Advisory Group Paper (00)03
OutputWorking Group Paper (00)03

CENSUS METADATA STRATEGY

This paper outlines the Census metadata strategy and seeks to:

- define the concept and scope of Census metadata from the 2001 Census;
- propose how Census metadata will be collected and collated; and
- describe possible ‘products’.

OWG members are asked to:-

(a) note the paper, and

(b) forward comments within two weeks of the meeting to:-

Dave Blythe, room 4300S, ONS, Segensworth Road, Titchfield, Fareham, PO15 5RR.
Email: dave.blythe@ons.gov.uk

2001 CENSUS

CENSUS METADATA STRATEGY

Introduction

1. 1991 Census metadata comprised over 70 separate user guides. These user guides were mainly hardcopy publications produced after the statistical data were disseminated. Although they provided a vast amount of information, problems with timeliness and a lack of co-ordination and cohesion of these products caused some difficulties for users in accessing the information they required to understand and effectively use Census data.

2. During the Census output roadshow meetings in spring 1999, there was a general consensus that metadata should be timely, comprehensive, coherent and easy to use. To this end, a more streamlined and co-ordinated approach is proposed. Technological developments and effective planning will enable us to provide more integrated and coherent products.

3. This paper outlines the Census metadata strategy and seeks to:

   - define the concept and scope of Census metadata from the 2001 Census;
   - propose how Census metadata will be collected and collated; and
   - describe possible 'products'.

4. Other types of metadata that will be developed (see paragraph 7), and technical aspects such as storage and dissemination tools are not covered in detail in this paper.

2001 Census Metadata

5. Metadata is basically 'data about data', that is, information to support the statistics being presented. Metadata should enable users to fully understand and make appropriate and effective use of Census results.

6. Metadata should be :

   - timely - available before or at the same time as the data;
   - comprehensive - covering all topics and areas that a user might be interested in;
   - coherent – a single source for all Census metadata; complementary 'products' which are consistent in design, format and presentation; and
   - easy to use.

7. There are likely to be four types of metadata produced for the 2001 Census.

- *Geographical metadata* - information on geographical hierarchies used in the 2001 Census and the Output Area Production System (OAPS). This type of metadata will be the responsibility of the Output Production team. GROS will be responsible for providing geographical metadata for Scotland.
- *Technical metadata* - information of a technical nature, such as how to load and run products and the ‘platforms’ needed. For example "This CD-ROM product requires as a minimum requirement a 433MHz PC with a Windows 3.1x. Once you have put the CD-ROM in the drive go to the start menu, select run…". This type of metadata will be designed to enable users to make efficient and effective use of the software products.
- *Product metadata* - information aimed at assisting users get what they want from Census outputs in the most efficient way. For example, this could take the form of a help facility within a CD-ROM product. This facility could advise the user of the suitability of the product for different purposes, guide the user on how to access the particular information they require, and highlight other related outputs that may be of interest.
- *Census metadata* - information on statistical aspects of the Census, providing background information on the results, clarifying the meaning and context. This type of metadata is the main focus of this paper.

Scope and content of Census Metadata

8. Census metadata covers a wide range of information, from background on the selection of topics and the development of questions asked in the Census, through to data collection and processing methodologies and assessment of data quality. A full list of topics that are proposed for inclusion as Census metadata is provided at Annex A.

9. Note that Annex A has been presented from an organisational, or *process*, perspective in order to assist in developing the strategy for the collection of this information. However, it is equally feasible to present this list from a *topic* or *subject* perspective.

Customer's Perspective

10. In order to decide what type of information should be included, and how it should be stored and disseminated, it is important to think from a customer perspective about how Census metadata might be used and for what purpose.

11. There are two main types of users of Census information - the 'general' user and the ‘specialised’ user. This distinction has implications for the structure of metadata products. General users are likely to require summary type information, whilst specialised users will seek more detailed or specific information. Users will want to access metadata in a variety of ways. For example, one user will want all available information on students, whilst another will be looking for detailed information about processes for addressing under-enumeration.

12. The framework used for storing metadata needs to be flexible enough to cope with these different demands. Links need to be created between operational processes, themes, topics etc, and information should be available on a *drill down* basis. This means giving summary or general information on a subject to start with, for example a definition, and then offering more detailed information if required.

13. A sub-group of the Output Working Group (OWG), the Dissemination Special Interest Group (DSIG), which comprises key census users, has been created to provide a user perspective on proposals for content and dissemination of 2001 Census metadata. The DSIG met on 8 December 1999, and the strategy set out in this paper was discussed and broadly agreed.

Potential Products

14. A broad outline of possible types of metadata products is provided below.

♦ *Census Metadata Warehouse* - The CMW should be a core product in its own right. It could be used as a rich source of comprehensive statistical information on the 2001 Census, and would avoid the problem of duplication of effort in producing and maintaining the information in different places.

Links between different topics in the CMW would allow users to access related information seamlessly. For example, it would be possible for a user to access all relevant information on a particular Census question. The information would cover development of the question, quality of responses, edit and imputation procedures, output classifications etc. Using drill down tools, the CMW would meet the needs of both general and specialised users.

♦ *Tailored views of information on the Census Metadata Warehouse* – 'Views' could be created in the CMW where interest in a particular topic or type of information has been identified in advance. For example, a predefined view of selected information on quality could be provided.

♦ *Stand-alone metadata products* –Stand-alone products could be developed where specific user requirements have been identified. To ensure that consistent information is presented, and to avoid duplication of effort, these stand-alone products should be sourced from the CMW.

Specific examples of stand-alone products are the *2001 Census Quality Report* and Data Dictionary (these are described in more detail in Annex B), for which there are already clearly established user requirements. Consideration will need to be given to whether a stand-alone product is required along the lines of the *General Report* which was produced for1991 and earlier censuses, it's scope, or whether this report can be replaced by the Quality Report and the CMW itself.

♦ *Integrated products* – Metadata from the CMW could be combined with statistical information to provide powerful, integrated products for the user. These products would allow users to view statistical information and metadata simultaneously. An example of this would be where metadata is provided with standard tables, so that a user could click on a term appearing in a table, such as 'Employment Status', and be given a definition of that term, perhaps with links to more information on that subject.

Timetable

15. The aim is to make Census metadata available before or with the statistical output. To ensure that this work is co-ordinated and completed on time a provisional timetable is set out below.

- Creation of metadata product prototype/s - *July 2000*
- Beta-testing with sample of customers - *August 2000 to October 2000*
- Completion of initial metadata product/s for users - *January 2001*
- On-going development and updating of metadata products - *January 2001 to April 2003?*

Summary

16. A summary of the key points set out this strategy is provided below.

- A single electronic repository of Census metadata - the *Census Metadata Warehouse (CMW)*, should be created and populated
- The CMW would be populated with Census metadata in a modular way, i.e., built up over time and updated as information becomes available.
- The CMW would form a core product, and be used as the source of material for stand-alone products such as the Census Quality Report.

ANNEX A – CENSUS METADATA CONTENT

1. The list below attempts to identify the scope and coverage of Census metadata categorised by the projects that will be required to contribute material. Many 'topics', such as 1991 comparability, recurring themes across projects. Metadata should be written from a UK perspective with differences in parts of the UK being outlined and explained.

Data Needs

- Questionnaire design (question and small-scale testing, business cases etc...)
- Classifications (including standard derived variables)
- Coding indexes
- Analysis of 1997 Census Test
- Analysis of 1999 Census Rehearsal
- Comparability with 1991 and earlier Censuses
- Harmonisation with other government social surveys
- The Census Forms (copies for reference) and information leaflets
- Population bases
- Concepts and definitions

Edit and Imputation

- Edit and Imputation rules

Disclosure and confidentiality control

- Legal aspects
- Methodology (different levels of detail for different users)
- Record swapping
- Thresholding
- Evaluation
- 1991/2001 comparability

One Number Census (ONC)

- % rates of imputed people by area
- Methodology aimed at various levels (e.g. easy, technical)
- Evaluation (matching, accuracy of population estimates, admin. data, "correction" process)
- Comparability with 1991/1981

Census Coverage Survey

- Methodology
- Evaluation

- CCS forms

Data Quality

- Census Quality Survey (CQS)
- Validation processes (fixed errors, not fixed errors)
- Validation rules and tolerance checks
- Coding quality

Processing

- Clerical vs automatic recognition (OCR and OMR)
- Scanning quality
- Methodologies
- 1991 / 2001 differences
- 10% to 100% processing
- Steps to address systematic errors

Data Collection

- Field procedures (e.g. details of field force, procedures for contacting people, use of different forms, treatment of communals etc.)
- Postback rates/call back
- Effectiveness of difficult to enumerate procedures
- Field quality checks
- Non-compliance

Output

- Lists and descriptions of products
- Formats, access paths, etc.
- Descriptions of measures followed to check tables produced (‘table acceptance in 1991’)
- Timetables

Geography

- Description, methodology, 1991/2001 comparability (to 1971 in Scotland), country differences etc.

Legal basis

- White Paper, Census Orders, Census Regulations, etc.

Community Liaison

- Background
- Impact on the Census

Links to sources of external evaluation of the Census

- Papers on 2001 Census from academics

Uses of Census Information by Users

- Census Offices could analyse this information to provide justification of undertaking a Census.

ANNEX B – OUTLINE OF PROPOSED STAND-ALONE PRODUCTS

2001 Census Quality Report

1. The Census Offices plan to produce a *Census Quality Report.* Quality will be considered in its widest sense and the report will draw together information from a range of sources to assist users to make informed decisions about the appropriateness of 2001 Census data for particular applications.

2. The aim is to publish this report, in electronic and hardcopy formats, either before, or at least simultaneously with, the statistical output. The proposed content is outlined below.

- ♦ *Population base, definitions and field methodology* – information on the 2001 population base, as well as issues associated with comparing 2001 Census output with that from the 1991 Census or other sources. The report will also cover field methodology as well as other definitions, such as households and communal establishments.
- ♦ *Information from the testing programme* – drawing on the wealth of information already available on the extensive programme of question testing.
- ♦ *Coverage* - information on the Census Coverage Survey and the One Number Census process, and about the coverage achieved.
- ♦ *Comparability with data from other sources* – information which will assist users to assess the suitability of the 2001 Census data for comparison with data from other sources, including the degree to which Census questions are harmonised with questions in other Government surveys.
- ♦ *Information from processing* - information about processing methodologies, including the comprehensive quality assurance strategy being implemented.
- ♦ *Edit and imputation* - a full account of edit and imputation, describing the edit rules applied, the level of non-response and the imputation methodology.
- ♦ *Validation of results* – description of the validation process and any unexpected results.
- ♦ *Census Quality Survey* – information on the methodology and outcomes of the Census Quality Survey.

2001 Census Data Dictionary

3. The Census Offices plan to produce a Data Dictionary that will provide customers with definitions of concepts used in the 2001 Census, in a dictionary format for ease of reference. Details of the full range of basic data classifications, including derived variables, will also be provided. This information will assist users who wish to better understand Census output, or who wish to define and commission their own customised output. The aim is to publish this report either before the statistical output.

4. This product would be based on the draft 2001 Census Classifications document produced for the Output Roadshows held in 1999, and would replace the 1991 Census Definitions publication.