national STATISTICS

Census Advisory Group

Advisory Group Paper (07)02

2011 Census Output: possible mechanisms for data distribution

Purpose

1 This paper describes two very similar suggestions for data distribution that are currently being put forward independently by users of Census data. The first arose at a workshop on 20 March 2007 attended by representatives of ONS, GROS, NISRA, and a selection of data distributors with long-term experience of managing Census data. The second was presented in April 2007 by Eurostat to a Task Force formed of representatives for Member States to consider options for providing future Census data to Eurostat.

2 At this stage no commitment is being made by the Census Offices. However, these suggestions will be taken into account in the analysis of user requirements, after wider and fuller consultation, and in the development of a strategy for output. **Members of the Census Advisory Groups are invited to comment by 31 August 2007.**

Background

3 Significant advances were made in the distribution of Census data from the 2001 UK Census. All three UK Census Offices established web-based systems that enabled data users to access a vast range of statistics and employ various mapping and analytical tools. In addition, a number of established data distributors were able to extend and enhance their services, systems, and software to better meet the needs of their user communities. The period after the 2001 Census also saw an increase in the distribution of census data via the internet by other organisations, in particular local authorities.

4 The UK Census Offices supplied all 2001 Census data on CD and DVD, and a large amount of this was provided in a format designed specifically for loading into database systems. However, users raised concerns that the format was not consistent across the three Census Offices, and data distributors who managed UK data had to reformat some data themselves. This reflects a general criticism of the 2001 Census in that the delivery of UK data was not as coordinated as it might have been.

5 The Registrars General for England and Wales, Scotland, and Northern Ireland have published an Agreement that states their intention to promote UK harmonisation in the 2011 UK Census, and to produce consistent and coherent outputs for the UK and for each component country. In particular reference is made to the final product being "*consistent, coherent and accessible statistics for the UK and for each component country, a joint database (and/or a common data schema) being a desirable way of facilitating that outcome, with a common approach taken to output specifications, quality, data format and timing of releases*."

Ideas from UK data distributors

6 The three UK Census Offices are therefore considering ways in which 2011 Census data might best be made available to systems that distribute data to users. The first stage of this process has been to consult with a small group of data distributors to establish some basic principles for the supply of data to these and similar organisations, and consider possible dissemination models and associated roles. This consultation is being conducted in association with a wider consultation of census users on output requirements, the strategy for which was described in Census Advisory Group paper AG(06)13.

7 It is important that the Census Offices understand the technical and cultural environment that is likely to exist when the 2011 Census data is released over the period 2012-2014 and subsequent years. No-one can know for certain but there are trends developing, particularly in the use of the Internet, that are likely to strengthen in the next few years. These will need to be taken into account in developing dissemination models.

8 A workshop was therefore conducted in March 2007, to establish basic principles for the supply of data to organisations intending to distribute it through their own systems. This brought together the UK Census Offices with a number of distributors of Census data to discuss, without prejudice or commitment, how 2011 Census data might be accessed and used by the different user communities. The meeting provided an opportunity to share thoughts about future dissemination and identify the issues involved in developing an efficient and cost effective delivery of 2011 Census data.

9 The workshop concluded that 2011 Census dissemination should make significant use of the Internet, as at present, but that the development of new systems would most likely rely on open source approaches that share software and applications. This would require data to be supplied from the Census Offices in a format that is compatible with a range of software. It was suggested that the Internet practice of providing live data feeds could be used allowing all data to be stored at the three Census Offices and avoiding the need for distribution on physical media.

10 There was much support for the development of prototypes and pilots using 2001 Census data, and the three UK Census Offices will be considering how best to be involved in such projects, taking cognisance of available resources. Some data distributors are carrying out further discussion and investigations and will be contacting the Census Offices with their findings.

11 A note of the workshop is included in an Annex to this paper. A set of basic principles were established at the workshop and these are listed in the note according to the level of agreement at the meeting. Those principles that received broad agreement covered: common data formats; embedded metadata; UK consistency; improved licensing arrangements; consistent and effective disclosure control; and targeting customer needs.

Eurostat proposal: a Census European Hub

12 ONS, GROS, and NISRA operate within the wider context of national and European government, and there are a number of initiatives that may influence the eventual dissemination model used to supply 2011 UK Census data. One of these is the Census European Hub, which could have a significant impact on the way in which the Census Offices meet their obligation to supply data to Eurostat. The initiative shares many features of the type of dissemination model discussed at the UK workshop.

13 The Census European hub is a concept, developed by The Statistical Office of the European Community (Eurostat), of a new system to publish European Census data on the Eurostat website. Although the system is still conceptual, prototypes and pilots are planned to take place soon.

How the hub would work

14 Eurostat are proposing that each national statistics institute (NSI) creates a set of non-disclosive 'cubes' of data. These would be available for Eurostat to use as the base for their Census dissemination system. The delivery of this information would be via an information hub that enabled data sharing on the Internet. Each NSI would provide web access to their data according to standard processes, formats and technologies.

15 Data providers would be able to make data available directly from their systems through a querying system managed by the hub, notifying the hub of any new sets of data and corresponding metadata. Users of the dissemination system would browse the hub to define a dataset of interest and receive the dataset from the hub, the latter having retrieved the data from the appropriate NSIs.

16 Eurostat envisage the Census European Hub using standards set out by the Statistical Data and Metadata Exchange (SDMX) initiative, launched in 2001 by Eurostat together with six other sponsors: the Bank for International Settlements (BIS), the European Central Bank (ECB), the International Monetary Fund (IMF), the Organisation for Economic Co-

operation and Development (OECD), the United Nations Statistical Division (UNSD) and the World Bank. The stated aim was to develop and implement standards and guidelines for a more efficient transmission and dissemination of statistics, including both data and metadata.

17 The SDMX standards are based on a common information model, and include formats based on Extensible Markup Language (XML). SDMX provides guidelines and tools to support the 'pull' method of data sharing, where the collecting organisation retrieves the data from the provider's website. The data may be made available for download in an SDMX-conformant file, or they may be retrieved from a database in response to an SDMX-conformant query. In both cases, the data are made available to any organisation requiring them, in formats which ensure that data are consistently described by appropriate metadata, whose meaning is common to all parties in the exchange. The transaction is supported by the use of an SDMX metadata registry. This is an application that can accept SDMX query messages and return the locations of SDMX compliant information. Eurostat is currently deploying the first version of a metadata registry to be used by any SDMX-based application.

18 Eurostat are intending to start the development of the Census European Hub by involving three NSIs in a pilot phase throughout 2007 and 2008. A complete implementation is planned for late 2010.

Chris Ashford

2011 Census Outputs Branch

8 July 2007

Annex

2011 Census Data Distributor Workshop

RSS
12 Errol Street
London EC1Y 8LX

20 March, 9.30am-4.15pm

Note of Meeting

Attendees

Rob Lewis (SASPAC)
Eileen Howes (SASPAC)
Alan Lewis (SASPAC)
Justin Hayes (CASWEB)
Luned Jones (National Assembly for Wales)
Sinclair Sutherland (NOMIS)
Mark Ireland (NOMIS)
James Reid (EDINA)
Hugh Neffendorf (Katalysis/SASPAC)
Barry Leventhal (MRS)
Peter Sleight (ACD)
Keith Dugmore (DUG)
David Martin (ESRC)
Richard Adams (Chemistrygroup)
Oliver Duke-Williams (WIKID)
Valerie West (GROS)
Alan Fleming (GROS)
Sarah Conroy (GROS)
Brian.Green (NISRA)
Jil Matheson (ONS)
Simon King (ONS)
Judy Hawkins (ONS)
Stephen McIntyre (ONS)
Nicola Tromans (ONS)
Ian White (ONS)
Callum Foster (ONS)
Clive Jerome (ONS)
Alistair Calder (ONS)
Chris Ashford (ONS)
Angele Storey (ONS)
Colin Lloyd (ONS)
Linda Williams (ONS)

1. Background

The workshop brought together the UK Census Offices with a number of distributors of Census data to discuss, without prejudice or commitment, how 2011 Census data might be accessed and used by the different user communities. The meeting formed part of the current consultation on user requirements for output, and provided an opportunity to share thoughts about future dissemination and identify the issues involved in developing an efficient and cost effective delivery of 2011 Census data.

The main purpose of the day was to establish basic principles for the supply of data to organisations intending to distribute it through their own systems. In so doing, to have taken the first step in exploring possible dissemination models for 2011 Census data. This involved examining the role of the Census Offices and the services being offered to the clients of data distributors. The meeting also provoked thoughts about the technological and cultural environment in which data will be disseminated, e.g. what the web infrastructure might look like in the future and what it may be like to use.

2. Presentations

The meeting began with a welcome from Simon King who gave a brief introduction outlining the purpose of the day. To begin the series of presentations, ONS had engaged Professor Richard Adams of Chemistry Group, an internet specialist, to describe the technology currently in place that will most likely influence future development in internet infrastructure and use. Professor David Martin of Southampton University then described a vision for 2011 Census dissemination that makes use of the internet, in particular open source approaches that share software and applications. Current developments and future plans for CASWEB and SASPAC were demonstrated by those organisations, and representatives of the commercial sector made brief presentations that outline their plans and views. On behalf of the three Census Offices, Chris Ashford made a brief presentation that outlined current thinking, focussing on consultation, aspirations, constraints, and possibilities.

Each presenter, apart from Richard Adams and Chris Ashford, had been asked to list their four most important requirements for the distribution of 2011 Census data. These were then used in a facilitated discussion to attempt to establish a set of basic principles that will underpin future options for dissemination models.

3. Discussion

The set of wishes were presented on flipcharts and attendees were invited to add wishes that they thought had not been included in the presentations. The full set is listed in an appendix to this document. Duplicate or overlapping wishes were identified, consolidated, and then listed separately. Each speaker, and one representative from each distributor who had not made a presentation, were asked to vote for the wishes that they considered to be the most important to address. A facilitated discussion of these identified the following as significant basic principles for data distribution.

4. Basic principles

Broad agreement

The data to be delivered/available in common data formats. (E.g. CSV (non proprietary)).

The metadata to be embedded with data.

UK consistency throughout – Questions, processes, etc. (A core set of questions/outputs. Consistent questions, quality, and metadata with any differences explained)

Data feeds. (Online data, open access at item level with a mechanism for interacting with data.)

Workable licensing model. (Including Ordnance Survey licensing arrangements)

Output planning should be weighted towards real serious census users not to casual ones that just want a number.

UK disclosure control needs to be smarter. (Must be additive and applied UK wide).

Ability to beta test systems. (includes getting the specifications right and managing changes)

Limited agreement

Flexible outputs – (not limited to previous concepts such as tabulations)

QA of data (by organisations outside of the Census Offices)

Error management mechanism. (accurate and timely reporting and correction)

Encouraging 'The next 5 million users' (citizens & intermediaries).

Free access to Census software. (As an ideal & to some extent dependant upon business models of the Census Offices)

Other suggestions

Although these wishes did not attract a high number of votes, these, and all other wishes listed on the day, will be taken into consideration by the Census Offices

Sort data 'issues' first – all else follows.

Information derived from the census database that users would be unable to create for themselves.

Pre-published timetable for data processing & data delivery

Bulk supply on removable media.

Encourage external innovation by providing application ready data.

5. Issues noted

Some wishes raised issues that could not be addressed on the day and were hence marked up for further consideration at a later date.

Imputation should involve modelling where data is sparse rather than donor records (e.g. matrix data), similarly disclosure control should be clever instead of adjusting numbers. Tables about the same things should add up to the same numbers.

Data architecture should be hidden from user who doesn't want or need to know.

6. Conclusion

The Census Offices found the day extremely helpful in developing an understanding of the requirements of data distributors, and we appreciated the contributions made by all those who attended.

Presentations throughout the day reinforced the need to have significant user engagement and consultation. This was very much the case in the wider world and was illustrated in the references to non census organisations, such as Lego, who were encouraging individual customers to specify products rather than be passive recipients of stock items. Huge user communities develop via the internet, propelled by super users, and census data distributors can be seen as super users who will influence the use of census data on the web.

The overriding view of the meeting was that data and metadata should in general be available to distributors via internet feeds, in common formats, though a more conventional delivery on portable media was still possible. There was also general approval of proposals to conduct a co-ordinated development of prototype systems using 2001 data.

Appendix : Wishes, votes, and notes

Flip chart wish lists:

- All wishes in black type, verbatim from flip charts
- All 'voted' wishes in **bold** black type
- Number of votes accrued, shown in red
- Additional notes shown in green
- Text deleted on the day shown as "strikethrough"

ESRC (David Martin):

1. 2011: One database Census.

2. Data architecture hidden from user who doesn't want or need to know. **ISSUE**

3. Quality assurance and de-duplication of effort.

4. Present users with 'Stones in a quarry – dug out and shaped and built up in an intelligent manner'.

CASWEB (Justin Hayes):

1. Get outputs used by making them useful and usable.

2. UK consistency.

3. Information QA and data management.

4. **Encourage external innovation providing application ready data. (2)**

SASPAC (Alan Lewis):

1. Bulk delivery in popular formats (csv/saspac) – web is unlikely to meet all re-supplier demands.

2. Controlled early release of data to re-suppliers for pre-processing and to trusted/experienced users for QA (rehearsal of 2011 output streams using 2001 data).

3. Agreement in advance of nationally comparable Standard Tables.

4. Improved error tracking/management systems.

SASPAC (Hugh Neffendorf):

1. There should be no differences between England, Scotland, Wales & Northern Ireland censuses except in cases where the need is overwhelmingly clear.

2. OS data embedded in boundaries should be available on consistent terms to ONS data.

3. Imputation should involve modelling where data is sparse rather than donor records (e.g. matrix data), similarly disclosure control should be clever instead of adjusting numbers. Tables about the same things should add up to the same numbers. **ISSUE**

4. **Output planning should be weighted towards real serious census users not to casual ones that just want a number.** Need for balance. **(6)**

Demographic User Group (Keith Dugmore):

1. 'Click-use Licencing', for statistics, boundaries and directories.

2. Simple topic packages at output area level in popular formats.

3. Some stats packs to have same variables for each of England, Scotland, Wales & Northern Ireland, plus whole of UK.

4. Implementation of the new UK disclosure control policy, inc attribute disclosure, additivity & consistency.

Value Added Resellers (Peter Sleight):

1. Beta testing & pre-release of data.

2. **Pre-published timetable for data processing & data delivery (1)**

3. **Bulk supply on removable media. (1)**

4. Foolproof mechanism for error reporting & version stamping of any revisions.

New wishes and commonality(ALL):

Common data formats. E.g. csv (non proprietary). **(9)**

Metadata embedded with data. (8)

UK consistency – Questions, processes, etc. (Core set of questions. Consistent questions & quality & O/Ps & metadata). **(8)**

Data as data feed. Online item level open access data & mechanism for interacting with data. **(6)**

Review licencing model (including OS – concern**). (6)**

UK disclosure control – smarter. (Additive, better method properly applied – UK wide). **(5)**

Ability to beta test systems. (Pinning down specs, DDI or similar). **(4)**

Flexible outputs – not based upon (not being limited to) **previous concepts** (tabulations). **(3)**

QA of data. (3)

Error management mechanism. (Common understanding - Reporting/getting out updates). **(3)**

Encouraging 'The next 5 million users' – citizens & intermediaries. (2)

Free **access to Census software.** (As an ideal dependant upon BUS model). **(2)**

Sort data 'issues' first – all else follows. (1)

Information derived from the Census database that users would be unable to create for themselves. (1)

Reduce commissioned tables/outputs.