CONSULTATION QUESTIONS

Are you responding *primarily* as a data custodian, data user or data subject? *(We recognise all people are data subjects and many organisations act as data guardians and data users, but please tick only one box)*

Data Custodian ☐

Data User (e.g. researcher) ☒

Data Subject (e.g. member of the public or group representing citizens) ☐

1. Are there any benefits of data linkage for statistical and research purposes that are not sufficiently described here?

Yes, there are further benefits ☒ No, the benefits are described fully ☐

If you ticked 'yes', please describe the further benefits of data linkage for statistical and research purposes.

The data linkage concept is simply one manifestation of a growing recognition that ever increasing volumes of data, if linked and analysed has the potential to bring many benefits. USA Obama announcement (http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf). The Obama Administration has announced a "Big Data Research and Development Initiative" with the goal of "improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some the Nation's most pressing challenges". Areas of application include environmental and biomedical research, education, and national security amongst others. Challenges to realise this initiative include new means to manage, analyze, visualize and extract useful information from large and diverse data sets, to develop new scalable software tools for analyzing large volumes of structured and unstructured data in distributed data stores and to create human-computer interaction tools to facilitate "rapidly customizable visual reasoning."

This also is a view shared by the World Economic Forum (http://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development) -"The amount of data in the world is exploding - Researchers and policymakers are beginning to realize the potential for channelling these torrents of data into actionable information that can be used to identify needs & provide services for the benefit of low-income populations." This is equally true for other groups and more generally.
It is also a distinct though related to the current debate and discussion concerning 'open data' and this is mentioned under further challenges.
Data linkage is an approach that is already in use, so that some of the methodology and technology already exists.

The consultation document is very much about data research, but is perhaps rather low-key, partly because the focus is on administrative and health data predominantly and is very much people oriented, with linkage of population characteristics. These are likely to be the highest quality data sets (individuals with unique identifiers), but other official data exist where linkages might also bring benefit, although the linkage process may be more difficult (not based on unique individual identifiers eg linkage based on geographical location). Data quality is key and not addressed within the document. Poor and variable data quality needs to directly addressed. Along with the linked data sets, there needs to be consideration given to the meta-data that for both historic and more current data provides information about the collection method and any underlying attributes that will contribute to data interpretation.

Particular comments on some of the benefits expressed in the consultation document:

Benefit 1 – Policies tend to be quite general and lack specificity, but it is clear that there are strong inter-connections so that data linkage will offer the opportunity to consider the drivers, pressures and impacts. The examples given are very simple, pairwise linkages but the full benefit will stem from a more holistic view.

Benefit 2 – this is very much a key benefit but data quality will ultimately determine the usefulness of the linked sets.

Benefit 3 - There is a common saying regarding "data rich and information poor", the power of official statistics comes from timeliness, robustness, and reliability driven by their analysis. Linkages will offer the opportunity for a more complete analysis. A key point here is that data need to be freely available to third parties where confidentiality constraints allow it (e.g. it will have to be anonymised before release). We need to ensure that those external to government can bring their creativity to bear to find new linkages and apply analyses. Inadvertent disclosure of individual information is often a key concern.

Benefit 4 – A longitudinal study is based on a temporal design and to achieve success here requires that there is consistency and comparability of the data frames over time. This is not a foregone conclusion. There needs to be more thought given, particularly by government, when trying to 'improve' the data collection process about the potential for a resulting loss of comparability.

Benefit 5 – It would be helpful to consider here the benefits of third party cost/benefit analyses to complement 'official' ones. In scientific communities, there are frequent discussions about data accessibility, where individual researchers often need to ensure that their research interests are not compromised. For example, NERC, as a funder of research that generates data, has created a number of data repositories where researchers are expected to lodge their data after a certain length of time. A model of this type could help to avoid future 'climate-gate'-like problems. Open data and data access is increasingly welcomed.

2. Are there challenges or barriers preventing more effective and efficient data linkages for statistical and research purposes taking place that are not sufficiently described here?

Yes, there are further challenges ☒ No, the challenges have been identified □

If you ticked 'yes', please describe the challenges or barriers.

The first challenge - the uncertainty about the legalities and public acceptability of data sharing and linkage – there is a large amount of uncertainty which needs to be addressed although many individuals would still find the idea of data linkage unacceptable even with additional guarantees. It may help if assurances about anonymity etc are independently policed to increase trust.
Satisfying public concerns and raising public awareness of the substantial benefits that may accrue are both essential and challenging.

The second challenge relating to incomplete data or data that cannot be linked will be difficult as without unique identifiers there are always going to be issues regarding incorrect matching and data quality is potentially highly variable. Mismatching and the effects this may have on any inferences are important and as more data sets are linked, it will become increasingly challenging.

To achieve success here requires that there is consistency and comparability of the data frames over time. This is not a foregone conclusion. There needs to be more thought given, particularly by government, when trying to 'improve' the data collection process about the potential for a resulting loss of comparability. Over a 10-15 year time frame, technology will likely have moved on exponentially and new systems introduced, so that some thought should be given to ensuring that data holdings are and remain linkable.

One additional challenge concerns handling the potentially large volumes of data and the software needed both for matching and for visualisation. This is important since in large complex data sets, it is essential that inconsistencies can be spotted and checks made. Automated techniques for internal consistency are now available and even automated repair although this is obviously more controversial. Automation is vital given the intractability of manual inspection of large data sets.

3. Are the guiding principles sufficient and appropriate? Please explain your answer fully and make suggestions for improvement.

Yes, they are sufficient and appropriate ☒ No, they are not □

Please explain your answer fully and make suggestions for improvement.

The principles seem sound and extensive. However, one further challenge concerns the prevention of inadvertent identification of individual information and security. Disclosure control and security are not new challenges but are ones that need to be addressed.

There was little mention of issues around the misuse of data – perhaps some clarification of the principles under which researchers and others would be granted access to linked data would be helpful.

While very supportive of the principles, we should also be conscious of the difficulties specifically in terms of interpretation of linked data and the evidence (population level) base for policy and interventions - sometimes described as the ecological fallacy. This fallacy describes how it may be difficult and potentially misleading to make generalisations from data about aggregations (e.g. populations) to disaggregated sets (e.g. individuals). The property of the population may not be the property of the individuals within the population and similarly the property of the individuals may not be reflected in the property of the population. The extract below, modified to be slightly more general than in its original form, explains further
“As a result there is often no readily available means whereby users of *aggregate* data can determine whether the results, hypotheses, and conclusions obtained from the analysis of *aggregate* data are applicable at the individual level ..... This is an important problem because individual-level inferences tend to be implicit in many applied uses of *aggregate data*; for example, the identification of problem areas for planning purposes, the use of a spatial classification to identify particular client groups in marketing, and the use of areal data by sociologists to generate hypotheses at the individual level. (S Openshaw, Environment and Planning A, 1984, volume 16, pages 17-31)”

There needs to be discussion round linkages to other data sources held more widely (eg in the UK rather than Scottish context). There is clearly extensive work being done within ONS and at national and European levels.

4a. Are the objectives set out for a Privacy Advisory Service in Section 3c the right ones?

Yes, the objectives are right ☒ No, they are not □

Please explain your answer fully and make suggestions for improvement.
The FOI commissioner and data protection legislation very clearly impact on this service, and should not be in conflict.

4b. Do you wish to be consulted on firmer proposals for a Privacy Advisory service as and when they are developed?

Yes ☒ No □

5a. Are the functions that will be led by the National Data Linkage Centre set out in section 3d the right ones?

Yes, they are the right functions □ No, they are not ☒

Please explain your answer fully and make suggestions for improvement.

There should also be a role for the research community to be involved in the NDLC. It currently also has a limited (in one sense) remit involving ISD and NSS and National Records of Scotland.

The NDLC may also need to have a key screening role concerning the validity of requests for data linkage. There may also need to be some form of oversight role concerning the uses of the linked data- this might require some form of a) oversight group (since each individual data source may not be sensitive, but when linked may introduce unexpected sensitivities and b) very careful wording of the terms and conditions of data use.

5b. Do you wish to be consulted on firmer proposals for a National Data Linkage Centre as and when they are developed?

Yes ☒ No □