Office for National Statistics

CENSUS ADVISORY GROUP

AG (13) 03 – 2011 Census microdata

2011 Census: Microdata Products

1. Introduction

This paper provides current proposals for microdata products for the 2011 Census. It provides proposals for products, sample sizes, the sampling strategy and draft content. The proposals presented are draft as development continues in some areas.

Advisory group members are asked to comment on the proposals outlined.

2. Microdata products

There are two types of microdata products:

- The **individual** level products are a sample of individual records, showing for each record a selection of variables. Some of these variables will relate to the household –for example socio-economic classification of head of household.

- The **household** level products are a sample of households and include individual level records for all members of the sampled households. These records will include a selection of variables, but crucially they will include variables about the relationships to other members of the household. The household (unlike the individual sample) are limited to private households and exclude residents of communal establishments.

Table 1 provides an overview of the proposed microdata products and mechanisms for accessing the different products. 2001 Microdata products are provided at Annex A for information.

Table 1: Proposed products and sample sizes

	Individual		Household	
Access	**Product**	**Draft Sample Size**	**Product**	**Draft Sample Size**
Internet	Public use	1 per cent		
Hard media (on request)	Safeguarded microdata sample	5 per cent	Safeguarded microdata sample	5 per cent
Virtual Microdata Laboratory (VML)	VML microdata sample	Initially 10 per cent	VML microdata sample	Initially 10 per cent

The microdata product specifications are based on a draft user specification compiled by Jo Wathan at the University of Manchester.

2.1 Access to microdata products

There will be a number of different ways of accessing microdata and the access route will depend on the level of utility (or risk) associated with the product. The higher the utility (and hence risk) the more stringent the access controls. The following sections set out current thinking for access arrangements.

Although not detailed in the access arrangements in sections 2.1 to 2.3, we are seeking to continue our relationship with the UK Data service, as a potential distributor of census microdata products. This access route is currently under consideration and any products disseminated through the UK Data Service would be at the higher end of utility.

2.1.1 Public use microdata dataset

The public use microdata will be published on the website for use under standard Open Government License (OGL). Publication will be in an accessible format with relevant materials to aid use, such as user guides.

2.1.2 Safeguarded products

Safeguarded datasets will be available subject to the user signing an end-user agreement. Distribution will be by physical media (DVD, data stick, etc.) or password protected email. The user agreement is likely to include at least the following conditions:

- users must have an established identity
- user must not seek to identify individuals
- user must not claim to have identified an individual
- detailed data and results must only be shared with other individuals who have signed up to the user agreement
- published outputs of research must be aggregate, and must not be disclosive (for example guidance on minimum cell counts in tables)
- some guidance on physical and electronic security of the data

2.1.3 VML Microdata products

Microdata products available in the VML will, as a minimum, be subject to the same conditions of access currently applicable to products in the VML. However, further consideration is currently being given to the need for:

- any variations to standard approved researcher application process required.
- any variations to standard terms, conditions, or methods of access required.
- any variations to existing output clearance standards and procedures required.

2.2 Microdata: draft contents

Some early assessment of the disclosure risks associated with the specified requirements have been undertaken. The proposed content outlined in the following sections reflects that early assessment, but is still subject to further assessment for confidentiality.

2.2.1 Public use microdata dataset

Table 2 shows the proposed specification for the public use microdata set. The detail and the number of variables are less than that specified in the draft user requirement. The reduced content is based on an early assessment for confidentiality using 2001 Census data; similar work is underway for 2011 data. It is likely that the specification for the public use file will be similar to that in Table 2.

Table 2: Draft contents for public use individual microdata sample

Variable	Approximate number of categories	Notes
Unique Reference No.		Unique ID for each record
Age (grouped)	8	10 or 15 year groupings
Country of Birth	7	Country within UK, EU, Elsewhere
Country of Residence	4	UK Country
Economic Activity	3	Working, unemployed, economically inactive
Ethnicity	5	Categories may be different in Northern Ireland and Scotland; white, black, Asian, mixed, other
Household Composition	6	Married; cohabiting; male lone parent; female lone parent; other related; other
General Health	5	As census questionnaire: Very Good, Good, Fair, Bad, Very Bad
Hours Worked	4	As census questionnaire: <=15, 16-30, 31-48, 49+
Industry	12	Broad groupings of main activity undertaken by employer
Marital Status	5	Single, Married, Separated, Widowed, Divorced, (including same sex equivalents)
Occupation	9	Major groupings of occupation
Usual Residency	2	Usual resident or student living away from home
Region	9	Government Office Region (England only)
Religion	8	Categories may be different in Northern Ireland and Scotland; Major religious classification
Residence type	2	Household or Communal
Sex	2	Male or Female
Student	2	Yes/No

2.2.2 Safeguarded products

The content for safeguarded products is currently being developed. Draft specifications for these products will be circulated for comment.

2.2.3 VML products

The draft specification for the individual microdata product to be made available in the VML is shown in Table 3. The specification for the household product will be circulated for comment.

Table 3: Draft contents for VML microdata product

Variable	Approximate number of categories	Notes
Unique Reference No.		Unique ID for each record
Accommodation designed or adapted for health conditions		
Address one year ago	4	Same address, same LA, elsewhere in UK, outside UK. *Should we extend to give a more precise geographical location?*
Number of adults in employment	numeric	
Armed forces indicator for household reference person	2	Could replace with industry of household reference person
Age	numeric	
Age group for communal establishment	4	As census questionnaire
Available to start work	2	Yes/No
Mother of baby under 1	2	Yes/No
Number of Bedrooms	Numeric	
Form completed on behalf of someone else	2	Yes/No
Bedrooms required using the bedroom standard	Numeric	
Provision of care	4	As census questionnaire
Number of cars and vans	5	0, 1, 2, 3, 4+ *Should this be extended to a full numeric?*
Carstairs poverty index	numeric	Index based on 2001 Census
Central heating	6	As census questionnaire
Passport held	204	Full country classification
Establishment client group	21	As census questionnaire
Country of Birth	204	Full country classification
Country of Birth of household reference person	204	Full country classification
Concealed family	2	Y/N
Country of residence	4	UK country
Number of care providers in household	numeric (0-30)	
Dependent child indicator	2	Yes/No
Education deprived	2	Yes/No
Employment deprived	2	Yes/No
Health and disability deprived	2	Yes/No
Housing deprived	2	Yes/No
Deprivation indicators	5	No. of dimensions 0-4, combination of individual deprivation factors
Long-term health problem	3	As census questionnaire
Family dependent children	19	0, 1, 2, 3+ dependent children by age group of youngest
Household dependent children		
Distance travelled to work	Numeric	

Notes – hh composition, type include alternative as well

3. Sampling to create the microdata products

3.1 Sampling Strategy

It is intended that household and individual samples will be non-overlapping samples. The use of non-overlapping samples means that, individuals cannot be paired between the two samples in order to draw further variable information about these individuals.

It is planned that the public use sample will be an exact sub-set of the safeguarded sample, which is an exact sub-set of the VML sample. In this manner the sample cannot be combined to provide added cases. Variables for the less detailed products will be collapsed from variables at the more detailed level. This will prevent users from finding out additional information about the members of the sample – even if they are able to identify the same individuals across samples.

3.2 Sample stratification

The current proposals are to stratify by local authority only. This will ensure that the sample is geographically representative at the local authority level (where this level of geography is included) and regional level. A review of the sampling strategy is currently underway by ONS methodology.

Annex A – 2001 Census microdata products

In 2001, there were a number of different microdata products that were available through three different access channels depending on the level of detail and the risk of disclosure.

Via the Cathie Marsh Centre for Census and Survey Research (CCSR)

Individual Licensed SAR (Sample of Anoymised Records) - 3% sample at GOR level.
SAM (Small Area Microdata) - 5% sample at local authority level.

Microdata within these products was de-identified and did not contain name, address or date of birth variables.

These products were available to researchers who signed an End-User licence. Researchers must agree not to attempt to obtain information about an identifiable individual and must not claim to have done so. These products were available to commercial researchers.

Via the UK Data Archive

Special Licence (SL) Household SAR (Sample of Anoymised Records) - 1% sample of households and individuals in those households (England and Wales)

Under Special Licence, researchers must also apply to ONS to be Approved Researchers before access is granted. Researchers agree to keep data under secure conditions and the institution they belong to is responsible for ensuring the conditions are met. The SL dataset is not available outside the UK.

Via the Virtual Microdata Laboratory (VML) from ONS sites

Individual CAMS (Controlled Access Microdata Samples) - 3% of sample person records. (More detailed classifications than the Individual SAR. Geographic information at Local Authority level.)

Household Controlled Access Microdata Samples (CAMS) - 1% sample of household and individuals in those households. (More detailed classifications than the Special License Household SAR. Geographic information at Local Authority level.)

These microdata products contain the same sample persons and households as the matching Individual and Household SARs (Samples of Anonymised Records) but they contain more detailed classifications and geography.

Researchers wanting access have to apply to ONS for Approved Researcher status. Researchers are allowed to carry out statistical analyses within the safe setting (VML) but are not allowed to take any notes or output away with them. All outputs are then checked by ONS to ensure that they are non-disclosive before being sent on to the researcher.