Archival of research data from Secure Data Research Environment projects

Description of the problem

To comply with the regulation or accessing and use data it is typically expected the data acquired, with approval from the data controller deemed 'personalised' or 'de-personalised' must be stored and accessed in a secure data environment, owing to the risk of identification or re-identification of individuals, according to appropriate information security standards.

Such data are approved for use and release by the data owner, and usually acquired under a Data Sharing Contract or Data Access Agreement. These contracts provide for use of data under a fixed length of time.

The data must be destroyed at the end of the term specified by the contract, or upon completion of the project, whichever occurs earliest. A contract may be extended with permission from the data owner.

Impact

The consequences of destroying data under the above arrangements are as follows:

- The statistical results generated from the data cannot be replicated and independently verified.
- There is no possibility of 'revisiting' or 'revising' the work following completion/contract expiry.
- While it might be possible to make a new application to the data owner for a new extract, it is highly unlikely to have the same extract as the original, especially if the data are extracted from administrative sources of data which are continually updated.

Research data retention / archiving requirements

It is considered good practice to archive a research project in such a way that the project can be audited, or an analysis can be reproduced. Research institutions, funders, and publishers will typically have guidelines or policies covering retention of data after a project has ended.

Guidelines from Wellcome Trust states:

> "*Organisations should publish standard procedures for signing off and archiving laboratory records and notebooks*"

The Cancer Research UK guidance states:

> "*All researchers should consider at the research proposal stage how they will manage and share the data they will generate"*

The Medical Research Council guidance recommends:

> "*The research community must foster and support a culture of transparency and honesty which promotes good practice, recognises relevant interests or conflicts and deals with these openly and explicitly. This applies across the whole range of research activity from study and experimental design, generating, analysing and recording (including archiving) data, sharing data and materials"*

The Economic and Social Research Council guidance states:

> "*We believe that a structured approach to data management results in better quality data that is ready to deposit for further sharing. Grant holders must formally deposit all data created or repurposed during the lifetime of a grant in a responsible data repository*."

Research data destruction requirements

Data providers that provide datasets for SDE projects usually have strict data destruction requirements regarding the raw data. It may be possible to request an extension to allow completion of an analysis, but this will not extend to long-term archiving.

NHS Digital require the data they provide to be destroyed when the Data Sharing Contract or Agreement has expired. These can be renewed annually, but require re-approval.

Public Health of England PHE provide data for one-year periods. After a year, the researcher must request an extension. At the end of the project the raw data and any sub-datasets created during the analysis must be destroyed.

ONS and UK Data Service provide access to control data only users who have been trained and accredited and their data usage has been approved by the relevant Data Access Committee. Access is provided for maximum 12 months but can be extended

Solution Required

The data owner currently doesn't seem to have facilities that allow analysts and researchers who have received data from them to deposit their original data extracts, and amended datasets. For this reason, a third-party organisation which can host sensitive data could use for archiving purposes if the data owner agrees.

Choosing third-party organisation that could securely archive these data is desirable for the following reasons:

- Data Processor would not continue to hold on to these data themselves, even if access to the original project team, or any analysts, was prohibited
- The access to this data could be monitored by a third independent organisation within a 'circle of trust'
- The data to be deposited with another institution could benefit existing or future research programmes. Validation of results or revision of analysis would be possible
- Would encourage the generation of metadata about the data deposited. This would make health data easier to use and more discoverable.

In practice, the process described below would be likely to be accepted from the both parties:

1. Upon completion of the project, or expiration of the Data Sharing Contract, the data processor will securely transfer a copy of the original extract data, in addition to other data sources, to the third pre-agreed organisation
2. The Data processor should then destroy all copies of the data it has acquired, issuing any associated proofs
3. The data owner will be notified of this transfer and destruction of data
4. The third organisation will curate and keep the data securely; No one will have access to these data files without the approval of the data owner
5. Should you or any member of your stuff in the future wish to access the data again, they will need to contact the data owner for permission, stating the reason for requiring access.
6. Upon approval, the third organisation would provide access to your organisation for a pre-agreed fixed period.
7. Any modified data created by the new process would be sent for archiving and any copies will be destroyed (repeating steps one and two above).

Here it is important to point out that the data owner will continue to act in the role of 'Data Controller' thought the process. Both your organisation and the third party (archiving organisation) will act as 'data processors'. Of course, and agreement should be in place among all the player which will specify the needed requirements.

Recommendations

As until now there is no possibility of archiving data outside of SDE and as in theory", the data provider can extract the same dataset upon request for another party people should focus on archiving documentation about their datasets instead.

The ideal documentation that could allow the secondary use of data should include Study and Data level information which can be archived as metadata for each project.

A comprehensive Study-Level Documentation usually includes:

- General information about the project: Project history, aim, objectives, hypothesis (protocol)
- Data Collection/application methods: Sampling Design or the data items on an application

- Quality Assurances: Data validation check that you are aware of
- Overtime problems that can affect the data: Methodology changes, updates on data files
- Information on Data confidentiality: what is needed to access the data and how
- Publication: previous studies that have used the same data

A comprehensive Data-Level Documentation usually includes:

Quantitative

- Variable Names
- Variables Labels
- Variable Description
- Value code Labels
- Derived or Constructed Variables

Qualitative

A Data List including information on the participants

All in all, it is considered good practise if each analyst who is working with the data to carefully document all of the data processing steps (starting from the instructions to the Data Provider) until their final results.

Metadata Template

When the archiving of the real data is not possible it is sensible to focus on archiving metadata about the datasets. Below is a template of metadata that it can be used for this purpose.

Title Details

General Number: Unique acquisition number
Title: title of the dataset
Principal investigator(s): Data owner
Data collector(s): Responsible for the data collection
Data Classification in CRUK: how the acquired data is classified inside your organisation

Background Information

General information about the study you have used the data and the data itself.

Data Generation Process / Reliability

A description on how the data has been created and reasons considered reliable.

Time coverage – Timeliness

How often the data are updated. From when there are available.

Variables

The names of the variable that can be found in the dataset. What the dataset measures in general.

Potential Problems

Previous known problems that exist with the data. Your comments when you were using the files.

Research Use

Why the data have been created. What they measure and how they have been used. Any publication related with the data files.

How to Access

Information of how a third party can access the raw data. General information on what is needed

Coverage, universe, methodology

Time Period:
Country:
Observation units:
Kind of Data:
Universe:
Data Classification:
Time Dimensions:
Sampling procedures:

Method of Data Collection:
Number of Units:
Frequency of Release:
Data Updated:
Last Update:
Next Update:
Latest Year Data:
Data held from CRUK: