Recommended requirements to develop discovery and access of Longitudinal Population Studies: Population Research UK

Purpose

ESRC and MRC recently commissioned an audit of the state of Longitudinal Population Studies (LPS) data. The report will be published by CLOSER in Spring 2023, a version is available within the documents for the Population Research UK Coordination Hub call (PRUK). A number of points were raised by the audit team for the attention of the funders. These are presented below as recommendations for PRUK to address, through the delivery of the PRUK Hub.

Why do we need data standards to be used?

- For data to be re-used widely and maximise the value of the original investment in collecting data, good quality documentation and metadata should be created.
- This enables the data to be curated and discovered by researchers in the future.
- Documentation should include details about the methodology used to collect the data, such as sampling frames, weighting, response verification etc.
- Documenting these processes helps to provide context and ensures researchers are aware of potential data collection biases/errors that they need to take into account when undertaking research with the data to avoid producing research outputs that could mislead.

UKRI-funded LPS studies and data standards

Requirement: The need to explore who is best placed to document linked study data: the studies funded by research councils themselves, or the data linking service which has linked the data.

- Linking data together (e.g. studies to each other, studies to administrative or other forms of data) require particular methodologies and assessment.
- These methodologies may be complex; they require understanding and therefore should be sufficiently documented to enable researchers to understand the effects of linkages on data samples to control for linkage characteristics and again avoid producing statistical errors when undertaking research.
- Which entity is best placed to understand how these linkage methodologies have been developed and therefore the most suitable candidate for producing documentation and depositing this with a data service for inclusion in a catalogue?
- Study staff will have in-depth knowledge of how the data have been collected and curated at an individual study level, and those staff are also likely to have prepared study data for linking to other data sources. Third-party data linking services may undertake data linking activities and therefore could be best placed to describe the linking methods. But they may be at a disadvantage if they are not familiar with the studies, and the make-up of the sampling frames in which data for the studies have been collected.
- Therefore, in cases where a third-party service has linked data, close collaboration between the data linking service and the studies is needed to adequately document the linked data. This collaboration and the resulting documentation should be included as objectives of the data linking exercise.

The role of existing data services

Requirement: Explore how data services can support documentation of linked LPS, and the consequences for researchers and studies where studies are not adequately documented and discoverable.

- Data services can assist because of their expertise creating systematic catalogue records describing data and enabling greater discovery of such data.
- They have experience in developing workflows that involve sophisticated data quality checks in parallel with documentation audits.
- They essentially provide a 'one-stop-shop' for researchers: find the data and explore the data using good quality data documentation and metadata.
- Using a systematic workflow (for documentation, metadata creation and cataloguing) such as those deployed by data services increases discoverability and reduces the cost burden for studies undertaking these activities themselves.
- This cost burden may arise if researchers ask questions of the studies which could be addressed through documentation and metadata. ESRC funds data services to document data and make data discoverable in an efficient way which reduces researcher-query burden on studies.
- A data infrastructure which provides a comprehensive set of services (ingest, curation, discovery, access, user support and training) effectively 'reinvests' to ensure that data are FAIR: Findable, Accessible, Interoperable, Reusable [this might be a principle which UKRI wishes to encourage].
- Although a number of data services, particularly Trusted Research Environments, operate in the landscape, they vary with respect to the range of services offered due to the extent of their resources and maturity. This variability may affect the potential value of the LPS, and the extent to which studies bear the cost of support researchers when studies, and linked studies, are not sufficiently documented.

The need for dedicated, skilled and experienced support

Requirement: define the skills and experience needed to support effective data curation, user support and training for LPS. Which organisations can best provide LPS support?

- A well-functioning and mature data service can provide effective data curation, user support and training service, as a 'one stop shop' if staff have sufficient technical and subject matter expertise.
- Understanding statistical sampling frames, data quality and confidentiality aspects, based on experience in e.g. epidemiological or social science disciplines, as well as data science, enables efficient and fit-for-purpose data pipelines to be established to create research ready data for researchers to make the most of.
- Services without such expertise are likely to struggle: the result is that demand for support from researchers (and the burden of servicing that demand) will fall on the studies.

Data Access

Requirement: Explore the most appropriate means of providing transparency in decision-making about where linked studies and linked studies will be stored and accessed by researchers. Determine a proportionate risk-management strategy to determine how data are accessed.

- Transparency in the decision-making framework used by data owners when deciding where data is to be accessed can vary. ESRC's Future Data Services (FDS) initiative will look at the role of Data Access Committees (DACs) which are commonly created to advise on applications from researchers to access study data.
- A consequence of the lack of a framework for deciding where data will be stored, accessed etc., is that these decisions are made asymmetrically, i.e. without coordination and without awareness and consideration of available options.
- Sub-optimal decisions about how data is stored and accessed, including linked study data, may result; researchers may not experience the best data access routes or the best support for using the data, and funders may not be achieving value for money as a result.
- Another consequence of a lack of framework is that data access decisions may be made by 'default' without broader consideration of the data access landscape. For example, studies could be made available in the particular settings without consultation or considering how other data infrastructures could enhance the quality of the data, leading to sub-optimal outcomes for data discovery and leading to burden costs for studies. This opaque 'data access by default' should be avoided.
- A data access framework should:
  - Establish the dimensions by which data access decisions are taken, particularly with respect to the access licence type (e.g. Open, Safeguarded or Controlled).
  - Ensure all involved understand and agree these dimensions, and can apply them consistently in decisions that determine how data are to be accessed.
  - The focus of the dimensions should be on the data and the privacy characteristics of the data;
  - Present examples of best practice and efficient decision-making processes.
  - Mandate clarity and transparency in decision making

Accreditation

Requirement: Assess the distribution of access to LPS and other data throughout the network of DEA-accredited data services.

- One of the thematic areas which Future Data Services will investigate as part of its strategic review, is Data Access, User Support and Training. As part of this we will examine how various data access accreditation systems work, coupled with examination of the decision-making process followed by data owners to decide where and how their data will be accessed.
- Accreditation schemes such as the Digital Economy Act scheme (managed by ONS on behalf of the UK Statistics Authority) certify that a data service can provide access to sensitive data providing that a number of criteria are met.
- Our understanding is that any DEA accredited data service should be able to provide access to any data where the legal basis for accessing and using data is the DEA.

- Accreditation of data services, and in turn, data access, should apply universally. Otherwise there is a potential risk that data access will not be equal.

Should all studies be curated?

Requirement: develop a decision-making framework to determine which LPS should be curated for the long-term.

- The LPS Data Audit revealed that some studies have been created for very specific purposes which are unlikely to be replicated.
- Expending effort on curating these data for wider discovery creates an opportunity cost: the resources that could have been devoted to enhancing studies widely demanded by our communities.
- A model may exist whereby smaller bespoke studies could be archived in perpetuity. But some studies' Data Access Committees (or equivalent) mandate that access takes place on a 'variable by variable' basis; this adds complexity to the data ordering system which conflicts with the process of archiving and providing any sort of access.

Creating a cross-disciplinary data catalogue

Requirement: work with funders and the existing data landscape to develop requirements for a cross-disciplinary and cross-institute approach for managing data discovery.

- To address the issues raised above, significant coordination is required to tackle asymmetries in data curation and discovery standards, and data access variability. This could be at the level of each study; a group of studies; or an entire data catalogue. This needs further consideration.
- One approach may be to determine a number of measures, or targets. RAG ratings could be developed and applied. Funders could mandate that such measures/targets are established as part of funding for individual studies, groups of studies or entire collections. Active monitoring and evaluation could be undertaken to review progress and risks.
- Without such coordination and monitoring there is a risk of poor practice prevailing, leading to poor outcomes (including poor value for money) for communities in the future.