UNITED NATIONS **E**

Economic and Social Council

Distr.
GENERAL

ECE/CES/GE.42/2007/SP/8
2 May 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)

IMPLEMENTING CODING TOOLS FOR A NEW CLASSIFICATION

Submitted by United Kingdom

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

SUMMARY

The United Kingdom (UK) introduced automated coding tools for coding descriptions of business activity as a means to move from its 1980 to its 1992 industrial classification system, the latter being an extension of the European NACE Rev. 1 standard. The tools provided consistency of coding of source data for the first time. A minor change to the classification system in 2003 resulted in an adaptation of the coding tool and guidelines for future work. Closer working with the administrative departments and modernisation of the ONS computing systems have resulted in further development of coding tools. This has provided a good basis for using such tools in the introduction of the new European classification, NACE Rev. 2, in 2007.

I. INTRODUCTION

1. This paper examines the development and use of automatic coding tools in the UK Office for National Statistics, focussing particularly on their application to coding the economic activity of businesses. It describes how automatic coding has become the key to implementing new versions of the UK Standard Industrial Classification (SIC), showing how a combination of coding stored business activity descriptions and the use of probabilistic conversion matrices has removed the requirement to conduct specific surveys when classification systems change.

II. BACKGROUND

2. The UK has been using automatic coding tools to convert text descriptions of activity provided by businesses into economic activity codes for over 15 years. This started with the use of text searching within electronic versions of published classification indices in the early 1990's, but soon progressed to the use of specialised software with the introduction of "Precision Data Coder"[1] (PDC) software in 1993.

3. In the early 1990's, UK Employment Department the redesigned its Census of Employment questionnaire to make use of developments in optical character recognition (OCR) technology that allowed free text business descriptions to be captured electronically. The PDC was then used to generate industry codes from these descriptions. This involved processing batches of text descriptions automatically and following this by interactive coding of those that the PDC could not code without intervention. The clerical process involved the selection of the best of multiple choices offered by the PDC. Where no choice was given or none of the choices was regarded as appropriate, the clerical coder modified the description to achieve success.

4. Following the merger of many of the key statistical functions within the UK government in 1996, responsibility for the Census of Employment passed to the newly created Office for National Statistics (ONS), along with the business register and other economic statistics.

5. The 2001 UK population census made use of automatic coding of business activity and occupation for the first time. The processing of this census was contracted out to a private company, Lockheed-Martin, who used the ACTR[2] coding tool developed by Statistics Canada.

6. Around the same time the ONS led a consortium of statistical agencies and private companies in the CLAMOUR project.[3] This project was funded through the European Union fifth framework programme for research and development. It considered new approaches to classification and the construction of statistical units, including work on developing automatic coding tools and principles. This project helped to establish the value of automatic coding as a tool to assist the development of classifications.

[1] A commercial product developed by Inference Group Pty Ltd, Australia – see www.inferencegroup.com.au
[2] Automatic Coding by Text Recognition.
[3] See http://www.statistics.gov.uk/methods_quality/clamour/default.asp

III. THE 2003 CHANGES – A DRESS REHEARSAL

7. A minor update of European economic activity classifications in 2003 provided an ideal opportunity to test the value of automatic coding software as a tool to facilitate classification system changes.

8. The business register had been accumulating free text business descriptions for local units from the Census of Employment and from other surveys. The project to implement the new classification within the UK took place during 2001 and 2002, partly funded by a grant from Eurostat. It started by identifying the local units affected by the change.

9. When the electronic versions of the agreed descriptions and associated metadata for the new codes became available, they were incorporated into a modified version of the PDC. A sample of free text descriptions already on the business register formed the basis of comparison of PDC and clerical coding.

10. The automatic and clerical coding rates and quality were compared. The results were used to inform the creation of probabilistic conversion matrices, which were used to re-classify businesses for which no economic activity description was held.

11. The main conclusions of this study were:

(a) Text descriptions of business activity are the key to changing the classification of units in statistical business registers. They allow the automatic coding of a number of units, provide a test of relevance for proposed classification changes, and help with the modification and improvement of coding tools. It is therefore important to capture business descriptions electronically and store them for future use.

(b) Many of the coding issues found in this project could be traced back to the quality of data supplied by contributors. Questionnaire design is therefore key to ensuring that information captured is of sufficient quality to meet automatic coding requirements.

(c) An automatic coding tool needs comprehensive metadata to build index entries. Time is needed to incorporate the metadata for a new classification into a coding tool and to test that the results meet the required quality standards.

(d) Automatic coding gave better results than clerical coding both in terms of accuracy and consistency. This is the result of investment over a number of years in refining and enhancing the coding tool used.

IV. A NEW CODING TOOL – ACTR – AUTOMATIC CODING BY TEXT RECOGNITION

12. In 2003 the ONS evaluated a number of automatic coding tools, with the aim of choosing one to become the standard for all coding functions across the office. This assessment was part of the wider Statistical Infrastructure Programme, which aimed to standardise the software used for a wide range of statistical processes and promote best practice within the ONS.

13. The outcome was a decision to choose ACTR from Statistics Canada as the office standard coding tool. ACTR was seen as providing acceptable results across a range of coding applications, initially for industry and occupation coding but with a view to wider application in the longer term. The flexibility to adapt and tune both the coding engine and the "knowledge base" files (which are used to allocate codes) was seen as a strong advantage of this tool.[4]

14. Following this decision, ACTR is being rolled out for coding applications across the office. It replaced the PDC as the tool to code the economic activity of businesses on the ONS business register in April 2006, ready to assist with the implementation of the new UK Standard Industrial Classification (SIC 2007), which is the UK fifth digit extension of NACE Rev 2.

15. ACTR is essentially a text matching engine. To be effective, each coding application needs a specific "knowledge base". This is basically a reference file of standard text descriptions and their associated codes. Incoming text is matched against the entries in the knowledge base, and each match is given a score. If the highest scoring match exceeds a pre-determined threshold, the code associated with the text is automatically allocated to the record.

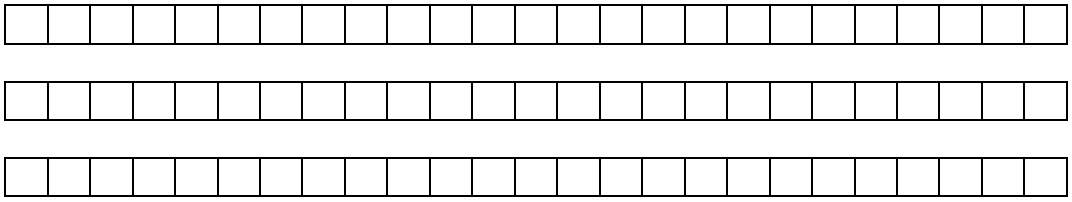
16. Figure 1 shows how business activity descriptions are captured on the current Business Register Survey questionnaire. If a description is already held, this is played back to the business to confirm or correct. If no description is held, the standard description relating to the current SIC code of the unit is used.

Figure 1
Excerpt from the Business Register Survey Questionnaire

2. **Your workplace business activity** (see guidance note 5)
Our records show your business activity as:

2a. If your current business activity is different, please describe it fully in the boxes below (one letter per box with a space between each word).
Show the main goods or services involved. Include details of products and materials used (e.g. Manufacture of UPVC windows). If you provide a service, describe it fully (e.g. licensed hotel and restaurant).

[4] For more information on ACTR see: http://www.census.gov/srd/papers/pdf/rr94-2.pdf



17. ACTR can be used in batch and interactive modes. Typically incoming text descriptions are fed through in batch mode. Those cases that are not assigned an automatic code are referred for clerical checking in interactive mode. For business activity descriptions, we use a threshold that delivers automatic codes for approximately two thirds of records, with an accuracy rate of more than 90%. "Accuracy" is difficult to determine because there is often an element of subjectivity in a coding decision. To arrive at this figure, we compared output from ACTR with codes independently assigned by our most experienced coder. Cases where the codes were not the same were referred to a classification expert for a decision on which was correct.

18. At the time of purchasing the previous tool (PDC), part of the evaluation was a comparison of clerical and automatic coding. For the clerical coding several coders were assigned the same task and the coding decisions compared with each other and with the outputs of the coding tool. Agreement between coders and with the coding tool was achieved in 85% of the cases.

19. Language is often rich and meanings can differ depending on the viewpoint of the coder. As an example "supermarket sub post office" would appear to have the meaning of a postal service that happens to be located in a specific building, a supermarket. However, it could be that the activity of food retailing is integrated with the postal service and the ordering of words with "supermarket" means that the primary activity is food retailing. Variations in the business description can alter the perception of the coder. This is a major issue for clerical coding. An automatic coding tool will replicate a decision for each sufficiently similar description.

20. The knowledge base is key to this approach and represents the greatest investment in the coding process. The ACTR knowledge base has a text format, which makes it relatively portable. For this reason, we have decided to make the knowledge base a "public good", available to anyone who wants to use it, either with ACTR or with some other matching engine. The reason for this is that if everyone wanting to code automatically to the SIC has access to the same free knowledge base, there would be much greater consistency of SIC coding within the UK. The current plan is for a new release of the knowledge base every six months to reflect improvements based on case-law decisions and user feedback.

21. In the context of business activity coding, the main suppliers for the UK statistical business register are HM Revenue and Customs, who provide VAT trader and PAYE employer records, and Companies House, who supply list of business incorporations. The VAT trader system requires businesses to supply a business activity description, although this is currently held only as a scanned image. Discussions with the administrative department have resulted in their recognition that a knowledge based coding tool such as ACTR would have a benefit in the area of VAT collection.

22. The PAYE employer system holds business activity descriptions for some seven million businesses. Agreement has been reached to provide these to the ONS in the near future. Coding within the administrative department has been a clerically intensive process. While the administrative department acknowledges that a tool such as ACTR would assist in automation, the quality of the descriptions and the sheer size of the data set would still require substantial clerical input to code interactively those that would fail batch processing.

23. The company registration system takes a different approach to coding, by supplying a copy of the industrial classification for companies to select a code when completing their annual returns. The outcome is that there is a heavy predominance in coding to property and business services that is not reflected in the activity coded from supplied business descriptions. However, changing to coding from business descriptions would be a major task and is not foreseen in the short-term.

V. IMPACT

24. Changing coding tools will inevitably result in changes in coding outcomes, affecting the quality of outputs. Resources have been invested in developing the SIC knowledge bases for ACTR, resulting in a throughput rate similar to that for the PDC, with an accuracy that is comparable. However, the difference in approach between the two tools (ACTR uses a text matching approach, whereas the PDC relies more on linguistic engineering techniques) means that there are some significant differences in the results.

25. Annex 1 shows the impact of changing the coding tool at the ISIC Rev. 2 section level. There are two components to this impact, the effect of changing the coding engine, and a quality effect linked to the development of the ACTR knowledge base. The only way to separate these two components would be to apply the new knowledge base to both coding tools and analyse differences in outputs. This was not possible in this case due to cost and licensing restrictions. Thus the movements shown in Annex 1 are a combination of the introduction of ACTR and the difference in quality between the current ACTR knowledge base and the reference data used by the PDC.

26. The impact needs to be considered in different ways for different groups of users. The impact on business counts will have the greatest relevance for users of business demography statistics, whereas the impact on employment, turnover and other economic variables will be of more relevance for those interested in economic aggregates. In terms of the potential use of ACTR by administrative departments and agencies, measures such as efficiency gains and impacts on burden and revenue are likely to be more important.

VI. USING ACTR FOR THE TRANSITION TO THE SIC 2007

27. Most countries introducing new economic activity classification systems over the next few years are conducting specific surveys to assist with the reclassification of businesses. The ONS has taken the decision not to follow this approach, partly due to cost and business burden constraints, but mainly because we believe we already have the necessary information in the

form of stored business activity descriptions to allocate SIC 2007 codes which meet acceptable quality standards.

28. Those businesses for which we do not have a stored business activity description are mostly small, and thus relatively rarely selected for statistical surveys. The SIC 2007 codes for these businesses will be determined using probabilistic conversion matrices, based on the proportions of businesses for which we do have descriptions that are allocated to the various allowable combinations of old and new codes.

29. The use of probabilistic conversion matrices inevitably means that a proportion of businesses is allocated the wrong code at the most detailed level of the classification. However, the error at unit level is much reduced at higher levels of the classification. Survey sample designs in the UK are increasingly using broader economic activity groupings to allow more detailed regional stratification, and to reflect the reduced sample sizes required to meet government targets on cost and business burden reduction. This means that the impact of unit level classification errors is much reduced.

30. The accuracy of aggregate data for smaller businesses depends heavily on the quality of the conversion matrices. If these matrices are constructed correctly, aggregates should be unaffected by errors at unit level. This approach therefore places considerable emphasis on the quality of automatic coding, thus the development of a suitable knowledge base for ACTR is critical to the success of the transition to the SIC 2007.

31. A potential additional benefit of the knowledge base approach is that it could be used to help inform the development of future versions of a classification. A knowledge base can be seen as providing a reference frame of business activity descriptions. If it is linked to stored business activity descriptions, it can be weighted by occurrence. The result is a series of weighted building blocks, which can be grouped into categories to give the different levels of the new version.

32. Another important feature of the transition to the SIC 2007 is the length of the transition period. Different statistical outputs will make the transition at different points in time between 2008 and 2011, based on a timetable largely driven by European Union requirements. The transition dates for several key outputs are set in European statistical legislation.

33. The timetable in the UK started with the publication of NACE Rev 2 and the UK SIC 2007 in January 2007. Only then was the final structure confirmed and implementation of the SIC 2007 on the statistical business register could commence. The NACE Rev 2 Regulation required the business registers to hold the NACE codes by 1 January 2008. Structural business statistics outputs are required on a SIC 2007 basis from the 2008 reference year, with short-period statistics following in 2009 or 2010, and finally national accounts in 2011.

34. This long transition will require a lengthy period during which businesses will have to be coded to both the SIC 2003 and the SIC 2007. One solution would be to have two separate knowledge bases for ACTR, one for each version of the SIC, and to pass business activity descriptions through ACTR twice, once with each knowledge base.

35. Not only would this approach be inefficient, particularly in terms of clerical resources to code cases that fail to code automatically, but it would inevitably lead to inconsistencies in coding. The approach taken has instead been to allocate a bridging code to each index entry within the ACTR knowledge base and to hold a relationship between the code and both the SIC 2003 and the SIC 2007 code. This means that business descriptions need to be passed through ACTR only once. Any decision to change the SIC 2003 or SIC 2007 codes for any index entry can then be reflected on the IDBR without resorting to recoding from the business description. This approach delivers further benefits if the coding system itself changes. All that is required is the creation of a relationship between the index entry and the classification system.

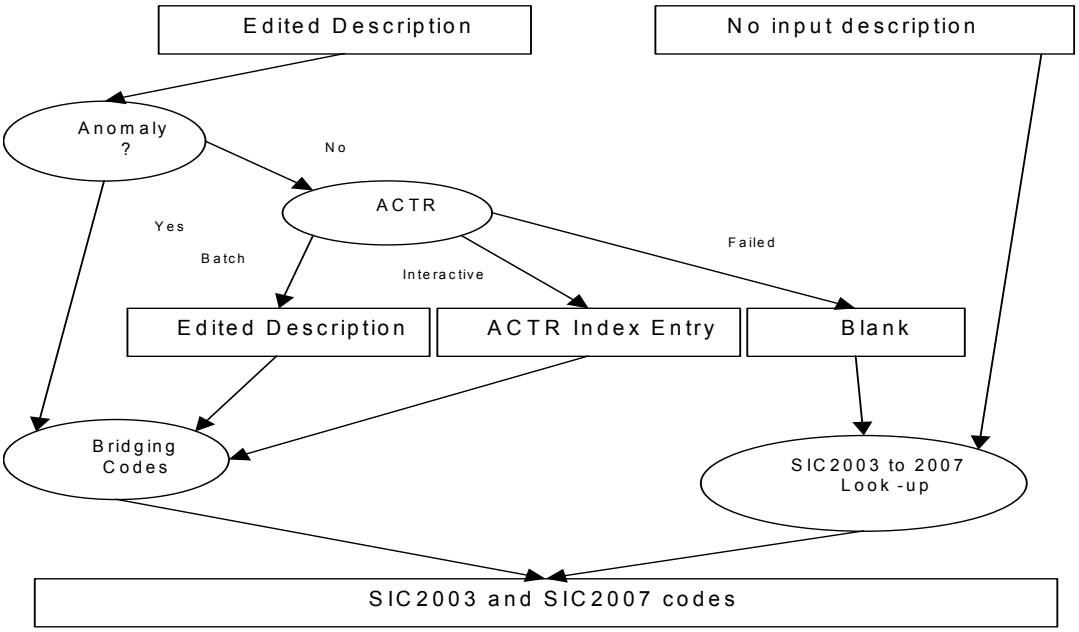
36. Figure 2 shows how this process works in practice. The starting point is one of two sources:

(a) Business descriptions already on the IDBR – these are edited descriptions based on previous input from businesses, usually through previous business register surveys.

(b) New business descriptions from the latest business register survey, as supplied by the business – these are scanned and keyed and may contain keying or spelling errors.

The first stage of the process is to copy the input business description to be used as the editable business description. The original input is always retained.

Figure 2
Logical Model of the Bridging Codes Approach



37. Where there is no information from the business, which is the case for most small enterprises and for non-responders to the business register survey, a SIC 2003 code is available but there is no business description. There are a range of sources for SIC 2003 codes for such enterprises:

(a) VAT Traders. The VAT registration form and the VAT business contact systems collect business descriptions but these are held only as images with coding from VAT done clerically and the ONS receiving only the resulting SIC 2003 code. The VAT trader system will deliver the SIC 2007 code from 1 January 2008.

(b) PAYE Employers. Coding is based on trade classification numbers (TCN) that predate the SIC 2003. The ONS receives the TCN and converts it to the SIC 2003 using probabilistic look-up tables. Conversion to SIC 2007 will be through the look-up table from SIC 2003 to SIC 2007.

(c) Company Registrations. The company registration system requires a business to allocate a four-digit SIC 2003 code. This is done at registration or through the annual returns. Where a code is not provided through an annual return the previous code is retained and may relate to an earlier version of the SIC. Companies House is proposing to introduce the SIC 2007 at the full five-digit level from January 2008. Existing codes four digit codes will be converted to the SIC 2003 Conversion to SIC 2007 will be through the look-up table from SIC 2003 to SIC 2007.

(d) Farms Survey System. This uses its own farms type classification that will not change. A look-up table from farm type to SIC 2003 will be used followed by conversion to SIC 2007 through the look-up table from SIC 2003 to SIC 2007.

(e) Construction Statistics System. This system provides SIC 2003 codes but not descriptions. It is expected to move to the SIC 2007 but in the interim conversion to SIC 2007 will be through the look-up table from SIC 2003 to SIC 2007.

38. For existing business descriptions, it is necessary to conduct a minimum of pre-cleaning, removing extraneous characters and correcting spelling errors. This improves the throughput of ACTR, which is essential when more than one million business descriptions are to be coded. For new codes, volumes are lower and quality is higher and no pre-cleaning is done. ACTR is then run as a batch process and generates three files:

(a) Unique code;
(b) Possible to code but multiple coding choices;
(c) Not possible to code.

39. Units for which there is a unique bridging code choice update the IDBR automatically, storing the SIC 2003, SIC 2007 and bridging code. Where there are several bridging code choices with the same score in ACTR, the first is chosen and the same procedure applies. There is a risk that this could lead to bias when the set of bridging codes cover more than one SIC 2003 to SIC 2007 combination and how to deal with this is under investigation.

40. Those that fail batch coding because there is no score above the coding threshold, the poor matches, are then subject to interactive coding. In these cases, the clerical staff make a choice from possible matches with relatively low scores, modifying the business description where necessary to achieve a good correspondence between the knowledge base description and the supplied business description. This is the element of the process that is most open to question as it might lead to bias. Although there is the possibility to telephone the business at this stage, this is an expensive process. Once the choice is made, the bridge code description is fed back to the IDBR, not the edited description. This means that for except the smallest businesses the possible clerical bias is present only until the following survey period when the description is presented to the business for checking.

41. Where descriptions are too poor to achieve a match and contact cannot be made with the business, the edited business description is not held on the IDBR. The SIC codes are derived automatically from other, generally administrative, sources and the standard SIC heading description is returned to the business as part of the next survey contact.

42. In some instances, the description is precise as far as the business is concerned but not sufficiently informative to generate a SIC code. An example would be “retailing”. The initial decision was to create a business rules knowledge base but experience with ACTR demonstrated that this could result in reductions in quality of coding of other description. Instead a separate list of these coding anomalies is held. Where such a description is provided, a precise character for character match to the list of anomalies results in blocking of the ACTR coding decision from updating the business register.

43. As a result of this process more than two-thirds of business descriptions are coded automatically, with the remainder being coded interactively with minimal clerical input. Consistency of coding is improved and the feedback to businesses provides a continuous improvement to the data held on the business register.

VII. CONCLUSIONS

(a) Automatic coding tools are essential to the efficient implementation of new versions of classifications, delivering considerable savings in terms of cost and burden on businesses compared to traditional survey approaches.

(b) Creating a knowledge base that is portable (i.e. independent of the coding engine), and sharing this with any interested parties, e.g. administrative data suppliers, considerably increases the consistency of coding.

(c) The knowledge base approach can help to inform the development of future versions of a classification, by providing a reference frame of business activity descriptions, which can be weighted by occurrence, to give the building blocks of the new version.

(d) The use of bridging codes permits simultaneous coding to multiple classification systems, essential if periods of dual-coding are required.

ANNEX

Analysis of the impact of introducing ACTR (initial results)

1. The table below shows movements in and out of NACE Rev. 1.1 sections (equivalent to SIC 2003 at this level). The counts are of local units, for which business descriptions are held and could be automatically coded by ACTR. The current stock reflects classifications derived from the PDC, and the new stock reflects the position according to ACTR.

NACE Rev 1.1 Section	Current Stock	In	Out	New Stock	In %	Out %	Stock change%	Churn%
A	29157	548	1031	28674	1.88%	3.54%	-1.66%	5.42%
B	891	44	32	903	4.94%	3.59%	1.35%	8.53%
C	1569	39	255	1353	2.49%	16.25%	-13.77%	18.74%
D	50511	9301	7953	51859	18.41%	15.75%	2.67%	34.16%
E	1507	331	274	1564	21.96%	18.18%	3.78%	40.15%
F	117972	4437	3654	118755	3.76%	3.10%	0.66%	6.86%
G	210659	8739	9410	209988	4.15%	4.47%	-0.32%	8.62%
H	70868	4627	1965	73530	6.53%	2.77%	3.76%	9.30%
I	40472	2835	1551	41756	7.00%	3.83%	3.17%	10.84%
J	33367	2823	1010	35180	8.46%	3.03%	5.43%	11.49%
K	165346	8187	10152	163381	4.95%	6.14%	-1.19%	11.09%
L	22862	2742	4186	21418	11.99%	18.31%	-6.32%	30.30%
M	43101	1482	3565	41018	3.44%	8.27%	-4.83%	11.71%
N	82316	4697	6027	80986	5.71%	7.32%	-1.62%	13.03%
O	68848	6885	6652	69081	10.00%	9.66%	0.34%	19.66%
P	0	0	0	0				
Q	7	7	7	7	100.00%	100.00%	0.00%	200.00%
Total	939453	57724	57724	939453	6.14%	6.14%	0.00%	12.29%

Key to NACE Rev 1.1 sections

A Agriculture, hunting and forestry
B Fishing
C Mining and quarrying
D Manufacturing
E Electricity, gas and water supply
F Construction
G Wholesale and retail trade; repair of motor vehicles, motorcycles and personal and household goods
H Hotels and restaurants
I Transport, storage and communication
J Financial intermediation
K Real estate, renting and business activities

L Public administration and defence; compulsory social security
M Education
N Health and social work
O Other community, social and personal service activities
P Private households employing staff and undifferentiated production activities of households for own use
Q Extra-territorial organisations and bodies
