables will be needed for back-coding new units in the future.
- 1 to n: it is not clear to which new code an entity should be attributed.
- n to m: a number of old codes have to be translated into a number of new codes.

Each individual relationship between an old and a new SIC-code can be expressed in a transition code, a code which is the relationship between the old and new SIC and which allows a comparison between old and new to make time series. Both versions can be derived of this transition code.

B4. Estimating changes and creating a research environment

When NACE Rev 2 and the SIC are finished and correspondence tables are made, the impact of the classification change can be seen. The changes can be communicated within the NSI.

An assessment can be made concerning the numbers of units that can be related automatically and those that have to be checked.

Often it is a prerequisite to create a research environment. Because all information concerning recoding has to be stored, a database with coordinated information should exist as well as a processing tool to assign codes.

B5: Critical timepath (Steve Vale)

Implementation plan	a.s.a.p.
NACE Structure ready	September 2005
Structure of national SIC ready, soon followed by explanatory notes, indexes, coding tools, conversion tables etc.	March 2006
Register implementation plans, sources (in NSI, external) to use, changes required in surveys, where are probabilistic models appropriate	April 2006
Changes to the register database	July 2006
Information gathered for new SIC	October 2007
Implementation of codes in business register	January 2008
Dual coding in business register	From January 2008 to December 2009

C: Sources

In many countries the (administrative) sources play an important role in updating the business register. They may also play an important role in implementing NACE rev. 2.

C1. External sources (Steve Vale)

Because of existing dependency on external registers (administrative like taxes, social security, Chambers of Commerce) or because of lack of capacity, it often will be necessary to use and thus try to adapt these external sources to statistical needs. This includes:

- Early warning: this will enable sources to implement changes at a convenient time
- Provision of instruments like explanatory notes, indexes, transition schemes (correlation matrixes) to the external sources;
- Assistance with the adaptation of computer systems and coding tools
- Dissemination of information to the employees of the external source operating the system, assign codes or audit (presentations, guides, training).

The ideal situation would be to have the administrative sources convert to the new classification **at exactly the same time** as the business registers of NSI's

Other 'external' sources can be important as a source of information:

- Chambers of Commerce, taxes (VAT, wages etc.), social security.
- Umbrella organisations, trade or telephone directories.
- The internet may be an interesting source both for individual enterprises and for the more aggregated level.

C2: Internal sources

Like external sources, internal sources can be important to prevent unwanted approaches to businesses. Of course internal sources should be coordinated in implementing and using NACE rev 2.

- What information is already available within the NSI? There may be enough information already within systems about enterprises and, even when a little outdated, it may suffice.
- Information from profiling of the large and complex enterprises.
- The existing national (5^{th} digit) SIC-codes may suffice to classify to NACE rev. 2.
- It may be that a kind of automatic coding system is used, which registers descriptions of the activities of a unit. A realistic description of activities may suffice to make a split following NACE rev 2. For this purpose it shouldn't be more than 5 years old. In case of using business descriptions a confidence rating should be applied. Low confidence means that the entity should be approached (Statistics Canada).

D: Tools for survey

As far as possible, tools should be used. Some are mentioned here.

D1. Use of indexes

Generally speaking using structural standardized information allows the use of indexes. A list of activity descriptions helps assigning codes. So when interpreting existing information on activities or evaluating that information from inquiries, indexes may be of great help to all countries.

Indexes have to be developed by the member states themselves at national SIC-level. This means that several countries have to do some of the same tasks, especially if they share the same language. We suggest some coordination by Eurostat (TFI) helping each other.

A search engine may be operated to consult indexes.

D2. Computer aided coding systems

As was evident from the TFI questionnaire sent to the member states that not many countries use coding tools or even more advanced systems. ACTR from Canada at the moment seems the most outstanding performance system.

Next to the fact that indexes with connected search systems are sometimes described as a coding tool, there are basically two types of systems.

First there are systems based on *linguistic engineering.* In this case text descriptions are related, recognising words which may be in a different order etc. This approach is very language dependent, so with 20 languages not very suitable for our TFI-purpose.

Secondly there are systems which use string matching. These work by storing descriptions and codes. If a new description scores more than a preset value the match is accepted and the code is used. It is possible to set score levels to give different trade-offs between quality of codes and quantity of descriptions. These systems will work in all languages as long as the index is translated.

D3. Automatic Coding by Text Recognition as example (ACTR)

ACTR is an example of a string matching system. In the survey of February it was concluded that ACTR might make a valuable contribution to the implementation of NACE Rev. 2.

As stated before it will be a huge task to recode all required units by sending questionnaires. Because of limits in capacity of statistical institutes and their sources and the burden on entrepreneurs, as much as possible existing information should be used.

ACTR assigns codes to descriptions. So there should be descriptions of activities and they should be as correct as possible. In order to be able to code a parsing strategy is operated. Parsing means that rough input-text of the respondent is adapted to text in the ACTR Database in order to be able to match and to assign codes.

ACTR requires a context, a *reference database,* before anything is possible. This context – possibly not only on activity but also on e.g. profession, CPA-categories etc. – is created by ACTR using specifications and data provided by users in form of plain text files. The larger the database, the more possibilities ACTR has. This is the information ACTR needs to standardize input and assign codes. A *transformation file* to code and recode and sufficient *disk space* are also required. It is also possible to set score levels between the quality of coding and the quantity of descriptions coded.

Conclusions about ACTR

- ACTR needs a lot of databases. This need can only be met when a register with descriptions is operated (either by the Statistical Institute or at the sources) or all units are surveyed by questionnaires with open questions.
- For the matching it is necessary that the Database is fed with full explanatory notes of the country's own SIC
- ACTR can be used for other classifications if suitable indexes are created;

Although ACTR may be a good instrument to code units from our point of view it is not useful to oblige countries to use such a system. Some countries will not be able to meet standards (elaborating the database causes a lot of work), others will already use indexes or a coding tool that do not exactly meet the standards of ACTR. Besides, emphasis should be on introducing new codes not on the development of new instruments.

E: Theoretical issues (in surveys)

E1 Determination of the principal activity in theory and in practice (Arto Luhtio)

As stated in the NACE introduction, the principal activity of statistical units should be determined according to value added, where possible. If this is not possible, a proxy can be used. Several proxies are discussed in this document. There are no clear priority rules for the use of different proxies. The use of a certain proxy depends on several issues, like the information available for the statistical unit concerned, cost-benefit considerations and the country practices. Although the basis situation is the same: the principal activity – and secondary activities, where applicable – of all statistical units recorded in the business registers need to be coded according to NACE Rev 2 from 1 January 2008 onwards. There is great variety how this is done in practice.

Value added can be used most often for enterprise groups, possibly for enterprises, but hardly for local units. Value added can be used to determine the principal activity of a truncated enterprise group in the following way: the annual turnovers of the resident members of the group are multiplied by the value added index of the respective activities. The multiplied turnovers are summarized by NACE at 2-digit level and the activity that contributes most to the total value added is identified as the principal activity of the group.

Some activities like insurance and financial intermediation don't have turnover in the traditional sense. The principal activity of groups, whose members are active in these fields, can be determined based on employment.

E2 Use of production data for assigning the new codes (Joachim Weisbrod)

The central problem of the implementation of a new classification is the efficient and reliable recoding of the activities of the statistical units. There is a variety of methods available, a mix of which will be used depending in the special circumstances in the different domains. Since information for recoding the principal activity is limited but recoding has to be done before statistical information becomes available, other available sources of information have to be used efficiently.

A possible source of information is the production survey according to the PRODCOM-list or a national version of it. Of course this source of information can only be used for units in the PRODCOM-survey, i.e. NACE Rev. 1.1. section C to E, but the manufacturing sector is still a large part of the industry. In Germany data are collected for local units, in fact 80% of the enterprises only have one local unit and one kind of activity unit.

The PRODCOM statistics produce production data (volume and value) according to a Europe-wide harmonised product list. The coding of this list is directly linked to NACE and CPA, as the first four digits of the PRODCOM code correspond to the respective NACE class and the first six to the CPA code. Units are obliged to report their production according to the most detailed headings of the PRODCOM survey. So prouction volumes and values are obtained which can be condensed to the NACE class level by aggregation.

If the new CPA is already implemented in the PRODCOM survey, the production values of the new NACE Rev. 2 classes can easily be aggregated. If not, the production values first have to be converted to the new coding system. Therefore it is very important that the 2007 PRODCOM-list is alreay double coded according to the old and the new classification.

The principal activity is to be determined using the 'top down method'. To receive a first indication of the new principal activity this method could be applied to the production values of the local units and the enterprises.

Theoretically the pincipal activity should be determined by the value added, second best would be the employment or turnover. This information is not available when NACE Rev. 2 is implemented in register. With additional information and assumptions, the net production values and/or persons employed may be estimated for the new NACE Rev. 2 classes before the 'top down method'is applied.

Production Statistics only cover NACE Rev. 1.1. section C to E, but this is still a large part of enterprise statistics. The method can't be used mechanically but in combination with other methods as automatic recoding and individual research. Yet it is a good method because a large part of the work can be done by computer.

E3 Use of structural survey data for assigning the new codes (Emmanuel Raulin)

The use of structural survey data for defining the principal and secondary activities in the business registers is increasing, although some countries argue that this may cause a bias in the register. However, there is no danger of bias in cases when all units above a certain threshold are surveyed. Also concerning the use of sample survey data some countries (e.g. Canada) have reconsidered that the benefits outweigh the negative effects of the caused bias and have started to use them.

These data can also be used for backcasting annual statistics series. Generally spoken there are two methods, the 'macro'and the 'micro'-method. The macro consists of converting series in the new classification directly. The micro is about individual information and the principal activities. The advantage of mcro is that it will be easy to recalculate data series in the new classification. The decisive variable can be either the added value or the number of employees. It is proposed to apply the micro-method to handle future classification change.

It is thought that it is necessary to have double coding for at least two consecutive years, with only the second year being used for carrying out retropolations.

Double coding requires the branches to be observed in a intermediate classification. Activities can then be given in both the old and the new classification. So only a single breakdown of turnover is needed. For secondary activities transition matrices are used, either individually or as 'average'.

E4. Assigning new codes to individual units (Micro-level)

Not only enterprises have to be classified. It will also concern local units and enterprise groups, but also (Local) Kind of Activity Units (KAU/LKAU) if these are recorded separately in the registers.

From our point of view regular updating at the level of the individual unit should have first priority.

The recoding can start from different units in different countries. Some countries (e.g. UK) start the recoding at local unit level, then determine the enterprise and enterprise group using bottom-up approach. Other countries start at the enterprise level and recode the local units after the enterprise.

At the units one main activity or one main and one or more secondary activities can be determined.

For the "1 to n" and "n to m"- cases an evaluation should be made to see if these units are important enough to spend much effort to get the perfect main activity-code on individual level. If no, new codes may be deducted.

E5. Assigning new codes at group level (Macro-level)

May be there is not sufficient information to prevent asking by surveys, on the other hand it is impossible to send too many units a questionnaire. So a solution has to be found.

It is possible to make an estimation of the number of units to be assigned to new codes by:

- Research. Based on expertise or by means of separately treated and investigated large entities it is possible to guess how a split could be made. Another possibility is sample survey in 1 to n or n to m areas. From this a kind of probabilistic model may be developed.
- Probabilistic models (see also E7). It is interesting to see which results these methods will have.

In the UK probabilistic correlation tables are used. There are several versions, depending for what purpose they are used. They may also be different for enterprises and local units.

A disadvantage of these methods is that it is impossible to assign accurate codes to individual units. Therefore a margin of uncertainty remains and will grow if this method is used too often. Our advise is to use these models carefully and if necessary only in statistically less relevant areas or for small units.

As a result on individual level, many of these (small) units may have the wrong codes, but aggregates will be more correct. Because they are typically single-site units, the impact on statistical output will be relatively low.

E6. Assigning new codes to large units

Large units are more often made to measure than the small ones. For large units there often exists some personal contact e.g. account-management. Here personal contact and the high level of updating make it possible to assign codes on factual information of activities (see also C2).

For larger units not involved in account management the recoding process can be realised by introducing an additional or changed question in the questionnaire.

E7. Constructing probabilistic correlation matrixes (Steve Vale)

It may be necessary to use correlation matrixes for probabilistic recoding. Next steps should be made:

1. Recode all units for which the information needed to do this with an acceptable degree of certainty is available
2. Cross tabulate these units by old and new code
3. Remove any invalid combinations (as determined by the look-up table supplied by classification experts)
4. Calculate percentages based on the remaining data (taking care when counts are particularly low)

5. Take into account any other relevant data, e.g. is the correlation affected by the size, or some other attribute of the units
6. Create a correlation matrix based on the above. This can be a simple matrix, or a more complex based on multiple variables.

The matrix should be tested to ensure that the results seem reasonable. It should be remembered that the correlation probabilities are likely to change over time.

F: Realisation by surveys

F1. Preparing for survey

A population should be made including new entries and ignoring disappeared ones.

Questionnaires have to be made. A general questionnaire applicable to all activities may ask a lot of work of the respondent in providing the information needed. On the other hand a coding tool may be operated to determine the code based on this information.

A questionnaire with closed categories makes the collection of information much easier but may not necessarily result into the optimal answer. Therefore the questionnaires should be tested by laboratory research.

In this case instead of a coding tool an optic reading system could be effective.

In the UK both open and closed questions are used: a broad open question to determine the broad sector, then a closed question to determine the precise activity. Sometimes both are needed to code a description. E.g. if a business writes ‘wooden doors’ the closed question is used to determine if it concerns manufacturing or selling.

F2. Surveying

Questionnaires should be sent to all selected units for which it is impossible to get information from other sources. A strategy should be chosen to deal with non-respondents. This includes decisions on the number of reminders, the medium used (mail, telephone, e-mail) etc. Units which can not be reached or refusals can be classified by probability based models. To prevent non-response a motivating letter should be added strengthening the importance for the respondent

Of course internet can be of great importance. Some arguments are:

- It is less expensive;
- It lowers the administrative burden;
- It is easier to process the information.

Of course the result may differ because it is dependent on the levels of penetration of internet in the economies of the member states.

F3. Quality controls

Information from surveys should be complete and correct. Especially where respondents have to specify the information it is necessary to notice that this full information. On the other hand it is necessary to control the codes that are awarded manually.

For both reasons a specialist should re-consider a sample of the survey population in order to establish its quality. This should be done as a regular audit.

F4. Recoding the Business Register

After having dealt with all units to be transformed, old information about these units in the register has to be adapted. Some facts to note:

- In the BR the new code system should be introduced in time.
- The BR should have both NACE-versions operational for one or more years. Of course this has to be possible for the register. Double coding is a requirement for statistical follow up and successful implementation.
- Because of the time needed to carry out research (2 years for many national statistical institutes) some units may end their activities and more important, some new ones may have been registered. In the survey population this should be taken into account.
- Units that are registered after closing the surveys have to be contacted or should be dealt with theoretically. Without information a code should be given based on experience or models. Margins of quality should be taken into account.
- Audits about quality of codes should be held. Corrections needed because of technical failure should be made.

F5. Treatment of corrections

Although the number of units to be surveyed should be minimised, it will always lead to correction of codes next to changes of codes which are described in the transition code-scheme.

Because changes in NACE and SIC will take place in areas that possibly couldn't be described in an optimal way before, it is to be expected that a relatively large number of questionnaires will cause corrections. Because only part of the register is surveyed, bias is created.

There are two ways to treat corrections. As discussed it is important for Statistics and National Accounts to be able to back-cast and make time-series. Therefore there could be arguments to ignore corrections. They could be given a code, knowing it is not the correct one, and have statistics adapt them in time. So no extreme effect occurs.

The more optimal way is to implement the corrections. In this case not only transition schemes should be made but also *input-output schemes of corrections* to alert statistics. The importance can then be assessed taking into account the size and importance of the unit. Of course it may have (extensive) consequences for the statistical domains.

If we allow quality improvements and don't investigate "1 to 1" and "n to 1"-correlations we introduce bias in the register (units will only move out of 1 to n and n to m classes and not into these classes).

Constraints in capacity, resources etc. may prevent the improvement of quality.

G. Follow-up

As follow up the next activities are necessary:

- Plausibility-checks on the new codes in the GBR;
- Possibly: improvement of transition and correction schemes;
- Investigation whether the sources of the GBR provide correct information;
- Adaptation of statistical domains and samples;
- The national SIC should be evaluated looking at the results of the surveys. Missing items may be added in the indexes, explanatory notes may be modified. It can also be decided to adapt the 5th digit structure based on the results of the surveys;
- Coding tools and other typifying systems should be adapted again;
- Registers using the SIC should be informed about all new changes;
- Evaluation: information on the results of the surveys should be given.

Annex 1: Members of TFI

Eurostat:

Alice Zoppé

Arto Luhtio

Leila Anupold

Michael Mietzner

Petra Sneijders

Isabelle Remond-Tiedrez

Paul Konijn

ECB

Heinz Christian Dieden

NSI's

Michel Blanc (Fr)

Michel Lacroix (Fr)

Thierry Lacroix (Fr)

Emmanuel Raulin (Fr)

Joachim Weisbrod (Ge)

Ana Isabel Sanchez-Luengo (Sp)

Raquel del Rio Paramio (Sp)

Ole Black (Uk)

Mark Williams (Uk)

Steve Vale (Uk), temporarily OECD

John Perry (Uk)

Norbert Rainer (Au)

Zsolt Völfinger (Hu)

Márta Rónai (Hu)

Hans van Hooff (Ne)

Ton Bonné (Ne)