Research Data Management

Research Data Evaluation Guide

Version 2.0 March 2023

University of Bristol

Research Data Service

Image: Checkmark, Pixabay, Public Domain.

University of BRISTOL

1. INTRODUCTION

The Research Data Evaluation Guide aims to assist researchers in the task of sorting research data according to its relative re-use value. The guide is intended to be of use when preparing a Data Management Plan or similar document, or when sorting data at the close of a project for deposit with a subject-based repository.

The guide does not provide a definitive answer as to whether or not a dataset should be retained or disposed of; rather it provides support for the researcher who is to make such a decision.

1.1 Why not keep everything?

While it is true that the cost of data storage tends to decline over time, the cost of organising, describing and then maintaining data in a usable form has a considerable cost. In practice, this data 'curation' cost is often far greater than the cost of data storage, even for very large datasets.

1.2 Policies and requirements impacting on data evaluation

The role of the researcher in deciding which data should be retained and shared is likely to be informed by several different policies and formal requirements. Those described below should not be taken as an exhaustive list and attention should be paid to institutional guidelines[1] and to any information governance policies relating to a particular discipline.[2]

1.2.1 Research funder policy

Most major research funders now have recommendations or even formal requirements as to what data should be retained and shared at the close of a research project. For example, UKRI require "Data with acknowledged long-term value should be preserved and remain accessible and usable for future research".[3]

1.2.2 Data centre policy

Where data is destined for deposit into a subject-based data centre, subject-specific evaluation criteria may apply. Where this is the case researchers are advised to follow the guidance provided by the data centre in question.

[1] Bristol's own Research Data Management Policy can be found at: http://www.bristol.ac.uk/research/environment/governance/research-data-policy/

[2] Such as the Human Genome Project's Bermuda Principles: http://web.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml#2

[3] UKRI Common principles on data policy https://www.ukri.org/manage-your-award/publishing-your-research-findings/making-your-research-data-open/#contents-list

1.2.3 Academic publisher requirements

Increasingly, academic publishers also require data which underpins a publication to be retained and shared. For instance, a condition of publication in a Nature Journal is that "authors are required to make materials, data and associated protocols promptly available to readers without undue qualifications".[4]

Ultimately though, it is often the researcher or research team responsible for the data, as subject specialists, who are best placed to decide if data has lasting value.

1.3 Evaluation criteria

Criteria for retaining data can be categorised as follows;

1. Data has special scientific or historical value. The data is scientifically, socially, or culturally significant. Assessing this involves inferring anticipated future use, from evidence of current research value.

2. Data is unique. A dataset is the only or most complete source of the information that can be derived from it. This information would be at risk if the dataset were lost.

3. Data has a high re-use potential. The data is likely to be of broad interest and its reliability and provenance have been assured. E.g. the data relates to a longitudinal study, is in a technical format which is widely supported, sufficient metadata is in place and any ethical issues have been addressed.

4. Data cannot be easily reproduced. It would not be feasible to replicate the data, or doing so would not be financially viable.

5. There is a strong economic case for data retention. Costs have been estimated for managing and preserving the data and are justifiable when assessed against potential future benefit.

One criterion may outweigh another. For instance, a medical dataset may be impossible to anonymise but have a very high scientific significance. In such cases data should be retained but the issues which challenge re-use must also be addressed (for example by depositing the data with a repository which has a proven mechanism for granting controlled access).

A dataset which is the product of scientific simulation software presents another example. The data generated may be easily reproduced, however, it may be extremely costly to re-run the simulation, or doing so may require very specialist software or hardware. In this scenario, the simulation software and its parameters would be retained but

[4] https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards

there may also be educational or scholarly value in retaining the products of a simulation. The decision would be informed by estimating the cost for managing and preserving the resource against evidence of potential future value.

Where large numbers of files (or datasets) are to be evaluated, the selection process should be undertaken at as high a level of data aggregation as will give a justifiable outcome.

2. DATA EVALUATION CHECKLIST

Mandatory criteria Answering 'Yes' to any of the questions below automatically results in selection for retention.		
Legal/statutory considerations	**Yes**	**No**
Is there a legal or legislative reason to retain the data?		
Is there any reason to believe the data may be used in litigation, public enquiries, police investigations, FOI requests, or any report or paper that could be legally challenged?		
Is the data the product of UKRI funding **and** underpins a published research output?		
Are there any other contractual obligations that require the data to be retained?		
Policy	**Yes**	**No**
Does the Research Data policy of the research funder call for the data be retained?		
Will the data be cited within a publication with a policy which requires underpinning data be made available?		
Do any discipline-specific guidelines apply which call for the data to be retained?		

Important criteria Answering 'Yes' to at least one of the questions from each section below should probably result in selection for retention.		
Reuse value	**Yes**	**No**
Is the data unique and/or impossible for others to reproduce?		
Does the data have broad appeal and is it likely to be of interest to others (e.g. a broad geographical or temporal range or an inter-disciplinary focus)?		
Is the data likely to have special academic value (e.g. does it represent a landmark discovery) or does it set an important new precedent likely to be followed by others (e.g. involve a new data processing technique)?		
Research context	**Yes**	**No**
Is the data likely to be cited/referenced within an academic publication?		
Does the data add value to any significant *established* data collections?		
Does the data align strongly with *current* research trends (i.e. do separate but parallel research activities exist)?		
Is the data likely to align strongly with *future* research trends? This should be inferred, based upon evidence of current value such as existing citation rates.		

Supporting criteria Answering 'Yes' to the majority of the questions below should result in selection for retention.		
Origin	**Yes**	**No**
Would the data be costly or difficult to reproduce?		
Does the data have its original integrity? (e.g. is unprocessed, and has been stored securely since it was generated)		
Will this become the reference (definitive) copy of the data?		
Condition	**Yes**	**No**
Does the data have sufficient metadata? (e.g. a catalogue-level description, a description of how the data is organised, documentation of how and why data was created, and a guide on how to use the data)		
Is the data of suitable quality for deposit into a Data Centre or other repository? (i.e. data is quality controlled, well organised, readable and uncorrupted)		
Is there proportionally more valuable data than non-valuable data within the dataset?		
Storage and preservation requirements	**Yes**	**No**
Can the data be stored (i.e. archived) without any exceptional requirements?		
Can the data be preserved in a usable form (i.e. remain fit for purpose) without any exceptional requirements?		
Is funding in place to fund the preservation (either by the research team, a host institution or data centre) of this particular data?		
Access limitations	**Yes**	**No**
If personal data is involved, was informed consent obtained from the research subjects for archiving and re-use of data? If 'Yes' is it feasible for a host repository to adhere to any terms of re-use?		
If approval by an Ethics Committee was required, is there evidence that this procedure has been followed?		
Does the nature of the data suggest any other restrictions on sharing, access and re-use? (e.g. data set involves sensitive health or political data)		
Is the data free from any terms and conditions which would limit access? (e.g. IPR restrictions, database licence requirements, commercial agreements which prohibit re-use)		
Technical limitations	**Yes**	**No**
Is the data in an acceptable technical format for deposit into a data centre?		
Is the data usable without any specialist software/hardware?		
If 'No' to the question above, is the required specialist software/hardware readily available?		
Is it feasible to generate different versions of the data to increase reuse value (e.g. create alternative file formats)?		

2.1 Post-evaluation

Wherever possible, valuable data should be deposited permanently with an institutional or national data repository. It is the responsibility of the researcher or research team to organise the data and provide data in a repository's preferred formats. Also, to provide the metadata requested by the repository and to provide enough information for repository staff to assess the research data's compliance with legislation.

Where data is not retained, the decision process and criteria for justifying disposal should be recorded, so that future researchers can understand why particular datasets were kept and others disposed of. Disposal decision records should be held by the repository which provides access to retained data.

2.2 Acknowledgments

Information in this guide is based upon material found in the Digital Curation Centre's How to Appraise and Select Research Data for Curation[5] and the Natural Environment Research Council's Data Value Checklist.[6]

[5] https://www.dcc.ac.uk/guidance/how-guides/appraise-select-data
[6] https://www.ukri.org/publications/nerc-data-value-checklist/