The data editing process within the new statistical infrastructure of the Office for National Statistics

Pam Tate, Methodology Group, ONS

national STATISTICS

Outline

- ONS modernisation programme
- Managing editing & imputation within the survey process
- Metadata on process feedback & interfaces
- Data & metadata structures
- Design & management implications
- Conclusions

national STATISTICS

ONS modernisation programme

Aim: to deliver standard technical infrastructure, methodologies & statistical tools

Components:

- Statistical Infrastructure Development Programme (SIDP)
- Re-engineering projects
- Information Management Programme
- Central ONS Repository for Data (CORD)
- Technical Web Development Programme

national STATISTICS

SIDP & the Statistical Value Chain

- ⇓ Decision to start a collection or analysis
- ⇓ Collection design
- ⇓ Accessing admin data
- ⇓ Sample design
- ⇓ Implementing design
- ⇓ Implementing collection
- ⇓ Editing, validation & imputation
- ⇓ Weighting & estimation
- ⇓ Analysis of primary outputs
- ⇓ Index number construction
- ⇓ Time series analysis
- ⇓ Further analysis
- ⇓ Confidentiality & disclosure control
- ⇓ Dissemination of data & metadata
- ⇓ Data archiving & ongoing management

national STATISTICS

Managing editing & imputation within the SVC

Simplification - editing & imputation (E&I) assumed:

- after Data collection &
- before Weighting & estimation

In new SI, methods should incorporate best practice and be applied in a standard way - so should be possible to manage process by metadata:

E&I tool needs to be able to:

- recognise dataset due for E&I;
- recognise which methods to be applied;
- and with what options & parameters

And afterwards needs to be able to:

- indicate dataset due for weighting;
- and with what methods etc.

national STATISTICS

Managing editing & imputation within the SVC

- Some of this information derives from E&I process
- Some from earlier processes - so needs to be 'carried' with dataset - directly or indirectly
- So need 2 kinds of process metadata:
  - managing progress through SVC
  - indicating options & parameters within individual processes
- These metadata depend primarily on what outputs are to be produced, and with what quality attributes

national STATISTICS

Information on E&I for rest of survey process

Relationship between editing process & output quality:

- editing changes affect accuracy;
- extent of imputation affects accuracy;
- time for editing affects timeliness of outputs;
- nature of edit checks affects comparability & coherence of outputs;
- imputation methodology affects comparability & coherence of outputs.

Hence these processing measures contribute (directly or as proxies) to quality indicators for outputs.

national STATISTICS

Information on E&I for rest of survey process

We also need measures of quality of E&I process itself, possibly suggesting ways of improving it.

Other E&I process measures may suggest ways of improving other survey processes.

And management information on operation of E&I process can contribute to management of survey process as a whole.

national STATISTICS

Role of metadata in interfaces between processes

Using these relationships between E&I and other processes to improve quality requires creation and use of metadata:

Collection design needs metadata on how effectively the data collection process has functioned in the past.

Implementing collection on mode of collection, and whether CAI used

Editing & imputation on what edit checks applied, what proportion of records failed each edit, and what editing changes were made

national STATISTICS

Role of metadata in interfaces between processes

Weighting & estimation on whether data were imputed, and whether data identified as implausible but confirmed

Analysis of primary outputs to support assessment and evaluation of quality of outputs, including reasons for implausible data

For each key output, the quality indicators need to include:

- % of records with data changed by editing;
- % of records with imputed data;
- difference made to output by editing;
- % of output derived from imputed data.

national STATISTICS

Data and metadata structures

This wide range of types and levels of metadata has implications for how data and metadata are held and managed.

Central ONS Respository for Data (CORD) will hold all forms of ONS data, from all surveys and sources, at all levels of aggregation.

It will incorporate CORM - repository for metadata about entities such as methods, surveys, datasets, data items, classifications, questions.

CORM will be able to ensure that metadata are available to users of outputs; and in parallel, microdata and associated unit level metadata will be available internally for analysis.

national STATISTICS

Data and metadata structures

We need to distinguish between:

- unit or record level metadata - ‘micrometadata’
- summarised or aggregated metadata

Microdata created as individual record passes through survey process - summary metadata derived from micrometadata but relate to various higher level entities

Micrometadata needed for monitoring process itself, and best held together with data - summary metadata better held in CORM together with other (e.g. descriptive) information relating to whole dataset

national STATISTICS

Data and metadata structures

Hence:

- CORD design needs to take account of micrometadata;
- CORM design needs to take account of summary level metadata;
- need to be processes for deriving summary level metadata from micrometadata.

Also, some micrometadata relate generally to the unit, and need to be accessible across sources - these may need to be held with a frame or register - which would imply a need for linkages between frame and CORD.

national STATISTICS

Managing the process interfaces

Interfaces between E&I and adjacent processes managed by 2 types of metadata:

- micrometadata on history of data passing through SVC;
- information on processes to be applied to dataset, and options & parameters.

Choice of options & parameters based on knowledge and analysis, and taking into account interactions with other processes.

national STATISTICS

Conclusions

Many linkages between E&I and rest of SVC:

- management of E&I process within SVC;
- information on E&I contributing to other processes.

These relationships can improve quality of outputs, and efficiency and quality of survey process - but this depends on creating right metadata and using them effectively together with survey data.

This involves 3 elements:

- specify right metadata at unit & aggregate levels;
- specify right structures to support use of data & metadata in managing survey process;
- and to support analysis of data & metadata to optimise survey process in future.

national STATISTICS

Thank you - any questions?

national STATISTICS