

**Improving Data Quality:
Application of Computer Assisted Coding (CAC) and
Automatic Coding (AC) for Occupation and Industry
Codes for Survey/Census of the Department of Statistics
Malaysia**

Masrol Hafizal Ismail

Abstract

Coding is one of the main activities in censuses or surveys such as census of population, census of businesses, labour force surveys and other socio-economic surveys. This paper attempts to study the application of Computer Assisted Coding (CAC) and Automatic Coding for Occupation and Industry Codes for the Department of Statistics, Malaysia (DOSM). The paper highlights the development of DOSM's CAC which includes defining the 'ideal system'; research and development in developing the prototype of the System Computer Assisted Coding (SCAC); issues and limitations; challenges and future direction of CAC.

Introduction

Coding is one of the main activities in censuses or surveys such as census of population, census of businesses, labour force surveys and other socio-economic surveys. Manual coding has been burdensome in surveys as well as censuses with its impact on timeliness and quality of data. Within the Department of Statistics, Malaysia (DOSM), Computer Assisted Coding (CAC) is not new where, at a smaller scale, a few state offices have taken the initiative to develop their own CAC for occupation and industry codes, using Lotus 123 or Microsoft Access.

Realizing the importance of an alternative to manual coding, the Research and Development Division (R&D) has embarked on a project to develop a standard computer assisted coding

system of occupation and industry (SCAC) for DOSM. In 2003, working together with the Information Technology Division, a prototype model using Visual Basic and Microsoft Access 2000 was developed. Version 1 of the SCAC was distributed to subject matter divisions and state offices in 2004. This paper highlights the development of SCAC in DOSM, which includes defining the 'ideal system'; issues; challenges and recommendations for the future direction of CAC.

What is computer assisted coding?

Various terms are used to describe such system, which include automated coding, documentation and computer generated coding. As for DOSM's SCAC, it is defined as stand-alone computer software that generates related codes for occupations and industries.

In principle, the objective of CAC system is to divide the tasks of coding between the computer and the coder on the basis of comparative advantage. The computer is adept at exact matching of the responses by searching large files quickly and accurately. The display on the computer screen will enable the coder to make interpretation or judgement as to when a match is close enough to be acceptable. The CAC system is able to minimize the amount of interpretation that is required, provides as much guidance as possible and still leaves some space for the coder to make his final decision. For this reason, coders will still need appropriate experience and training¹.

Literature review

Advances in computer technology have resulted in computer applications going a step further. Nowadays, many statistical organizations are either doing research on automated coding or using some type of automated or computer assisted

¹ Report of The Workshop on Computer Assisted Coding, ESCAPE New Zealand, 17-21 April 1989

coding². The experience of three statistical agencies in implementing AC or CAC is outlined as follows:

Statistics Canada

Statistics Canada is constantly testing programs that will automate repetitive and routine tasks³. For example, the Statistics Canada Labour Force Survey data files are collected from the Regional Offices of Statistics Canada and are run through an automated coding system that assigns industry and occupation codes based on the Standard Industrial Classification (SIC) System and the Standard Occupation Classification System. The advantages of the automated coding system are that the process has become increasingly faster, consistent and more economical.

Reference to the SIC System, was first introduced in 1978⁴. The procedure involved two steps: the first (known as the Keyword system) is a screening of descriptions based on the presence of certain key words or part of words to detect descriptions which might not be correctly coded in the second step. The second step is a simple exact matching of descriptions of SIC code assignment against the description or code combinations in the library of descriptions coded in the previous month surveys. If a description requiring SIC code assignment exactly matched one of the descriptions in the library, then the code for the matching description is applied.

² Lyberg, L., and Kasprzyk, D. (1997). "Some Aspects of Post-Survey Processing," in L. Lyberg, P. Biemer, and M. Collins (eds.), *Survey Measurement and Process Quality*, A Wiley - Interscience Publication, United State, pp. 359

³ Statistics Canada. (2003). *Statistics: Power from Data*, Retrieved from the World Wide Web: <http://www.statcan.ca/english/edu/power/ch3/intro.htm>

⁴ Rowland, J.H., and Kinack, M.D. (1994). *A Comparison of Two Methods of Automated Industry Coding*, Statistics Canada, Retrieved from the World Wide Web: http://www.amstat.org/sections/srms/Proceedings/papers/1994_197.pdf

In 1986, the automated coding system was enhanced with the use of Hellerman algorithm. The Hellerman matching routine searches a file of previously coded descriptions for one, which best matches the description requiring coding. Both the coded description and the respondent's description are standardized such that trivial words have been removed and suffixes taken off words left in the description. The matching is performed using a formula based on common words and heuristic weights of the words. The heuristic weight of a word measures how specific a word is to a code, i.e., if a word only appears in descriptions with the same code, this word would have high heuristic weight; conversely if a word appears in a number of descriptions all with different codes, the word would have a low heuristic weight. The phrase with the highest score will be the one which best matches the respondent's description. In practice, before a phrase is used to assign a final code, it must exceed a threshold score and must be at least a certain percentage higher than the phrase with the next highest score⁴.

Automatic Coding by Text Recognition (ACTR)⁵ is developed by Statistics Canada based on the Hellerman algorithm⁶. The characteristics of ACTR are as follows⁷:

- a. Automatic and interactive coding;
- b. Consist of coding engine and maintenance tools; customer builds and tunes the coding index;
- c. Generic: Can code a range of classifications;
- d. Flexible: Allows different coding strategies, and thresholds;
- e. Has (API) and has been ported to UNIX/Windows;

⁵ Technically, ACTR can accommodate any language. Statistics Canada intends this system to perform in English and French, but currently only English is functional. In DOSM, research on two languages, i.e., English and Malay for CAC is being done

⁶ See also Hellerman, E. (1982), *Overview of the Hellerman I&O Coding System*, draft memo, US Bureau of the Census

⁷ Swier, N. (2004). *ONS Classification Coding Tools Project*, Occupation Classification Workshop RSS, London

- f. Multiple coding database;
- g. Dependent coding using filters; and
- h. Powerful parsing capabilities.

US Bureau of the Census

At the Bureau of Census, ACTR's coding abilities are being tested to determine its applications.

Australia Bureau of Statistics

Australia Bureau of Statistics (ABS) has taken an effort to introduce CAC system, changing the way it codes industry, occupation and status in employment data in Labour Force Survey (LFS). These changes were introduced in the LFS February 2000 survey.

From February 2000, respondent's descriptions of their occupation will be assigned an appropriate Australian Standard Classification Of Occupation (ASCO) unit group code using a CAC system. The same system is also adopted for industry coding, and used in the National Census. The CAC system is expected to result in more consistent coding than the current method⁸.

Coding of occupation and industry

Occupation code

In Malaysia, the Malaysia Standard Classification of Occupation (MASCO 1998) is being used for occupational coding scheme for the Labour Force Survey, Household Income Survey and Population and Housing Census. The MASCO 1998 was published in 1998 by the Ministry of Human Resources, Malaysia and, this edition is an adaptation

⁸ Australian Bureau of Statistics. (2002). *Forthcoming Changes: Industry, Occupation and Status in Employment Data*, Retrieved from the World Wide Web: <http://www.abs.gov.au>

of the International Standard Classification of Occupations 1988 (ISCO-88) published in 1990 by the International Labour Office, Geneva⁹. The occupational classification structure has four main levels, namely, major groups, sub-major groups, minor groups and unit groups.

The system is hierarchical, for example, unit group 2122 Statistician is in minor groups 212 Mathematicians, Statisticians and related professionals, which is in sub-major group 21 Physical, mathematical and engineering science professionals and falls into major group 2 Professional.

2 Professional

21 Physical, mathematical and engineering science professionals

212 Mathematicians, Statisticians and related professionals

2122 Statistician

The following example depicts the finer details of several occupations:

Group	Title	Code
Major Group	Clerical Workers	4
Sub-major Group	Office Clerks	41
Minor Groups	Secretaries and keyboard-operating clerks	411
Unit Groups	Secretaries And Verbatim Reporters	4111
	Confidential secretary	41111
	Personal assistant	41111
	Private secretary	41111
	Secretary	41111
	Secretary, confidential	41111
	Secretary, word processing	41111

⁹ Manpower Department, Ministry of Human Resources Malaysia. (1998). *Malaysia Standard Classification of Occupation 1998 (MASCO 1998)*, Kuala Lumpur

Industrial code

The Malaysia Standard Industrial Classification 2000 (MSIC 2000) serves as a common framework for the collection and compilation of industry statistics from different sources, thus facilitating comparison of data. The MSIC 2000 conforms closely to the International Standard Industrial Classification (ISIC Rev. 3), published by the United Nations, thus enabling international comparisons¹⁰.

In principle, the MSIC 2000 provides four hierarchical levels. Each level is constructed and defined in terms of the categories of the next lower level that is each Division comprises Groups, which in turn consists of Classes where each subsequent class is made up of several items.

The first three levels, namely Division, Group and Class conform closely to the ISIC so that data compiled for these levels are internationally comparable. In the construction of the fourth (lowest) level, the underlying principle adopted is that the item should reflect, as realistically as possible, the way in which activities are actually organized within establishments. This is also to ensure that the items reflect recognizable segments of the Malaysian economy and that data classified according to the MSIC are as relevant as possible for analytical purposes.

The system is hierarchical. An example given is the information of the five digit codes.

Level	Title and description	Codes
Division	Fishing, Operation of fish hatcheries and fish farms; Service activities incidental to fishing	05

¹⁰ Department of Statistics, Malaysia. (2000). *Malaysia Standard Industrial Classification 2000 (MSIC 2000)*, Kuala Lumpur

Level	Title and description	Codes
Group	Fishing and Service Activities Incidental to Fishing	050
Class	Fishing, operations of fish hatcheries and fish farms; service activities incidental to fishing	0500
Item	Ocean and coastal fishing	05001
	Inland fishing	05002
	Aquaculture	05003
	Fishing	05009

Coding operations

Coding of industry and occupation is the process of transforming textual description of industry and occupation of respondent into code numbers of the industry and occupation¹¹.

Paper and Pencil Interview (PAPI)

For the Paper and Pencil Interview (PAPI), industry and occupation codes, are generally the only two items in economic and household surveys such as the Labour Force Survey, Household Income Survey and Population Census that are not pre-coded. Therefore, one has to proceed into the next process where a special coding operation is required.

In DOSM, coding operations require experienced and trained coders, who on the basis of the survey responses e.g. textual

¹¹ Hossmanns R., Mehran F. and Vijay Verma. (1990). *Survey of economically active population, employment, unemployment and underemployment: An ILO manual on concepts and methods*, International Labour Office Geneva.

descriptions of industry and occupation and other ancillary information given on the questionnaire, will assign for each item (industry and occupation) numeric codes from the corresponding coding index and record the assigned code in designated place of the questionnaire.

The coding operation also involves query to be resolved by supervisors. It may, in addition involve a formal verification procedure by expert coders and on an occasional basis, an in-depth evaluation using quality control procedures.

Computer Assisted Personal Interview (CAPI)

As a result of the rapid advances in computer technology, several national statistical offices (NSOs) have dedicated some of their efforts towards optimising computer technology in data collection and processing processes. Computer Assisted Personal Interviews (CAPI) is a tool where an interviewer would key in responses from their respondent or interviewee directly into a purpose-built computer program on a small device or laptop.

CAPI allows the interviewer to search through a large database of responses and select an appropriate code during the interview. The interviewer will show the respondent the related description for occupation and industry description.

Several findings indicate that data quality is improved due to a reduction in data entry and consistency errors. The data collected is more likely to be an accurate reflection of the interview as the data entry and editing stages are carried out in the presence of the data source e.g. the respondent.

Criteria of an “ideal system” for CAC for industry and occupation code

The criteria of an ideal system are needed to provide a guideline for first stage development and also as a basis for further study. It also means that certain characteristics will

provide some indications of basic requirements of prototype CAC.

Base on the paper 'Computer Assisted Occupation Coding' written by Diane Bushnell¹², an ideal system should include as many of the following characteristics as possible.

The system should:

a. Be easy to use

Users need a minimum of special instructions

b. Display exact matches first

If an exact match for the entered job title is found in the coding index then this match should always be displayed at the top of the suggestion list.

c. Be at least as reliable and accurate as manual coding

As one of the main aims of introducing the computer assisted occupation coding is to improve reliability and accuracy of occupation coding this criterion is a very important one. We know that one of the problems with computer assisted coding in general is that users favour codes at the top of the list and may select the top code even it is clearly inappropriate. Thus it is important that, where an exact match does not exist in the index, the agreement between the code presented at the top of the suggestion list and the 'best' code is high.

¹² Bushnell, D. (1995). *Computer Assisted Occupation Coding*, Office of Population Censuses and Surveys, UK, Retrieved from the World Wide Web: <http://www.blaiserusers.org/lbocpdfs/1995-1998/busne95.pdf>

d. Be fast

The process of assigning a code should be at least as fast as manual coding using a proper index.

e. Be able to pick up information from the interview

The system should be able to pick up information, which has already been collected during the interview. This is useful for two reasons: coding may not always take place at the exact moment that the information is collected, e.g. initially the system will be tested by interviewers coding at home so information from the interview will be passed to the coding system at a later stage of the questionnaire; several pieces of information may be added together and used for coding so these would need to be picked up after all the data have been collected.

f. Have a simple mechanism for creating and updating the coding frame

Although the Occupation and Industry Code does not change very often, it is essential that minor amendments can be made easily. The process of constructing the code frame should be fairly straightforward and need no special technical expertise.

g. Be able to restrict search by auxiliary fields

It would be useful if the codes suggested could be restricted to a subsection of the coding frame by filtering on information other than the job title. For example, if the respondent were self-employed, then only valid codes for self-employed occupations would be displayed initially (although it would be necessary to have the option to extend the search to show all codes).

h. Include the possibility of displaying supplementary information on the occupation

It would also be useful if information pertinent to particular codes could be displayed. For example, synonyms for job titles (gaffer, foreman, supervisor etc.) or summary of typical tasks carried out by jobholders in those occupations.

i. Allow interviews to take advantages of hierarchal structure of the coding frame

Job titles will be assigned a code and this code will have a general name, known as the unit group heading. This heading can be used as a check that the correct code has been assigned. It might also be useful for the interviews to step through the hierarchy, as a check on accuracy and also to increase awareness of the coding frame structure.

j. Include the possibility of automated and/or semi-automated coding

Automated coding can be defined as a system, which assigns a code to the job title without any interaction with the coder. Semi-automated coding, in this context, is defined as a system where coding is automatic if there is one match with a goodness-of-fit above a certain cut-off point but where the coder assists with coding if the fit is not good enough. Both types of coding would obviously reduce the burden on users/interviewers and increase reliability, but bias and accuracy may suffer unless the system is very good.

k. Include the option of excluding selected words from the matching process

To improve the performance of the matching algorithms, it would be useful to exclude some words from the matching process. For example, if the job title contains the word 'and' this will introduce superfluous information into the match.

l. Be cost effective

The cost of developing and using a computer assisted coding system is obviously one of the key factors in our final decision making process. However, it is also one of the most difficult criterion to evaluate and so far, the cost implications of the various systems have not been explicitly analysed.

Development of DOSM's SCAC

Since the traditional method like PAPI is widely used for census and economic surveys in DOSM, we have decided to build a **stand-alone** CAC system for occupation and industry codes. The three possibilities are as follows:

- a. PAPI - The SCAC will assist the coder in the office/home on the coding and editing process.
- b. CAPI – The SCAC is built in, and attached to CAPI. The SCAC will assist field enumerator to search through a large database, and select an appropriate code during the interview.
- c. PAPI and CAPI – The SCAC will support the two methods and should be user-friendly.

For the Department of Statistics, Malaysia, the system will provide an alternative for users to select the language that they prefer most, either the Malay or English Language.

The prototype model of DOSM's SCAC

The prototype model of DOSM's SCAC has the main characteristics of the actual system. The prototype was distributed to all state offices and subject matter divisions in December 2004. Having the prototype model, it is expected that the following objectives could be achieved:

- a. Able to assist users and system builder (IT Division) to evaluate the completeness of the overall application.
- b. Able to identify any weaknesses and find the solutions.
- c. Ensure the development of an ideal system¹³.

Process of developing DOSM's SCAC

Figure 1 shows the process of developing DOSM's SCAC prototype system. Several sources, with a combination of ideas from a previous study and research from major users (Subject Matter Divisions) were used to create the prototype system.

The manual code of MSIC 2000 and MASCO 1998¹⁴ were analysed to identify the actions need to be taken.

Research for additional sources, especially on findings, discussions, paperworks, journals and papers of workshops, survey was conducted by staff of DOSM.

The research confirmed the importance of developing the prototype SCAC system chart. It was drawn-up so as to facilitate the Information Technology Division to develop the prototype system.

¹³ Refer to criteria of an "ideal system" for CAC for industry and occupation code, page 66-70

¹⁴ See also Occupation Coding and Industrial Coding, in page 62 and 64

Figure 1: Development of the prototype model of DOSM's SCAC

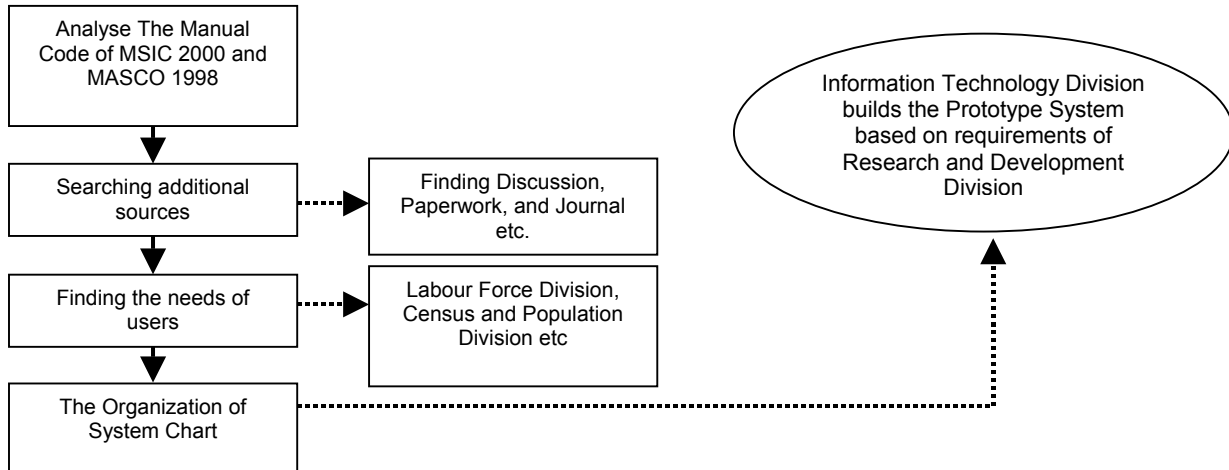
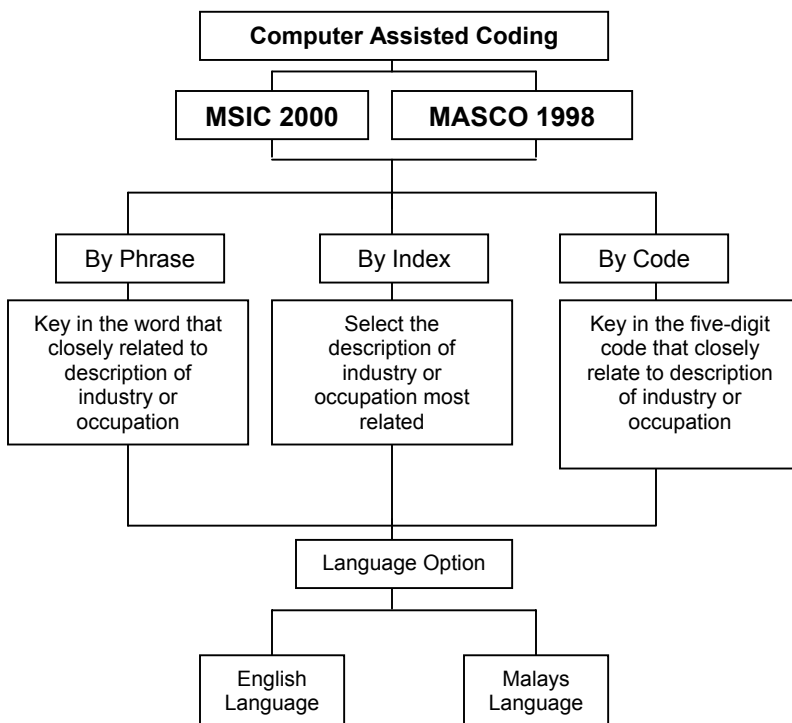


Figure 2: The System Organization Chart



The characteristics of prototype system

a. Major code

The prototype consists of two major coding systems e.g. Malaysia Standard Industrial Classification 2000 (MSIC 2000) and Malaysia Standard Classification of Occupation (MASCO 1998).

b. Language option

In term of language, there are two options i.e. Malay and the English Language.

c. Type of searching

Users select three types of search: by phrase, by index (with by alphabet) and by code. Most users of prototype prefer searching by phrase. Search by code allows users to search from one digit to the detailed 5 digit code.

d. Software

The prototype was developed based on Microsoft Access and Visual Basic.

Figure 3: Main menu of CAC prototype

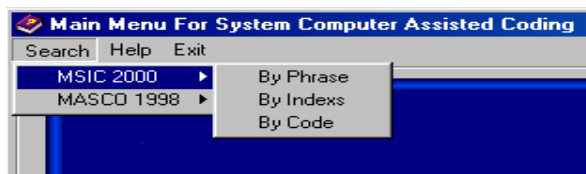
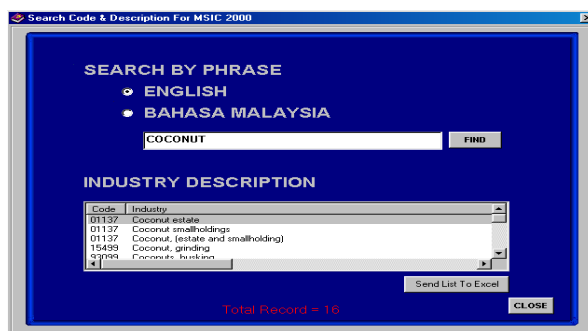


Figure 4: Template for search by phrase



Issues

Although the development of automated coding and computer assisted coding (CAC) has advantages and benefits in term of practical applications, several issues with respect to its limitations need to be resolved.

a. Manual code - experienced coder versus new coder

The experienced coder prefers manual coding, especially when detailed description (from respondent) is needed for five-digit coding.

For the new coder, the simple description in SCAC will help them to code up to five-digit. However, due to changes in the economic structure some of the descriptions have to be referred to the manual code. If a description is not available or no record was found in the SCAC, the new coder might conclude that the item code does not exist, where as it could be found in the manual. SCAC could be improved by enabling it to recognize synonyms or words with approximately the same meaning. For example, coconut in Malay Language is “kelapa” and its synonym is “Nyior”.

b. System

SCAC is a prototype system, which is still in the process of development and testing. CAC system in most countries have gone through several phases of development with the final product able to provide uniform or standardised coding operation compared to manual coding. Problems do exist when the manual code does not give a complete description and a unique code to SCAC database. Issues of completeness of manual code will have an impact on quality of the SCAC.

c. Facilities requirement

In recent years, DOSM has also embarked on improving its facilities e.g. computer hardwares and softwares. Maximising usage of computers is a dynamic action towards SCAC development and implementation.

d. Other issues

Lars Lyberg and Daniel Kasprzyk¹⁵ stated that issues needed to be observed more closely include the differences in quality between cases coded by exact match, inexact match, CAC, and pure manual; the effects of coding errors on data analysis; quality and cost issues related to dictionary size and dictionary contents; and better articulated decision rules to help program areas determine their best strategy to implement, that is, what is the combination of approaches that yields the best data at lowest cost.

The challenge

The high expectation of users is a big challenge to the R&D Division in DOSM to develop an effective and efficient system. Most of the users thought that coding jobs would be easily solved by the system. Statistics Canada¹⁶ with advanced solution of Automated Coding still provides space in the questionnaire for manual coding.

CAC was introduced¹⁷ with the intention of reducing the burden on coders and improve timeliness of industry and occupation coding. This is the real challenge in creating 'ideal system' for the Department of Statistics, Malaysia.

¹⁵ Lyberg, L., and Kasprzyk, D. (1997). "Some Aspects of Post-Survey Processing," in L. Lyberg, P. Biemer, and M. Collins (eds.), *Survey Measurement and Process Quality*, A Wiley - Interscience Publication, United State, pp. 353

¹⁶ Refer to Issues, pg 75

¹⁷ In July 2003, suggestion is to introduce CAC in Census 2010 due to time consuming on manual coding.

Recommendations

The following is suggested:

- a. A specific division or unit in DOSM concentrates in updating and standardising the industry and occupation codes and works closely with other related subject matter divisions.
- b. Statisticians should be computer literate and have to be aggressive in developing softwares and work closely with Computer Analysts.
- c. The use of computers for data capture should be optimised. It should be noted that sufficient number of computers should be provided to coders.
- d. Several economic surveys in DOSM used IMPS 41 Software, a freeware supplied by the US Bureau of Census. Recently, the US Bureau of Census introduces a free software called CSPRO 2.5. It maximizes the usage of Computer Assisted Coding for data capture. Usage of CSPRO 2.5 will complement the development of SCAC.
- e. For the next revision of MSIC 2000 and MASCO 98, it is suggested that CAC will be developed so as to optimise its usage.

Conclusion

In the future, the system will be able not only to match a simple job or industry description but also give a hint to the coder perhaps in terms of weighting like in the ACTR tool. Further study on SCAC is recommended and it consists of two parts, as follows:

Analyse information / study from several agencies that have developed Computer Assisted Coding; and

- b. Analyse the Prototype CAC for Department of Statistics Malaysia.

In the short-term, it is expected that the code lists and description of industries and occupations be updated. This is to ensure that the CAC database is complete and up-to-date. For the long term, it is suggested that the above immediate steps should be taken to implement the above recommendations.

Acknowledgement

The views expressed in this paper are those of the author and do not necessarily reflect the opinion of the Department of Statistics Malaysia. The writer acknowledges the support and cooperation of the Information Technology Division for creating the prototype CAC and Ms Nazaria Baharudin for helpful comments on an earlier version of the paper.

References

- Australian Bureau of Statistics. (2002). *Forthcoming Changes: Industry, Occupation and Status in Employment Data*, Retrieved from the World Wide Web: <http://www.abs.gov.au>.
- Bushnell, D. (1995). *Computer Assisted Occupation Coding*, Office of Population Censuses and Surveys, UK, Retrieved from the World Wide Web: <http://www.blaiserusers.org/lbocpdfs/1995-1998/busne95.pdf>.
- Department of Statistics, Malaysia. (2000). *Malaysia Standard Industrial Classification 2000 (MSIC 2000)*, Kuala Lumpur.

- Economic and Social Commission for Asia and the Pacific (ESCAPE). (1989). *Report of The Workshop on Computer Assisted Coding*, New Zealand.
- Hossmanns R., Mehran F. and Vijay Verma. (1990). *Survey of economically active population, employment, unemployment and underemployment: An ILO manual on concepts and methods*, International Labour Office Geneva.
- Lyberg, L., and Kasprzyk, D. (1997). "Some Aspects of Post-Survey Processing," in L. Lyberg, P. Biemer, and M. Collins (eds.), *Survey Measurement and Process Quality*, A Wiley - Interscience Publication, United State, pp. 353-370.
- Manpower Department, Ministry of Human Resources Malaysia. (1998). *Malaysia Standard Classification of Occupation 1998 (MASCO 1998)*, Kuala Lumpur.
- Rowland, J.H., and Kinack, M.D. (1994). *A Comparison of Two Methods of Automated Industry Coding*, Statistics Canada, Retrieved from the World Wide Web:http://www.amstat.org/sections/srms/Proceedings/papers/1994_197.pdf.
- Statistics Canada. (2003). *Statistics: Power from Data*, Retrieved from the World Wide Web: <http://www.statcan.ca/english/edu/power/ch3/intro.htm>
- Swier, N. (2004). *ONS Classification Coding Tools Project*, Occupation Classification Workshop RSS, London.