Five steps to decide what data to keep

DCC Checklist for Appraising Research Data

Please cite as: DCC (2014) ‘Five steps to decide what data to keep: a checklist for appraising research data v.1 Edinburgh: Digital Curation Centre. Available online: www.dcc.ac.uk/resources/how-guides

Digital Curation Centre, October 2014
This work is licensed under Creative Commons Attribution BY 2.5 Scotland except Section 4, which is adapted under licence CC-BY-NC-SA, from UK Data Archive (2013) Data management costing tool and checklist. Available at: www.data-archive.ac.uk/create-manage/planning-for-sharing/costing
Preface

This guide aims to help UK Higher Education Institutions aid their researchers in making informed choices about what research data to keep. The content complements other DCC guides: How to Appraise & Select Research Data for Curation, and How to Develop Research Data Management Services. The guide will be relevant to researchers making decisions on a project-by-project basis, or formulating departmental guidelines. It assumes that decisions on particular datasets will normally be made by researchers with advice from the appropriate staff (e.g., academic liaison librarians) taking into account any institutional policy on Research Data Management (RDM) and guidance available within their own domain. As such, the guide should also be relevant to staff with responsibility for defining such policy in a Higher Education Institution, a Professional or Learned Society or similar disciplinary body.

The guide assumes that part way through their research the Principal Investigator, or other researcher responsible for data management, will want to choose what data to keep, informed by commitments already made to share or retain data (e.g., in a Data Management Plan). The unit of appraisal is a ‘data collection’ and this may include different files carrying different access permissions and/or licence conditions.

The text also assumes that the institution will provide the following capabilities:

- institutional catalogue/registry of publicly-funded data of long-term value, enabling potential users to find out what data exists, why, when and how it was generated, and how to access it

- facilities to keep selected data of potential long-term value if no external repository is available and help to digitise any non-digital material if there is a valid external request

No assumption is made about how either of the above capabilities will be provided; for example, they might be repository or managed storage services, distinct from or integrated with a publications repository or a CRIS (Current Research Information System). In either case the capability could be provided in-house, or outsourced e.g., through Janet Cloud Services. The guide may be adapted to reflect local services and guidance on selecting external repositories for data deposition. DCC can provide help with this customisation to institutions’ needs and visual design.

Angus Whyte, Digital Curation Centre

---

3 Details of Janet Cloud Services available at: www.ja.net/products-services/janet-cloud-services
4 A DCC Checklist for Evaluating Data Repository Services will be available 2014.
5 Further information at www.dcc.ac.uk/tailored-support, or contact info@dcc.ac.uk
Five steps to decide what data to keep

Introduction - what are your choices?

As a researcher you will probably select from the data available at various points in the research cycle. You will select from the data sources available to work with at the outset of your study, select from the data assembled for analysis and then select analysed data to make further statements about what has been found, some of which may be included in a publication.

With more digital technologies being used in research there is a growing need to make further choices about what to keep for the long-term, selecting what data to make available or to dispose of. The best time to do this data appraisal is well before the end of the project, or periodically, if it’s a longitudinal or reference data collection.

This guide aims to help you make what may be quite difficult choices around what data to keep in order to meet your own purposes and satisfy your institution and external funders. You may have a number of choices about who will look after your data:

1. Use an external archive or repository already established for your research domain
2. Use a data sharing platform such as Figshare.com or Zenodo.org
3. Offer it to a publisher as supplementary material to a research article
4. Use your research group’s established data management facilities to preserve the data according to recognised standards in your discipline
5. If available, an institutional based research data repository (see right)

The choice may be straightforward if you have an established data management facility in your domain, or even within your research group or department. Your research funder may recommend a data centre or self-deposit archive. For example, the UK Data Service offers social scientists the ReShare archive (reshare.ukdataservice.ac.uk).

When choosing it is important to consider factors such as whether the repository:

- Gives your submitted dataset a persistent and unique identifier
- Provides a landing page for each dataset, with metadata that helps others find it, tell what it is, and cite it
- Helps you to track how the data has been used
- Responds to community needs and/or is certified as a ‘trusted data repository’
- Offers clear terms and conditions that meet legal requirements e.g., for data protection and allow reuse without unnecessary licensing conditions

A forthcoming DCC guide offers further help in selecting external repositories. Your institution may offer Research Data Management support to help you deal with these issues and get the most out of the investment put into your research. This could involve:

- Registering datasets with the institution’s Data Catalogue to help make the research more visible
- Depositing the dataset with an institutional repository to maintain a long-term record of its safekeeping and, if it is publicly available, the access and download statistics.
- Keeping selected data safe in the dedicated storage your institution offers for long-term retention
- Recommending external repositories that may be appropriate

*Lists of data repositories are maintained by Re3data (www.re3data.org!). The University Library may be able to offer more discipline-specific guidance on data repositories.
Thi guide takes you through five steps:

1. Consider potential **reuse purposes** - what aims *could* the data meet?

2. Check for indications that it *must* be kept considering **legal or policy compliance risks**

3. Identify which data *should* be kept as it may have **long-term value**

4. **Weigh up the costs** - which data management costs have already been incurred and therefore contribute to its value, and how much more is planned and affordable? Where will the funds to pay these costs come from? Considering these questions will give you the cost element of your data appraisal and should help identify any need for external advice, e.g., on how to deal with any shortfall in the budget.

5. **Complete your data appraisal** - this will list what data must, should or could be kept to fulfil potential reuse purposes. The appraisal should also summarise any actions needed to prepare the data for deposit, or the justification for not keeping it.

This guide draws mainly on the existing DCC guide, *How to Appraise and Select Research Data for Curation*[^7], NERC Data Value Checklist[^8], and University of Bristol Research Data Evaluation Guide[^9]. Section 4 is adapted from the UK Data Service’s Data management costing tool and checklist.

What data and for how long?

This guide uses a broad definition of research data: "representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship". When selecting what data to keep you will need to consider which of the following broad types is suitable for reuse:

a) Source data; data collected, created, or held elsewhere that the research has used

b) Assembled datasets; data extracted or derived from (a) above

c) Referenced data; data reworking a subset of (b) above in order to take the analysis further or draw conclusions. Consistent with whatever is considered ‘supplementary material’ to research findings in your domain.

The guide also uses the term data collection for any collection of objects that would be needed to access and interpret the above e.g., notebook, protocol, software or set of instrument calibrations, whether wholly or partly digital. In some cases there may be a justification for keeping the software used to manipulate and analyse the data generated. That justification may be as strong as for keeping the data itself, e.g., if the software would be required to enable the results to be reproduced. In other cases it may be even stronger, e.g., if the results could be reproduced just from the algorithm that was used to carry out an analysis.

A single research project could easily produce a number of ‘data collections’, each matching a different potential use. Just as research may be written up for several audiences and publishers, data collections could be produced and deposited in different repositories. Each data collection may itself comprise various digital files that need different access permissions and/or licence conditions. You will need to plan how to package these up into collections, taking account of your intended repository’s terms and conditions for depositing data. If you will need to deposit different kinds of outputs with different services, you may need to plan to organise them in multiple data collections.

The phrase long-term is used in this guide to mean ‘beyond the end of your research project’. Or, if the research data is contributing to a reference collection or longitudinal study, its long-term value could be assessed periodically, e.g., every 3 years. Specific guidance on how long data should be retained may be available in your institution or funder’s Research Data Management Policy. The DCC also provides guidance on funding body policies.

Other sources that can help you assess what to keep if they exist for your project include a Data Management Plan produced when the research was conceived - this may identify possible long-term uses of the data while there may also be a Pathway to Impact statement holding some ideas about longer term objectives for research outputs.

Step 1 Identify purposes that the data could fulfil

Consider the purpose or ‘reuse case’ that the data could serve beyond the research context in which it was created or collected. Any one of the following 7 reasons could justify retaining data for long-term access. Many, though not all, involve making it accessible beyond your research group, at least once you have had the opportunity for first use.

- **Verification**: enable others to follow the process leading to published findings and potentially reproduce or verify these
- **Further analysis**: increase opportunities for further analysis of the data e.g., using new methods, integration with other sources for meta-analysis, whether through new collaborations or third-party analyses
- **Building academic reputation**: data that is discoverable has greater visibility, which can boost citation rates for the published findings
- **Community resource development**: publish a data resource of value to a known user group, e.g., a reference dataset, methods test-bed, or domain database
- **Further publications**: the publication of a data article will contribute to scholarly communication and debate about data management or reuse in your domain
- **Learning & teaching**: embedding data in a learning/teaching or public engagement resource to enhance its interactivity, engage users in learning about or participating in the research
- **Private use**: find the data more easily in years to come to exploit other potential uses
The first two reasons in particular overlap with funding bodies’ policy aims, as these typically focus on ensuring the integrity of published research findings and on maximising return on the investment in data. The funders’ main concern is to preserve the ‘data behind the graph’\textsuperscript{15}, but the onus is on researchers and those who directly support them to translate policy into meaningful guidelines. Contractual and other legal obligations may also come into play. We return to these under ‘What data must be kept’ below.

To help your decision making you could match up the reuse purposes most relevant to your research against types of data likely to be needed, as in Table 1 below.\textsuperscript{16}

### Table 1. Example preservation guideline

<table>
<thead>
<tr>
<th>Reuse case</th>
<th>Preservation guideline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Further publications</td>
<td>Referenced data with additional documentation</td>
</tr>
<tr>
<td>Learning &amp; teaching</td>
<td>Samples of source &amp; assembled data with analysis scripts</td>
</tr>
<tr>
<td>Verification</td>
<td>Referenced data plus analysis scripts</td>
</tr>
<tr>
<td>Further analysis</td>
<td>All source data plus software used to collect</td>
</tr>
</tbody>
</table>

---

### Step 2: Identify data that must be kept

Generally the decision on what ‘must’ be kept will depend on the data creator’s priorities, i.e., on how valuable the data is for the purposes identified above, considering the costs of preparing it for long-term use. But the decision will also need to account for legal, regulatory or policy compliance issues. At the point of deciding what to keep these mostly concern whether data should be publicly available or have restricted access, on what terms and conditions it should be accessible, and ensuring that risks of non-compliance are addressed.

In this step, consider the basic questions below to help identify these. Seek further advice from your institution’s Research Data Management service, or similar support staff e.g., Records Manager, if you are unsure whether risks are best addressed by keeping the data or disposing of it and, in either case, how securely this needs to be done.

---

### Are there research data policy reasons to keep it?

UK Research Council research data policy principles emphasise that data with “…acknowledged long-term value” should be retained.\textsuperscript{17} Journals, learned societies and professional associations are active in defining what this means in individual disciplines.

Decisions on what ‘must’ be kept will need to take account of any relevant funder or institutional policies.\textsuperscript{18}

- Will the data underpin an article submitted to a journal that has policy requiring it to be available?
- Will data produced through RCUK funding underpin a published research output?

A ‘yes’ here will indicate you should keep the ‘data behind the graph’ (discussed above). Step 3 below gives more help on working out anything else that may be of ‘long-term value’.

---

### Do regulations require the data to be available?

The main questions here are;

- Does the data need to be retained to comply with Freedom of Information or Environmental Information regulations?
- Are there disciplinary regulations that require data to be retained as part of the research record e.g., for health and safety reasons?

---

\textsuperscript{15}See, for example, ‘Data policy for the Shelf Sea Biogeochemistry Programme’. Available at: www.uk-ssb.org/data/policy/


\textsuperscript{17}RCUK (2011) Common Principles on Data Policy. Available at: www.rcuk.ac.uk/research/datapolicy

\textsuperscript{18}E.g. PLoS Data sharing policy www.plosone.org/static/policies#sharing and the JoRD project at: jordproject.wordpress.com/project-data/social-science-journals-that-have-a-research-data-policy
Legal regulations covering Freedom of Information and Environmental Information require research data to be made available on request, if the research is complete and data relating to it is still available. This implies that any data that is kept should be clearly identified according to an information security classification scheme e.g., public access/internal/confidential/secret. For general guidance on any exemptions that may apply to the data, consult the Information Commissioners Office (ico.gov.uk) and Scottish Information Commissioner (www.itspublicknowledge.info) websites.

Are there other legal or contractual reasons?
These are likely if the research has public policy implications, it involves a commercial partner, or has potential spin-off applications.

☐ Does the data provide information of commercial value, or is it used in a patent application?

☐ Do contractual terms and condition state or imply that the data must be retained?

☐ Is it reasonable to believe that the data may be used in public enquiries or police investigations, or in any report that could be legally challenged?

Does it contain personal data relevant to the reuse purpose?
The Data Protection Act defines personal data and sets out criteria for deciding how long it should be kept, how it must be stored and requirements for disposal. If you answer ‘yes’ to all of the questions below, the next step is to follow guidance available from the Information Commissioner’s Office, including how to anonymise data if needed. The UK Data Archive also offers guidance and provides a Secure Lab, allowing personal data to be used in academic research under strict controls.

☐ Does the data contain details that directly identify an individual or can be used to infer their identity, either in isolation or through linking it to another data set?

☐ Does your institution’s ethics approval allow the data to be retained for further research?

☐ Does the consent agreement allow data to be reused for the purpose that you are now envisaging?

☐ Did the data subjects give their informed consent to its archiving?

☐ If so, is it feasible to adhere to any conditions of their consent e.g., any commitment to anonymise the data?

☐ Can the data be securely stored and actively managed to recognised information security standards e.g., ISO27001?

If your response is ‘no’ to any of the above you may be able to get help to resolve any issues from your institution’s Research Data Management service or Records Manager.

Step 3
Identify data that should be kept
Bearing in mind the potential reuse purposes you identified earlier, consider the criteria and questions below to help decide which data should be retained and for what reason. As a general rule the data should be kept if you have already identified a compliance reason, or you can answer ‘yes’ to at least one of the questions under any two of the headings (criteria) below.

Tick any of the criteria that you expect the data to rate highly on, as far as this can be estimated. You can weight criteria differently according to the long-term aims the data needs to meet and how certain you are about its value in relation to those aims.

Is it good enough?

☐ Description: is there enough information, e.g., from an up-to-date Data Management Plan, about what the data is, how and why it was collected, and how it has been processed, to assess its quality and usefulness for the aims you identified?

☐ Quality: is the data quality good enough in terms of completeness, sample size, accuracy, validity, reliability, representativeness or any other criteria relevant in the domain?

---

19See e.g. Wikipedia at: en.wikipedia.org/wiki/Information_security#Security_classification_for_information
20Information Commissioner’s Office (n.d.) Anonymisation. Available at: ico.org.uk/for_organisations/data_protection/topic_guides/anonymisation
21UK Data Service (n.d.) Secure Lab. Available at: ukdataservice.ac.uk/use-data/secure-lab.aspx
Is there likely to be a demand?

- **Known users:** are there users waiting for this data, or is there past evidence of a demand e.g., will this add value to an established resource or series?

- **Recommendation:** does the funder, or a learned/professional society or equivalent body in the research field recommend sharing data of this type or on this research theme?

- **Integration potential:** does the data describe things that fit standardised terms or vocabularies in other research domains, such as geographic locations and time periods?

- **Reputation:** was the data produced by a research group or project that is highly rated on the originality, significance and rigour of previous research outputs? Will making the data available be likely to significantly enhance a group or project’s reputation?

- **Appeal:** could the data have broad appeal e.g., as it relates to a landmark discovery, a significant new research process, or international policy and social concerns?

Is there likely to be a demand?

- **Non-replicable:** would reproducing the data be difficult/costly (or impossible as in the case of unrepeatable observations)?

Is there likely to be a demand?

- **Cleared:** is the data classified according to its sensitivity and free from privacy/ethical, contractual, license or copyright terms and conditions that restrict public access and reuse? Are any restrictions normal for the study domain?

- **Open format:** is the data in a format that does not require license fees or proprietary software/hardware to reuse?

- **Independent:** if any specialist software/hardware is needed to use data, is that widely used in the field of study and readily available?

Is it the only copy?

- **Unique:** is this the only and most complete copy of the data?

- **At risk:** is the data held somewhere that cannot guarantee long-term storage?

### Step 4
**Weigh up the costs**

This step helps consider the economic case for keeping the data. It is important to consider the data management cost impact on your research commitments and your organisation’s budgetary constraints. If you have recently done that and can give an unequivocal ‘yes’ to each of the following questions you can skip this step.

- Is funding available to pay data management costs arising during the research, including those of preparing the data for archiving?

- Is funding available to pay any charges for storage and curation beyond the research period?

You can use this section to estimate any shortfall in the time or other costs budgeted for data management. Any costs that have already been incurred will count on the ‘value’ side of the economic case for keeping the data, while any shortfall will count against it. Your institution’s Research Data Management Service, Research Office, Library or IT service may be able to advise on how to meet commitments for data in the ‘must keep’ category.

Use the headings below, or any cost categories used in the Data Management Plan, to estimate how much has been spent on staff time, equipment/hardware, or software and service charges, and how much still needs to be spent in these categories.

The table is only for your own purposes (figures do not need to be disclosed to anyone who would not otherwise have access to them). It should serve two purposes: firstly, to help identify the value accumulated in your data and, secondly, to identify any areas where you may need to seek external help to avoid the risk that this value cannot be realised.

---

22RCUK has given general guidance on what costs may be included in grants, at blogs.rcuk.ac.uk/2013/07/09/supporting-research-data-management-costs-through-grant-funding/
<table>
<thead>
<tr>
<th>Creation, collection &amp; cleaning</th>
<th>Spend to date</th>
<th>Needed to complete</th>
<th>Budgeted</th>
<th>Likely shortfall?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creating a suitable consent form and obtaining consent for data sharing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data transfer or transcription from sites, media or instruments</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description and documentation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validation, checking or cleaning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formatting and file organisation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digitisation of paper or physical objects</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Short-term storage &amp; backup</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storage space for all working data for duration of project</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Backup of all data for duration of project</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Short-term access &amp; security</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Providing access and authentication for external collaborators or participants</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online and physical protection of data from unauthorised access or disclosure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Team communication &amp; development</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data management meetings</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Online collaboration, virtual research environment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data management training</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Preservation &amp; long-term access</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Copyright clearance, licensing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classifying data sensitivity and anonymising personal data (if required)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preparation for archiving, conversion to open file formats</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metadata for data citation, discovery and reuse</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data deposit charges</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long-term storage costs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Staff time (person hours)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Equipment (£)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Software or service charges (£)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Step 5
Complete the data appraisal

The final step is to weigh up the value and any costs still to be incurred, considering the long-terms aims, the qualities you identified, the time and money already invested in it and the risks of being unable to prepare any ‘must keep’ data for preservation. Completing a table, as illustrated below, may help you do this. The table can be used for communication with others involved in your research, or with support services in your institution or selected data repository.

To complete the data appraisal table, summarise the outcomes of steps 1 to 5 as follows:

1. Identify each data collection that in Step 1 you considered to have a potential reuse purpose (e.g., as in Table 1 earlier) along with its creator(s)

2. Identify the relevant reuse purposes from Step 1

3. In the ‘value’ column, insert relevant keywords from the headings in Step 2, i.e., any compliance reason either for keeping the data, or for disposing it; and also from Step 3 (reasons the data may have a reuse value)

4. Under ‘Risk of Budget Shortfall’, identify this as high, medium or low depending on the likelihood that costs will not be met by available funds and support

5. Record your decision in the ‘Keep?’ column. If it helps, use the ‘MOSCOW’ acronym, i.e., Must, Should, Could or Won’t keep, to agree priorities. If “Won’t Keep” is the decision, then you can justify your reason for non-retention in the box below the table

.....continued overleaf
## Example

### Data appraisal for project title

<table>
<thead>
<tr>
<th>Data collection title, creator and description</th>
<th>Reuse Purpose</th>
<th>Value</th>
<th>Risk of Budget Shortfall</th>
<th>Keep? MOSCOW</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 DCC 2014 RDM Survey Responses &amp; Analysis</strong>&lt;br&gt;Angus Whyte, Diana Sisu&lt;br&gt;Responses from survey of UK Higher Education Institutions managers involved in decision-making on Research Data Management with survey questions, documentation, anonymised responses and analyses</td>
<td>Verification, Re-analysis Visibility</td>
<td>Description, Quality Cleared, Open format, Known users</td>
<td>Low</td>
<td>M</td>
</tr>
<tr>
<td><strong>2 DCC 2014 RDM Survey Unanonymised responses</strong>&lt;br&gt;Angus Whyte&lt;br&gt;Unanonymised responses from survey of UK Higher Education Institutions managers involved in decision-making on Research Data Management</td>
<td>N/A</td>
<td>Compliance</td>
<td>Low</td>
<td>W</td>
</tr>
<tr>
<td><strong>3 Etc.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

### Summary of action required (Referring to row numbers in table)

- Document the sources and variables used in analysis (1)
- Label graphs clearly in spreadsheet (1)

---

### Justification for not retaining data ('won't keep') (Referring to row numbers in table)

- Unanonymised responses will be deleted as stated when survey introduced to respondents (2)

---

### Completed by
Guidance on Research Data Retention Policies


Jisc Infonet (n.d.) Higher Education Business Classification Scheme and Records Retention Schedules. Available online: bcs.jiscinfonet.ac.uk/he/


Acknowledgements

Many thanks to the following for comments on earlier drafts: Laurence Horton, London School of Economics and Political Science; Gareth Knight, London School of Hygiene and Tropical Medicine; Debra Hiom, University of Bristol; Helen Young and Susan Manuel, Loughborough University.