

TABLE OF CONTENTS

PART FOUR: DATA PROCESSING, TABULATION, ANALYSIS AND DISSEMINATION.....167

IX. DATA PROCESSING, TABULATION, ANALYSIS AND DISSEMINATION.....	169
A. INTRODUCTION.....	169
B. DATA PROCESSING ASPECTS FOR PARTICULAR TOPICS.....	171
1. Economic activity status.....	171
2. Status in employment.....	172
3. Occupation.....	172
4. Place of work.....	173
5. Industry.....	173
6. Institutional sector.....	174
7. Informal sector.....	174
8. Informal employment.....	175
9. Working time.....	175
10. Income.....	175
C. DATA QUALITY AND VERIFICATION.....	176
D. DISSEMINATION.....	176
1. Printed publications.....	177
2. Electronic dissemination.....	177
3. Microdata.....	178
4. Confidentiality and data security.....	178
5. Graphical and related output.....	178
E. METADATA.....	178
F. EMERGING TECHNOLOGIES FOR DATA PROCESSING AND OUTPUT.....	179
G. DATA CAPTURE AND CODING METHODS.....	179
H. OUTPUT.....	180

PART FIVE: CODING OF OCCUPATION AND INDUSTRY183

X. PREPARATIONS FOR CODING OCCUPATION AND INDUSTRY.....	185
A. OBJECTIVES.....	185
B. STRATEGIC CODING AND PROCESSING OPTIONS.....	186
1. Process all cases or a sample only.....	186
2. Field or office coding?.....	187
3. Level of coding.....	190
4. Coding of vague and difficult responses.....	191
C. PLANNING AND ORGANIZING CODING OPERATIONS.....	193
1. Finance and resources.....	193
2. Expertise, experience and rehearsal.....	193
3. Estimating coding rates.....	194
4. Coding staff.....	194
5. Coding teams and supervisors.....	195
6. Coding tools.....	196
7. Coding problems and queries.....	197
8. Quality assessment and quality control.....	197
9. Premises, infrastructure and equipment.....	198
10. Processing in one location or in several.....	198
11. Handling of documents.....	198
12. Use of automatic or computer-assisted coding.....	199
13. The problem of different languages.....	201
XI. THE DEVELOPMENT AND USE OF CODING INDICES.....	203
A. WHAT IS A CODING INDEX?.....	203
B. DEVELOPING AND UPDATING THE OCCUPATION CODING INDEX.....	205

1.	Sources of information for construction of the index.....	205
2.	Organization and structure of the index	207
C.	USING THE OCCUPATION CODING INDEX.....	209
1.	Using the occupational response.....	209
2.	Using ancillary information on industry or name and type of employer.....	210
3.	Use of other ancillary information for coding of occupation.....	211
4.	Inadequate occupation responses and queries.....	212
D.	DEVELOPING AND UPDATING THE INDUSTRY CODING INDEX	213
1.	Types of industry coding index.....	213
2.	Lists or registers of establishments	214
3.	Indices reflecting the answers given in response to industry questions	215
E.	USING THE INDUSTRY CODING INDICES.....	217
1.	Using the industry response	217
2.	Using ancillary information on occupation.....	218
3.	Use of other ancillary information for industry coding.....	219
4.	Inadequate industry responses and queries	219

PART FOUR

**DATA PROCESSING, TABULATION, ANALYSIS
AND DISSEMINATION**

IX. DATA PROCESSING, TABULATION, ANALYSIS AND DISSEMINATION

A. INTRODUCTION

484. Data processing relates to the activities involved in converting census responses into a computer-readable data file of records that have data that are as free of errors as possible and that are ready for tabulation and further analysis.

485. Unless advanced technologies are used for data collection (for example online Internet completion and return of electronic questionnaires, automated telephone interviewing, hand held devices, and so forth), most countries will have paper questionnaires that must be converted by key entry and/or optical recognition into computer data files for analysis and tabulation. The data entry process also applies to other census topics and is covered in *Principles and Recommendations for Population and Housing Censuses, Revision 2*. Consequently, those topics are not mentioned further in the present chapter.

486. When pre-coded responses are not given for a particular question, clerical, computer-assisted or automated coding may also be required. That issue relates especially to occupation and industry and is covered in chapters X and XI.

487. The classifications used in each census should, to the extent possible, be comparable with previous censuses in order to facilitate continuity and comparability over time. Similar considerations should apply for comparison with other sources of data.

488. Owing to the size and complexity of census operations, it is likely that errors of one kind or another may arise at any stage of the census. The present chapter outlines specific points that will help to identify and correct as many of those errors as possible (data editing⁶⁵ and amendment), the derivation of selected data from census responses (for example, whether a person is employed in the informal sector), and the output phase (tabulation and dissemination of census results).

489. Every national census organization should establish a system of quality assurance and improvement as an integral part of its census programme. The data processing to detect and correct errors is only part of that programme. There is no single standard quality control and improvement system that can be applied to all censuses or even to all steps within a census. Census designers and administrators must keep in mind that no matter how much effort is expended, complete coverage and accuracy in the census data are unattainable goals. However, efforts to first detect and then to control errors should be at a level that is sufficient to produce

⁶⁵ *Principles and Recommendations, Revision 2*, para. 1.311 refers to critical edits that have the potential to block further processing as well as to non-critical edits that do not have that potential. Some responses are impossible (outside valid ranges or totally inconsistent with other answers, such as data recorded on economic activity when the person is below the minimum age). Some statisticians also refer to fatal edits (impossible responses) as opposed to query edits (answers that need to be checked or confirmed, but which may be possible but unlikely).

data of adequate quality for the main uses, while offering the opportunity to significantly reduce costs and at the same time improving the quality of the census within the constraints of the budget and time allotted.

490. Automated data capture, repair and coding systems (see chap. IX below) increase data quality greatly and also introduce a different set of risks compared with traditional census processing approaches. If not properly monitored and managed, data quality problems can remain undetected until late in the process when cost and timing constraints limit the options for any corrective activity.

491. Some methods of measuring data quality from data capture processes, such as substitution rates or measures of key entry errors, are inadequate, as those forms of monitoring simply measure the overall incidence of errors but not the significance of the errors. Indeed that approach could lead to considerable extra expenditure for the correction of trivial errors that lead to no appreciable gain in quality. For that reason, data quality should be measured at the topic response level rather than at the individual character or numeral level. It should be done in two ways: independently processing a sample of records using manual processes and comparing the results for each of the records with those obtained through the automated systems; and in aggregate by comparing the overall data for an area with the expected results based on other information for that area (for example, from the previous census or other data sources). Where manual or computer-assisted processes are used, it is useful to establish a quality control system in which a sample of records is coded independently by a second operator. Any discrepancies in the results can then be resolved by an expert or experienced operator. That will allow calculation of estimates of error rates as well as the identification of problem areas and cases where re-training may be necessary.

492. The above mentioned process should be undertaken continuously during processing with a focus on early detection of quality problems and an understanding of any systems or processes that have contributed to them. The amount of error that is acceptable and the degree of intervention and systems or process change undertaken will depend on the assessment by the census agency of the overall fitness of purpose of the output and the overall cost and timeliness impacts. This will vary from topic to topic.

493. In chapter II (concerning planning and design for a population census) the need to involve data processing staff at an early stage of census planning was already mentioned, as well as the importance of appropriate form design (to facilitate data entry and with sufficient space for codes to be entered). Chapter II also mentioned the possibility of processing a sample to produce advance results. The present chapter covers more detailed aspects of data processing for particular topics and the tabulation of census results.

B. DATA PROCESSING ASPECTS FOR PARTICULAR TOPICS

1. *Economic activity status*

1.1 *Current and usual economic activity status*

494. The usual activity reference period and the current activity reference period do not necessarily overlap, for example, if the period for usual activity covers to the end of the previous month and the period for current activity is defined as the previous seven days from the enumeration day. In the gap between the two reference periods the person may have changed activity to become, for example, employed, unemployed or retired. More generally, a person's economic activity status during the longer reference period for usual status may be different from that person's current status over the shorter period used for determining their current status. Thus, for those countries collecting both current and usual economic activity status, it is not recommended to force any consistency between those data items during data processing.

495. In addition, most of the not economically active categories can change status from the usual measure to the current measure and consistency in that instance should not be forced either. For example, a person could be a student during all months of the usual measure but be engaged solely in housework or have started work in the seven days of the current measure.

496. If retired or aged is given as a category for the not economically active, it is strongly recommended that a lower age limit be put on the category (such as 45 years of age) for the editing procedures. The limit should not be set too high as some people are retiring or being retrenched from wage employment earlier in recent years. It would be particularly wise to remove persons in the younger age groups from the category, as their inclusion may lead to incongruous results, such as having persons "aged or retired" that are 20 years old.

497. Cross-tabulations of usual by current activity are a useful method to check for large numbers of unusual cases before publication. It may be necessary to go to small area data to find explanations or possible errors of interpretation. For example, a large new factory or a natural disaster may explain odd-looking differences between usual and current activities in particular areas. Such cases are worth explaining in the presentation of the results.

1.2 *Employment*

498. The determination of whether a person is currently employed may depend on the answer to more than one question (especially if there is a separate question to identify those who are temporarily absent).

499. The determination of whether a person should be classified as employed or not is almost always through a derivation process. All possible combinations of answers to the questions on employment should be allowed for. The derivation should be automatic (by computer program) and will also check that the proper sequence of responses has been followed. Inconsistencies between responses to different questions (including questions being answered when they should be skipped over) must be resolved.

500. It is useful to do an age distribution of the categories of the employed. Extreme occurrences, such as employed persons over 80 years of age, and any unusual values at unusual

ages in general should be investigated.

501. Tabulations of the characteristics (age, occupation and so forth) of the temporarily absent should be done as part of evaluation of this category.

1.3 Unemployment

502. Given the wide variations in the reference periods used, especially when two criteria are used in the questions to derive the unemployed, and in the way either one or two of the criteria may be applied, information about the definition and the question(s) used needs to be provided as part of the results. In addition, when questions on both “seeking” and “available for work” are used it would be useful and would facilitate comparisons across countries if some of the tables showed the two components, as well as the resulting estimates when they are combined to identify the “unemployed”.

2. Status in employment

503. It may be necessary to derive the standard status in employment categories from some of the responses given. This is relatively simple with modern tabulation software, and most statisticians and data processing staff are now accustomed to much more complex derivations from survey data.

504. There are no viable consistency checks for the International Classification of Status in Employment categories using age, sex or education level since the categories can be found among employed persons of all ages, educational levels and sex. Some checks (for example, employers aged under 18 years) may demand investigation but may still be correct.

3. Occupation

505. As stated in chapter XI, it is strongly recommended that countries code their occupational data in a way that makes it possible to produce statistics in accordance with the International Standard Classification of Occupations. ILO has provided considerable assistance in the use of occupational classifications, with the major documentation being contained in chapters X and XI below; see also Hoffmann and others (1995) and Hussmanns, Mehran and Verrna (1990).

506. Data consistency edits can be made between occupation groups, industry groups and age and education levels: for example, most persons classified to a “professionals” category will have educational attainment which is higher than the “no education” or “primary school” level.⁶⁶ Care should be taken in the use of such edits, however, so as to avoid the risk of biasing the relationship between such variables as the level of educational attainment and occupation (for example, mismatches between occupations and education levels to study inadequate employment situations). Analysts may be particularly interested in identifying the number of persons who do not have the level of education normally expected for particular occupational groups. Checks are also possible between occupation and sector of employment: for example, “government tax and excise officials” should be in the government sector. Note, however, that in the government sector there will be a large number of jobs given occupation codes other than that for “government

⁶⁶ Designers should be aware that there is a tension between occupation and educational attainment, and that checks should not be overdone.

administrative worker”. Occupational information should be coded at the highest level of detail made possible by the recorded responses (see chap. X for further details).

507. As previously indicated, any unusual or implausible combinations should be investigated.

4. *Place of work*

508. Checks of the consistency between the type of place of work and other selected variables (such as occupation and institutional sector) may identify possible errors to be investigated (for example, doctors or public sector employees operating from a market stall).

509. Similarly, the geographical location of place of work would normally be within reasonable commuting distance of the place of residence, and large discrepancies would need to be explained when disseminating results. An exception concerns commuting workers who work far from home and come home only for the weekend. In that case, the use of the “usual” concept to define place of residence may not be appropriate for some analyses.

510. In respect of tabulations, the main categories of type of place of work are useful information in themselves for labour market analysts. The cross-classification of those location categories by status in employment and industry and occupation will give very valuable information on the structure of employment. That will be the case particularly for the informal sector, and also for the identification of other categories that often are of special interest, such as “paid employee home-workers”. Those target groups should be clearly identified at an early stage of census preparations, and the processing steps necessary to extract information about them should be tested to ensure that the target groups are being correctly identified.

5. *Industry*

511. The coding of industry in population censuses and surveys is covered in some detail in chapters X and XI. There are useful suggestions also in Hussmanns, Mehran and Verma (1990) as well as in other ILO papers and documents. Although industry coding is an established topic, there may be problems with industry data collected in censuses, and the points presented below are particularly noteworthy.

512. The industry description and code should refer to the establishment where the person actually works and not to the legal unit to which the establishment may belong. For example, a major company in a country, “XYZ Ltd”, may be well known but may engage in a variety of activities at different establishments around the country. Generally, industry should be coded for each separate establishment (usually at separate locations), not only for the main activity of XYZ Ltd. When there is access to a reasonably up-to-date register of establishments, and where industry is coded for each unit, the industry question should aim to collect the detailed name and address of the person’s place of work (see para. 338). Then coding can be done by reference to the register and will be compatible with employer surveys based on the same register. The issue is explored further in chapter XI (see paras. 680-681).

513. Unfortunately, few countries have a comprehensive updated register of employers with separate identification of all establishment locations. Thus, in most countries the detailed name and address of the establishment can be used as a source for the correct industry code for only some of the responses, and a description of the activity at the establishment has to be used as the

basis for assigning codes for the rest.

514. However, even if a register of establishments (or of employers) is very poor, it is often useful to have a limited list of major companies and enterprises with their industry codes when creating the coding indices described in chapter X. For example, XYZ Ltd may be dominantly an iron and steel mill but may also have XYZ Coal Mines and XYZ Iron Ore Mines as separate establishments. Entries in the coding index for industry reflecting that structure will help to ensure that respondents working at the separate establishments are given their relevant industry codes and not simply coded to the code given to XYZ Ltd. A similar listing of major government departments and organizations is also very useful in coding respondents working for the government to their correct industry codes: for example, to code those working with the country's health administration separately from those working in government hospitals.

515. There are a limited number of consistency checks of industry codes with the results from other questions. They include checks that jobs in establishments coded to industry classes related to public administration are in the government sector (but note that many government sector workers will work in other industry groups), and that only certain occupations are associated with the industry category for "activities of households as employers of domestic personnel". Other edits are generally not definite and depend on the situation in a given country. Particular attention should be paid to any codes identifying separately "production for home consumption", including subsistence agriculture.

516. Informal sector derivations should be carefully checked to ensure that only relevant industries are shown. Some limited checks are possible of industry against status in employment: for example, self-employment cannot occur for "activities of households as employers of domestic personnel", or for "industry classes related to public administration". A full range of edits should be prepared before data processing commences, but cross-tabulations with other topics should be done early and progressively to locate improbable occurrences that will need explanation or possibly further edits. It is recommended that this not be left to the final tabulation phase since the resolution of such queries can delay the results considerably.

6. *Institutional sector*

517. If the question on institutional sector is combined with the status in employment question(s) as presented in some of the examples shown previously, it will not be necessary to do a consistency check between status in employment and institutional sector (for example, government sector workers being classified as self-employed or unpaid helpers). However, in cases where both status in employment and institutional sector data are collected separately, it is important that they be crosschecked. There are no possible checks against sex, age or education level. Checks against the industry and occupation classifications are also possible as mentioned above.

7. *Informal sector*

518. The design of a procedure to measure employment in the informal sector should be clearly thought out and well tested. Some computer derivation may be necessary and will need very careful consideration and prior testing.

519. Careful checks and cross-tabulations of status in employment categories by industry and

occupation groups will be essential to edit any informal sector data, and re-coding is likely to be necessary to remove absurdities.

520. If the derivation procedure is properly designed, then consistency checks with the institutional sector will not be needed; otherwise consistency checks will be required to ensure that there is no public sector or company employment in the informal sector.

521. Such preparations will prove to have been well worth the extra effort when the demands for small area and regional estimates of informal sector employment are received.

8. *Informal employment*

522. The procedure to identify those in informal employment should also be carefully designed. There are no viable consistency checks since those in informal employment may be in any occupation, industry or institutional sector. If information is collected to derive whether a person is in the informal sector, then it would be useful to check that all who are engaged in the informal sector are also classified as informally employed.

9. *Working time*

523. If a person is recorded as temporarily absent from his or her main job then the actual hours recorded for that main job should be zero. If a person has more than one job reported then working time should be calculated for all jobs for both hours actually worked and usual hours, if the data was so collected. If hours actually worked are recorded separately for all jobs then it would be reasonable to expect that usual hours would also be recorded separately for all jobs.

524. Consideration should be given to compiling statistics of average hours actually worked as well as distributions of the employed in five-hour groups of hours actually worked. In presenting the distributions of employment by working time, care should be taken to ensure that the distribution shows both low hours and excessive hours.

10. *Income*

525. When income is collected for the same reference period as the measurement of economic activity status, then only persons classified as employed should have income from employment. Unemployed and inactive persons may receive other types of income, but not income from employment. Persons with secondary paid or self-employment jobs should have income from the second activity as well.

526. In cases where the reference period for the measurement of income differs from the reference period for the measurement of economic activity status, then no viable consistency checks are suggested.

527. Total household income should always be equal to or more than the sum of the individual job-related incomes in the household. Codes for "not stated" must be allowed for this question in particular, with a clearly defined strategy on how to deal with such cases in the processing of the question, for example, by imputation or by exclusion from averages.

528. Extremely low and high incomes should be checked since they can distort averages. Those checks can be done by status in employment and by occupation.

C. DATA QUALITY AND VERIFICATION

529. Before results are disseminated, it is important to ensure that economic activity estimates are checked and confirmed. That requires an evaluation of the data quality and comparisons with other data sources, where available. As previously stated, such evaluation and verification will be facilitated if concepts, definitions, reference periods, scope (including age cut-off) and the like are comparable between sources. Post-enumeration surveys are also an important part of the evaluation (see also *Principles and Recommendations, Revision 2*, paras. 1.274-1.277 and 1.379-1.400.) Quality assessment techniques or strategies include checking coefficients of variation, the involvement of subject-matter experts in the evaluation of data and the production of test tables at various geographical levels to identify where additional checks may be required. Such preliminary tables should be prepared and evaluated early in the output-generation phase. If significant concerns about data quality are identified, it may be necessary to consider reprocessing to detect and amend unforeseen errors and re-running the early tables.

530. The disseminated results should include a summary of the evaluation, an assessment of the data quality and an explanation for users on the differences between estimates from different sources.

D. DISSEMINATION

531. A census is not complete until the information collected is made available to users in a form, and to a timetable, suited to their needs. Thus in disseminating the results of the census much emphasis should be put on responsiveness to users and on high standards of quality in the production of statistics. A major strength of the census is that it supplies statistics for small areas and small population groups, and the tabulation and dissemination systems should make full use of that strength by generating detailed tables of results.

532. Annex II of the *Handbook* lists the tables relating to economic characteristics that are recommended in annex II of *Principles and Recommendations for Population and Housing Censuses, Revision 2*. That annex also includes other recommended or additional tables of interest, for example on the economic activity of persons with disabilities and on the economic activity of the foreign-born population.

533. The tables in annex II of the *Handbook* have generally been recommended for production for national, regional and provincial levels with an urban/rural dissection, and even for production at the locality level. Some have been prepared in considerable detail (for example, those cross-classified by occupation and industry), and therefore it would be more appropriate to release the more detailed tables electronically (see below) or in publications and abstracts that relate to only those small geographical areas.

534. The large size of census files can affect the timeliness of the results. It may not be necessary to process and tabulate all census results before dissemination. Rather, the data can be released in stages, so that some census results are released before others. If that is done, it is suggested that the results on economic characteristics be released at a time close to the time when other related topics, such as education, are released.

535. As indicated in *Principles and Recommendations, Revision 2* (paras. 1.208 to 1.209), there are several ways of making the results of a census available to the user:

- (a) As printed reports containing standard and pre-agreed tabulations, usually at the national, regional or local district area level, that may be obtained from government agencies or directly from booksellers;
- (b) As unpublished reports (often referred to as abstracts) comprising standard tables but produced for either smaller geographical areas or population sub-groups not otherwise included in the published reports; they may often be requested by users, who may have to contribute towards a proportion of the marginal costs of their production;
- (c) As commissioned output produced from a database, comprising customized cross-tabulations of variables not otherwise available from standard reports or abstracts;
- (d) As microdata, usually available in restricted format only and supplied under strictly controlled conditions.

536. A range of products and output media should be available to meet the changing requirements of users. Major users should be consulted in advance to determine their needs. Users will also need supplementary metadata covering definitions, classifications, and coverage and quality assessments.

537. Each of the different forms of output is presented below.

1. Printed publications

538. Owing to production costs, printed publications may eventually play a secondary role in the dissemination of the main census results, though paper still provides a medium that does not readily deteriorate and does not require the user to have any necessary hardware, software or technical skills.

539. It is suggested that only the basic and recommended set of tables appear in the printed publications that are prepared for broad user and community use, and that even then those tables do not need to include statistics for small geographical areas. National, regional and provincial statistics would generally suffice.

540. Those printed publications (and output via the Internet, CD-ROM and other media) should include graphs, maps and textual analysis as appropriate.

2. Electronic dissemination

541. The release of some outputs (especially of detailed cross-classified results and results for small population groups, and some of the optimum set of tables) may be possible only by distribution through the use of high-capacity electronic media. However, when data are provided in electronic form, special attention should be given to providing users with easy means of data retrieval.

542. The outputs and relevant metadata should be accessible in standard formats as well as in common database and spreadsheet format for easy retrieval and manipulation.

3. *Microdata*

543. National statistical agencies should also consider the possibility of releasing data files of unidentifiable unit records so that advanced users can undertake more detailed analysis. Such releases should be subject to careful controls on data security and confidentiality. For more information, see “Principles and guidelines for managing statistical confidentiality and microdata access” (United Nations, 2007).

4. *Confidentiality and data security*

544. The computer systems handling census data should have strict safeguards to prevent unauthorized access to the information (see also United Nations, 2007, as mentioned above). Care should be exercised to avoid the inadvertent disclosure of information (for example occupations or income) in respect of identifiable individuals through the statistical results of the census. Special precautions may apply particularly to statistical output for small areas.

545. Measures to ensure disclosure control may include some, or all, of the following procedures:

- (a) Swapping of some records between households of similar demographic characteristics within the same higher geographical area;
- (b) Restricting the number of output categories into which a variable may be classified, such as aggregated age groups rather than single years of age;
- (c) Where the number of people or households in an area falls below a minimum threshold, suppressing statistical output—except, perhaps, for basic head counts—or amalgamating with that for a sufficiently large enough neighbouring area;
- (d) Randomly modifying or rounding data before the statistics are released;
- (e) In the case of microdata or public use samples, removing all information from databases relating to name, address and any unique characteristics that might permit the identification of individual respondents.

5. *Graphical and related output*

546. Products should be developed that allow statistical and geographical information to be delivered together with geographical information systems to meet as widespread an interest and with as much flexibility as possible, commensurate with assurances on confidentiality. When databases are provided with associated graphing and mapping capabilities, their usefulness will be greatly increased.

547. Ideally users should, themselves, be able to generate graphs and/or maps easily, and then to print or plot them or make the images available for other users. Several census organizations have produced such a product, sometimes in cooperation with a commercial company.

548. Printed and electronic output should include summary graphs and maps.

E. METADATA

549. The provision of adequate metadata is essential to allow users to understand the meaning

of statistical output. A metadata system contains, in particular, definitions of terms and variables, classification schemes and an assessment of the quality of the estimates. For variables for which international standard classifications have been used, those classifications may be included in the metadata system. For variables that are not classified by international standards, the definitions, concepts and classification schemes actually used in the census should appear in the metadata system.

550. A metadata system may provide supplementary information on characteristics of surveyed and published data. Such information is essential for users to understand the strengths and limitations of the statistics and to use the data properly in more detailed analysis.

551. The population census for the 2010 round of censuses has to ensure comparability with the data from the previous population censuses, to the extent possible. Nevertheless, it should include new elements relevant to development that has taken place during the time since the previous census. Thus, the metadata for the population census around 2010 should follow the metadata system from the previous population census with an update in line with the needs resulting from development since the previous population census. With the rapid changes in information technology that are taking place, the design of the newer metadata system may also need to change.

552. When disseminating the results of the census, the associated metadata should also be disseminated. This disseminated metadata should include an evaluation of the quality of the census results, including at topic level, and an evaluation of its comparability with other sources of statistics on economic activity.

F. EMERGING TECHNOLOGIES FOR DATA PROCESSING AND OUTPUT

553. Technology has been used to assist in all phases of population censuses for many years, and particularly in data processing, analysis and dissemination. Well-established approaches and technologies might well be the most viable option for many countries, and adoption of a new technology or approach should be considered only where there is a sound understanding of the new approaches and technologies and the developments can be managed. There should also be a clear understanding of both the risks and the benefits.

554. The following sections cover only some of the current technologies. For more information, see *Principles and Recommendations for Population and Housing Censuses, Revision 2* (United Nations, 2008) and *Handbook on Census Management for Population and Housing Censuses* (United Nations, 2001). Further technological development is certain and will affect future census taking.

G. DATA CAPTURE AND CODING METHODS

555. Several data capture technologies have traditionally been used, such as key entry and optical mark recognition (OMR). The last decade has seen significant improvements in intelligent character recognition (ICR), data repair, imaging and automated coding technologies that have reduced the cost of census processing and the time it takes to do it. They have also improved data quality. Those trends are likely to continue.

556. For most countries, the most cost-effective option is likely to be a combination of digital

imaging, intelligent character recognition, repair and automated coding whereby the census forms are first processed through scanners to produce an image. Recognition software is used to identify tick box responses and to translate handwritten responses into textual values. Confidence levels are set to determine which responses require further repair or validation. Automated repair typically involves the use of dictionary look-up tables and contextual editing. Operator repair can be undertaken on images not recognized. This is cost-effective only for those questions where there is a high probability that the repaired data can then be automatically coded.

557. Automatic coding uses computerized algorithms to match captured responses against indices. The responses that cannot be matched are then passed to a computer-assisted coding process. That methodology eliminates inconsistencies caused by human error inherent in manual and computer-assisted coding. The quality of the automated capture and coding needs to be carefully monitored during processing to ensure that the system is functioning as specified. However, the process does offer the opportunity to improve the quality of the data and has the potential to reduce processing costs significantly.

558. In computer-assisted coding, a coding clerk typically keys a truncation of selected words from responses written on the form or captured in some other way. Matching index entries are displayed on the computer screen and, when the index entry most closely matching the response is selected, the appropriate classification code may be written automatically to a data file or database.

559. Census agencies also need to consider how the data are going to be held through the processing stream. Traditionally, census processing has been conducted using a flat file that gets progressively updated, with the earlier version of the file retained for backup and recovery processes. Typically that procedure has been allied with batch processing where a discrete group of forms (typically for an enumeration area) are processed together. Thus data from the forms will be entered, edited and coded as a group, allowing a high degree of workload control. Databases allow information to be held and processed at the individual field level. That provides a greater degree of flexibility: once census data are electronically captured, it is possible to maximize both processing efficiency and quality, as similar responses can be readily grouped and coded together. However, holding the census data in a database requires more complex systems to manage and deliver work. Consideration also needs to be given to backup and recovery mechanisms.

560. Systems such as those described above typically require far more extensive development and testing than a traditional census processing system. A number of factors need to be considered throughout the design process and integrated into the development of the system, such as the work organization of the remaining clerical processes.

H. OUTPUT

561. Traditionally, census output has been conceived in terms of generating tabulations, usually for sophisticated clients well acquainted with the census data, its structure and content, and other metadata. Less sophisticated users have traditionally relied on static products, such as publications, that generally contain a limited range of outputs.

562. Internet dissemination allows for much improved service to users by permitting the design of products to meet the needs of different types of census data users, from the novice to the sophisticated; the cost-effective dissemination of a much wider range of census data; and improved usability of the census data.

563. The main strength of a census in a developed statistical system is to complement the information provided by other data collection methods such as surveys with a focus on small domain statistics, that is, information for small geographical areas and for small population groups (both social and economic). Internet dissemination can support both uses of the data. For small geographical areas, Geographic Information System (GIS) technology can be used as means for defining areas of interest in searching for data as well as for mapping of the outputs of the search. A range of packages can be used to focus in on populations of interest from large predefined matrix tables.

564. The Internet dissemination system should provide freedom for clients to specify the form of the output, whether as hard copy or as a data file that can be exported into a range of commonly available statistical analysis, tabulation or mapping packages.

565. Some countries may wish to consider providing access to clients to submit tabulations directly online to be run against the census unit record file. Again, protecting the confidentiality of the census data is a prime consideration in such systems. In addition to implementing confidentiality procedures (such as random rounding), there may be a requirement to limit the size of tabulations that can be submitted through this method.

PART FIVE

**CODING OF OCCUPATION AND
INDUSTRY**

X. Preparations for coding occupation and industry

566. Chapter X explores the various strategies and preparations that are necessary to ensure effective and reliable coding of industry and occupation responses, and also considers the costs and advantages of the different alternatives. It outlines the objectives and main strategic choices to be made and presents the main organizational factors determining the effectiveness and success of the coding and processing tasks. The chapter describes the development of the major tools, the coding indices and how to use them effectively.

567. The basic tools include such international classifications as the International Standard Classification of Occupations (ISCO)⁶⁷ and the International Standard Industrial Classification (ISIC)⁶⁸, which are available on the Internet. Since requirements will vary from country to country, the chapter covers common issues that must be addressed in any census operation when questions related to occupation and industry are included in the census questionnaire.

568. At the time of release of the present manual, the most recent versions of the classifications were ISCO-08 (forthcoming) and ISIC, Revision 4 (United Nations, 2008a), endorsed by the United Nations Statistical Commission at its thirty-seventh session in 2006. In general, it is preferable to adapt the international classifications for national or regional use rather than to use them directly in the collection of national statistics. In many countries such adaptations may already exist (or are under development) for use in economic and labour statistics. Wherever possible such adaptations should also be used for the census. In their economic and labour statistics, some countries may still be using national classifications based on earlier versions of the international standards (for example, ISCO-88 or ISIC, Revision 3.1) and will need to make a decision whether to update their national classifications in time for use in the 2010 round of censuses.

A. OBJECTIVES

569. The main aim of the coding and processing of the information given by the respondents about their place and type of work (about the industry and occupation in which they work) is to determine and record correctly to which of the categories in the respective classifications the jobs belong, at the most detailed level of the classification possible on the basis of the information provided in the response. That task has to be completed within an overall processing plan for the census, to a pre-specified timetable and either within pre-specified cost limits or in a fashion that will minimize cost, given the specified data requirements.

570. The development of a census-processing strategy with the preceding aims needs to consider many aspects and requirements of the processing task, including the following:

- (a) The existing data processing capacity and infrastructure;
- (b) The type and format of the information to be processed;
- (c) The volume of data to be processed and the throughput rates required;

⁶⁷ Available from <http://www.ilo.org/public/english/bureau/stat/isco/index.htm>.

⁶⁸ Available from <http://unstats.un.org/unsd/cr/registry/isic-4.asp>.

- (d) How the processing of the information about industry and occupation is embedded within the total data processing task for the census;
- (e) The level of detail required to satisfy important user needs in the national context, as well as for international reporting.

571. The precise impacts of those aspects will depend on the choices made with respect to some strategic aspects of the coding and processing of census forms, as explained in the following section.

B. STRATEGIC CODING AND PROCESSING OPTIONS

1. *Process all cases or a sample only*

572. Coding of industry and occupation is typically one of the most expensive and time-consuming operations in the processing of a census. To reduce costs, to make the management and quality control of coding easier and to enable earlier production of results, consideration should be given to obtaining and/or coding the information for a sub-sample only.

573. A policy to code a sub-sample only may be implemented at the data collection stage by fielding longer and shorter versions of the census form, so that only a sample of the population answers the questions designed to capture the information needed to code industry and occupation. To avoid error-prone field sampling procedures and administration of different forms or schedules, one may alternatively collect the information from everyone but process it for a sample only. Either form of sampling will significantly, almost proportionately, reduce the operational costs of coding but not the costs of preparing for the coding operation. However, sampling means the introduction of sampling-related errors (imprecision in the resulting estimates), and that will be an important concern when estimates are to be produced for small population groups or small geographical areas. As the provision of small area and small group data on a nationally consistent basis is seen as one of the major functions of the census in most countries, many users may see the processing of industry and occupational information for a sample as contrary to the objective of the census, unless the sample size is sufficient to meet those requirements. That perception may be especially prevalent if there is a regularly conducted, annual or quarterly labour force survey that already provides sample-based national statistics for employment and unemployment, classified by occupation and industry.

574. Other considerations relevant to the coding of a sample or sub-sample are the following:

- (a) The use of a sample for the coding of industry and occupation will mean that sampling imprecision will make comparisons between categories and over time more difficult;
- (b) Sampling may involve a serious loss of descriptive power, particularly for small areas, small industries and small occupational groups;
- (c) It may hinder the production of summary socio-economic indices for small areas or groups, as they are based on the occupation variable.

575. If the decision is made to collect industry and occupation information for only a sample of the population, then attention must be given to the design of the sample, balancing the requirements of statistical precision against those of operational simplicity and robustness. The sampling fraction should in principle be determined by balancing the precision of estimates

required for the smallest aggregates of the population for which separate figures are to be produced against the saving of cost and time resulting from reducing the processing load. For operational convenience, sampling has often been done by selecting all households in a whole enumeration area to be included in the sample, thus sending whole bundles of questionnaires into one processing stream or the other. Such an approach may, of course, result in very high clustering of particular occupations and industries, and will give valid statistics only at high geographical levels, thus contributing to further loss of precision.

576. If the sampling within enumeration areas is done manually it is also important that the sampling procedure be simple (for example, by selecting every n^{th} case, with $n = 5$ for a 20 per cent sample, 10 for a 10 per cent sample and 20 for a 5 per cent sample, as appropriate). The sample selection should be done on a probability basis. In practice, it can be conveniently approximated by taking every n^{th} census household in each data collection area (enumeration district) and including all members of those households in the processing sample. In that case, owing to variable household size and to the clustering of individuals within households, there will be a departure from a true simple random sample of individuals (the units of analysis for household-based occupational statistics). An effort should be made to estimate the sample design effect, but guidance on how it should be done is beyond the scope of the present publication.⁶⁹

2. *Field or office coding?*

577. With respect to census coding, the following choices are available:

- (a) The respondent codes himself/herself to a predefined category;
- (b) The enumerator codes in the field, either during the interview or before the questionnaire is forwarded for further processing;
- (c) Specially trained coders code in connection with consistency checks of the questionnaire and data entry.

578. The choice is a strategic one with respect to the balance between costs, quality of coding and resulting statistics, and control of the coding process.

2.1. *Coding by the respondent*

579. In practice, coding by the respondent means that respondents are requested to place their job in one of a set of predefined categories presented to them in written form on a questionnaire, or on a card read or handed over by the enumerator.

580. The main advantage of this approach is that it is the least expensive of the possible coding procedures. It can easily be adapted for use with optical mark recognition and intelligent character recognition technologies resulting in even greater cost savings. The main disadvantage is that it results in lower quality data, in terms of reliability, validity and detail.

581. The lower level of reliability results from the difficulty of assuring consistency in the ways the respondents relate the pre-defined categories to the job about which they are asked to give information. The content of each pre-defined category has to be described with a limited

⁶⁹ For details on design effects, see for example Leslie Kish (1995).

number of words, normally in the form of a category title, and it may be very difficult to convey to respondents the intended understanding of the ways in which their jobs relate to the different categories. The fact that the number of possible categories to choose from has to be severely limited will also limit the number of possible mistakes to be made. This is of little comfort, however, when many users of industry and occupation statistics need to make much more detailed distinctions and to work with much more homogeneous categories than those obtainable when using this approach.

582. Nevertheless, the cost savings are such that some national statistical agencies have been asking respondents to code in their census, at the expense of one of the main advantages and purposes of the census: the possibility of providing statistics for (relatively) small groups in a consistent manner for the whole country. The quality of self-coded occupation and industry data is so restrictive that it would probably be better to augment the sample size of existing sample surveys as an alternative method for obtaining more detailed occupational data.

2.2. *Coding by the enumerators*

583. Coding by enumerators can take two forms:

- (a) The enumerator assigns the response to a pre-coded alternative during the interview, on the basis of the respondent's answer(s) to the standard questions; or
- (b) The enumerator writes down keywords of the respondent's answer(s) and then codes the response after the interview, but before the questionnaire is forwarded to the processing centre.

584. The cost advantages of the first possibility are almost the same as for respondent coding. The main difference is that the interview is likely to take a little longer because of the need for the enumerator to understand and "translate" the information received to the appropriate category. The procedure may be used if the respondents themselves are not expected to read the questions and write the answers. The effect of the procedure on the validity and usefulness of the resulting statistics is similar to that described in section B.2.1, "Coding by the respondent". On the one hand, the reliability of the coding may improve relative to coding done by the respondent if the enumerator has been well trained and has received more detailed instructions on what types of jobs the different pre-coded categories are supposed to cover and where there are ambiguities that will require probing. On the other hand, the enumerator may misunderstand what the respondent reports and therefore not select the correct category.

585. The advantage of the second approach, in which the enumerator writes down the response that is to be coded later, over the use of categories re-coded by respondents or enumerators, is the possibility that it offers for much more detailed coding and thus for much more valid and useful results. An enumerator coding outside the immediate interview situation may be given complete coding indices to assist with the coding process, as well as other coding aids—including the possibility of conferring with supervisors. Another advantage is that, as the enumerators gain experience with coding, they will become more aware of the type of information that is required to code correctly and improve the recording of those characteristics with interviews. The main disadvantage is that, since they are more numerous and geographically scattered, the enumerators cannot be given the same amount of training, supervision and support as that given to specialized coders, to assure coding reliability. Narrowcasting of satellite transmissions has been successfully used to provide interactive training of census field staff in remote locations and should be considered as a solution to the training of geographically scattered enumerators.

586. A statistical consideration with the second approach is that all coders have irreducible idiosyncrasies that produce biases in the distribution of codes, which they allocate, relative to the mean of all coders. It can be shown that, if biases of similar magnitude are imparted by all coders, whether in the field or in the office, the effect on the overall variance of the results will be less if each of a large number of fieldworkers codes a relatively small number of cases than if each of a relatively small number of office coders code a large number of cases. Depending on the relative sizes of the work quotas of field and office workers, the total variance of estimates with field coding may still be lower even if fieldworkers code in a more variable (less consistent) way than office coders. Note, however, that the above reasoning is valid mainly for national aggregates. For smaller areas the biases of individual field coders may not cancel each other out, and it may be possible to distribute the workload among central coders to prevent clustering of their biases. The best solution is, however, to provide the coders with training and tools, which can minimize systematic errors and biases.

587. It may be possible to save time and reduce costs and operational complications by using field coding in circumstances where industry and occupation are the only items requiring office examination and coding. In that case, the field-coded questionnaires could be passed straight on to computer data entry. However, fieldworkers may be thought unsuitable in terms of background and training to act as coders. In addition, field coding sacrifices to a large extent the important advantages of a controlled and supervised coding environment that can provide direct feedback on coding quality and an incentive for correct coding. Such direct feedback can be provided to office coders in a census through quality control procedures but will be much more difficult to provide to enumerators, as field operations are likely to have been finished by the time serious problems are discovered, and field supervisors will normally not be equipped to give high priority to quality control of coding results.

588. The arguments in favour of using field coding, as compared to specialized coders detached from the data collection process, are closely linked to having a permanent field staff that can be trained cost effectively in coding and can accumulate experience. Field coding may therefore be a realistic option for continuous labour force and similar surveys, as well as for local administrative offices, but the ad hoc and much larger-scale operations like population censuses should normally include the coding of industry and occupation in the central processing operation.

2.3. *Office coding (including computer-assisted coding)*

589. Specialized coders, located in a few processing centres, may specialize entirely in the coding of one variable or in coding in general, and/or they may carry out the coding as one part of an integrated data entry, coding and data control operation for the census. The exact context will depend on the whole organization of the processing operation, but the concerns outlined below should be given serious consideration.

590. With a large coding operation, such as for the census, the coding staff will become thoroughly accustomed to the practical routines of the task, the rates of work required and so on. Specialized coding units may become strongly production-minded and isolated from the rest of the organization, with little incentive to assess their product in terms of external validation criteria. It is therefore necessary to establish rigorous measures of quality control in coding operations. In addition, individual coders may find short-cut methods that reduce the laboriousness of their task but may also incorporate errors or unjustified assumptions, if not actual

violations of the coding instructions.

591. Although it may also occur when coding is done by enumerators in the field, such unofficial departures from, or additions to, the specified coding procedures can be a particular problem in a specialized office coding environment. They may become institutionalized to the point where no distinction is perceived between them and the coding rules derived logically from and designed to support the system of classification in use. Supervision and routine systematic audits of the coders must be employed to ensure compliance with the established coding rules and procedures. More generally, issues of external validity and internal consistency of coding may fall into abeyance unless the routine procedures of the unit include specific and properly designed checks on the levels of coding validity and reliability achieved.

592. Coding may be one element in a larger processing task for each operator. This may make the coders' tasks more interesting and can be a useful method of using employee resources efficiently. An additional argument for this solution is that fewer persons will handle the census forms, and that simplifies the control task. (Note, however, that to simplify the discussion the larger processing context is mostly ignored in the rest of chapter X.)

593. The large-scale coding exercises mounted for national censuses of population generally rely on special recruitment and training of inexperienced coding staff. The performance of staff that are inexperienced may limit or delay the acquisition of the adequate coding habits described. Nevertheless, the coding results are likely to be poor unless recruitment, training and coding procedures are well-planned, well-executed and supervised. A particular problem is that, owing to resource limitations or practical difficulties at a time of heavy stress during the preparations for census processing, there may be insufficient time to establish rules and routines and generally test the coding procedures before production coding has to start. If this happens the organization may quickly become overwhelmed by the sheer volume of documents and data to be handled, with a consequent loss of control and a severe drop in standards, which may prove difficult to recapture and reverse later on.

594. In general, existing coding procedures used in regularly conducted labour force surveys should be adapted for use in the census. This avoids the need to "reinvent the wheel" and helps to ensure that coding is consistent among collections.

3. *Level of coding*

595. The purpose of the coding process is to determine and record, from the responses obtained from the respondent, the code for the category in the classification to which the job of the respondent belongs. The coding can be seen as a translation process, whereby the coder translates the industry and occupation responses into the correct categories; only in a minority of cases will the words of the responses be the same as those used to designate categories in the classifications. The raw materials for the process are the responses, and the tools the coders can use are the coding index, the coding instructions and the persons responsible for answering queries.

596. The coding process should therefore be designed to find and record the most detailed codes supported by the responses. The more information retained after the coding, the more valid and therefore valuable the resulting statistics can be for the users. A compromise option, often overlooked, is coding at different levels within different parts of the structure. However, traditionally, the most common procedure has been to decide that coding should be done at a

particular level of the classification structure, for example the three-digit level, no matter what information has been provided in the response. The arguments for this have commonly been as follows:

- (a) That it would be too costly to code to a larger number of categories, both in terms of coding errors and in terms of staff hours required;
- (b) That the responses would not support coding to more detailed categories;
- (c) When coding only a sample, it would not be possible to publish results for the more detailed categories owing to a lack of observations.

597. However, closer examination of those arguments in the light of the experience gained by statistical agencies has revealed the following:

- (a) The marginal costs of coding to a larger number of categories in the classification, in other words, to a lower level of aggregation, are rather small in terms of increased error rate as well as in terms of work hours needed for coding. The error rate for aggregate categories does not seem to increase. The most significant cost associated with coding to a lower level in the aggregation structure is the associated increase in the required sample size;
- (b) Experience clearly shows that the industry and occupation responses are very uneven in the level of detail they will provide. Many responses will support detailed coding, especially if the questions are formulated along the lines outlined in part three. At the same time, a significant number of responses will not support the level conventionally chosen. By insisting on a predefined level, the coding process may therefore both lead to an unnecessary loss of information for a large part of the returns and to a misrepresentation of the data quality for other parts;
- (c) The similarity criteria used to define the categories in an industry or occupation classification are mostly defined with reference to the nature of the production process and the type of work performed, without much regard for the number of employed persons in the resulting categories. Consequently, the number of jobs that can be found in categories defined at a particular level in the classification may differ greatly. Even in the few cases where statistical balance has been adopted as an important consideration when constructing the classification structure, the number of jobs in a category such as “customer service clerks” defined at an aggregate level may, for example, be smaller than that of a category, such as “shop sales assistants”, defined at a lower level in the structure of the classification, but within another high-level category;
- (d) In addition, the tabulation of industry and occupation statistics may involve both the merging of categories and cross-classification with other variables such as age, sex and region. Even in cases where the data cannot be published, the microdata can be extremely useful for internal analysis. Consequently, one should not restrict tabulation possibilities during the coding process.

4. *Coding of vague and difficult responses*

598. Most industry and occupation classifications specify residual categories of “type x industry/occupations not elsewhere classified (n.e.c.)”. “Not elsewhere classified” categories are designed to take care of activities and jobs that belong to the more aggregate category but that are

not similar enough to the specified subcategories within it to belong in any of them, and which in themselves include too few cases to warrant separately specified categories. Responses should be coded to an n.e.c. category only when they match index items of that category. The not elsewhere classified categories should not be used to code responses that the coders cannot assign to any of the specified categories. Such responses can exhibit the following characteristics:

- (a) Be too vague and imprecise to allow the coder to determine to which category the job belongs;
- (b) Indicate that the establishment (or job) in question produces a combination of goods or services (or the job involves tasks and duties) that cut across the distinctions made in the industry (occupation) classification;
- (c) Represent a type of production or work not covered by the classification.

599. The proper way of handling such responses will depend on the type of case, as follows:

- (a) Vague and imprecise responses should be coded to the level in the aggregation structure supported by the information contained in them; they should not be forced into any particular detailed category where it is likely that only a small proportion of the jobs would fall if one had an adequate response. For example, in one census, 15 per cent of the jobs coded to the major group “clerks” could not be coded to any of the more detailed categories within that major group. It would have represented a significant distortion of the results if they had all been placed in one particular more detailed category together with the jobs that properly belonged to that category. A common method of dealing with this type of response is to provide entries in the coding index for commonly occurring vague responses. Those entries are assigned the code for the relevant higher category, followed by trailing zeros. For example, in Canada the census was coded to the level supported by the responses and then the responses were allocated proportionally to the more detailed categories in a transparent manner;
- (b) The classification of establishments with an uncommon mix of activities or jobs with an uncommon mixture of tasks or duties should, as far as possible, be made on the basis of the general priority rules of the classification. Such responses should preferably be left to expert coders or raised as queries for the classification experts. Disruptions of the coding process can be minimized if the responses are given a special code and the questionnaires put aside for later examination by the experts. The same treatment should be given to responses that seem to represent establishments with activities or functions not covered by the industry classification and jobs with tasks or duties not covered by the occupation classification. The reporting of these types of difficult cases is an important input to the process of updating, maintaining and possibly expanding and revising the relevant classification and coding index.

600. The information to be used as a basis for the coding of industry or occupation is first and foremost that contained in the responses to the respective industry and occupation questions. However, in a significant number of cases where that information is not sufficient for the coder to choose between possible alternatives, ancillary information provided by the responses given to other questions may provide the basis for making the choice. (For occupation the most important type of such information is often that pertaining to the industry of the workplace.) However, it must be emphasized that the use of such ancillary information should be specific and strictly controlled, to avoid an undermining of the descriptive and analytical use of the variables, in

particular when they are being used together or for cross-classification. This means that rules for proper use of such ancillary information—when and how to use it—must be incorporated into the coding indices and the coding instructions.

C. PLANNING AND ORGANIZING CODING OPERATIONS

601. The remarks that follow are based on the assumption that coding is to be carried out by specialist coding staff in the context of the final processing of the census, since that is the solution most commonly chosen for census operations by national statistical agencies. In most cases it is easy to deduce, from the information that follows, the guidelines that should apply to field staff coding or to other decentralized arrangements.

602. Assembling the right resources to process industry and occupation information in the right places at the right time and managing those resources efficiently are fairly complicated tasks. They require anticipation and cooperation between different parts of the census organization and must be coordinated or integrated with the other processing tasks. Large volumes of documents and data have to be handled and, because of the interdependence of different stages within the overall processing plan, the penalties for failures of operational or quality controls, in terms of delays and cost increases, can be heavy. The main outline of the processing plan for a population census, particularly as it affects requirements for finance, staff, equipment and premises, may need to be worked out long before the start of the actual processing. Managerial staff who will be involved in the planning and supervision of the coding operation and professional staff responsible for designing the classification and coding procedures, training coders, updating classifications and coding indices, and interpreting results need to collaborate closely at the planning stage.

1. *Finance and resources*

603. A substantial amount of money is required to support the processing of a census, and it will have to be estimated and provided for under appropriate budgets. Estimates for each part of the processing task often need to be made several years in advance and integrated into the financial planning and procurement procedures of the responsible agency to ensure adequate provision. Such budget considerations demand early decisions about resource requirements, which may in turn precipitate strategic processing decisions that have resource implications (for example, staff numbers and pay rates, number of processing offices and use of computer-assisted techniques). It is important to ensure that financial, resource and operational planning are coordinated, so that the technical assessment of requirements determines bids for resources, rather than vice versa.

2. *Expertise, experience and rehearsal*

604. The coding of industry and occupation demands special expertise in those who plan, manage and supervise the operation. The processing of each census relies heavily on the technical information, expertise and experience gathered at the previous similar exercise. Information and experience from that exercise may be documented in detail, but practical expertise is likely to reside very largely in the heads of a small number of experienced staff. Staffing continuity in key positions is therefore extremely desirable. However, circumstances change, and it is not possible to rely entirely on documented or undocumented experience from the last census. Staff working

with the labour force survey or other related surveys where information about industry and occupation is collected on a regular basis will often have very relevant experience and tools, and they should be consulted. If and when outside advisers are used it is also extremely important to verify the validity of their experience for the local context and to have the opportunity to modify their estimates of costs and time requirements, based on concrete experience. A processing rehearsal is very important as an aid to planning and estimation, and the results of such a rehearsal need to be available early and at a stage when it is still possible to adjust plans for the main operation in the light of them.

3. *Estimating coding rates*

605. Some key planning parameters, for example, coder work rates and effective coding throughput rates, can be estimated reliably only from well-documented previous experience or well-planned tests. Certain problems arising from the scale of the full operation, such as its effect on problems of recruiting, maintaining and controlling staff, may be hard to test in advance. Experience also shows that performance rates vary significantly over time during the processing period. Coding rates are much lower and query rates much higher early in the process than later. There is also the danger that relaxing controls can cause a drop in standards towards the end of the process, beyond what is warranted by the improvements in the coding operation that result from the early improvements in the coding tools used and from on-the-job learning by the coders, supervisors and classification experts.

4. *Coding staff*

606. It is important to make good estimates both of the number of coders required and of the numbers of first-line supervisors needed to control the coding process, as well as of the number of specially trained staff needed to resolve queries.

607. In a census operation the large volume of work to be processed within a limited period will require special staff recruitment and training, both for initial recruitment and for anticipated staff turnover in the course of the task. Thorough enquiries and consultations should be undertaken about likely sources of suitable staff recruits, since financial constraints are likely to prevent actual recruitment until the last moment. There may be external pressures to employ particular groups of persons, even when their suitability cannot be guaranteed. Common criteria should be defined and applied in the selection of all staff. The terms on which staff are to be employed, including the minimum acceptable level of education, pay rates, grading, disciplinary guidelines and rules for hiring and firing, need to be carefully defined. It is important to provide adequate time and resources for staff training at both the coder and supervisor levels, and to recognize that the specialists, who are to advise the supervisors and resolve the more difficult queries, normally cannot, and should not, be recruited and trained for a particular census operation but be part of the permanent competence of the organization.

608. Persons with the following characteristics are best able to perform the tasks of the coder:

- (a) Literate, with an aptitude for the job and a suitable temperament;
- (b) Willing and able to speak up when a problem is detected;
- (c) Clerically accurate and careful;

- (d) Willing and able to follow detailed instructions conscientiously, with an understanding that he or she should not alter or improve upon such instructions prior to consultation with and clearance from supervisors; prepared, however, to raise queries in cases of genuine doubt;
- (e) Honest and trustworthy;
- (f) Persistent and willing to work steadily for long periods;
- (g) Able to work reasonably rapidly and to maintain a steady level of productivity.

609. Persons responsible for recruiting and selecting coding staff should have the above-mentioned characteristics in mind. Several of them, such as (a), (c) and (d), are best assessed through an objective screening test (which may also be applicable to other types of routine clerical tasks). Success at the initial recruiting stage will help to ensure retention of qualified staff and to limit the need for recruitment and training of replacements while production coding is in progress.

610. Proper training of coders is very important. Coding of industry and occupation is best learned through practical instruction in specific procedures, such as document handling routines, use of the coding indices and so on, interspersed with supervised practice on appropriate, specially designed exercises. It is to be expected that trainees will learn at different rates; in this case, retention of what is learned and the ability to stay on task are key attributes. The training period can also be used to identify persons who are the most qualified and able to follow the coding instructions precisely and those who may be better suited to other types of census processing work. It is important to remember that staff recruited to replace coding staff who leave before the end of the coding operation will also need to be trained.

5. *Coding teams and supervisors*

611. Production coding on large jobs is best organized by allocating coders to teams, each under a first line supervisor. The supervisor's role and work tasks need to be carefully specified and are likely to include the following: controlling work flows; monitoring and maintaining work rates; enforcing work discipline; motivating staff; resolving and recording coding queries; applying quality control procedures; and so forth. For first line supervisors, the need to have previous experience in coding operations depends on their role in query resolution. In principle the coding operation may be organized in a way that gives the operational supervisors a very limited role in query resolution. Then it is not essential for them to have had prior experience of industry and occupation coding. However, in most cases it will be preferable to give supervisors the responsibility for first line query resolution owing to their close contact with the coders and the shorter response time and higher capacity relative to classification experts. Supervisors with responsibility for query resolution should be given a good understanding of and training in the classifications and coding systems.

612. The number of coders allocated to each supervisor is important. Typical ratios are between 6 to 1 and 12 to 1, but the appropriate ratio needs to be assessed in each case, taking account of the flow of work that supervisors will be required to manage. Maintaining a reasonable workload among supervisors will help to avoid bottlenecks, improve the reliability of coding and ensure that staff are able to report problems and queries in a timely fashion. It is also important to maintain staff morale and discipline and to sustain productivity rates. Particular problems may arise where coders expect that it will be difficult to find a new job after the end of the coding

operation; in such circumstances, special bonuses may be effective in maintaining productivity rates. It may also be possible to retain to the end of the process only those staff with a potential for more permanent and long-term employment with the statistical service. Their experience will be valuable both for other surveys and for the documentation and explanation of census procedures for the benefit of the users, including those who will be preparing the next census.

6. *Coding tools*

613. It will be necessary to provide appropriate documentation, procedures and training materials not only to coders, to guide the coding process itself, but also to supervisors. The basic tools required by coders include the following:

- (a) **Coding instructions.** They should cover all operations that the coder is required to carry out. The procedures and instructions for handling all relevant items and operations should be integrated. The instructions relating to the coding of industry and occupation will need to be particularly clear and specific with respect to (i) the order in which checking, coding and editing tasks are to be carried out; (ii) the procedure for analysing verbatim material for significant terms; (iii) the use of the coding index or other coding documents; (iv) the circumstances and procedure for using ancillary data; and (v) the circumstances in which to refer a “difficult to code” response to a supervisor for query resolution;
- (b) **Coding index.** This is the key coding document through which verbatim terms incorporated in job titles, descriptions of tasks and the like, are translated into codes. Coders should not be encouraged to interpret verbatim responses in terms of their own conception of the purpose or criteria of classification, but rather to follow in a conscientious way the instructions laid down for consulting the index. For those reasons it is essential that the index be clearly set out, explicit and easy for coders to use. The use of the coding index, instructions and procedures should allow for updating in the light of decisions made in resolving queries and problems that arise and are dealt with in the course of coding;
- (c) **Query resolution procedures.** A query occurs when coding clerks cannot assign a code using the specified procedures and indices. There should be clear instructions on when and how the coders should raise queries, how to record and report them, and their resolution. Queries are the most useful inputs to both immediate and future work to update the coding index and the classification itself. Early in the census processing, it may prove necessary to carry out such updating frequently if the coding indices, or even the classifications, prove to be out of date or inadequate for other reasons. The updating of coding indices should be done by, or in close collaboration with, those responsible for the standard classifications so that consistency between collections can be maintained and any problems with the classification itself can be identified;
- (d) **Legal and administrative forms.** Coders should sign a legally binding undertaking to maintain the confidentiality of census data. Other documents used by both coders and supervisors are likely to include forms for recording queries and their resolution; for controlling the flow of work and reporting progress; for quality monitoring; and so on. To ensure that target throughput rates are achieved, the productivity of coders and coding teams needs to be monitored and progress charts maintained. Special measures for motivating coders should be used, for example, the posting of productivity and error rates for coding teams or for the best individual coders.

7. *Coding problems and queries*

614. A coding query occurs when a coder is unable to assign a classification code to a census or survey response using the standard coding procedures and tools. No matter how carefully coding instructions and the coding indices have been prepared, it can be guaranteed that large numbers of detailed queries will be raised in the course of a major coding operation. This happens if the indices prove to be out of date or incomplete in some respects. Another reason for queries is that actual responses will be more varied than anticipated by the index constructors, even with the most carefully analysed pre-census tests and well-designed questions and instructions to enumerators. Any revision of the structure of the classifications since the last census or survey may also lead to a new crop of problems in the treatment of vague and inadequate responses at the borderlines between categories.

615. Evidence of shortcomings in the documentation that comes to light in the course of production coding will need to be rapidly and consistently addressed and fed back to the coders and their supervisors in the form of amendments to their tools. Appropriate procedures for reporting and recording queries need to be established in advance. In addition, decisions need to be made to resolve them and to incorporate any consequent amendments in the coding documentation and procedures. The roles of supervisors in processing queries and amendments need to be defined, for example, how and when they should communicate with the coding experts, and how new versions of the tools should be distributed to the coders. Particular care is needed in the coordination of query reporting and document amendment where coding is being carried out in several different locations, for example, if different provincial or local offices have responsibility for processing the census returns. All coding sites and teams need to be informed and given updated tools as quickly as possible.

8. *Quality assessment and quality control*

616. To provide adequate information on quality of output, explicit allowances for the resource and time costs of formal quality control, rather than casual observation by supervisors and ad hoc checking of coded output, need to be built into the processing plan. Those costs will cover the establishment and staffing of a quality control unit responsible for acceptance testing of coding during the start-up of the coding operation and for assessment of the reliability and consistency of the operation as a whole. A procedure for sampling and sub-sampling the work of the coders for quality control purposes needs to be defined, and the quality control unit should be staffed at a level that will enable it to keep pace with the main coding operation. Coding schedules must make allowance for corrective action (for example, 100 per cent checking) in the case of batches that fail a quality control test.

617. Built-in quality control procedures will also be required. It is necessary to design separate procedures to handle (a) on-line acceptance testing of coders' work, and (b) overall monitoring and assessment of performance.

618. Coders' performance must meet criterion levels of accuracy in following coding instructions. The aim of acceptance testing is to identify deficiencies in performance rapidly so that corrective measures can be taken. The aim of overall monitoring is to estimate average levels of coding accuracy and inter-coder consistency for the entire coding exercise. Estimates of coding reliability need to be supplemented by estimates of the validity of coding if a balanced overall

assessment of the quality of the statistical output is to be made. Estimates of validity may be obtained from a post-enumeration study in which the whole data collection, coding and editing process is repeated for a sample of census cases. On the basis of such quality assessments it may be possible to separate the contributions made to total variance by source of variance, that is to say, due either to data collection or to coding.

9. *Premises, infrastructure and equipment*

619. The large volume of questionnaires and of work entailed in census processing requires suitable office space and the entire infrastructure necessary for properly supervised clerical operations, as well as for easy movement, storage and retrieval of forms. A special requirement is for security of documents bearing personal details. Proper attention must also be paid to the functionality and capacity of the desks, chairs, shelves, filing cabinets, lighting, heating and ventilation and to the adequacy of paper, pencils and other stationery. Attention to such factors can boost the morale of the staff, prevent high staff turnover and encourage attention to work quality and speed. Suitable premises need to be specified, identified, costed, approved and booked well in advance. If coding staff are to operate with special equipment, such as computer terminals or optical readers, special arrangements may be needed to estimate the requirement, identify suitable and reliable equipment and carry out tests, estimate and provide for capital expenditure and depreciation, provide for replacement in case of breakdowns, go through procurement procedures and so forth.

10. *Processing in one location or in several*

620. Census processing creates a substantial but temporary demand for suitable staff and premises. Recruitment of extra staff and other managerial tasks may be more complicated if all processing is carried out in a single location. There may also be other cost and logistical arguments for carrying out processing at one or several locations other than, or in addition to, the central census processing office. In any case, it needs to be borne in mind that for such relatively complex tasks as census coding and processing, maintaining consistency in coding between coders and coding teams, while difficult, is important. One reason for this is that census coding inevitably generates large numbers of queries. For example, in the 1981 census of the United Kingdom of Great Britain and Northern Ireland, more than 30,000 coding queries were processed, and it is thought that many more were dealt with informally. A large number of queries may lead in some cases to amplifications or changes to coding indices and instructions, which then need to be applied in a consistent fashion (the task is much easier if all documents are in electronic format). The maintenance of consistent production and quality control standards is generally more difficult to achieve when coding is carried out in several locations than if it is centralized in one place. It will be important to establish possibilities for communication by telephone, fax and/or electronic mail, especially if the operations are not centralized.

11. *Handling of documents*

621. The coding of occupation and industry will normally be an integrated part of the total processing of the information on the census questionnaires. In that case the main concerns will be (a) how to receive the forms; (b) how to store them; and (c) how to allocate them to staff so that one can control and ascertain that forms have been processed. It should be easy to find individual forms that for some reason need to be rechecked. If each form has to be handled by more than one person, for example, because different persons carry out the coding of different variables or

because data entry is done by special operators, then the flow of documents must be planned to avoid bottlenecks and loss of forms. All movements of questionnaires from one location, such as a workstation or office, to another and to storage should be carefully recorded.

622. Capturing the responses by intelligent character recognition to create electronic records is an option that has the potential to significantly reduce processing costs and problems caused by the flow of forms from one process to another. A less expensive and technically simpler option is to scan the questionnaires and to work with image files of the documents. When considering the cost of either intelligent character recognition or scanning, it should be noted that an archive is created as part of the process.

12. *Use of automatic or computer-assisted coding*

623. Until very recently most census operations reserved computer usage for “downstream” applications. That is, after data had been collected, manually transposed to data collection sheets and keypunched onto cards or paper tape or entered on magnetic tapes or disks, the machine-readable data would be fed into a computer (usually a large mainframe computer) and verified through a series of computer edit checks. That process would involve passing the data through a computer program to ensure that any out-of-range codes or non-allowable combinations of codes were detected and reported for further investigation. Those types of controls are still valid and important, but computer edit checking cannot detect invalid coding within the range of valid codes if errors in data collection and coding do not generate non-allowable combinations of data. Apart from edit checking, data processing by computers was usually reserved for the final stage of the whole operation, the production of statistical tables.

624. More recently, data processing has benefited from the development of database management systems as well as a wide range of applications for the vastly enhanced capacity of personal computers, standing alone or linked through internal networks. For large-scale data processing applications there are both mainframe and personal computer versions of the systems, which allow the non-specialist to design screens and forms and to develop a system for the processing of data at all stages, including data entry, edit checking, management of the processing facilities and the presentation of statistics for publication. Previously, the manipulation of census information or large-scale social survey information via such systems was prohibitively expensive in terms of programming and the hardware requirement for online data storage. Technical advances have now made large-scale database management a feasible proposition for such applications.

625. It is also only relatively recently that real progress has been made in bringing computers to bear on the tasks carried out, in an inherently slow and laborious way, by census coders. However, the situation is changing rapidly, and systems for automatic coding (AC) or computer-assisted coding (CAC) of industry and occupation are now used in a number of countries. Automatic coding and computer-assisted coding improve the consistency of coding and reduce its time and cost. Both types of system read captured responses. The responses can be captured by intelligent character recognition or input manually. Manual input requires operators with appropriate keyboard skills that may not be available in some countries. The success of intelligent character recognition varies by language, alphabet or characters used and requires responses to be written clearly and inside the designated boxes. Successful intelligent character recognition operations may require the use of special quality paper and ink, as well as handling procedures for

the questionnaires before, during and after the data collection that will protect them from humidity, sun and other spoiling influences. That may prove difficult to ensure in many countries.

626. So far no workable system has been developed that can fully automate the human decision-making task in coding occupation and industry. Some automatic coding systems allocate codes automatically to more than 70 per cent of the cases, but the development costs have often been high and there have been problems in making the systems sufficiently “intelligent” to simulate reliably the performance of trained human coders. The reported error rates for responses coded by many of the systems are of the same order of magnitude as those of human coders. Moreover, the residual need for human intervention in the coding process in a substantial proportion of cases reduces the benefits that such automation could have on simplifying, speeding up and reducing the cost of data processing. Both manual and automated systems will have difficulty with vague and contradictory responses. Those limitations do not, however, eliminate the gains that can be achieved using an automatic coding system, in particular if it is integrated into a data entry and processing system that starts with optical reading of the questionnaires.

627. Computer-assisted coding uses the computer to assist the coder to find and assign the correct code. Typically the coder inputs a keyword or the first few characters of the keyword. The index items appropriate to that keyword appear on the screen with the appropriate coding instructions. The process uses few keystrokes, as typically the coder can scroll through options presented on the screen. Computer-assisted systems maintain the advantages associated with manual coding but at the same time can enhance the quality, consistency and speed of the coding.

628. Countries that have adopted automatic coding for such classifications as occupation and industry typically adopt a three-stage coding process. The first stage involves attempting to code all responses using the automatic coding system. In the second stage, responses that have not been assigned a code automatically are coded using computer-assisted coding. The cases that cannot be coded using standard coding procedures, usually a small number, are then dealt with by specially trained staff using query resolution procedures.

629. The process described above presents a number of opportunities to introduce quality control measures. First of all, it is good practice to pass some or all of the captured response data through the automatic coding system before finally assigning codes to the data file. Coded data from the test runs can be sorted in code order together with the captured response, and the index entry matched, so that expert staff may identify any systematic errors or new response types. Adjustments can then be made to coding indices and, if necessary, to tolerances in intelligent character recognition systems. The procedure can be repeated several times before making a final pass of the data through the automatic coder and can lead to improvements in both accuracy and the number of responses coded automatically.

630. Additionally, a sample of responses assigned a code by the automatic coding system can be selected for coding by computer-assisted coding, and any discrepancies analysed and corrected. Similarly, a sample of responses for computer-assisted coding can be coded by two coders and any discrepancies referred to expert coders. This will make it possible to monitor errors. Coding staff who make a large number of errors can undergo further training or be assigned to more suitable duties. The procedures also allow the identification of problem response types and the establishment of error rates and inter-coder consistency rates.

631. Although computer systems have the potential to yield dramatic improvements in the quality of data, to control the environment within which data are collected and to reduce the

amount of time that elapses between data collection and data dissemination, the true costs of implementing the system must be explained. The costs must:

- (a) Include a realistic estimate of the rate of depreciation of hardware, although it may frequently be put to good use for statistical surveys after the completion of the census processing;
- (b) Take account of the dependence on specialist programming and systems analysis skills for the implementation of the required software.

632. If it is intended to introduce an automatic coding or computer-assisted coding system, trials of the hardware and software and of the machine/operator interface need to be conducted well in advance. If there are other surveys (perhaps a labour force survey) that collect occupation and industry variables in a similar manner it would be advantageous to develop and test an automatic coding or computer-assisted coding system with those surveys. Until the feasibility and operational robustness of the machine-based system have been established, it is prudent to make parallel plans for reversion to a manual and/or clerical system as a fallback position.

633. To minimize the risks of development, and probably the cost, one should seek to acquire the rights to operate a system that has already been tried and tested. In selecting a system, ease of operation and adaptation to national circumstances (for example, the language used by the operators, the classification and the coding index) should be given priority. The operational advantage of an automatic coding or computer-assisted coding system is likely to depend more on the type and cost of data registration and the suitability of the coding index than on any particular feature of the search and decision-making algorithms of the system. However, fast response times and an easy-to-understand interface between the computer and the operator should also be important considerations when selecting a computer-assisted coding system.

13. *The problem of different languages*

634. In the preceding commentary little reference has been made to the problems encountered in countries where the population uses more than one language in daily life, although the issue of multiple languages in questionnaires was addressed in chapter II (see para. 76). On the assumption that enumerators will know the language of the respondent and therefore can write down correctly in that language the answers to the industry and occupation questions, the best solution will be to make sure that the coding indices can reflect those answers, as follows:

- (a) Separate coding indices for each major language will provide the best solution where it is operationally feasible;
- (b) Otherwise it may be possible to create multilingual coding indices to allow the coder, whether the enumerator or a specialized coder, to find an index entry that corresponds to what was written down as the response of the respondent. A multilingual coding index will be larger than a single language one, but not necessarily dramatically larger; in many countries the terminology reflecting modern sector activities and jobs will be common to many languages, and it will mainly be the terminology reflecting traditional activities and jobs that will differ. Such activities and jobs are normally less varied than those in the modern sectors, and there may therefore be fewer terms in each language describing them.

635. If the coding index can be constructed in only a single language, then someone, usually the enumerator, will be required to translate from the response to the language of the index. The problem with this is that the correct translation of occupational terms will require not only a good general knowledge of the two languages in question, but also knowledge of the particular area of work in order to understand precisely how particular terms for activities, products, services and jobs are used in the local context. Very few persons will normally be able to satisfy that requirement over the whole range of work situations covered by a population census or labour force survey.

XI. The development and use of coding indices

636. The present chapter is based mainly on experience from English-speaking industrialized countries, as the limited documentation readily available on the development and use of coding indices and coding procedures has originated mainly in such countries. It is difficult to assess the extent to which the documented experience is transferable to other languages and cultures. While that caveat should be kept in mind when reading the following text, the experience from those English-speaking countries may provide a good starting point for work and experiments in other languages.

A. WHAT IS A CODING INDEX?

637. The process of coding industry and occupation information involves the task of matching responses to questions in a census or survey with index entries to find the appropriate classification codes. The coding index is the key instrument for the matching process. The index can take the physical form of a durable printed publication, a loose-leaf binder, a computer printout or a machine-readable file within a computer system. The matching can be carried out by a person (the coder), by a computer or by means of interaction between the coder and a computer.

638. Detailed industry and occupation coding has traditionally been carried out using clerical (manual) procedures. Relevant information is recorded verbatim in the field by the respondent or by enumerators and the resulting raw data are brought to one or more central offices. There, clerical staff scrutinize each case, decide to which industry or occupation category to allocate it and record an appropriate code on a document, or directly onto a computer-readable medium, for further processing. It should be noted, however, that automatic coding and computer-assisted coding systems are emerging rapidly. Users of manual coding should consider the eventual adoption of such systems in the long term and begin preparing for them far in advance of their implementation.

639. The coding index is the principal instrument for linking the words used in the various parts of the response to the numerical code that represents the allocation of that response to the corresponding category of the classification. The coding index guides the coder by listing information, for example, keywords that can be found in the responses, and indicates how different responses are allocated to the detailed or more aggregate categories of the classification, depending on the nature of the information in the response and on the instructions for the coding process. Thus the census coding index must be a reflection of the types of responses that the respondent writes on the census form or that the enumerator writes on the basis of information received from the respondent. It should contain the words and expressions that the respondents will use when asked to give the information about their place of work and their job in the census enumeration.

640. It is important to recognize that a coding index is different from the index of categories specified in the classification, which is usually just a list of the titles given to those categories that are separately defined in the classification. The titles chosen for the categories in the classification are designed to be as descriptive as possible for the category content, given that only a few words may be used. Only a few of the titles will correspond to the terms used by individuals when asked about the activities of their place of work or about the tasks and duties of their job. The coding index is also different from a list of titles or terms that may have been chosen to be descriptive of the content of the categories. That type of list may also contain entries that may never be used as a title by any person describing their job or place of work. When such a list exists, it may,

however, serve as a useful starting point for the construction of a coding index.

641. Since it has to be in place before census coding operations start, the coding index has to be constructed in anticipation of what the census responses will look like. The basis therefore will have to be actual responses to similar questions in the last census, in household surveys carried out after that time and in census pre-tests. Terms and expressions concerning the types of economic activities in which people participate and the jobs that they have can also be found in advertisements for products and services (for the industry index) and job vacancies (for the occupation index), and in registrations of vacancies and job seekers in employment offices.

642. The collection and coding of the elements to be included in the coding indices should be carried out by experts on the respective classifications to ensure that they are done correctly. The work will be painstaking and time consuming, but the investment involved in the collection and coding of up to five, ten, twenty or even thirty thousand index entries ahead of the census will prove well worth the effort in terms of the speed and reliability with which hundreds of thousands or millions of census forms, depending on the size of the work force, the complexity of the economy and the coding strategy chosen, can be coded during census operations.⁷⁰

643. At the start of the census operations the coding index must be considered as incomplete. While provisions must be made to update it during the whole period of the census coding operation, it will probably be necessary to do so much more frequently and with more new items early in the process than later. The updating process should be an extension of the query resolution process in the sense that the nature and outcome of resolved queries should be made available to all coders as soon as possible, in case they encounter the same situation. The best approach is to issue a new version of the coding index. It is better to issue a completely new version of the coding index than to issue additions to an existing index, since the new entries will belong in different places in that index. Furthermore, if the coders have to transfer the information from a note on additions into the main coding index, there is a danger that they will make mistakes. New, complete versions of the coding indices issued frequently in the first weeks of the coding operation will also reduce the danger that individual coders will assign codes on the basis of their own notes on the coding of particular responses. Such notes can easily be the source of systematic differences between coders in the coding of responses not reflected in the initial version of the coding index.

644. During census operations, the physical form of the coding indices must reflect the temporary nature of the version currently in use at any point in time. Each issue should be precisely dated and, when choosing the form of the paper versions, the main consideration should be the speed of reproduction. Both paper and electronic versions should be issued with orders to

⁷⁰ It is in many ways correct to regard the coding index as the ultimate manifestation or embodiment of the classification. However, it should be kept as a working tool and not be given the status of being an official part of the classification, for the following two reasons: (a) it will be necessary to update the coding index during the census operations and when it is used later in other coding operations. That flexibility may be difficult to achieve if it is a formal part of the classification; (b) in order to reflect actual responses the index may need to include terms, e.g., brand names, which are commonly used to describe places of work or types of jobs, but which may be trademarks protected by copyright or may be difficult to include for other reasons in an official publication. Such problems will not usually arise when the coding index is used only as an internal working document and is made available to other organizations only for use as such.

destroy earlier versions.⁷¹ If using paper, a binder is very adaptable if space is initially left on each page for several additional entries. It is then possible to reprint and replace one page without requiring changes to other pages. If using an electronic index it is best to keep the index on a server. Changes to an index on a server are immediately reflected at all workstations. If, for reasons of efficiency, it is necessary for copies of indices to be held on workstations, arrangements need to be in place to ensure that updates to all workstations are made automatically for the server.

645. Only when it is clear that further significant additions to the coding index will not be forthcoming during the remainder of the census coding operation should it be issued in a form more suited for dissemination to other users, such as government departments, survey agencies and academic users. Then a well-bound quality publication may be considered, but a format more compatible with its status as a working tool and with regular, although less frequent, updating may be preferred: for example, a ring binder or computer printout may be the appropriate format.

646. Up to this point the presentation of the coding operation and coding tools has been relevant to both industry and occupation coding, as the issues are the same for the two variables, and when both are present in the census those issues and the preparations should be addressed jointly. However, as the information sought for the two variables is different, the responses obtained will also be different, and therefore the content and construction of the two coding indices will be different. They will therefore be presented separately in the remainder of chapter XI. Sections B and C concern the occupation coding index and its use, while sections D and E concern the industry coding index.

B. DEVELOPING AND UPDATING THE OCCUPATION CODING INDEX

1. *Sources of information for construction of the index*

647. The process of updating a coding index for occupations should be one part of the general process required to maintain the national standard classification of occupations. As new ways of organizing work between or within enterprises or new technologies are introduced, new jobs will appear with new combinations of tasks or new types of tasks associated with them. The new jobs may be given new job titles by their incumbents or their employers, or may be referred to under existing job titles. At the same time, existing jobs may be given a new title without their tasks and duties being changed in any significant manner, as a result, for example, of reorganization of the enterprise or of a change in their placement in a wage hierarchy. Thus, there is a need to keep track of job titles and the associated job descriptions, monitoring the relationship between that information, index entries and the associated occupational codes. If the preceding tasks have not been undertaken by the custodian of the national standard classification of occupations or by any other party or entity, then it will be necessary either to update the occupation coding index or to create it from scratch for the census.

648. A full-scale job content monitoring exercise cannot be used for census preparations as that would be too time consuming and costly. The most realistic alternatives would be to carry out

⁷¹ Copies of each version of the index need to be retained by management and by the team responsible for the classification, however, to assist in analysis of data quality and to allow for recovery in the event that an error introduced into the index results in systematic errors in the coded data.

(a) post-coding reviews of recent survey operations; (b) job vacancy reviews; (c) consultations with job placement services; and (d) a job monitoring exercise for the industries known to be either restructuring or undergoing technological change.

649. The first procedure makes use of the tools and members of the coding team(s) used for recent survey(s). Coders are a good source of information on the adequacy of the index that they have used. Their suggestions for improving the index and for additional or revised entries must be recorded and investigated. Ideally preparations for the use of post-coding review procedures for the development of index material should have been made an integral part of the design of the data processing operation for the respective surveys, as it is essential that information that may be useful in updating the classification and index be carefully collated and retained. Such information typically consists of records of problems encountered, queries raised and decisions and amendments to the working instructions adopted in the course of coding. A sample of the actual coded responses captured in machine-readable format can be very useful in classification development, index development, development of automatic coding or computer-assisted coding systems and coder training. Such procedures should be made part of the normal routine for any continuous or regular surveys, such as labour force surveys, as well as for the registration of occupations, which takes place in the local offices of the employment services.

650. Job vacancy descriptions available from job advertisements in newspapers and journals, on bulletin boards and the Internet, or from registrations at local offices of the employment services may provide a useful source for constructing or updating information on job titles and detailed job descriptions, and therefore also a coding index. The descriptions may be especially helpful where they have been coded to the occupational classification as part of the job vacancy recording process in an employment service. From such announcements of job vacancies it should be possible to find a reflection of the impact of technical and organizational change on the allocation of tasks to jobs, and to develop proposals for new entries for the index (and the classification). The advantage of that approach is that it does not require an expensive initial search for contacts, as follow-up inquiries to a vacancy of interest can use the name, address and contact person of the employer found in the vacancy notice. The main disadvantage is that job vacancies that have been advertised in newspapers or on bulletin boards, or submitted to and recorded at an employment agency, normally cover job vacancies for only a limited range of occupations and industries.

651. In cases where no national index of occupational titles exists, the index developed for the International Standard Classification of Occupations (ISCO) may be a useful starting point for the development of a national index. It should be stressed, however, that even for English-speaking countries using a national classification based on ISCO, the ISCO English index would need to be modified for use in national circumstances, as the words in the index need to reflect national usage and there is great variation between countries. Direct translation of the English, French or Spanish ISCO indices into other languages will not yield good results. Those indices are useful sources of information concerning the content of categories in ISCO and on the correct placement of some internationally used occupation titles.

652. In some countries employment agencies have established standard procedures for the collection of job vacancy materials. For example, when employers contact an employment agency, the agency may create a computerized record of information containing the job title and a brief description of the main duties or tasks associated with that job title. Those records may be used within a word-processing system or may form part of a database of vacancy information. In some cases they may have been linked, or coded, by the national employment service to the national standard classification of occupations.

2. Organization and structure of the index

653. When organizing the material for the coding index of occupations, the first issue concerns the structure of the index itself. Basically the choice is between two different approaches, which can be described as “all inclusive” and “structured”.

654. When the all inclusive approach is adopted, every distinct type of response found in the process of coding should, in theory, have an entry in the index, although allowance may be made for misspellings and inversions of words that are without consequence for the meaning of the response. An advantage of this approach is that it may be possible for coders, when faced with an obscure job title and/or task descriptions, to find that title and/or those tasks listed in the index. A significant disadvantage is that an all inclusive index may become very large and its sheer size may slow down the process of searching for the appropriate entry in the index, and thereby slow down the coding, whether the coding is done manually or is computer assisted.

655. Furthermore, large verbatim indices may create the impression that coding is a simple task, involving a straightforward matching between a response and an index entry. However, no matter how large the index (and some occupation indices have over 30,000 entries) it will always be the case that a significant proportion of responses fail to match the index entries exactly. In those cases the coder has to be given rules and/or use judgement to make the best match. Consequently, in a number of statistical agencies the approach has been to develop a structured index.

656. A structured index does not try to reflect every possible response because it is accompanied with instructions to the coder on how to break down the available response into keywords and qualifying nouns or adjectives. The primary entries in the index are the keywords. If a keyword in itself is not sufficient to uniquely identify the category, an appropriate qualifying word (or phrase) must be added to distinguish between the possible alternatives having the same keyword. If that is not sufficient to resolve all ambiguities, a second or higher-order qualifying word should be used. The following examples may illustrate the system for transforming an occupational response into an entry in a structured coding index according to the following format:

Response: \Rightarrow keyword/first qualifying word/second qualifying word
Cost accountant: \Rightarrow accountant/cost
Drilling machine operator: \Rightarrow operator/machine/drilling
Aircraft instrument maker: \Rightarrow maker/instrument/aircraft
Room maid: \Rightarrow maid/room
Marine biologist: \Rightarrow biologist/marine
Capstan lathe setter-operator: \Rightarrow setter-operator/capstan lathe

657. The following examples, based on the coding index used for coding occupation in the 1986 population census in Australia, will illustrate the use of the qualifying words as well as the way instructions about the use of the index can be incorporated with the index entries. The codes given are from *Australian Standard Classification of Occupations—ASCO Coding System: Unit Group Level* (Australian Bureau of Statistics, 1992):

5999 Researcher/market/interviewing

2909 Researcher/market/statistician
 2907 Researcher/market (except above)
 2701 Researcher/accountancy
 2107 Researcher/agricultural
 2907 Researcher/anthropology
 2999 Researcher/assistant to parliamentarian
 2107 Researcher/biological sciences (except medical)
 2101 Researcher/chemistry (except medical)
 2109 Researcher/medical
 3103 Researcher/toxicology
 2000 Researcher (no additional information about type of research)
 8919 Restaurateur/assisting in kitchen
 4705 Restaurateur/cooking
 1503 Restaurateur/supervising staff and administration
 6505 Restaurateur/waiting on tables
 1503 Restaurateur (no additional information about specific tasks)
 3999 Retoucher/photographic
 4503 Retoucher/printing
 1311 Secretary/assistant/senior government officer/computing division
 1307 Secretary/assistant/senior government officer/distribution division
 1313 Secretary/assistant/senior government officer (except above)
 6503 Secretary/club/tending bar
 1599 Secretary/club (except above)
 1201 Secretary/trade union
 5601 Secretary/receptionist
 5101 Secretary (no additional information about specific tasks)
 5101 Secretary (except above)
 4405 Signwriter⁷²
 4921 Silverer/glass
 4923 Silversmith
 2815 Singer

658. The much smaller number of index entries in a structured coding index than in an all inclusive index is a result of the restriction of the index to keywords and essential qualifying words when possible and of the use of the instructions “except above”. Consequently, it is possible to exclude from the index words used in responses that are immaterial to the selection of the correct category.

659. In a structured index for coding occupation, the keyword is the single word in the relevant response that alone can serve as an occupational title, however imprecise. The qualifying words will usually indicate some form of specialization of tasks. Sometimes the keyword may be precise and in itself suffice as an index entry, such as, “silversmith” in the examples above. However, the keyword may also be very ambiguous, as in the “researcher” and “secretary” examples above. Note that the qualifying words in some of the “secretary” examples do not serve to distinguish between specializations but between occupations that are very different in nature.

660. The construction of a structured coding index must reflect and support the coding rules to

⁷² Note that although the term “signwriter” may not appear in dictionaries as a single word, it needs to be included in the coding index if it is given as a single word in census responses. An entry “Writer, sign” would also need to be included in the index.

be used for assigning occupation codes on the basis of the responses to the relevant questions, using, when necessary, permissible ancillary information given in other responses and indicated in the index as qualifying words. This means that one should organize the index alphabetically first in terms of keywords, then in terms of the first qualifying word. The entries that also have a second qualifying word should be listed before those that do not, and the “except above” instruction should follow the entries with qualifying words. The keywords listed in the index must reflect those which can be selected from the permissible parts of the responses, and the qualifying words must reflect those that can permissibly be selected, in the order in which they should be selected.

661. In English it will be normal to use as keywords those that can be found in either (as first priority) the title component of the occupational response or (as second priority) the task component. First priority for qualifying words will be those normally found in the title or task components of the response. Second priority, and based on the rules for using industry information in the coding of occupation, should be given to words found in the response for industry or name/type of employer.

662. The advantages of having a structured coding index are threefold. First, it causes the coder to search for index entries in a way that is consistent with the coding rules. Second, it speeds up the task of coding by restricting the coder’s search through the index owing to the smaller number of entries. Third, when the index is searched either by computer or human, it reduces the risk of finding matches with words that are irrelevant for the coding decision and therefore reduces error.

663. Some words may be found to be in common usage in job titles, but can be ignored for the purpose of creating a coding index. For example, such words as “senior”, “junior” and “executive”, when given as qualifying words, may not carry information about the tasks that constitute a particular job. It is possible to exclude such words from the index and to instruct coders to ignore them.

C. USING THE OCCUPATION CODING INDEX

1. *Using the occupational response*

664. Coding can be seen as a process in which the task of the coder is to translate the information provided by the recorded responses to the appropriate code in the occupational classification structure. The main tools for the translation are the coding index and the coding instructions—including instructions on when a response should be treated as a query to be resolved by supervisors or expert staff. The instructions should specify the following:

- (a) How the translation process should be carried out;
- (b) What items to look for in the occupational response and in what order;
- (c) What type of ancillary information to use from other responses;
- (d) When such ancillary information can permissibly be used and how to use it.

Ideally the coding index should have been constructed to reflect and support the use of

these instructions.

665. The starting point should always be the response given to the occupational question(s), specifically, question(s) on the type of work in the person's own job and on the usual or main tasks and duties carried out in the job. The question should normally elicit a job title and a few words on main tasks. When using an all-inclusive coding index the coder should start by marking the words that are relevant for the search in the index. When using a structured index the coder should identify the keyword, normally part of the response pertaining to the job title, and look for it in the index. The part of the response relating to tasks should then be used where necessary to supplement or modify the information provided by the job title. Alternatively, those tasks may need to be transformed to a title, for example, "baking bread" to "baker, bread"; "cleaned school" to "cleaner, school". Transformation of a task response to a title should be performed when there is no proper title response or when there is no index entry corresponding to the title given, as may be the case for "civil servant", "helper" and other non-informational titles. If the title is inconsistent with the tasks or if the occupational responses are not sufficient to determine a detailed occupational category, then the coder should choose one of the following three alternatives:

- (a) Look at the form for recorded ancillary information of a specified type for further clarification;
- (b) Use an appropriate code for inadequately specified responses;
- (c) Refer the case to supervisors as a query.

The coders should be given clear instructions and training on the circumstances that would allow any of the alternatives to be chosen. Coders should not be expected to make judgements that require an understanding of the classification structure.

2. *Using ancillary information on industry or name and type of employer*

666. Most modern occupational classifications, including ISCO-88 and ISCO-08, are designed according to the principle that "occupation", meaning a particular pattern of work tasks and skills that constitute an individual's job, should be kept conceptually separate from "industry", meaning the sector of the economy in which the job is performed. Thus, an "electrical maintenance fitter" may work in any of a range of different industries, and that person's occupation cannot be validly deduced from a knowledge of the industrial category of the employing organization. Without violating the above-mentioned principle, it must nevertheless be recognized that certain occupations or occupational titles are to be found solely or predominantly in particular industrial sectors. In such cases, knowledge of the industry may clarify an occupational title or description that is inadequate for coding purposes. For example, a "face worker" working in the coal mining industry may be deduced to be a miner engaged in coal cutting.

667. In other cases the descriptions of work activities used to identify an occupation are best formulated in terms of, for example, the nature of the material worked with (such as wood, rubber or leather). That information may be deducible from knowledge of the industrial sector in which the job is located and may again help to clarify a vague occupational description. For example, the occupational term "coil winder", used on its own, is ambiguous since coding depends on whether what is wound is some form of metal wire, some form of textile product or some other material. Knowledge that the job is located in a textile-manufacturing establishment may be sufficient to resolve the ambiguity with a reasonable degree of certainty.

668. Thus, since some interrelationships between occupation and industry are inherent in the industrial structure of the economy, they can be made use of to improve the accuracy of occupational coding. However, there are costs as well as benefits in that practice. In the first place, there is always some danger that inferences from industry to occupation will be based on incorrect or out-of-date assumptions about the distribution of occupations across industries. In the second place, when coding is being done on a large scale, coding work rates and inter-coder consistency are important considerations. Coding rates are likely to be slowed if the coder routinely considers extra sources of information to arrive at a code, particularly if the extra information is itself hard to interpret. In such circumstances there is also a danger of increasing inter-coder variability, since different coders will tend to interpret the information in different ways. The latter two problems are minimized if:

- (a) Industry is coded in advance of, or at the same time as, occupation so that no further interpretation of verbatim responses is required for the occupation code;
- (b) Coders are allowed to use data on industry only where the responses to the specific questions on job title and activities are inadequate to determine an occupational code;
- (c) The choices that an occupational coder can make on the basis of the industry code are exhaustively specified through index-referenced instructions.

669. A simplified example may clarify the above. A coder encountering the job title “coil winder” would be instructed in the coding index under “winder, coil” to look first at the description of job activities for information on the type of material wound. In cases where no indication of that was found, the coder would look next at the pre-allocated industry code. If it was code “x”, standing for “textile manufacturing”, the occupational category would be determined as “textile yarn winder”; if the industry code was “y”, standing for “electrical machinery manufacturing”, the occupational category would be determined as “wire winder”. If the industry code were other than x or y, the occupation would be placed in an appropriate “inadequately described” category by the coding index.

3. *Use of other ancillary information for coding of occupation*

670. Some countries permit coders to use information about the educational and vocational qualifications of respondents as ancillary information to determine the appropriate occupational code. That approach should be based on detailed knowledge of the relationships between training and qualifications, on the one hand, and the corresponding occupations, on the other. In all countries the relationship varies between occupations, and in most countries the relationship is close for only a limited number of occupations. Even when the relationship is close, the fact that a person has a particular qualification does not mean that his or her job will include the corresponding tasks. For example, a person with a medical degree who is working in a hospital may not have tasks corresponding to his or her degree. That may be because the person has been promoted to a job that consists mainly of management tasks, or it may be because that person could not get a job corresponding to the type and level of his or her technical training (for example, owing to a lack of necessary language skills).

671. For the preceding reasons, education is most useful in an exclusionary function. A person working in a hospital without a medical degree can reasonably be excluded from being a medical doctor. The exclusionary function also can be employed as an edit to resolve problems associated

with distinguishing between levels of the same discipline. The use of information on qualifications or educational attainment as ancillary information should therefore be very carefully controlled and probably restricted to query resolution by expert coders or to computerized edits of coded responses.

672. The use of ancillary information of any kind in an occupation coding process may have a tendency to bias the relationship between occupation and other census variables collected and may also have a negative impact on the efficiency of the coding process. Coders should not, therefore, be allowed to use responses to questions on age, income, hours worked or other socio-economic variables to determine an occupation code.

4. *Inadequate occupation responses and queries*

673. Some responses simply cannot be coded to a detailed occupational category. That will normally be for one of the following reasons:

- (a) The response may be vague: it does not contain enough information to be coded according to the coding index and coding rules;
- (b) The response may be precise, but may use a title and/or indicate types of tasks or combinations of tasks that do not correspond to any of the index entries. Supplementary material should be used by experts to code such precise responses and then make appropriate changes to the index.

674. Unfortunately, the number of cases of type (a) is likely to be substantial, even with well-formulated occupational questions and well-trained enumerators. In order to keep to manageable proportions the number of queries that the supervisors and expert coders must handle, the coding index and the coding instructions should be designed to guide the coders with respect to the most common of such cases.

675. The simplest solution will be to specify that the response should be coded to a default category, as in the examples of “researcher”, “restaurateur” and “secretary” in the above-mentioned example from Australia (see para. 657). The default category may in some cases be a specific detailed category since it reflects the dominant usage of the terms found in the response: for example, “1503 Restaurateur” and “5101 Secretary” as default categories. However, the default category will often have to be one of the aggregate categories in the classification, since it is not possible to identify one particular detailed category as dominant within the aggregate category indicated. In the example from Australia above, “2000 Researcher” indicates that a response giving only “researcher” as information could be coded validly only to major group 2. Similarly a response like “manager, running a business” would normally have to be coded to the aggregate category for “managers”, unless the industry response gave very clear information.

676. There is a real danger that coders may use default categories as dump categories for difficult-to-code responses before they have tried to find a precise code. Some countries have therefore tried to avoid their use by first line coders, preferring that such codes be used only by supervisors. However, that strategy may create a very large query burden for the supervisors. It may therefore be better to allow and monitor carefully the use of default codes by first line coders.

677. The fully specified responses that are not adequately covered by the coding index should always be handled by expert coders and recorded carefully. This is done because it is important to

ensure consistent treatment of similar cases and because those cases represent an important source of information for the updating of both the coding index and the classification. During the coding operation, such cases can be handled either by using priority rules specified for the classification or by assigning them to one or several categories (or supplementary codes) for occupations not adequately covered by the classification.

678. Priority rules can be applied to some of the responses that indicate task combinations that cut across the categories defined in the classification, for example, “baker, baking/selling/managing shop”. Most classifications based on ISCO-88 or on ISCO-08 will specify priority rules in terms of tasks performed for the allocation of such jobs to occupational groups. In ISCO-88 it is specified that priority should first be given to the tasks that require the highest skill level, and secondly that production-oriented tasks should be given priority over managerial or administrative tasks. “Main tasks”, in terms of time spent, for example, are not to be given priority unless they completely dominate, both because an employer is likely to be concerned that a worker can carry out the most skilled tasks required, even if they are only seldom activated, as in emergencies, and because information about time allocation of tasks is normally not available. Thus in the above example the code to be specified in the coding index should be the code for “baker” according to the priority rules of ISCO-88. However, this may change in ISCO-08.

679. Precise responses that cannot be resolved by priority rules should be assigned special “inadequately described” or “not defined” codes placed in categories created for coding purposes within the aggregate categories to which the jobs clearly belong. Steps should also be taken to ensure that these cases can be closely examined outside the coding operation to determine what, if any, contribution such cases can make to the updating of the coding index and to the classification. Note that these categories are not the same as the “not elsewhere classified” categories of the classification. Great care must be taken to avoid confusion between the two types of category.

D. DEVELOPING AND UPDATING THE INDUSTRY CODING INDEX

1. *Types of industry coding index*

680. Most census coding operations will find it useful to have two coding indices for the coding of industry, as follows:

- (a) A list of as many as possible of the establishments which, at the time of the census, are/were operational in the geographical region covered by the coding operation, where each establishment has been given the correct industry code by those who are specialists in establishment surveys and in the coding of establishment activity. In practice such lists (or business registers) may often cover only large, formal sector establishments as they have been created from lists kept in tax offices, licensing offices and/or chambers of industry and commerce. They may nevertheless cover significant proportions of the work force, and their use for census coding will eliminate one possible source of inconsistency in employment statistics between the census results and the results of establishment surveys;

- (b) A list of significant word combinations reflecting the answers given in response to industry questions, that is, an index of the same type as that created for the coding of occupation.

681. The industry coding process will therefore usually attempt first to match the name and address of the person's employer with those in the list or register of establishments. If a match cannot be made using the register of establishments, then an attempt is made to match the description of the industry with the index of type (b) above.

682. The process of updating the coding indices for industry responses should be viewed as part of the general process required to maintain the industry classification. As new ways of organizing work between or within enterprises or new technologies are introduced, new functions, end products and services with separately identifiable units (separate establishments) will appear. For example, with the introduction of movies on videotape, units of production providing rentals of the tapes were created. New industries will be created and some old industries may disappear. However, those responsible for census preparations will frequently find that while work to maintain establishment lists of type (a) above may have been carried out by those responsible for an establishment register or for establishment surveys, that work has not been undertaken with respect to lists of type (b), neither by the custodian of the national standard industry classification, nor by anyone else, and that in fact not only the updating but also the creation of this type of industry coding index must be undertaken from scratch for the census.

2. *Lists or registers of establishments*

683. Where lists of establishments or business registers already exist for use in economic statistics or are used for administrative purposes, they should be adapted for use in census processing. Since information provided by households in censuses and surveys may be less detailed and precise than that available in economic collections, it may be necessary to make adjustments to the list to reflect the type of information typically provided in response to questions on census forms on name and address of place of work or employer. If no such list exists it may be useful to compile a list of known large establishments and pre-code and assign an industry code to each establishment.

684. For each unit, the list of establishments should give both a name and the physical location, indicated by a street address if possible, otherwise by naming the (smallest) district in which the unit is located or is operating as one unit. If alternative forms of the name, such as abbreviations, initials or old names, are used or have recently been in use, they should also be included in the list as separate entries, owing to the possibility that persons working for them may use those variants in their answers. The entries in the list should be organized alphabetically, with clear rules for where to find entries consisting of initials and abbreviations.

685. Notices and advertisements in newspapers, journals and on the Internet about their products and services, creation and expansion, or about vacancies, will be used by many formal sector businesses, and they may provide a useful source for constructing or updating information on establishments and their activities. Telephone directories, trade directories and similar publications are most useful. They often classify industries (the yellow pages, for example) and in some cases provide short descriptions for some entries. The main disadvantage is that small, informal sector establishments are not normally well covered by such reference materials.

686. Experience suggests that in most countries only a minority of census responses will be

coded successfully using a list of establishments. Some developed countries, for example, report that 30 or 40 per cent of industry responses can be coded in that way. Since assignment of the code is based on known characteristics of the establishment, the codes assigned will have a high level of accuracy.

3. *Indices reflecting the answers given in response to industry questions*

687. Most industry coding is likely undertaken using indices of type (b). It is important to note that indices of that type for use in population censuses and other household or person-based data collection activities are very different from those used in economic surveys of establishments. The words used in responses provided by individuals to questions about the kind of industry of their employers are likely to be very different from those provided in establishment-based collections. In countries where information on industry is regularly collected in household surveys, such as labour force surveys, it is possible that an index of type (b) may be maintained for that purpose.

688. A full-scale establishment monitoring exercise cannot be used for census preparations, as that would be too time-consuming and costly. The most realistic alternatives would be to carry out (a) post-coding reviews of recent household survey operations; (b) reviews of advertisements and notices in newspapers and other media; and (c) consultations with tax authorities, chambers of commerce and the like. In addition, as much help and information as possible need to be obtained from those responsible for establishment surveys.

689. Procedure (a) makes use of the tools and members of the coding team(s) used for recent survey(s). Coders are a good source of information on the adequacy of the index and other tools that they used. Their suggestions for improving the index and for additional or revised entries must be recorded and investigated. Ideally preparations for the use of post-coding review procedures for the development of index material should have been an integral part of the design of the data processing operation, as it is essential that information that may be useful in updating a classification and index be carefully collated and retained. Such information typically consists of records of problems encountered, queries raised and decisions and amendments to the working instructions adopted in the course of coding. Such procedures should be made part of the normal routine for continuous or regular surveys, such as labour force surveys. They should also be applied to the registration of establishments and activities, which takes place in the local tax offices, licensing offices or chambers of commerce, and which is also carried out by those responsible for establishment surveys.

690. When organizing the material for the coding index of activities (industries) the first issue concerns the structure of the index itself. Basically the choice is between two different approaches. In some statistical agencies the approach has been that the index should be all-inclusive, that is, every distinct type of response found in the process of coding should, in theory, have an entry in the index, although allowance may be made for misspellings and inversions of words that are without consequence for the meaning of the response. An advantage of that approach is that it may be possible for coders, when faced with an obscure type of activity or product, to find those terms listed in the index. The main disadvantage is that the index may become very large and its sheer size may slow down the process of searching for the desired entry in the index, and thereby slow down the coding, whether the coding is done manually or is

computer assisted.

691. Large verbatim indices also create the impression that coding is a simple task, involving a straightforward matching between a response and an index entry. However, no matter how large the index, it will always be the case that a significant proportion of responses fail to match the index entries exactly, and one has to use rules and/or judgement to make the best match. If an exact match is not found, the chance that an incorrect entry will be selected will tend to increase with the number of entries available. For those reasons, an alternative approach may be to develop a structured index.

692. A structured index does not try to reflect every possible response. Instead, it is accompanied by instructions to the coder on how to break down the available response into keywords and qualifying nouns or adjectives. The primary entries in the index are the keywords. If a keyword in itself is not sufficient to uniquely identify the category, an appropriate qualifying word (or phrase) must be added to distinguish between the possible alternatives having the same keyword. If that is not sufficient to resolve all ambiguities, second or higher-order qualifying words should be used. The following examples may illustrate the system for transforming an industry (type of activity) response into an entry in a structured coding index according to the following format:

Response: \Rightarrow Keyword/first qualifying word/second qualifying word:

Sheep farm: \Rightarrow sheep/farm

Car rental agency: \Rightarrow car/rental

Youth club: \Rightarrow club/youth

Tax assessment office: \Rightarrow tax/assessment office

Cleaning service: \Rightarrow cleaning/services

Cleaning products production: \Rightarrow cleaning/products/production

693. The following examples are from the coding index used for coding industry to the United Kingdom Standard Industrial Classification of Industrial Activities, 1992:

15.11/1	Abattoir
74.40	Advertising/agency
74.40	Advertising/agent
74.40	Advertising/campaigns/creation
74.40	Advertising/campaigns/realization
74.40	Advertising/consultant
74.40	Advertising/contractor
92.11	Advertising/film/production
31.50	Advertising/lights/manufacturing
74.40	Advertising/material/design
22.22	Advertising/material/printing
22.22	Advertising/newspaper/printing
22.12	Advertising/newspaper/publishing
74.40	Advertising (no further information)

694. In a structured index for coding industry, the keyword is the word in the relevant response that alone can serve as a designation of a service, a product or a function, however imprecise. The qualifying words will usually indicate some special form or variety and/or the type of activity

associated with the product or service. That sequence has been chosen because the number of different designations for activities is much smaller than the number of designations for different products, services and functions. Sometimes the keyword may be precise and in itself suffice as an index entry, such as “abattoir” in the example above. However, the keyword may also be very ambiguous, as in the advertising examples above.

695. If the coding of industry and occupation is to be done by the same staff, the coding rules and structure of the coding index for industry must not be in conflict with the rules to be used for assigning occupation codes. Industry codes are assigned on the basis of the responses to the questions on industry, using, when necessary, permissible ancillary information given in other responses and indicated in the index as qualifying words. Consequently, the index should be organized alphabetically, first in terms of keywords and then in terms of the first qualifying word, with those entries that also have a second qualifying word listed before those that do not. The “except above” or “no further information” instructions should follow the entries with qualifying words. The keywords listed in the index must reflect those that can be selected from the permissible parts of the responses, and the qualifying words must reflect those that can permissibly be selected, in the order in which they should be selected. Rules for identifying the key or functional words for industry may need to be quite different from those for occupation, however.

696. The advantages of having a structured coding index are threefold. First, it causes the coder to search for index entries in a way that is consistent with the coding rules. Second, it speeds up the task of coding by restricting the coder’s search through the index owing to the smaller number of entries. Third, when the index is searched either by computer or by human it reduces the risk of finding matches with words that are irrelevant for the coding decision and therefore reduces error.

E. USING THE INDUSTRY CODING INDICES

1. *Using the industry response*

697. Coding can be seen as a process in which the task of the coder is to translate the information provided by the recorded responses to the appropriate code in the occupational classification structure. The main tools for the translation are the coding indices and the coding instructions—including instructions on when a response should be treated as a query to be resolved by supervisors or expert staff. The instructions should specify the following:

- (a) How the translation process should be carried out;
- (b) What items to look for in the industry response and in what order;
- (c) What type of ancillary information to use from other responses;
- (d) When such ancillary information can permissibly be used and how to use it.

Ideally the coding indices should have been constructed to reflect and support the use of these instructions.

698. The industry response is the information given in response to a question or questions about the name and address of the person’s employer in a particular job, and the main type of industrial activity undertaken by the employer at that location. The starting point should always

be the response given to the first part of the industry question(s)—the question(s) on the name and geographical location, for example, the street address of the place of work. If the name list provides an exact match on both name and location, then the industry code given to that unit in the name list can be given to the response. If there is no exact match, then the coder should make use of the regular coding list for industry by selecting a word from the response that provides information about the type of products, services or function the unit produces or provides. If that procedure is not sufficient to determine a code, as with “advertising” in the example above, then the coder should identify supplementary words or qualifiers that may give more precise information about the product and/or the type of process involved. If the industry response does not contain sufficient information to determine a detailed industry category, then the coder should choose one of the following three alternatives:

- (a) Look at the form for recorded ancillary information of a specified type for further clarification;
- (b) Use an appropriate code for inadequately specified responses;
- (c) Refer the case to supervisors as a query.

The coders should be given clear instructions on the proper alternative to choose.

2. *Using ancillary information on occupation*

699. Most modern industry classifications, including the International Standard Industrial Classification, Revision 4, are designed on the principle that “industry”, meaning a particular set of productive activities resulting in one or more products or services produced by an economic unit, should be kept conceptually separate from “occupation”, meaning the type of work performed by a person working in the establishment. Since many different occupations may be represented in the same establishment, one cannot normally draw valid conclusions about the industry of the workplace from one person’s occupation, even if it happens to be an occupation that tends to cluster in a particular industry, such as bus drivers. However, there are a few exceptions to this, for example, “university teachers” are found only in the education industry and “taxi drivers” only in transportation. For some “own-account” workers there will be a direct link to a particular industry, for example, own-account plumbers should logically work only in the construction industry. Such cases can be identified and incorporated into the industry coding index.

700. Thus, since some interrelationships between occupation and industry are inherent in the industrial structure of the economy, they can be made use of to improve industry coding. However, there are costs as well as benefits in that practice. In the first place, there is always some danger that inferences from occupation to industry will be based on incorrect or out-of-date assumptions about the uniqueness of an occupation to a particular industry. In the second place, when coding is being done on a large scale, coding work rates and inter-coder consistency are important considerations. Coding rates are likely to be slowed if the coder routinely considers extra sources of information to arrive at a code, particularly if the extra information is itself hard to interpret. In such circumstances there is also a danger of increasing inter-coder variability, since different coders will tend to interpret the information in different ways. The latter two problems are minimized if:

- (a) Occupation is coded in advance of or at the same time as industry, so that no further interpretation of verbatim responses is required for the industry code;

- (b) Coders are allowed to use information on occupation only where the responses to the specific questions on name and address of place of work and activities are inadequate to determine an industry code;
- (c) The choices that an industry coder can make on the basis of the occupation information are exhaustively specified through index-referenced instructions.

701. A simplified example may clarify the above. A coder encountering the establishment name “Institute for Marketing Studies” might be instructed, in the coding index under “institute/marketing”, to choose between the code for “marketing” and the code for “education”. If the occupation is given as “manager”, “secretary” or “janitor” there would be no basis in the occupation information to make the choice, as all three types of occupations may exist in either industry. However, if the occupation is given as “account executive” then it is likely that the establishment is a marketing firm (with a fancy name), and if the occupation is given as “professor” or “lecturer” then it is likely that the establishment is a training institution.

3. *Use of other ancillary information for industry coding*

702. Among the information usually collected on census questionnaires, only occupation is generally relevant as ancillary information for industry coding. The only exception is that in the case of one-company locations such as mining camps, where all employed residents are employed in the same establishment. In such cases, it may be possible to use information about the geographical location of the person’s place of residence to assign an industry code.

4. *Inadequate industry responses and queries*

703. Some responses simply cannot be coded to a detailed industry category. That will normally be for one of the following reasons:

- (a) The response may be vague: it does not contain enough information to be coded according to the coding index and coding rules;
- (b) The response may be precise, but may use a title and/or indicate types or combinations of products, services or functions that do not correspond to any of the index entries.

704. Unfortunately, the number of cases of type (a) is likely to be substantial, even with well-formulated industry questions and well-trained enumerators. To keep to manageable proportions the number of queries that the supervisors and expert coders must handle, the coding index and the coding instructions should be designed to guide the coders with respect to the most common of such cases. The simplest solution will be to specify that the response should be coded to a default category. The default category may in some cases be a specific detailed category since it reflects the dominant usage of the terms found in the response, such as “74.40 Advertising” as the default category in the example above from the United Kingdom. However, the default category will often have to be one of the aggregate categories in the classification, since it is not possible to identify one particular detailed category as dominant within the aggregate category indicated.

705. As stated above, there is a real danger that coders may use default categories as dump categories for difficult-to-code responses before they have tried to find a precise code. Some

countries have therefore tried to avoid their use by first line coders, preferring that such codes be used only by supervisors. However, that strategy may create a morale problem among the first line coders and place a very large query burden on the supervisors.

706. The fully specified responses that are not adequately covered by the coding index should always be handled by expert coders and recorded carefully. That is done because it is important to ensure consistent treatment of equal cases and because those cases represent an important source of information for the updating of the coding index. During the coding operation, such cases can be handled either by using the priority rules specified for the classification or by assigning them to one or several categories for industrial activities “not adequately covered” by the classification.

707. Priority rules are difficult to apply to the responses that indicate activity combinations that cut across the categories defined in the classification, for example, “repairing cars and selling petrol”, since most industry classifications based on ISIC, Revisions 3 and 4, will specify priority rules in terms of contribution to value added of the enterprise, or number of persons employed. That information will not be available to the census coders or their supervisors. One solution may be to refer such cases to the classification specialists, who may be able to identify the establishment and on that basis determine a code.

708. Precise responses that cannot be resolved by priority rules should be placed in special “not adequately covered” categories created for coding purposes within the aggregate categories to which the activity clearly belongs. Steps should also be taken to ensure that those cases can be closely examined outside the coding operation itself to determine what contribution, if any, such cases can make to the updating of the coding index and of the classification itself. Note that the above categories are not the same as the “not elsewhere classified” categories of the classification, which include clearly defined activities that are not important enough in scale to be given a separate identity within the larger category to which they belong. Great care must be taken not to confuse the two types of category.