# Possible Futures for The ONS Longitudinal Study in a Post-Census Landscape

The LS is an approximately 1% sample of the population of England and Wales and contains census data from 1971 onwards (for as many censuses as sample members were alive, present (and in the census) in England and Wales, and successfully linked), together with vital events data such as births and deaths. The Centre for Longitudinal Study Information and User Support (CeLSIUS) facilitates research use of this sample, predominantly but not exclusively for UK based academic researchers; we support users in drafting their applications to use the LS, through giving expert advice and guidance during analysis and in training, and in disclosure checking users' output; this paper is written by CeLSIUS members, and is informed by many years of experience in supporting research users.

The LS in its current form relies fundamentally on decennial expansion through the linking of additional census data. In order to assess the immediate impacts of a cessation of census data additions, we have identified a set of variables and groups of variables that are based on census questions, for which reliable production of data from administrative sources might be difficult to achieve. Thus we ignore variables such as age and sex, despite being almost ubiquitous in analysis extracts, as administrative sources can be demonstrated to be usable. The variable selection is primarily informed by an illustration in ONS' consultation document[1] which classifies variables by both extent of coverage in administrative sources, and progress by ONS in researching their potential for providing a viable alternative to census sources. In Table 1 we give totals by variable, although most projects will use more than one of these at the same time; the analysis reflects 41 open projects in total. We have limited our analysis to those projects using the 2011 versions of these variables and have excluded completed projects; this reflects current usage and we note that interest in characteristics can change over time.

Looking at the five most commonly used variables in our selection, country of birth is the most widely used of these 'variables of concern', being used in over half of the projects we studied. Relationships within the household are also widely used in LS based research, but that the capture of this information in admin data is partial. Some data sources might usefully enumerate the numbers of people in a household by age and sex, but without giving comprehensive information on the relationships between people, thus leaving normative interpretations of family structures as the only

[1] https://consultations.ons.gov.uk/ons/futureofpopulationandmigrationstatistics/

likely classificatory option. Educational qualifications are again widely used in research. Here, we agree that linkage to pupil database and HESA records may produce good quality data – indeed, likely to be richer than census data - for recent qualifications, but we are concerned about capture of data for qualifications gained longer ago, or outside the UK.

Occupation is frequently used, and we are particularly concerned about limitations on collecting detailed data about occupation including not just job titles but the place of work and means of travel to work. We note that many researchers used variables such as NS-SEC to explore social mobility, and that this is inherently dependent on occupation data. The final variable in the 'top 5' is self-rated health. This is of value and should be seen as distinct from admin data arising from healthcare and clinical diagnoses; as with data on qualifications we note that successful linkage with health-related administrative sources such as Hospital Episode Statistics, or prescribing data, may on the otherhand offer rich new possibilities for research.

*Table 1 Open LS projects that make use of selected variables*

| **Variable** | **Projects** |
|---|---|
| Country of birth | 28 |
| Relationships in household | 26 |
| Educational level / qualifications | 24 |
| Occupation / social status | 23 |
| Self-rated health | 18 |
| Household tenure | 15 |
| LLTI or disability | 13 |
| National identity / ethnic group | 11 |
| Economic status / working activity | 10 |
| Welsh language use | 8 |
| Communal establishments | 7 |
| Caregiving | 6 |
| Transport available / distance travelled | 5 |
| Religion | 5 |
| Visitors staying over on census night | 4 |
| Passports held | 3 |
| Main language | 3 |
| Second address | 3 |

As we have noted above, the LS in its current form is dependent on the decennial census for extension into the future. The proposals refer to the LS as is, and also to a proposed LPD that linked the whole population. This paper will be expanded to include a full analysis of a number of scenarios, and we note that by the time of the conference we are likely to know much more about the future of the census, allowing scenarios to be more focussed.

Considering the LS as is, we can reflect on how a move away from a traditional census might be accommodated. The proposals are built around the Dynamic Population Model, which incorporates the Statistical Population Dataset enhanced with surveys etc., having gone through a de-duplication process to arrive at a 'correct' residential location for individuals. If we assume (although this does not seem guaranteed) that the DPM incorporates usable identifiers for linkage that include date of birth as well as address etc., then it would be feasible to extract a subset of those individuals who are LS sample members and to 'roll forward' the LS with admin supplements. However, this would only be comprehensive for the core variables in the SPD. Whilst surveys could quite usefully be exploited to model aggregate estimates of a variety of characteristics, they could not be attached to LS members except in the low-probability case that a person in the DPM was both an LS member *and* included in a survey. Where the DPM was linked to comprehensive records (rather than a survey), then there would be scope to extract records for LS sample members. This might provide a regular stream (perhaps inconsistent) of some variables such as address. It is unclear to us whether responses to census-like questions typically asked for equality and diversity monitoring (such as ethnic group, nationality and disability status) in healthcare, educational and other contexts could legally be used in linked assets. We do therefore see a possible future for the ONS LS which would continue to link events for sample members from various sources. However, a significant qualifier is that the variables listed in Table 1, widely used by many researchers, could not necessarily be replicated – this would remain a problem for an ongoing LS, as well as for ongoing aggregate data estimates.

The consultation document also describes the creation of a linked 'full population' dataset. This is an exciting prospect, and one which offers potential, but which also promotes a variety of questions. Some of these are practical – the modes of access, and user support – whilst others are methodological. Our understanding is that linkage would be done on an automated basis. Our view is that the existing 1% LS sample should be retained – as described above – with high quality clerically checked linkage – and that this be used to calibrate the automated full population linkage, in particular, to identify biases in linkage. This would permit a way forward for the LS that would enable continuing analysis of the sample members with replenishment of the sample, and at the same time lend strength to the broader automated sample. The LS would remain distinct in having

In summary, the proposals offer both opportunities and threats to research using the LS. The potential for new linkages is promising and would address users' wish-list items voiced over many years. However, any discussion of new linkages has always been in the context of them being supplementary to the usual array of census variables, rather than being – as proposed – a partial substitution. We have encouraged research users of the LS to submit their individual views to this

consultation, but in reviewing the range of variables for which there is currently no tested reliable alternative source, and the reliance of many users' research on them, it is difficult for us to fully endorse the proposals.