Best Practice Guidelines for Developing International Statistical Classifications

Andrew Hancock, Chair, Expert Group on International Statistical Classifications

Abbreviations

CPC	Central Product Classification
ICD	International Classification of Diseases
ILO	International Labour Organization
ISCO	International Standard Classification of Occupations
ISIC	International Standard Industrial Classification of All Economic Activities
ISO	International Organisation for Standardisation
UNESCO	United Nations Educational, Scientific and Cultural Organisation
UNSD	United Nations Statistics Division
WHO	World Health Organization

Contents

Introduction

The purpose of this document is to provide guidelines for best practice in the development, maintenance and implementation of international statistical classifications. The best practice guidelines contained in this document have been developed for use by international agencies, national statistical agencies, and other organisations that may develop, maintain and implement statistical classifications.

Best practice is defined in this document as the general principles and guidelines used in the creation of international statistical classifications. It includes:

- defining what is an international statistical classification;
- components of an international statistical classification;
- guidelines on principles to consider when developing an international statistical classification.

Background

Why have international statistical classifications?

It is a fundamental need for any statistical system to have standard concepts, definitions and classifications. International statistical classifications are developed and adopted by international agencies to ensure that there is a standardised and consistent approach to classifying statistical data. The aim is to provide a basis for:

- statistics that are reasonably comparable between countries;
- developing national classifications for the same variable/characteristics.

Statistical classifications group and organise information meaningfully and systematically, in exhaustive and structured sets of categories that are defined according to a set of criteria for similarity. A primary purpose of a statistical classification is to provide a simplification of the real world and to provide a useful framework for collecting, organising and analysing data from both statistical and administrative collections, as well as providing a framework for international comparability of and reporting on statistics.

Statistical classifications are also developed to support policy making and to facilitate the collection and organisation of statistics.

Statistical classifications can be used to:

- collect and organise statistical information in a standard way;
- aggregate and disaggregate data sets in a meaningful way for complex analysis;
- support policy and decision making.

To coordinate the international work on statistical classifications, the United Nations Statistical Commission mandates that an Expert Group on International Statistical Classifications be “the central coordination body for the current and future work on classifications that are the responsibility of the United Nations Statistics Division, and for the coordination and review of

other classifications that are the responsibility of other international organizations and that have been proposed for adoption by the Statistical Commission."[1]

It is important to note that international statistical classifications developed by international organisations other than the United Nations Statistics Division will also be officially approved and adopted through the relevant mechanisms of those organisations, for example the UNESCO General Conference of Member States, or an International Conference of Labour Statisticians authorised by the Governing Body of the International Labour Organization (ILO).

Defining a statistical classification

This part of the document defines a statistical classification and introduces the essential components and guidelines for its use.

What is a statistical classification?

The following definition should be used by national statistical agencies:

> "A statistical classification is a set of categories which may be assigned to one or more variables registered in statistical surveys or administrative files, and used in the production and dissemination of statistics. The categories are defined in terms of one or more characteristics of a particular population of units of observation. A statistical classification may have a flat, linear structure or may be hierarchically structured, such that all categories at lower levels are sub-categories of a category at the next level up. The categories at each level of the classification structure must be mutually exclusive and jointly exhaustive of all objects in the population of interest."

Guideline

A statistical classification is a set of discrete, exhaustive and mutually exclusive categories which can be assigned to one or more variables used in the collection and presentation of data, and which describe the characteristics of a particular population.

Practices and principles of statistical classifications

The United Nations Statistical Commission has endorsed a set of principles for statistical classifications and how they are developed or revised. (See 'Standard Statistical Classifications: Basic Principles' by Eivind Hoffmann, Bureau of Statistics, International Labour Office and Mary Chamie, United Nations Statistics Division, 1999). These principles should be read in conjunction with this document.

Essential components of a statistical classification

The essential components of a statistical classification are:

- a consistent conceptual basis;
- a flat or hierarchic structure;
- categories that are mutually exclusive and exhaustive;
- definitions that are clear and unambiguous, and which define the content of each category;

[1] Mandate for the Expert Group on International Statistical Classifications, 2011

- that it is up-to-date and relevant;
- that it is robust enough to last for a period of time;
- that it meets user needs;
- that it provides comparability over time and between collections;
- that it provides guidelines for coding and output of data collected using it.

A statistical classification is one that follows prescribed rules and guidelines which are generally recommended and accepted. They ensure that the information is classified consistently and can be developed by any international agency, who will also act as custodians of that classification. For example, the International Standard Classification of Occupations (ISCO) is developed by the International Labour Organization (ILO), which is also the custodian of ISCO; the International Classification of Diseases (ICD) is developed by the World Health Organization (WHO), which is also the custodian of ICD.

These rules and guidelines are also applicable for national statistical agencies in developing statistical classifications.

Principles to consider when developing an international statistical classification

This section provides guidelines on what principles to consider when developing an international statistical classification.

1. Custodians

The custodianship of international statistical classifications generally resides with the United Nations Statistics Division. A number of major international classifications are, however, under the custodianship of other international agencies e.g. WHO, ILO, UNESCO, ISO.

In consultation with international and multi-national agencies, including national statistical offices and/or thematic experts, the custodians of international classifications are responsible for the development and maintenance of the classifications they look after. The custodians are also responsible for supporting the implementation of international classifications in relevant countries, and in national statistical and administrative agencies.

Custodians are required to present international classifications to the UN Expert Group on International Statistical Classifications before their finalisation.

UN Expert Group on International Statistical Classifications

The role of the Expert Group is to advise on best practice classification principles; review the concepts and principles of existing and emerging international classifications; facilitate the harmonization of related classifications; review classifications against the criteria for inclusion in the international family; and facilitate or undertake classification reviews where appropriate. The Expert Group will also provide guidance to the United Nations Statistics Division and other custodians of international standard classifications on technical matters of classification revisions or development, as well as strategic planning for classifications work, if so requested.[2]

2. Conceptual basis

It is important that an international statistical classification is based upon sound and agreed concepts and principles. The conceptual basis of the classification should be detailed in the explanatory notes and explain why the conceptual approaches taken have in fact been undertaken.

The conceptual basis should be well defined and documented to enable users to understand what the classification is about and should be used for categorising, interpreting and structuring the classification.

The conceptual basis is an essential component in enabling the creation of a sound and logical structure and for making sense of the classification. It may be based on principles or concepts developed through international collaboration, and the production of an agreed international standard, and/or through stakeholder consultation or agreement between national statistical agencies.

[2] Mandate for the Expert Group on International Statistical Classifications, 2011

Guideline: The creation or development of a conceptual basis is mandatory for all international statistical classifications.

3. Classification Structures

Classifications are either structured as a flat classification (a simple listing of categories) or as a hierarchic classification (with a logical sequential hierarchy of categories ranging from detailed to broad levels).

There are no really hard and fast rules for when to use which type of classification structure. However, the structure should ensure that the most detailed categories are at the bottom or lowest level of the hierarchical classification.

(a) Flat

A flat classification has only one level, i.e. it is a listing of categories. Flat classifications are usually developed when there is no need to group categories into more aggregate groups. However, the categories must be mutually exclusive and the classification must be exhaustive.

An example of a flat classification is:

Sex

A flat classification may also be referred to as a linear classification.

Guideline: A flat classification structure should be used when a simple listing is required or when there is no requirement to aggregate or group categories into categories that may be useful for description or analysis alone in combination with other variables

(b) Hierarchic

A hierarchic classification is a classification with more than one level of aggregation. These are structured with the most general or broad categories at the top and the most detailed categories at the bottom. Depending on the descriptive and analytical needs, each level can be used when recording a value for the variable e.g., in a survey response or an administrative record. However, there are advantages to recording (coding) to the most detailed category supported by the information available.

Examples of hierarchic classifications are:

International Standard Industrial Classification (ISIC);
Central Product Classification (CPC).

Guideline: A hierarchic classification structure should be used when there is a requirement to aggregate or group categories into categories that are sufficient for the descriptive or analytical needs, alone or in combination with other variables, or when a classification is to be used by several statistical domains with different needs of aggregation. This would allow for cross-domain consistency.

4. Classification Types

There are two types of international statistical classifications: reference classifications, and derived/related classifications.

(a) International Reference Classifications

An international reference classification is one developed by an international agency such as the United Nations Statistical Division (UNSD), International Standards Organisation (ISO), International Labour Organization (ILO), United Nations Educational, Scientific and Cultural Organisation (UNESCO), or World Health Organization (WHO).

The aim of international classifications is to provide a common framework for collecting and organising information about a particular statistical system, concept or variable. Their use, either directly or through national adaptations, facilitates the exchange and comparability of statistics and other information between countries. These classifications have generally been developed through extensive international consultation, and have achieved broad acceptance and official agreement for use.

Countries should be able to report in international categories at least at the higher levels of the international statistical classification.

Guideline: An international or reference classification may require adaptation to meet country specific conditions as it is not always possible that the classification can be used as it was developed i.e. there may be categories defined for international use which do not apply in country specific circumstances, or there may be country specific circumstances which are not catered for in the international or reference classification.

The definition of correspondences (which map or link classifications together) are mandatory between international classifications within the same subject/topic/family, as they facilitate international reporting and enable time-series management, for example ISCO88 to ISCO08, or ISIC Rev 3 to ISIC Rev 4. However, this is not a prerequisite to the classification being recognised as an international standard.

Correspondences between international statistical classifications in different subjects/topics, such as ISIC to CPC, or ISCED to ISCO, are advisable but not mandatory.

(b) International Derived or Related Classifications

Derived or related classifications are usually based upon an international reference classification. They may be developed by:

- applying the concepts of the reference classification in a more rigid or alternative way to produce a different classification hierarchy or structure;
- adopting the reference classification structure and categories at the higher levels, and then adding additional lower level detail for regional or national purposes;
- rearranging or aggregating parts of one or more reference classifications to form a new variation of the reference classification.

National or regional classifications are often regarded as derived or related classifications.

5. Mutual Exclusivity

The categories in a statistical classification need to be mutually exclusive of items at the same level of the classification i.e. each member of the population of primary units should only be classified to one category; and it should be possible to classify all units to a category in the classification.

A classification with categories which are not mutually exclusive will confuse users and not enable the statistical classification to be accurately and consistently used.

Guideline: Mutual exclusivity is mandatory for all statistical classifications.

6. Exhaustiveness

A classification should only be exhaustive for all possible values that the variable can take for the primary units for which the classification represents. For coding (recording) it should be noted that surplus or unnecessary categories often hamper the effectiveness and usefulness of the classification.

7. Statistical Balance

In general, a statistical classification should not have categories at the same level in its hierarchy which are too disparate in their population size. Statistical balance allows a classification to be used effectively for the cross-tabulation of aggregate data. An issue is the need to ensure statistical balance but maintain homogeneity particularly in statistical samples to ensure that identical elements are being classified in the same way.

Guideline: Statistical classifications should in principle be balanced although for international classifications this is not always possible. Forcing classification categories to conform to size limitations can mean that the categories will not be meaningful or useful.

8. Statistical Feasibility

The statistical feasibility of a statistical classification means that it is possible to effectively, accurately and consistently distinguish between the categories in the classification on the basis of the information available, e.g. as responses to questions that can be reasonably asked in statistical surveys or on administrative forms. To ensure suitability of the classification for use in developed and developing countries the classification needs to be widely discussed and tested.

Guideline: Statistical feasibility is a fundamental aspect when considering how to use a classification in statistical data collections. With well-designed coding tools and procedures, it should be possible to code to the correct categories effectively.

9. Classification Units/Statistical Units

The classification unit is the basic unit to be classified in the classification (e.g. the job in an occupation classification, or the activities in an industrial classification such as ISIC.)

Statistical units are the units of observation or measurement for which data are collected or derived. Statistical units can be people, products, businesses, geographic areas, events, jobs etc. This may, or may not be the same as the unit of classification.

Statistical units may also be referred to as units of observation which are entities on which information is received and statistics are compiled in the process of collecting statistical data.

Note that the term reporting unit is used to mean a unit that supplies the data for a given survey instance.

10. Time-Series Comparability

In developing and using a statistical classification, consideration must be given to ensuring comparability over time between current and previous versions of the classifications. Important time-series breaks should be avoided but may sometimes be necessary when this reflects changes to the reality that the classification should mirror. Time-series can be managed through the use of correspondences (which map or link together different versions of classifications), or through back-casting or dual-coding.

Components of a Classification

This section outlines the components required when creating a standard classification.

Classification Title

The classification title is the formal title associated with the classification. Examples of naming are:

International Standard Industrial Classification (ISIC)
International Standard Classification of Education (ISCED)
Central Product Classification (CPC)
Standard International Trade Classification (SITC)

For display in a web environment where information is generally in a sequential list, examples to facilitate finding by a user, could be

Industry, International Standard Classification of
Education, International Standard Classification of
Products, Standard Classification of

Guideline: Title conventions are as listed below:

"International Standard Classification of...."
"Standard International ..... Classification"

Classification Identifier

This is usually the abbreviation associated with the classification. Examples would be:

ISIC
ISCO
CPC

Classification Version

A classification version is a set of mutually exclusive categories representing the classification variable for a particular period of time. For a hierarchic classification, each level of the classification needs to be mutually exclusive and stand-alone.

A version is valid for a given period of time. Usually new versions are created with a frequency that balances the need to retain comparability over time with the need for the classification to represent the underlying reality at the time when the observations are made.

Guideline: A new classification version will be developed when the scope, concepts or structure change, when there is the addition of new or a deletion of old categories, and/or sometimes, modifications to descriptive definitions.

Classification Levels

A classification structure is composed of one (flat classification) or several levels (hierarchic classification) of aggregation. The bottom level of a hierarchic classification is always the most detailed level i.e. has the most precise information (detailed values) for the variable classifying the statistical unit. Categories at this level are aggregated into broader categories in the classification.

There should be sufficient levels in a classification to meet the range of statistical needs which the classification is intended to fulfil.

Guideline: The number of levels defined should be kept to the minimum to give the users the detail they need for different types of description and analysis. Hierarchic classifications usually require no more than 5 levels but should not have more than 9 levels.

NB: It should be noted that to create effective and transparent code patterns becomes more difficult with more levels in the classification.

Coding Structure

Codes consist of one or more alphabetical or numerical characters assigned to a category in a classification. A code may consist of a combination of alphabetical or numerical characters.

There are no standard criteria for when to use alphabetical codes versus numerical codes. Numerical codes are more useful particularly when creating logical and sequential hierarchical classifications. However leading zeros might be required to ensure a standard code pattern can be stored within computerised classification management systems.

Code patterns need to be consistent and logical for each level they are used. For level one of a hierarchic classification the code pattern should be the first position indicating the most aggregate level e.g. 1 for the first most aggregate group; for level 2 it should be 12, for level 3 it should be 123 - i.e. a logical hierarchic structure. This does not preclude using other patterns but with these it may be difficult to link one level to another (e.g. the use of roman numerals followed by alpha characters followed by alpha-numeric characters would not be advisable.

Sometimes codes will have '.' in them. An example would be in the Harmonised System Classification. The '.' provides a delimiter at a specific level. It isn't required generally.

The code structure should be robust enough that the addition of new codes can be done in the future.

In a flat classification the codes may either be sequential numbers or a combination of letters that may serve as easily understandable initials for the category. For example, two-character alphabetic codes for representing the names of countries.

Guideline: Every category in a classification must have a code and the code structures need to be consistent and logical for each level they are used.

Descriptors

Descriptors are usually a one-line text describing the category of the classification. This text serves as the official title of the category.

The descriptor should be unique within the classification and meaningful, to illustrate with certainty the exact content of the category. Each descriptor should be meaningful on its own i.e. no further information should be needed to see that this category has content different from all others. When categories at different levels in a classification hierarchy have identical content (for example when a category at level 2 is not further disaggregated at level 3) they should generally have the same descriptor.

Definitional descriptions/Explanatory notes

Definitional descriptions provide supporting information about the classification category. Often, they are statements which clearly define the category or they may assist users in determining the boundaries of the category.

Explanatory notes may explain the content by giving examples of inclusions and exclusions, or provide rules or guidelines for how to use that category.

Guideline: Definitions are optional but are usually included in classifications where further definition of categories is required.

Coding index

A coding index reflects probable responses to requests for information in statistical surveys or administrative forms. These are stored with the classification category to which they belong. A coding index is created to process responses.

A coding index is used to allocate a classification code to a response. It usually contains descriptions obtained from a variety of sources which include survey responses and write-ins in administrative forms. Misspellings may be included. A coding index with precise rules on how it should be used in a coding operation should be constructed for computer assisted coding/automatic coding as well as for manual coding.

Alphabetic Indexes

An alphabetic index should be made available which provides a listing of the classification descriptors and related synonyms or phrases relating to the classification categories. The index can contain reverse entries if desired and may be helpful when constructing a coding index to be used for processing of responses to survey questions or administrative inquiries.

Residual categories

Residual categories are designed to classify units that do not fit into the other, fully specified classification categories. They may be required to ensure that all categories in a flat classification, or at all levels in a hierarchical classification, are jointly exhaustive of all units in the population. The need for such categories is dependent on whether or not the other categories specified are exhaustive, and is determined at the discretion of the classification developer.

When residual categories are used within hierarchical classifications, they usually inherit the name of a higher-level category qualified by a term such as ‘other’ or ‘not elsewhere classified’. For example, ISCO-08 Sub-major Group 22 (Health Professionals) includes a Minor Group 226 (Other Health Professionals), which in turn includes a Unit Group 2269 (Health Professionals Not Elsewhere Classified). It is good practice to use a consistent naming convention for residual categories at each level.

Vague and imprecise responses to questions in surveys should not generally be coded to residual categories, but should be coded to the level of detail in hierarchy supported by the information contained in the response. In such cases it may be necessary to create supplementary categories and codes that are not formally part of the classification structure.

Other Issues/Definitions

Correspondence

A correspondence provides a link between different versions of a classification or between different classifications. A correspondence details how a category in one classification relates, or links, to the new/other classification. Sometimes the category doesn't change across classifications, sometimes a category splits into several categories in the new/other classification, and sometimes there is no corresponding category. For the latter situation a decision needs to be made as to the presentation in such instances with possible options being to exclude altogether or include and map to 'no equivalent category'

A correspondence can consist of the following relationships:

One to One (1:1)
One to many (1:n)
Many to One (m:1)
Many to many (m:n)

A correspondence is a mandatory requirement between different versions of international classifications e.g. ISIC Rev 3 to ISIC Rev 4, or CPC V1.0 to CPC V2.0.

A correspondence between different international classifications for the same or related variables/characteristics is desirable where this is relevant.

However, the mapping of different international classifications to each other is also dependent on user need i.e. there is no point creating a correspondence unless there is a demand for it e.g. SITC to BEC.

Units of Measure

Units of measure are often associated with statistical classifications used for the production of trade and/or commodity data. Units of measure are a way of quantifying the units being classified, and are part of the basic category definition. The units usually correspond to international standard codes and definitions for weights and measure based on ISO 1000 or the International System of Units (SI). The units may be associated with the classification units or the data produced from the classification.

Coding Decisions/Case Law/Determinations

It is essential to record and make easily available any previous quality coding decisions, case laws or determinations which may assist users of the classification. There may exist a process whereby users can raise queries that need a resolution which may also assist in clarifying the scope of categories. These decisions should be incorporated into the rules for using the coding index and may provide an agreed interpretation of:

- how to classify new situations/responses (e.g. jobs with new combinations of tasks and duties, new types of production activities or new products or services) which have

emerged since the classification was released (and should be noted for inclusion at the next review), or have resulted from updates of explanatory notes;

- how to classify difficult or unusual situations/responses that the existing descriptive definitions do not easily resolve;
- how to classify categories for which there have been varied interpretations by users (i.e. to get consistent coding);
- relevant administrative or legal interpretations through case law or legislation in a specific country.

Appendix One: Checklist for Developing a Statistical Classification

1. Status of the classification
   Should the classification be a standard or collection specific classification?
   What should the name be?

2. Which data collections will use this classification?
   Is it specific to a particular collection?
   Will it be used in administrative data collections as well as sample surveys and censuses?
   What are the possible non-statistical applications?

3. What are the underlying concepts used in this classification?
   How are the concepts defined?
   What are the statistical units being classified?
   What are the reporting units?
   Are there other concepts which are closely linked to the classification?

4. Scope of the classification
   What population of units does the classification cover?

5. Primary uses of the classification
   Is the primary use of the classification as a collection and processing tool?
   Is the primary use of the classification as a tool for statistical analysis?
   Will the assigned classification category influence or determine administrative decisions affecting the unit being classified?

6. User Consultation
   Should a reference group of key users and/or subject matter specialists be established? (This group should be consulted on content, scope and structure).
   Should a statistical advisory group of stakeholders be established? (This group should be consulted on the relevance of the statistics that may be produced with the classification).
   How will conflicting user requirements or applications be resolved?

7. What are the classification/similarity criteria?
   Are they compatible?
   Why were they chosen?
   What compromises were made to meet special requirements of particular users?

8. Structure of the classification
   Does the structure have an appropriate number of levels?
   Are compromises necessary for reasons of statistical feasibility or statistical balance? May these concerns be better handled by data collection and processing instruments and/or when preparing the statistics than when structuring the classification?
   What is the compatibility with other statistical concepts and classifications, and the comparability with international standards?

9. Are the proposed categories well defined?
   Are they mutually exclusive and exhaustive when considering the descriptive definitions and explanatory notes, as well as the coding instructions?

Are the names chosen for categories precise and appropriate?

10. Appropriateness of code structure
    Is the code structure appropriate?
    Are there any special code structures, conventions proposed or required?
    Are supplementary codes required?
    Are residual categories specified and used appropriately?

11. Relationship with other classifications
    Are there any (other) relevant international standards?
    What is the relationship between the classification and any other classification?

12. Statistical Balance
    Will the classification produce output that is statistically balanced?
    Should the design of the classification include the setting of any ideal minimum sizes for categories at each level?