Working Paper No. 11
ENGLISH ONLY

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE-EUROSTAT Work Session on Registers and Administrative Records for Social and Demographic Statistics
(Geneva, 9-11 December 2002)

Invited paper – SESSION 1

THE ONS LONGITUDINAL STUDY, QUALITY ISSUES FROM 30 YEARS OF DATA LINKAGE

Submitted by Jillian Smith, Louisa Blackwell and Kevin Lynch,
Office for National Statistics, UK

Introduction

1. The ONS Longitudinal Study (LS) contains anonymised linked census and life event data for one per cent of the population of England and Wales. Life events include births, deaths, widow and widower-hoods, cancer registrations, migration. The LS sample was originally drawn from the 1971 Census of Population by taking all people born on one of four selected dates of birth in the calendar year. These four dates were used to draw the sample again at the 1981, 1991 and 2001 Censuses and to link routine event registrations prospectively throughout the period of the study. New LS members enter the study through birth and immigration. Exits from the study, through death and emigration, are recorded as events on the database. There are over 500, 000 LS members at each census and the size and scope of the study permit multi-cohort analysis of microdata over time. The LS uses routinely collected administrative data, so there is no respondent burden.

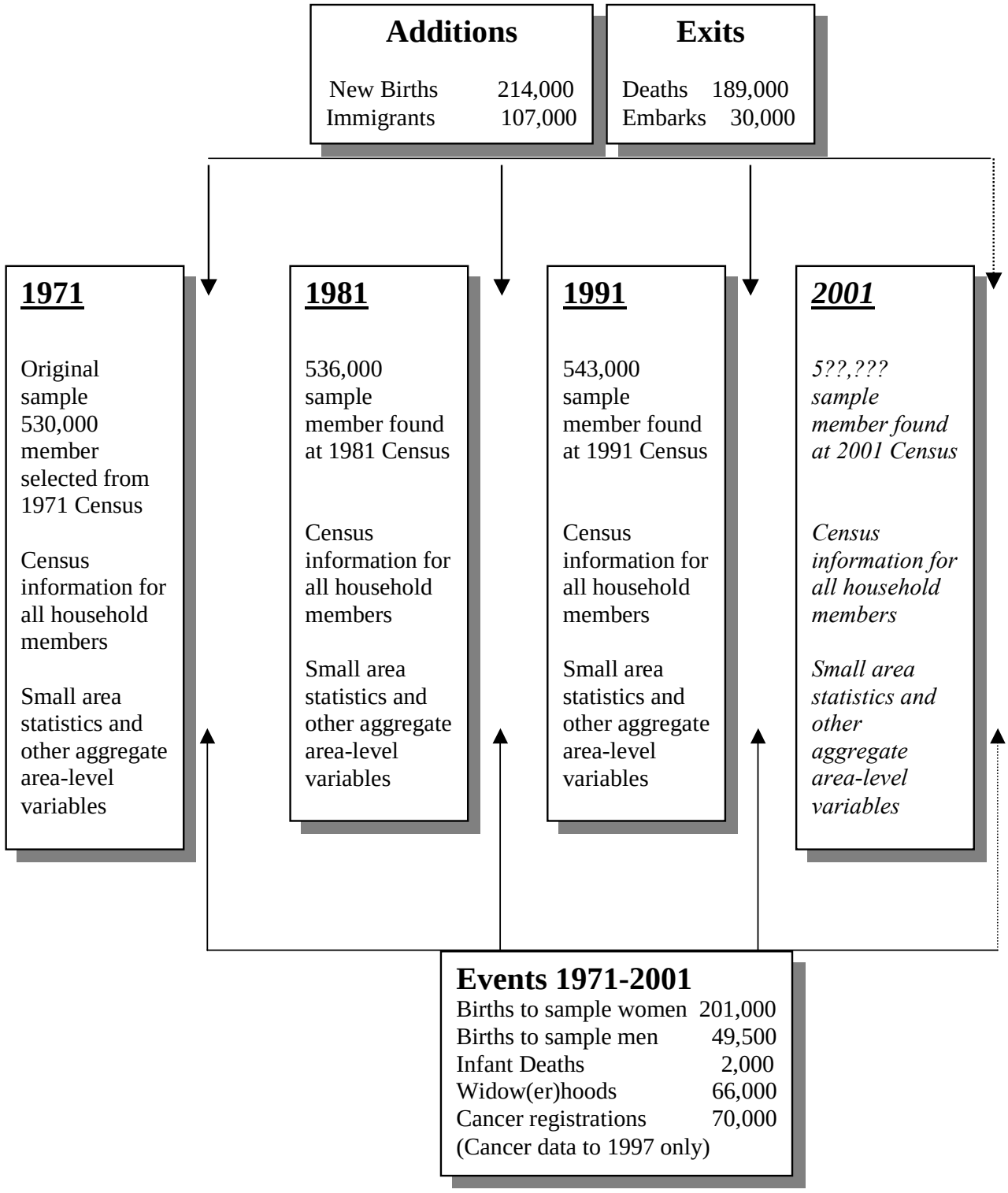
2. For thirty years this dataset has been prepared, maintained and developed to a high standard of quality to ensure maximum representativeness and analytical capability. The recent reconceptualisation of quality, including work within the European Statistical System (see for example Linden and Sonnberger 2002, Linacre, 2002), has produced a framework for managing, assessing, reporting and improving quality in statistics. The quality model has seven often overlapping dimensions: relevance, accuracy, timeliness, accessibility and clarity, comparability, coherence and completeness. This paper describes some of the ways by which the management of the LS addresses the different elements of this framework, using examples from the current linkage of the 2001 Census data. The 2001 data will be available to users in 2004. Quality covers both process quality and quality of outputs, and there is often a tension between the different quality dimensions within these two interrelated domains. We discuss ways in which they are being negotiated and resolved in the 2001 Link project. An outline of the LS structure provides the context for this discussion.

The data

3. When the 2001 Census link project is complete the LS database will include information from four census samples: 1971, 1981, 1991 and 2001. The entire census record for the LS member and all members of that person's household are included. This contrasts with other census datasets which are adjusted, aggregated or otherwise limited for confidentiality reasons. Vital event information from ONS registration systems has been linked to individuals in the sample since 1971. The LS sample is replenished with births on the four LS birth dates and immigrants with LS dates of birth. Members leave the sample through death or embarkation, though it should be noted that their records are not deleted from the study, but linked to an exit event. Figure 1 lists other major events that are linked, including births and infant deaths registered to women in the study, cancer registrations, widowhoods and widowerhoods.

4. For reasons of confidentiality, the LS analysis is conducted using an anonymised analysis database which is quite separate from the data capture systems. The linkage of source data at individual level into the LS is achieved through an intermediary register, the National Health Service Central Register (NHSCR), in which the LS members are flagged and numbered with a unique reference number. Event data are added regularly into the LS database by comparing the incoming event information against the register using name and date of birth. Where the records are found to be for the same person, they are said to be successfully 'traced'. The LS reference number is then appended to the event file. After this it is a straightforward task to add the new event data into the main LS database. For some events the method of linkage is by direct search, by date of birth, of the registration databases. Further details of the linkage systems are given elsewhere (Smith, 1999)

Figure 1 The structure of the ONS Longitudinal Study



5. Every ten years since 1971, a large linkage exercise has been carried out to link the appropriate census sample into the LS, again using the NHSCR. Currently (November 2002) the linkage of the 2001 Census data is underway, for the first time using automatic tracing techniques. A description of the process follows. The census sample is drawn from the census data on the four dates of birth. This sample is compared to the NHSCR data using key information such as name and date of birth. Data tracing is achieved through an automated process and then, for more difficult records, through manual tracing. New members are flagged on the NHSCR and those found to be existing LS members have their reference number copied from the NHSCR to the census file. The flagged census file is then incorporated into the LS database. The automation of the first part of the data comparison work has had a major impact on the efficiency of the whole exercise, cutting cost significantly. A further important development in 2001 is that staff who are tracing records manually can view an image of the related census form. This enables them to check for misspellings or variations in details within the form and trace people through other household members. As a result we now anticipate a final trace rate in excess of 99%. This compares with 98% in 1991 and 97% in 1981 and 1971.

6. The 2001 Census link project has required continuous liaison with our data users and suppliers and between three LS teams working across separate ONS office sites (located in England in London, Southport and Titchfield). We now demonstrate ways in which LS quality is managed in this process with reference to the seven dimensions of data quality identified in the proposed ONS quality measurement and reporting protocol (Full, 2002).

Relevance

7. The 2001 census link into the LS was initiated by a comprehensive review of the study which made recommendations for the next decade. This was conducted in 1998 (Review, 1999), with extensive user consultation, and set the scene for improved accessibility, greater efficiency, automation of the linkage process, greater involvement of users and a series of further reviews of security, and more. Appendix 1 sets out the recommendation of this review.

8. Originally set up to enhance the study of mortality, the LS has also been used to address a broad range of social policy concerns, including fertility, population distribution and migration, population and household analysis and forecasting, housing and equal opportunities. A list of publications that use the LS can be seen at http://www.celsius.lshtm.ac.uk/publications.html. Existing users include civil servants and academics from a range of disciplines and with varying research interests.

9. ONS ensures that the data are made as accessible as possible, given stringent confidentiality procedures, through the provision of a 'safe setting' environment. The complex nature of the data and confidentiality requirements mean that users access the LS through research support staff. ONS staff support government, other public sector users and occasionally private researchers. Academic users are supported by the Centre for Longitudinal Study Information and User Support (Celsius), based at the London School of Hygiene and Tropical Medicine. Both teams facilitate consultation and feedback between LS users and the LS staff responsible for data management and development. This helps users to recognise data strengths and limitations and interpret and use the data appropriately. It also provides a user perspective on issues of data processing quality. As well as formal user representation on, for example, the LS Steering group, the LS Research Board and the LS Development Programme, daily informal contact and issue-based consultations provide opportunities for this vital two-way process.

10. As an example, users were consulted over the degree of double coding required for 2001 births and deaths event data. Double coding to two classifications, in this instance the 1990 and 2000 Occupational Classifications, permits an analysis of data comparability over a period when the classifications used have changed. However the requirement for the dual coding work coincided with the intensive work period of the 2001 Census link processing. Through consultation with users who will use dual coded data to assess comparability we were able to limit the dual coded sample to pre- Census data only (that is, 2001 events records up to April 29). This creates a dual coded subsample closely matched to user requirements, which also limits the burden on the LS processing team at their busiest time.

Accuracy

11. It is essential that users have a comprehensive account of LS data accuracy so that they can evaluate their strengths and limitations, and use the data appropriately in their analyses. The LS data linkage that ONS carries out represents a secondary process on primary sources. The primary sources are census and vital statistics. We therefore have a duty to disseminate to users details of primary data source accuracy, which is routinely reported by primary data producers, as well as details of LS data linkage accuracy. In the 2001 Census Link project we are engaged in an ongoing process of quality management that aims to strike the correct balance between meeting users' data needs by minimising, where possible, non sampling error on the one hand, and resource costs on the other (Ruddock, 1999). These decisions are taken in the context of paramount attention, at all times, to issues of data security and confidentiality.

12. LS Data quality reports following previous censuses have involved calculating tracing rates, overall linkage rates and sampling fractions for different socio-demographic groups. These tables indicate the extent of census under-enumeration for different groups (Hattersley and Creeser, 1995). For example young males from minority ethnic groups are usually undercounted. Tracing rates vary and, for example, young women who have married are often difficult to trace if their change of marital status involves geographic mobility and a change of surname. Cross-sectional no-trace and under-enumeration rates are compounded in data that are linked over time. Overall linkage rates reflect both of these and indicate the representativeness of the longitudinal LS sample. They provide cautionary guidance for users, for example on the reliability of using the LS for studies of migration and ethnic minorities (see for example Blackwell, 2000). In addition to recalculating these standard quality indicators for the 2001 data, we shall provide an account of the implications that the edit and imputation rules have for LS data and analysis.

13. An important issue for the LS is that all the data added to the study is collected for a different purpose. The UK Census is first and foremost a periodic headcount of the population that provides a population base (denominator) for the distribution of public resources. The LS receives census information after its primary purpose has been achieved and has only small influence on its collection. It is therefore necessary to ensure adequate preparation of the data for its longitudinal function. The 2001 LS/Census sample has field-level imputation which replaces missing information with data for other individuals using a series of matching criteria to find an appropriate substitute. Clearly for a longitudinally linked dataset, such as the LS, such imputation would introduce spurious personal characteristics, introducing serious inconsistencies across time. As part of the 2001 Census link project we have therefore devised a series of flags to identify where fields have been imputed, and an important aspect of data quality reporting will involve quantifying the degree of data imputation and providing guidance to users on how imputed data should be treated in analysis. We envisage that users will need to take a reflexive approach to the use of imputed data. For example in the case of derived variables that are flagged as imputed, users may wish to interrogate the degree of imputation in the contributory fields to assess the relevance of imputation in those particular fields for their research question. A specific example would be the economic activity field, which distinguishes between full and part-time employees, self-employed workers and economically active students, the unemployed and different categories of economically inactive people. Researchers using this variable to identify LS members in employment may not be concerned if their hours worked are imputed: including records where only the number of hours worked is imputed will therefore boost the number of usable records.

Timeliness

14. Users require accurate data that are available at pre-established dates. There is a time lag between the availability of primary data and its appearance in the LS database because of the linkage processes and rigorous checking involved. The LS Development Programme is managed under formal project management procedures to ensure delivery against deadlines. Within the programme, work is planned and implemented as a series of discrete projects and workpackages.

15. For example, the operational tracing of the Link project has been managed using a bespoke monitoring system which allows weekly updates of tracing progress and rates. At every stage of the

linkage process the inherent trade-off between accuracy, timeliness and cost has to be decided. With the excellent reference systems available to the NHSCR it is possible to extend the time taken to trace each record with a (diminishing) chance of ultimately finding the person successfully. However the constraint of cost dictates that the time spent is limited.

16. Similarly, in drawing the LS sample from the 2001 Census, inevitably some LS records are missed due to errors in stated or scanned date of birth. A return to the Census for an additional sample to adjust the LS sample would be difficult, expensive, incomplete and time consuming. However, investigation and reporting for data quality purposes is important and can be achieved outside the critical time path of the Link project thus avoiding any delay to the delivery of the final dataset.

Accessibility and clarity

17. The LS is provided free at the point of use to analysts. This facilitates accessibility and supports ONS' aim to maximise the use of the dataset subject to the paramount requirement to protect confidentiality.

18. There is a stringent requirement for comprehensive documentation about the LS processes and data, both for internal ONS use and for users of the database. The cyclical nature of the census-link work makes it particularly important that processes and decisions are fully documented. It cannot be assumed that individuals' experience of the 2001 LS/CensusLink project will be available for future census link work, unless it is recorded. In addition, comprehensive documentation of our record linkage experiences can inform the methodology for further record linkage work that the ONS may conduct in future. Whilst the ten year cycle of the LS poses some risk to quality management in that there is considerable discontinuity of staffing over the decennial cycle, it also lends itself to a natural spiral of rising standards of quality. Many of these are related to computer systems, where the advancing ability to manage large and complicated datasets has revolutionised analysis in the last couple of decades. The LS started out as a mainframe based series of flat files, but is now maintained as a fully relational database on a PC network system.

19. Users need to understand the methodological issues that have some bearing on data quality. The LS documentation project aims to provide details of the source data, link methodology, output data quality and formal metadata for the LS database that is geared towards different levels of user need. In keeping with ONS' commitment to web dissemination we plan to make this information available through the National Statistics website. This medium will also permit the re-use and greater availability of existing documentation, as .pdf files linked to the web pages. There is also a case for producing a hard copy version of the new metadata, for those who do not yet have the necessary degree of access to web technology. Users are being consulted on quality measurement methodology and the design of the proposed web pages to ensure that the information is provided in a user-friendly way, that it is relevant to user needs, and that it interfaces with other sources including the Celsius and National Statistics websites.

20. In addition we plan to enhance users' awareness of the new data through their active involvement in beta-testing. A series of data tests is planned to ensure the integrity of the database before release to users. We also plan to involve experienced users in a series of data testing projects through which new data quality issues will emerge. We shall implement an infrastructure for data quality communication between the development team and users so that each issue will be assessed, addressed, documented and disseminated in an iterative way, as appropriate to the specific issue.

Comparability

21. The LS has to address several dimensions of comparability: comparability in the data collection process, comparability of data outputs over time, comparability with other data outputs in the UK and comparability with similar datasets in other national contexts. Internal comparability overlaps with the issue of coherence and is discussed below.

22. Comparability over time is a major concern that confronts users of all longitudinal and time series data that span changes in classificatory and collection systems. It is useful for users to be able to assess comparability by repeating their analyses on dual coded data. A current concern for the development team is the provision of 2001 occupational data coded to both the 1990 and 2000 Standard Occupational Classifications (SOCs). SOC 2000 coding has been provided by Census but the LS team has responsibility for SOC90 coding. In producing these codes we shall meet (and hopefully exceed) the minimum quality standard set by the Census coding Service Level Agreement. The project is in the final stages of preparing an automatic coding process, which among other techniques uses the existing census SOC 2000 code to verify the SOC 90 code. After automatic coding the residual will be referred for manual coding.

23. We also aim to enhance users' ability to compare the LS with other UK datasets. An integral part of the LS data accuracy measurement involves comparing LS sample distributions with Census distributions. These are reported as net sampling fractions:

$$\text{Net sampling fraction } \alpha_i = \frac{\text{number in sub-group i (traced LS population) x 100}}{\text{number in sub-group i in census population}}$$

(Hattersley and Creeser, 1995, p180).

24. Sub-groups that are difficult to trace (young women and men of working age, over-75s, the economically inactive and some minority ethnic groups) affect comparability of the LS sample, and this is compounded in longitudinal data (as described in relation to accuracy, above). The 2001 One Number Census estimate (Brown et al, 1999) has led ONS to recalculate mid-year population estimates as far back as 1982. To assess LS comparability we also plan to produce two sets of sampling fractions for 1991, one with the original Census count in the denominator and the other using the revised 1991 ONC estimate.

25. Similar datasets to the LS exist or can be prepared in a number of countries, including Canada, Denmark, Finland, France, Iceland, Italy, Israel, the Netherlands, Norway, Sweden and the USA (see the Review, 1999). These considerable statistical resources have the potential for new research on critical issues such as health inequalities, fertility patterns, social inequality and social exclusion on an international level. The task of documenting data comparability and issues of data harmonisation to facilitate this research requires resources that are beyond the scope of the LS Development Team. ONS has submitted an Expression of Interest in partnership with the University of London's Institute of Education to conduct this work under the European Commission's Sixth Framework Programme.

Coherence

26. The LS relies on data collected primarily for other reasons, as discussed in relation to data accuracy above. Census processes not only impact on LS data accuracy, but also on its internal consistency when compared with vital events data. The registration systems are in place to record key circumstances of peoples' lives, such as birth and death, required for legal and social reasons. For this reason the LS has a particular shape and content which facilitates some analyses and precludes others. For example, uniquely the LS links mortality to the extensive life style and circumstances information of the censuses, and it does this for a very large sample in comparison to most surveys. On the other hand, it is currently not possible to link marriage registration data into the LS because the key linkage information, date of birth, is not collected.

27. There are subtle differences in the data collection processes that can pose problems for users. For example the 2001 Census will provide information on the occupation of all household members aged 16-74 years if they worked. However, at birth registration until 1993 only the occupations of employed fathers, of wives of unemployed or economically inactive fathers or of lone mothers were collected. This non-comparability poses problems for researchers of maternal employment. It is overcome in fertility analyses, for example, by relying solely upon mothers' occupations at census. But this approach obscures

occupational mobility around childbirth and means that the data quality varies depending on the time lapse between the birth and census dates.

28. The ONS has initiated a number of re-engineering projects that aim to harmonise definitions, classifications and methodological standards across all its activities. Once fully operational, these programmes with substantially improve contemporary coherence within the LS. However assessing and reporting on comparability between historic and emerging data will continue. The conflict between data relevance, in terms of user needs for current, comparable and harmonised concepts and methodologies, and data coherence, is particularly challenging in a longitudinal dataset that encompasses thirty years of data. This issue emerges even within the structure of the documentation of the LS. On the one hand the LS team would like to follow metadata guidelines for harmonisation across all contemporary datasets. On the other hand we must maintain consistency with older data collections, significantly the previous censuses. We have resolved this issue by a pragmatic approach which follows contemporary guidelines if possible but adapts them where consistency across time is necessary.

Completeness

29. The concept of completeness is linked to relevance and requires that ONS statistical provisions should meet the needs and priorities of users, subject to the requirement to maintain confidentiality. The LS has plans to add further datasets to the study as and when this is acceptable and possible. In addition users are encouraged to bring aggregate data to add to the study to enhance analysis, within the safe setting environment. For example, many deprivation indices, life-style classifications and small area statistics have been added to the study.

30. The LS has great potential for use in conjunction with other datasets, where together their individual deficiencies can be offset. In a 1991 analysis, the 1958 cohort study was used to show that smoking patterns of mothers in that study were associated with similar socio-economic differences as those found for women's lung cancer mortality in the LS (Pugh, Power, Goldblatt and Arber, 1991). More recently the LS, the General Household Survey and national fertility statistics have been used in conjunction to model cohort specific fertility rates and trends and to predict parity progression ratios (Rendall et al, 2002).

Conclusion

31. Quality management in the LS involves the monitoring of processes and products in a cycle of continuous improvement. Ongoing liaison between the three LS teams in Southport, Titchfield and London highlights processing issues where there is scope for data quality improvement. This has been illustrated here by, among others, the automated tracing processes and new imputation procedures implemented by Census. Inevitably many of these improvements take place behind the scenes. It is therefore vital that the impact of changes made is assessed, documented and communicated to users. In turn continuous user liaison, on a formal, informal and ad hoc basis helps to inform processing decisions and the allocation of project resources. Comprehensive and accurate documentation of data quality is essential for both LS production and dissemination:

> ‘Interpretability is perhaps the one dimension of quality where the NSO *(National Statistical Agency)* should aim to do more than the user is asking. There is an element of user education in the provision of metadata. Spreading the message that all data should be used carefully, and providing the information needed to use data with care, is a responsibility of the NSO that goes beyond simply providing what users seek’. (Brackstone, 1999).

The cyclical and long term nature of the LS allow scope for a continuous virtuous circle of quality improvement. At every stage of the processing, opportunities for new techniques, new technologies and new transmission of information present themselves. This process of proactive, continuous change is demanding and relies crucially upon the contribution of a dedicated and expert staff, which we are fortunate to have in the LS.

Appendix 1 - Recommendations from the Review of the ONS Longitudinal Study 1998

Primary Recommendations

- Maintaining the confidentiality of the LS remains the first priority (ref. chapter 6).
- The 2001 LS/Census link is made and the following 10 years of vital events are added to the study (ref. chapter 11).

Full List of Recommendations

Confidentiality and Privacy

1. Maintaining the confidentiality of the LS remains the first priority (ref. chapter 6).
2. Explicit reference to the LS is made in the forthcoming Census White Paper (ref. chapter 6).
3. A provision for consultancy on confidentiality matters relating to the LS should be available to ensure best practise is being followed and to provide external affirmation that the confidentiality provisions are rigorous and will withstand scrutiny (ref. chapter 6).
4. All means of gaining the support of public interest organisations for the study should be explored, starting with an approach to the Data Protection Registrar (ref. chapter 6).

2001 LS/Census Link

5. The 2001 LS/Census link is made and the following 10 years of vital events are added to the study (ref. chapter 11).
6. Census planning for 2001 continues to recognise the LS requirements for consistent definitions and classifications over time, eg. for occupational classifications and household definitions (ref. chapter 4).
7. Decisions on an automatic link, using computer matching techniques for name and date of birth for the 2001 LS/Census link, recognise the estimated savings involved and are also taken as soon as possible (ref. chapter 8).
8. Funding decisions for the 2001 LS/Census link recognise the value of maintaining business as usual work at the same time, in order to maximise timely availability of the post 2001 link LS database and documentation for early analysis (ref. chapter 8).

Promoting Use of the LS

9. Increased use is promoted through a new programme of seminars and information dissemination (ref. chapter 11).
10. To ensure greater public use is made of the LS, a target should be set for increasing the use made of the LS data by government to support public policy analysis, development and/or evaluation (ref. chapter 3).

Reducing Costs

11. Planning proceeds for automation of interfaces in LS data capture systems in order to achieve long term savings (ref. chapter 10).
12. Efficiency gains in occupational coding on event data are explored (ref. chapter 10).
13. Recovery of academic computing costs is sought (ref. chapter 10).
14. The LS team explore options for reducing computing costs, to the benefit of both academic and non-academic users (ref. chapter 10).
15. In total, cost savings of 20 per cent are sought (ref. chapter 10).

Information and Access

16. To assist users in awareness of data deficiencies, the evaluation and documentation of immigration and emigration data is made a priority area for future work (ref. chapter 4).

17. Subject to confidentiality requirements, proactive work to improve accessibility to the LS by technological or other means is continued (ref. chapter 7).

18. Training activities aimed at LS users are continued and, resources permitting, increased. This will assist users to make full use of the LS access provisions already in place (ref. chapter 7).

19. The possibility of a freely accessible demonstration dataset should continue to be investigated (ref. chapter 7).

Further Data Links

20. A policy for links between data sources and their use and funding by government and others is defined in the context of the wider remit of ONS and the definition of an independent National Statistical Service (ref. chapter 7).

21. The statistical policy case for linking the JUVOS data into the LS is further developed as a pathfinder for other data links (ref. chapter 7).

22. Further linkage options of data to the LS should be discussed with the Department of Social Security and Department for Education and Employment (ref. chapter 7).

References

Blackwell, L. (2000) ‘Fragmented life courses: the changing profile of Britain’s ethnic populations’, Population Trends, 101.

Brackstone, G. (1999) ‘Managing Data Quality in a Statistical Agency’, Survey Methodology, 25.2:139-149.

Brown, J.J., Buckner, L., Diamond, I.D., Chambers, R. and Teague, A., (1999) ' A methodological strategy for a one number census in the UK', Journal of the Royal Statistical Society, Series A, 162:247-67.

Full, S. (2002) ‘Towards a framework for Quality Measurement and Reporting within the Office for National Statistics (ONS)’, paper presented to the GSS Methodology Conference, July 2002, London.

Hattersley, L. and Creeser, R. (1995) Longitudinal Study 1971-1991: History, organisation and quality of data, London: HMSO

Linacre, S. (2002) ‘Achieving Quality: a many pronged strategy’ paper presented to the GSS Methodology Conference, July 2002, London.

Linden, H. and Sonnberger, H. (2002) ‘Assessment of data quality and LEG on quality recommendation’, presented to the UNECE/ Eurostat Work Session on Metadata, March 2002, Luxembourg.

GSS (1997) Statistical Quality Checklist , London: ONS.

ONS (1999) Review of the ONS Longitudinal Study 1998, London: HMSO.

Pugh, H., Power, C., Goldblatt, P., and Arber, S., (1991) ' Women's lung cancer mortality, socio-economic status and changing smoking patterns' Social Science and Medicine, vol. 32, no. 10, pp 1105-1110

Rees, P., Martin, D., Williamson, P. (Eds), (2002), the Census Data System, London:Wiley.

Redall, M., Smallwood, S. and Joshi,H., (2002), 'Use of the Longitudinal Study in combination with the General Household Survey and National Fertility Statistics', paper presented at a Quality Issues in Social Surveys seminar, London, 31.10.2002.

Ruddock, V. (1999) Measuring and improving data quality, GSS Methodology Series no.14, London:National Statistics .

Smith, J. (1999) The history and future of record linkage in the ONS Longitudinal Study, Statistical Journal of the UNECE 16, Amsterdam: IOS Press.

- - - - -