Depositing Data

Why select and appraise?

It is neither possible nor feasible to keep all digital data forever. Long-term preservation and curation of data entails additional costs that would last beyond the duration of the research project. Besides the costs related to data retention, with a growing volume of digital data, discovery becomes harder as the noise-to-signal ratio increases.

Five steps to help you decide what data to keep

DCC: Five steps to decide what data to keep (CC-BY)

1. Consider potential **reuse purposes** - what aims could the data meet?
2. Check for indications that it must be kept considering l**egal 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

1. Step 1: Identify purposes that the data could fulfill

Consider which type of data is suitable for reuse

a. Source data: data collected or created that the research has used
b. Assembled datasets: data extracted or derived from (a)
c. Referenced data: data reworking a subset of (b) above in order to take the analysis further or draw conclusions

Data collections may also include any other objects necessary to access or interpret the data above, e.g., software, lab protocol, notebook etc.

? you may need to create multiple data collections as there may be different access permissions or license conditions

Consider which purpose the data could serve beyond the research context

- **Verification:** enable others to follow the process leading to published findings
- **Further analysis:** increase opportunities for further analysis of the data
- **Building academic reputation:** data that is discoverable has greater visibility and can boost citation rates for the published findings
- **Community resource development:** publish a data resource of value to a known user group
- **Further publications:** the publication of a data article will contribute to scholarly communication
- **Learning & teaching:** embedded data in a learning/teaching engagement to enhance interactivity
- **Private use:** find the data more easily in the future to exploit for other potential uses

2. Identify the data that must be kept

Your decision on what data to keep will also depend on **legal, regulatory or policy compliance issues**. You will need to check whether there is any institutional or a funder's policy that defines what data should be retained - these are mostly the 'data behind the graph' - and whether the data should be made available or have restricted access.

3. Identify data that should be kept

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 question under any two of the headings (criteria) below.

Is it good enough?

- **Description:** is there enough information about what the data is, how and why it was collected?
- **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?

Is there likely to be demand?

- **Known users:** are there users waiting for this data, or is there past evidence of a demand?
- **Recommendation:** does the funder or a learned/professional society recommend sharing data of this type?
- **Integration potential:** does the data describe things that fit standardised terms or vocabularies in other research domains?
- **Reputation:** was the data produced by a research group or a project that is highly rated on the originality, significance and rigour of previous research outputs? Will making the data available enhance the group/project's reputation?
- **Appeal:** could the data have broad appeal? (e.g., relating to a landmark discovery)

Are the data easily replicable?

- **Non-replicable:** would reproducing the data be difficult/costly?

Is there likely to be 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?
- **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 the 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?

4. Step 4: Weigh up the costs

Consider the economic case for keeping the data - how much has been spent on staff time, equipment, hardware/software and service charges, and how much still needs to be spent. Is funding available to pay for preparing the data for archiving and to pay any charges for storage and curation beyond the research period?

5. Step 5: Complete the data appraisal

Weigh up the value and any costs still to be incurred, considering long-term 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.

Useful Resources

Whyte, Angus & Andrew Wilson. 2010. How to Appraise and Select Research Data . Edinburgh: Digital Curation Centre