Setting Up a National On-line Census Data Service

by Virginia Knight [1]
Census Dissemination Unit
Manchester Computing Centre
University of Manchester

The ESRC, together with the Information Systems Committee of the Universities Funding Council, have set up a Census Initiative at a cost of some 3.1 million to make information from the 1991 Census available to academics for teaching and research. The Initiative has purchased data from the Census, set up units to support it, and funded programmes of research, training and deveolopment. As part of the Initiative, datasets derived from the Census are being held at Manchester Computing Centre and the Census Dissemination Unit has been set up from 1992 to 1997 to support them. Manchester Computing Centre, as well as supporting computing at Manchester University and UMIST, acts as a national computing centre for the dissemination and support of a number of large datasets, such as the General Household Survey and the Family Expenditure Survey. It was seen as an appropriate place to hold the 1991 Census datasets since they are large and complex and require specialist support.

A similar initiative existed for the 1981 Census. In this case, the data was distributed from regional computing centres at London, Bath, Edinburgh, Manchester, Aberdeen and Newcastle; this distributed scheme reflects the early developments in networks in the early 1980's. The data was held at the ESRC Data Archive, which, however, did not provide on-line access, although they had a national support post, whose holder ran courses produced documentation, supplied data to regional centres and gave advice. A software package, SASPAC, was developed by a team at Durham and Edinburgh Universities which could extract and manipulate the data. The data was held on the Manchester mainframe and Manchester Computing Centre produced a two-volume manual which described the SASPAC package, and explained how to run programs using the data from Manchester. It was apparent, however, that the support was required at the point of access and supply. The Computer Board (the predecessor of the ISC-UFC) funded an additional support post at Manchester to help users to access the data on-line and to develop the on-line service, for example by implementing specialist software for the Special Workplace and Migration Statistics, and to produce documentation and provide research support. The support post proved sucessful enough to establish Manchester as a national datasets centre with considerable expertise in hangling and analysing Census data.

For 1991, it was decided to combine the support posts which had existed at the Data Archive and at Manchester in a single unit.

The datasets supplied by the Census Offices under the Initiative comprise the following: the Small Area and Local Base Statistics, which are sets of tables covering all aspects of the Census, the Special Migration and Special Workplace Statistics, which are matrices dealing with migration and journey to work, a postcode/ED directory, and digitised boundary data (compiled by the EDLINE consortium) corresponding to various geographical zones for which Census data is being released. Although the sample of microdata from the 1991 Census is also being held at Manchester, it is being handled and supported by the Census Microdata Unit,which is entirely separate, though working in close collaboration with us. CHEST has also purchased the 1991 version of SASPAC, developed by the London Research Centre.

The timetable originally planned for the release of the datasets was a long way from what actually happened! Because of a misunderstanding in the coding of the term-time address of students, there was a six month delay in processing them at the Census Offices. The Small Area and Local Base Statistics should have been released from the beginning of 1992, so when I began in my present post in June of that year, we were expecting them to arrive within a few weeks. In July we received data for three counties, but on inspection we found that some of the tables were incomplete; for example, there appeared to be no people over 75! It turned out that the program which had written the raw data was defective, and OPCS had to produce a corrected version of it. The data produced by this started to come out at the end of August and arrived during the autumn at the rate of about two counties a week.

The Small Area and Local Base Statistics came out in two waves, because the tables are of two kinds; most tables cover the entire population, but some were only produced for a 10% sample. The 100% tables came out first, and were more or less complete at the beginning of this year; the 10% tables started to come out in January 1993, and we finished loading them at the end of April.

The 100% or 10% data for a county would typically

arrive on five magnetic tapes, making a total of over 700 tapes. After our tape librarian had allocated them a place in the tape library, I would copy the tapes on to a mini-disk. For England and Wales, the county and district level files were combined, and had to be split up. These would be raw data files, which would then have to be reloaded as SASPAC system files. This process could normally be completed within 24 hours. Inevitably, it would be held up if several counties arrived at once, and to avoid long delays, we devised a way of loading the data as system files offline. Thus I could be copying one lot of raw data files off the tapes while in the background another county was being converted into system files.

It was not all plain sailing, however. A number of errors in the raw data meant that the first attempt to convert it into system files failed. These were of two kinds; a displaced field in the header record at district level in nine counties, and impossible grid references in another ten. When this happened, the raw data would have to be emended. When the file concerned was an ED level file, it was too large for our screen editor to be used, and a FORTRAN program had to be written to change it. The Census Dissemination Unit had a valuable role to play in performing quality assurance on the data. The Census Offices are unable to do this themselves since they can only extract six areas at a time! There were also blank tapes, defective tapes, tapes with the wrong data on and files which were incomplete.

In addition, the sheer size of some files caused problems. The largest single file was the 100% ED level file for Outer London, which was 320 Mb. This was split across three tapes. Manchester Computing Centre is used to coping with multi-volume files, but a further problem arose when the file was too large to fit onto one of our backup cartridges. No one had ever needed to copy a file that large on to cartridge before, but eventually we worked out a way of splitting it between two cartridges.

SASPAC is at heart a PC package, and the mainframe version has had to be created from the PC version. This has been done at Manchester for our CMS system, and a similar process is going on here in Edinburgh.

At present Manchester Computing Centre is making the data available in two main forms; we are holding it on our mainframe both as raw data and as system files, and we are also supplying it as PC SASPAC system files.

The total raw data for Great Britain at all areal levels comes to some 6-7 gigabytes. We have been allocated a certain amount of minidisk space on the Amdahl at Manchester to hold it. This consists of 5 minidisks each with a capacity of 1.2 gigabytes. We have filled one and two-thirds of a second of these with system files, but there is not sufficient space on the others to keep all the raw data. As a compromise, we keep all the raw data files in prime space, except for the lowest level files (for Enumeration Districts/Output Areas) which are the largest ones. The complete, corrected raw data has been backed up on to cartridges which can be loaded by robot, and any files which are not online can be quickly re-called.

Our tape librarian has been creating PC SASPAC system files from the corrected raw data. These are not held on line; the only need to do this is if someone at another site wants to transfer the data in this form. If we receive such a request we can put the relevant file on line until it has been transferred.

We have a number of ways of sending the data to other sites. We encourage users to help themselves to the data they need where this is practical, firstly giving them access to the appropriate minidisks. Newcastle University have been successfully using TCP/IP coloured book to transfer all the raw data; FTP will also work, but is slower. We can send data to other sites ourselves, though the success of this depends on the bandwidth of the JANET link for the receiving machine. An attempt to FTP a large raw data file to Portsmouth University managed to bring their mainframe to a standstill. When it proves impossible to transfer data over JANET, we are prepared to send the data on floppy disks or on magnetic tape; for this service there is a media charge to cover the cost of the materials used and the time to copy the files. There is no charge for the data as such, because it has already been paid for under the Census Initiative.

Manchester Computing Centre has also developed a sideline in supplying data to Local Authorities in the form of PC SASPAC system files on floppy disk or tape streamer. This is because the Census Offices have been unable to cope with the demand for these datasets.

Most users are accessing the data on the Manchester mainframe. This requires that they get a Manchester userid, which can be obtained via a Manchester rep at the computing centre at their institution. Where there is no rep, we can deal with them directly. The advantages of this method are that all the data is quickly accessible on a powerful machine. Any problems with the data can be reported to me and sorted out directly; there is no problem with, for example, supplying a corrected version of a dataset to a remote machine. There is also a full range of support and a proper user note. Incidentally, it enables us to monitor use of the data easily, both by counting the number of users currently registered and by using the project accounting package VMACCOUNT to find out the amount of CPU time which has been used up and by whom.

The drawbacks, from the user's point of view, are that

they have to learn how to use the CMS system. Even a computer-literate user brought up on a different type of mainframe may find the nested screens in CMS and the XEDIT editor offputting at first. Moreover, some sites have difficulties in getting full-screen CMS on many of their terminals. One way of getting round this is to use JTMP to run jobs remotely. A job can be submitted from the user's own mainframe, sent to Manchester, and the output will be sent back along the same route. Once a workable JTMP framework has been set up, there is no need for the user to log on directly to CMS at all. Oxford and Bangor are using the JTMP method profitably.

A number of sites are holding the data on their own mainframe. A few are taking all of it; these include Newcastle (for the use of a specific project) and Edinburgh. Because of the size of the data this is a large commitment to space which is only worthwhile where a large number of people want to look at a lot of the data. Other sites are just taking a subset of the data, perhaps their own county and a few others nearby. If they are holding it on their mainframe, they have to devise their own implementation of SASPAC.

PC SASPAC is being made available under the CHEST deal to academic sites who want to hold it. The cost of a site licence is just under 3,000 for five years, and about 15 sites have taken up this offer. However, some sites which have have received the software under this deal have not requested any data, which suggests that they do not have the resources to run a properly supported 1991 Census data service. There are certain advantages to the PC version; it is menu-driven and people do not have to learn any computer system to be able to use it. Instead, they are guided through menus by selecting items. The drawbacks are that it is significantly slower than the mainframe version, and the user is limited to the amount of data which can be fitted on to a PC. Even allowing for the impressive compression which SASPAC performs on raw data, it still takes up a large amount of space. For example, the data for Cambridgeshire is reduced from over 100 MB of raw data files to 18 MB of PC system files. PC SASPAC is therefore not useful to someone who will want to study low-level data for a large number of counties, though it is suitable for someone who is only interested in one or two and not fond of computers! For this reason it has proved popular with Local Authorities.

There is another package for accessing the Census data, C91. This was developed by Powys County Council and runs on PC's only. We do not support C91 but we are prepared to supply raw data to people who have it. It is also possible to read the raw data directly using SPSS; this is not for the faint hearted!

Where sites are distributing the data locally, in whatever form, they must set up their own local registration system. We have produced guidelines for local registration forms and conditions of use, which should be based on ours, and approve the proposed forms and conditions before the system is set up.

The Census Offices in Great Britain are very concerned that the datasets should not be misused, and the registration procedure has been developed with this in mind. The conditions of use to which all users are required to agree insist among other things that users should not attempt to identify particular individuals or households, that the data should only be used for academic purposes, and that any publications which make extensive use of the data should acknowledge this. We act as 'gatekeepers' by advising on licensing arrangements for the use of the data. When a user is supplying data to commercial organisations, they must contact the Census Offices for permission and the status of their account at MCC (if they have one) will be affected. If they are using data on behalf of an organisation in the public sector, such as a local authority, the Sports Council or the Equal Opportunities Commission, there may be no need to pay for the data or to contact the Census Offices, but again they may have to pay for their use of the Manchester mainframe.

A typical user will contact me for a registration pack, or collect one from his or her institution. The registration pack contains a registration form, an order form for the SASPAC manual, the Conditions of Use and a description of the datasets by Professor Philip Rees, the head of the Census Initiative. In addition to the standard questions about name, address, etc. the registration form asks for a keyword description of the user's interests, the length of time for which the data is required and the source of funding. The completed form is then returned to me, and I add the user to the list of people who can access the minidisks on which the data is held. I return to them a copy of their form, together with a guide to accessing the data at MCC. Currently we have registered about 400 people by this method. Most of our users are in geography, social science and medical departments.

The Census Offices require that anyone who is using the data should assent to the Conditions of Use, and to cut down on paperwork, we have devised a class registration form, to be used when a class is accessing the data. The lecturer completes the first page, and then the students should all sign on the next one. Similarly, if one person is obtaining data on behalf of another, they should both be registered. This often happens, for example, when a research assistant pulls information off for a less computer literate academic.

The postcode to ED directory has been put on line as a set of raw data files, one for each county and one for the

whole of England/Wales, together with an online guide to them.

We also offer various kinds of support to users. I have devised a half-day course on the Census which I have given at Manchester and at about twenty-five other institutions. The audiences for these have been made up of people who are already registered as well as potential users. The course consists of an hour talking about the Census, the Initiative and the various datasets available, followed by another hour of demonstration of SASPAC programs and sometimes hands-on exercises at the end. The handout accompanying the course, which contains about fifteen sample programs demonstrating the functionality of SASPAC, has proved a useful supplement to the manual, where the examples are less geared to the needs of the typical academic user. I have also given presentations at workshops at Cardiff run under the Census Initiative and spoken at and attended other relevant meetings.

We run an email list on the Census datasets, which is to be transferred to the MAILBASE system at Newcastle; any message sent to the list will be sent as email to all subscribers. It is intended that any subscriber can post a message to the list; however, most of the messages have been from me, telling the list about the latest data to arrive and about future runnings of my course. This has proved an efficient way of informing users about latest developments, and was particularly important when the data was still arriving. There is a bulletin box which can be displayed on CMS using the command CENSUS91; this is also displayed whenever a SASPAC job is run. We have devised a 'help' system which will look up the identifying codes for districts and wards; like other help systems at Manchester, this works by presenting the user with a list of options, one of which can be selected with the cursor. We intend to develop another help system which will give sample programs, based on the examples in my handout. There are a number of helpful files online, which can be read and downloaded by registered users. There is also a section on the NISS bulletin board, which is occasionally updated as further data arrives; to date, it has been consulted over 300 times.

Manchester Computing Centre has reprinted the SASPAC manual produced by the London Research Centre, which comes in two volumes, one describing the language and one giving outlines of the tables. Roz Scott Huxley at MCC has also written a user note for SASPAC on the Manchester mainframe; in addition to explaining commands relating to SASPAC, it also assumes the user is not familiar with CMS and the Xedit editor, and takes them step by step through creating a command file, running it interactively, offline or remotely and collecting the output. I also write regularly for the local and national newsletters produced by MCC, and for the ESRC Data Archive Bulletin.

We are also prepared to answer queries from people who are having problems in using SASPAC, whether by telephone or email. There is not a very large volume of these, as many people get help from another user at their institution. Often the problem can be solved at once; other questions have shown up problems with the data or with the SASPAC software. In the former case these can be referred back to the Census Offices and in the latter to the London Research Centre.

Future work will include loading the Special Migration Statistics, the Special Workplace Statistics, the digitised boundary data and the appropriate registration procedures for them as well as developing the online help system. We will also have to transfer the data to the new machine at Manchester when it is installed towards the end of 1993 and implement SASPAC on the UNIX system which will be mounted on it.

1. Presented at IASSIST/IFDO 93 Conference held in Edinburgh, Scotland. May 1993.