FDS Briefing note 4

Orphaned datasets

Author: Felix Ritchie

In general purpose data archives such as the Secure Research Service (SRS) or the UK Data Service SecureLab, the provenance of datasets may become unclear. This can be for three reasons:

- the organisation that deposited the data no longer exists
- the records of the circumstances of data collection have been lost or destroyed
- the processes to transform the data for the source datasets into the researcher datasets have been lost or destroyed

We refer to the first two as 'unknown owner' and the latter as 'unknown process'. In these circumstances, the archive may consider whether it has the legal right or duty to make these available for research, and, if so, how should it manage them.

This briefing note reviews the considerations. We assume that the datasets have research value.

Who owns the data and can make decisions about sharing?

For the unknown owner case, this depends upon the level of detail in the data and the gateway through which the data was originally accessed. If the data was deposited in the archive with the intention of making it available for research in perpetuity (assuming this was consistent with the gateway used and/or any consent obtained from data subjects) ands with the archive having full control over access, then the absence of the original depositor should not prevent further use.

If however the deposit was time-limited, there are reasons to suspect that the gateway does not allow unlimited onward use, or the archive is required to ask for the depositor's approval for access, there is not an automatic right. It is quite likely that there is a lawful gateway, because data for research use can be relatively freely circulated under UK law, but it would be wise to get a formal statement. It is also worth considering the ethical case: would it have been reasonably expected that the data would (continue to) be used in this way? Again, if the data were collected for research used, this is likely to be a low bar to cross but it should be checked.

For the unknown processor case, this is much simpler. The archive is aware of what the data can be used for. The fact that has undergone additional processing does not affect the archive role as controller of the data.

For example, consider the case of the 'Harris/Robinson capital stock'. This was created around 2000 by two academics, Richard Harris and Kate Robinson, who carried out a clerical review of ONS business data records to match businesses across a change in ONS' business register. These were then processed in SPSS code to generate a 'capital stock' variable. This was of substantial value to researchers, both for the linking of old and new data and the use of a local-unit based capital stock (and distinct from the 'Martin capital stock' created by Ralf Martin at the enterprise level). It is not clear if the code still exists, but the clerical work that went into the initial linking is almost certainly no longer available. However, the dataset continues to be valuable to economists working on productivity models.

To share or not to share?

Assuming that the datasets have value to researchers, and that the right to share can be established, removing the dataset for the archive would damage research to no apparent purpose. The rationale, from the archive's perspective, for doing so would be that it feels uncomfortable with providing something to researchers which it cannot explain, or that it might be seen as giving a 'warrant of fitness'. Both of these could undermine the archive's credibility as a trusted home for quality-assured datasets. However, this assumes the archive is not able to explain its involvement, which is unlikely to be the case.

Making the data available does not mean the archive need to make any statement about the fitness of the data – 'buyer beware' can operate. Researchers are not obliged to use the data, and should make judgements based on the documentation available. The archive can help the judgement by making as much information available as possible. Researchers are, in general, poor at reading metadata which is not directly relevant to them, but highlighting the creators of the dataset – and providing contact details for any queries – should be noticeable.