

Classifications *Newsletter*

United Nations Statistics Division (UNSD)

Number 21
March 2008

Implementation on the horizon - ISIC Rev.4 and CPC Ver.2 are discussed in ESCWA region

The United Nations Statistics Division (UNSD) and the Economic and Social Commission for Western Asia (ESCWA) held a Workshop on International Economic Classifications in Cairo, Egypt, on 10-13 December 2007. The organization of this Workshop was carried out in cooperation with the Egyptian Central Agency for Public Mobilization and Statistics (CAPMAS) and the Gulf Organization for Industrial Consulting (GOIC).

Workshop participants came from countries in the ESCWA region and the Maghreb, mainly from national statistical offices and Ministries of Industry. A total of 57 people attended the Workshop.

The primary objective of the Workshop was to build on the existing understanding of the newly revised activity and product classifications, ISIC Rev. 4 and CPC Ver. 2, to provide assistance regarding their adoption to countries in the region and to foster regional cooperation and harmonization in classification matters. As such, this workshop built upon a previous classification workshop held in Beirut in 2004.

The deliberations of the Workshop were wide-ranging in addressing the topic of implementation, spanning such diverse issues as the regulatory and legal conditions needed for applying the new classifications and specific detail on the most effective strategies for numbering classification categories. A considerable part of the time was dedicated to the conduct of working groups where country and regional participants had the opportunity to discuss a number of classification-related issues and share experiences directly among themselves. One session of the Workshop was dedicated to presenting the current status of the revision of the International Recommendations for Industrial Statistics which has subsequently been approved by the United Nations Statistical Commission at its session in February/March 2008.

A large portion of the Workshop was used to introduce the final structure of the revised ISIC and CPC to participants and to highlight new concepts that have been used and guidelines that have been followed in the construction of the new classifications. A better understanding of this background information will help countries during the development of their own national versions of the classifications, making them consistent with the principles applied in ISIC Rev.4 and CPC Ver.2.

As part of the discussion on application issues, the treatment of outsourcing in ISIC was one of the topics addressed, since a consistent treatment of this phenomenon in the classifications is becoming more and more important with the growing share of outsourcing, especially outsourcing across borders. While the treatment in the case of services industries is rather straightforward, difficulties arise particularly in the case of outsourced manufacturing activities. The treatment of these cases in ISIC is aligned with the treatment of goods for processing in the System of National Accounts and the Balance of Payments. As such, the determination in ISIC is made based on ownership of input materials. The Principal in such an outsourcing arrangement is classified to Manufacturing if and only if he owns the input materials. In other cases, the activity of the Principal is essentially that of a wholesaler/retailer and he would be classified to Section G of ISIC Rev.4 (unless other activities carried out in the same unit would override this). The CPC provides the appropriate products associated with the manufacturing services provided and the goods produced. Specific emphasis was placed on the fact that, in practice, care should be taken to record the production of the resulting goods only once to avoid double-counting, which is particularly difficult to track if outsourcing happens across borders.

With regard to the CPC, the new concepts relating to the information economy were examined in addition to concentrating on overall changes to the scope, structure and detail in the classification. By way of implementation tools, correspondences were requested between the CPC and other classifications, namely ISIC, the Extended Balance of Payments

March 2008

(EBOPS) and the World Trade Organization's GNS/W/120 list. It was pointed out that these links already exist and plans are afoot to update and maintain them, even though some of them are by nature not as precise as might be desired, owing to the differences in scope and purpose of the classifications being linked.

With regard to the sequencing of implementation activities, the Workshop recognized the challenge of addressing implementation of the classifications contemporaneously with the implementation of the revised System of National Accounts. A particular question was raised in this context, regarding the combined influence on data by changes in classification and changes in concepts and methodology in the SNA. How could changes in time series be allocated/interpreted back to changes in these two systems to properly measure changes in economic performance of countries? This is an issue that will be brought to the attention of the revision process of the SNA, but will also be addressed in guidelines for backcasting of data that will be released as part of materials to support the ISIC implementation.

Other practicalities were discussed, such as the timeline recommended by the Expert Group on International Economic and Social Classifications for the implementation process at national level. In this regard, the most immediate advice to country participants was to aim to effect their national adaptation of ISIC by 2009.

Finally, the discussions of the Workshop resulted in recommendations for future actions for the countries in the region as well as for the United Nations. On the part of participating countries, these included: setting up of necessary laws and regulations to facilitate implementation of the classifications; establishing specialized units on classifications to facilitate the implementation of the classifications and coordination among relevant bodies within a country; encouraging a system of regional cooperation; drawing up timetables for applying the classifications and training of employees in statistical offices and ministries. Those for the United Nations included: organizing specialized training workshops to help in applying the revised economic classifications; making the international classifications available in electronic format via the UNSD website and providing necessary implementation material in Arabic.

All documents discussed at the workshop are available on the UNSD Classifications website, at <http://unstats.un.org/unsd/class/> under "Training and Workshops"

Adoption of the International Standard Classification of Occupations, 2008 (ISCO-08)

The work of the International Labour Office (ILO) on updating and improving the 1988 version of the International Standard Classification of Occupations (ISCO-88) continued throughout 2007. A draft classification structure, circulated in February 2007, was discussed at meetings of the UN Statistical Commission, the UN Expert Group on International Economic and Social Classifications and the ILO Technical Expert Group for Updating the International Standard Classification of Occupations. Based on feedback from these groups and on comments provided by countries and other interested parties, the ILO prepared a revised draft for consideration by an ILO tripartite Meeting of Experts on Labour Statistics, convened in December 2007. This meeting passed a resolution adopting the draft classification structure, with minor modifications, as the International Standard Classification of Occupations, 2008 (ISCO-08). This resolution was endorsed in March 2008 by the ILO Governing Body, confirming the status of the new classification as an international statistical standard. ISCO-08 was also presented to the Statistical Commission at its thirty-ninth session February/March 2008. At that session, the Commission also took note of the plans to implement the classification.

The ISCO-08 structure is available on the ILO Website in the three official languages of the ILO, that is in English, French and Spanish, at <http://www.ilo.org/public/english/bureau/stat/isco/index.htm>.

Definitions of categories, an updated index of occupations, and a correspondence table between ISCO-88 and ISCO-08 will be released on the Website during 2008. Draft definitions of categories are available in English for comment on the ILO Website. The new classification, with full documentation, will be published in book form in English, French and Spanish as soon as possible following release on the Website. All inquiries should be addressed to ISCO@ilo.org.

Industry Coding of Business Units

*John Crysdale and Michael Pedersen
Standards Division, Statistics Canada*

Industry coding of business units at Statistics Canada is an ongoing and multi-faceted activity centered on the Business Register (BR) and supported by Standards Division. This note describes the roles of the BR and Standards Division in coding business units, the coding tools used by coders, the coding process itself, and finally, methods of quality control for the coding.

Business Register

At Statistics Canada, the BR is the source of the frame used for business surveys. A “survey frame” is essentially a list of all businesses from which business surveys can draw their sample. The BR is a database containing information on the basic characteristics of all 2.3 million businesses operating in Canada. These basic characteristics include employment, revenue, expenses, assets, geographic location, type of legal entity, as well as the 6-digit North American Industry Classification System (NAICS) industry code.

The Canada Revenue Agency (CRA) is the BR’s primary data source for determining the population of businesses in Canada. CRA administers tax laws for the Government of Canada and for most provinces and territories. One of CRA’s responsibilities is to obtain a description of the business’s activity as part of the process of registering a new business. In order to reduce respondent burden, CRA shares basic information on each business unit with Statistics Canada. Approximately 25,000 records per month are received by Statistics Canada, from CRA, for coding to NAICS.

Standards Division

Industry classifications, concordance tables, coding tools, and a database with detailed business activity descriptions are developed and maintained by Standards Division. Rulings on industry coding are also provided. The work is done in collaboration with other national and international statistical agencies, with other divisions within Statistics Canada, and with outside users. Industry classifications include the various Canadian versions of NAICS, the different versions of the Standard Industrial Classification (SIC), the American and Mexican versions of NAICS, the International Standard Industrial Classification (ISIC) and the

Statistical Classification of Economic Activities in the European Community (NACE).

Coding Tools

There are a number of different tools used to perform industry coding or to assist in the coding process. A batch coding tool is used to code large numbers of business units using an automated process. There are two tools used for batch coding.

The first of these tools is called the “Direct Match”. Here, the descriptive phrase provided by CRA, for the unit to be coded, is matched against a large database of business activity descriptions and associated NAICS codes.

The second tool is referred to as the “Indirect Match”. In this case, and unlike the Direct Match, the descriptive phrase is parsed. Parsing eliminates extraneous text and reduces the phrase to its most significant components. Parsing is performed using a package known as “Automated Coding by Text Recognition” (ACTR). Parsed phrases are then matched against a database of business activity descriptions and associated NAICS codes.

There are also two interactive tools used in the industry coding process. These tools are used for units that cannot be automatically coded or in cases where there is only one or a small number of units to be coded.

The first of these tools is the Industry Classification Coding System (ICCS). With this tool, the user enters a short description of the activity they would like to code. A list of possibilities is returned and the user makes a selection from the possibilities. The tool is essentially an electronic version of the industry classification manual but with an expanded database of business activity descriptions, the ability to code to many industry classifications, and a facility for coders to submit queries to Standards Division. This tool is typically used in situations where the coder is trying to assign a code without having the business owner or an employee of the business on hand to answer questions. The user of this tool requires a reasonable knowledge of the industry classification in order to make an assignment.

Another interactive tool is the Decision Tree Coding Tool. This tool comprises a structured series of questions and answers that lead to a standardized business activity description to which a unique 6-digit

NAICS class is assigned. The advantage of the Decision Tree is that someone who is not necessarily knowledgeable about the industry classification, but who is knowledgeable about the activities of the unit being coded, can easily and accurately assign an industry code to that unit.

Typically, this tool is used when a representative of the business is available in person or over the telephone. A series of questions and answers can then occur. For example:

Q: What is the main activity of this business? **A:** It is a restaurant.

Q: Does it have table service? **A:** Yes, it does.

Q: Does it serve alcohol? **A:** Yes, it does.

There are two other sources to facilitate industry coding. The first resource is the classification manual itself – in hardcopy or electronic formats. A special attraction of the electronic manual is the availability of powerful search tools to facilitate the coding process. The second resource is the considerable expertise – accumulated over the course of many years of work – of the classification specialists in Standards Division. This expertise is made available to all those who would call on it, both within Statistics Canada and outside.

Industry Coding Process

The industry coding process can be said to start at CRA, where it centers on Business Number (BN) registration. Businesses must have a BN in order to operate in Canada. During the application process, business units must supply a description of their major business activity that allows them to be NAICS coded.

Businesses applying at a CRA office on a walk-in basis, and those which apply by telephone are coded through the Decision Tree. Those applying for BNs through other means — such as on the Internet — bypass the Decision Tree, and supply a free-text description of their business activity. In the event that there is any follow-up by CRA, the free-text description is revised through the Decision Tree. Units that are audited are also coded via the Decision Tree. The key information — the BN and the business activity description — is then transferred from CRA to Statistics Canada.

At Statistics Canada, the files are first run through the Direct Match process to deal with all cases having an exact match to a phrase database. Business units with descriptions that match are

assigned a NAICS code, the BR is updated and the unit is removed from the coding process.

Business units with descriptions that do not match exactly go to the Indirect Match process. This involves text parsing and assignment of a numeric score based on the degree of a match. Exact matches of parsed phrases and matches scoring above a critical threshold are assigned a NAICS code, the BR is updated, and these units are removed from the coding process.

For the remaining units, manual coding is performed. This involves using the ICCS. The reference file used in the ICCS comprises all the business activity descriptors found in the classification manual, as well as a large number of additional variants on those descriptions. Coders may also find supplemental information on business units by checking business directories, doing Internet searches and making phone calls to respondents. This almost always leads to the assignment of a NAICS code. However, in a small number of cases, no assignment can be made. In percentage terms, the Direct Match accounts for 50% of coding, the Indirect Match accounts for 10%, and manual coding accounts for the remaining 40%.

Quality Control Methods

A number of activities at Statistics Canada serve to validate/update the industry assignments stored on the Business Register. These include regular profiling of large businesses, and their business units, by Statistics Canada specialists, as well as ongoing feedback from subject-matter areas. Such subject-matter feedback can occur before a survey is sent out (based, for example, on a preliminary telephone contact by the collection area) or it can occur afterwards based on results received, including respondent identification of changed activities on actual questionnaires.

In addition, quality control on coding is performed by Standards Division and Business Register staff, working in cooperation. There are two types of quality control. One type involves experts looking at a sample of all cases where NAICS codes have been changed within a given time period. The other type of quality control involves experts looking at a sample of all cases regardless of whether or not they have been changed. For such an assessment, respondents are re-contacted by Statistics Canada staff. In this

context, the Decision Tree plays an important role. The advantages of the Decision Tree are that verification can be performed in a fashion that is potentially different from that in which the NAICS code was originally assigned and – perhaps more importantly – the verification can be performed by persons who are not expert in the NAICS classification.

Conclusion

The industry coding process at Statistics Canada is multi-faceted and changing. A large number of business units are coded each month in order to support Statistics Canada's survey activities. Many different tools are now in use and new ones are being developed to handle the many challenges posed by the need for high-volume, high-quality data collection.

Using an Automated Industry Coding Application for New Business Registrations in Australia

Michael Meagher

Australian Bureau of Statistics

Introduction

The Australian Bureau of Statistics (ABS) uses the Australian Business Register (ABR) as its primary source of information to identify new businesses. The information on new businesses flows through to the ABS Business Register which is used as a register or frame for business surveys run by the ABS.

During the process of registering with the Australian Tax Office (ATO), businesses are requested to identify their main industry from a pick list and also to provide their main business activity as a free text written response. The pick list of industries (equivalent of ISIC sections), is based on the industry divisions as defined by the Australian and New Zealand Industry Classification (ANZSIC). This information is then used by the ATO to code reporting units to their relevant ANZSIC Classes.

The automated coding process developed by the ABS in consultation with the ATO was developed to help improve on the quality of the ANZSIC coding and reduce manual coding and the time required for processing information received from clients. The automated coding application, or Autocoder, was implemented for automated ANZSIC coding of

employing businesses in 2004 and for non-employing businesses in 2006.

The Autocoder uses an 'exact word' matching algorithm through an index file with codes which are regularly updated. This process ensures better control of the links between codes and descriptions. However it should be noted that continued investigation is occurring in the ABS into using information retrieval technology for searching to improve match rates, as significant savings can be made at an organisational level.

Processing the information

The Autocoder uses both the main activity and main industry information from the Australian Business Number (ABN) registration form. The information from the registration forms is converted into a short description and a 3-character abbreviation code. For example; main activity description 'green olive grower' and main industry 'Agriculture' is formatted to the following text string: 'green olive grower; agr'.

The Autocoder matches the activity description (or part of it) against an index file of activity descriptions and assigns a four digit ANZSIC code. Where industry descriptions cannot be matched they are given a dummy ANZSIC code. Units with the dummy code are manually ANZSIC coded by the ATO staff.

Parsing rules

The first stage of the auto coding process involves running the activity description across parsing rules. Parsing rules enable the activity description to be changed into a more usable description to be employed by the Autocoder, thus improving the probability of the entry being matched to an ANZSIC code.

The parsing rules change the following in an activity description:

- Fix commonly known misspelt words (e.g. 'servise' to 'service')
- Remove trivial words (e.g. the, as, Ltd, Pty)
- Convert words to their synonym (e.g. 'breeder to farming' and 'grower to farmer')
- Convert symbols to words (e.g. '+' to 'and')
- Reword specific phrases to a more common description (e.g. 'repairing of automobile' to 'repairing automobile').

For example the activity description 'green olive grower; agr' would be changed to 'green olive farming; agr'. The process avoids the need for multiple entries of similar activity descriptions to be included in the index file of the activity descriptions.

Index file of activity descriptions

The index file contains word string entries of activity descriptions and is based on the primary activity descriptions in the ANZSIC publication. However, to increase the match rate, entries are added based on knowledge gained about how businesses describe their activities. The index file also includes entries for common activity descriptions which cannot be auto coded by giving them a dummy ANZSIC. This helps minimise mis-coding by the Autocoder.

Furthermore, activity descriptions can have a number of words therefore entries are formatted into coding levels to ensure that the correct ANZSIC code is applied to an activity description. The index entries are broken up into 3 coding levels consisting of: Basic words; Qualifying words; and Distinctor phrases. The 'Basic word' is the word on which the coder undertakes the first search using the main activity description field. While the 'Qualifying word' is the word on which the coder undertakes the second search using a combination of main activity and main industry fields. The Distinctor phrase is the last word on which the coder carries out the search and uses the main activity field only. Examples of these index entries are: 'Basic word'- tomato; 'Qualifying word'- agriculture; and 'Distinctor phrase'- greenhouse.

The entries mentioned above are formatted and ordered in the same pattern to ensure that the Autocoder program can logically and systematically apply the processes for matching activity descriptions against the index file of activity descriptions.

Reducing multi matches

Multi matches can occur when the Autocoder matches two or more basic words in an activity description with entries in the index file with different ANZSIC codes. For example, the basic words 'printing paper' matches to the index entry '1510 Printing paper, manufacturing'. However the Autocoder can also match the basic word 'printing' to the index entry '1611 Printing, manufacturing'.

Metadata is used to avoid common multi matches by identifying certain basic words (or combinations of basic words) and prioritising them to enable common combinations of activity descriptions to be coded.

Prioritising words assists the Autocoder to identify the words 'printing paper' together to have a higher weight and worth more than the words individually. Therefore the Autocoder will now only match the activity description 'printing paper manufacturing' to the index entry '1510 Printing paper, manufacturing' and will ignore the word 'printing' in the matching process.

The quality of the Autocoder

The maintenance of the quality and the accuracy of the index and associated coding files are important to the ABS, therefore quality checks are regularly performed on these index entries and files. In 2007, the Autocoder coded approximately 50% of new business registrations to a quality level of approximately 97%.

The ABS seeks to maintain a high quality ANZSIC coding by the Autocoder by undertaking 12 monthly data updates. Regular updates to the Autocoder files are necessary to reflect current ANZSIC coding methodology; increase match rates; allow for efficient coding and ensure that the quality of the coding does not deteriorate.

Editorial note

The Classifications Newsletter summarizes recent developments in the field of international classifications, announces upcoming events and draws attention to the availability of relevant classifications material in print and on the Internet. The Classifications Newsletter can be found at the United Nations Classifications Website <http://unstats.un.org/unsd/class> under "Newsletter".

To receive the Classifications Newsletter by e-mail, you can sign up for our Newsletter mailing list at <http://unstats.un.org/unsd/cr/registry/regmaillist.asp> or select "Mailing list" from the menu at the Classifications Website. For further information please contact the Classifications Hotline:

United Nations Statistics Division
Attn: Economic Statistics and Classifications Section
United Nations
New York, NY 10017, USA

E-mail address: chl@un.org
Fax: +1 212 963 1374

Classifications Website:
<http://unstats.un.org/unsd/class>