United Nations Regional Seminar on Census Data Dissemination and Spatial Analysis

Santiago, Chile

31 May – 3 June 2011

The role of metadata in census data dissemination

United Nations Statistics Division

Introduction

- increasing need to make statistical data understandable calls for comprehensive and accessible metadata
- increasing computerization of statistical processes and dissemination of data via the Internet requires a properly designed metadata system
- situation in many countries: much documentation in print form or outdated storage media; some information not digitized, lost or never captured
- increasing need to standardize metadata development; international standards are emerging

Outline of presentation

- What is metadata?
- Types of metadata
- Basic purposes of metadata
- Users of metadata
- Sources of metadata
- Metadata in the P&R
- Metadata standards
- Issues and challenges

What is metadata?

- "data that defines and describes other data" (ISO definition)

- Metadata is descriptive information (or documentation) about data which:
  - informs users about the content, quality and condition of data
  - guides on proper usage or interpretation of data; is instrumental in transforming data into meaningful information
  - provides information on the processes of statistical production
  - describes the structure of datasets, explains, locates, or makes it easier to retrieve, use, or manage data

Types of metadata

- **Reference metadata**
  - allow a thorough understanding and interpretation of the corresponding statistical data
  - Describe the concepts, definitions, methodology and quality of data; production and dissemination process, data access conditions, etc
- **Structural metadata**
  - provide information about the structure of the dataset
  - act as identifiers and descriptors of the data, making it possible to properly identify, retrieve, browse and further process the data

Basic purposes of metadata

- facilitate discovery of relevant data
- support correct use of data
- provide transparency in data
- archiving, preservation, institutional memory
- support statistical production oriented purposes:
  - planning, design, implementation, maintenance, evaluation

- assist in retrieving and processing data
- enhance interoperability

Basic purposes of metadata (cont'd)

- **Supporting correct use of data:** some of the explanatory and contextual information that can be provided include:
  - reference date
  - data source
  - descriptions of the data program
  - descriptions of the concepts, variables, and classifications
  - information about the data collection methods
  - information about the structure of the dataset
  - data access conditions, release policy, confidentiality
  - technical information (the computer system, software packages, medium of storage, list of data files, etc.)

)

Basic purposes of metadata (cont'd)

- **Providing transparency in data:**
  - documentation of pretesting of collection instruments
  - description of the sample design, size, weights, including adjustments for non-response
  - description of estimation procedures, weighting, editing, and imputation methods
  - discussion of potential non-sampling errors
  - measures and indicators of quality: response rates; coverage ratios; proportion of records with missing/invalid data; proportion of data items with edit changes; etc
  - comments on deviations from recommended or planned definitions/classifications
  - description of any substantial changes in procedures or methodology over time
  - descriptions of known data anomalies and corrective actions
  - methodology and results of evaluations

Users of metadata

The major users include:

- users of statistical data
- producers of statistical data
- researchers on the development of statistical systems
- the software tools using and producing metadata

Sources of metadata

- metadata items can be captured throughout the statistical production life cycle
- statistical processes which generate metadata include:
  - design processes
  - operation processes (collect, validate, analyze, disseminate)
  - maintenance, evaluation and redesign processes

Metadata in the P&R

- Metadata is a key element of census dissemination to ensure that the underlying concepts are well understood and that the results are well interpreted

- All tabulations should include the following metadata or references to where this information can be obtained:
  - census questions; reasons why they are asked
  - conceptual definitions (census dictionary)
  - geographic hierarchies used
  - changes since the previous census with regard to content, operational methods or geographic boundaries
  - quality indicators such as coverage rates and item non-response
  - if a long-form sample is used in the census, metadata should also provide information on the sampling design, size and variability of the results
  - when the census tabulations include suppressed data cells due to small numbers, the metadata should also include a methodological note on the rules and methods of suppression

Metadata in the P&R

- all recommended tabulations are presented with required metadata
- type of metadata items include:
  - population groups included
  - source of statistics (type of census)
  - type of population count
  - classifications
  - definitions of urban/rural, duration of residence, etc.

Metadata in the P&R

P1.4-R. Native and foreign-born population, by age and sex

Geographical division, sex and age (in years)	Total	Native	Foreign-born	Not Stated
Total country **Both sexes** **All ages** Under 1 year 1-4 5-9 10-14 15-19 20-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-69 70-74 75-79 80-84 85-89 90-94 95-99 100 and over Not stated **Male** (Age groups as above) **Female** (Age groups as above) **Major civil division** (as for "Total country") **Principal locality** (as for "Total country")				

Population included: total population

Classifications:

(*a*) Geographical divisions: (i) total country; (ii) each major civil division; (iii) each principal locality. Distinguish between urban and rural for (i), (ii) and (iii)

(*b*) Place/country of birth: native; foreign-born

(*c*) Age: all ages; under 1 year; 1-4 years; 5-9 years; 10-14 years; 15-19 years; 20-24 years; 25-29 years; 30-34 years; 35-39 years; 40-44 years; 45-49 years; 50-54 years; 55-59 years; 60-64 years; 65-69 years; 70-74 years; 75-79 years; 80-84 years; 85-89 years; 90-94 years; 95-99 years; and 100 years and over; not stated

(*d*) Sex: both sexes; male; female

Metadata for this tabulation:

(*a*) Source of statistics:

- Traditional population census
- Register-based population census
- Registers/Surveys systems
- Rolling surveys
- Civil registration

(*b*) De jure or de facto population or a combination with detailed description

(*c*) Definition of urban and rural areas

(*d*) Definition of age

Core topics:

- Place of usual residence or Place where present at time of census
- Sex
- Age
- Place of birth/country of birth

Note:

These data are the basis for assessing the net contribution of immigration to the age and sex structure of the population. In countries where immigration has occurred on a large scale, it is very useful to tabulate the data on age-sex structure separately for the native and the foreign-born population. Thus the effects of immigration on the growth and structure of the population can be examined and estimates of future mortality and fertility can be improved by taking into account differentials between native and foreign-born population. The provision of the category of infants under one year of age is useful for studying relative underenumeration of foreign-born and native infants.

Metadata standards

- current situation w.r.t. metadata:
  - lack of common metadata standards and guidelines
  - lack of consistency/harmonization in statistical processes
- standardization:
  - ensures consistency and comparability of content
  - avoids duplication and diversity of definitions
  - ensures reduction in cost of data development

Metadata standards (cont'd)

- two international metadata standards are becoming well established
  - SDMX
  - DDI
- a number of international agencies have endorsed SDMX; supported by the UN Statistical Commission
- IHSN's *Microdata Management Toolkit* uses the DDI metadata standard

Issues and challenges

- metadata should be an integral part of statistical collection and dissemination processes
- metadata strategy is needed to:
  - provide comprehensive and accessible metadata
  - achieve harmonization
    - use of a common set of terminology
    - adopt a set of common metadata items, structures
  - capture metadata automatically at source and produce metadata to serve multiple purposes
  - preserve metadata; establish metadata registry
- regional and international collaboration between NSOs is an important consideration
  - development of standards for metadata management