UK Data Service

Collections Development Selection and Appraisal Criteria

External

13 February 2018
Version: 05.00w

T +44 (0)1206 873546
E collections@ukdataservice.ac.uk
ukdataservice.ac.uk

Contents

1. Scope

This document sets out criteria for selecting and appraising data collections across the UK Data Service. This enables more transparent implementation of the UK Data Service's Collections Development Policy, approved by the Service's Data Infrastructure Strategic Advisory Committee (DISAC) in November 2012.

A coordinated and robust approach to appraisal criteria for the service is required to ensure that:

- acquisition decisions will be based on explicit procedures which can be justified;
- ingest activities can be prioritised and data collections will follow an appropriate ingest and access pathway;
- reporting can be conducted on collections development activity.

These criteria provide additional benefits to the UK Data Service and other data service infrastructures including their use for training in data selection and appraisal and checking and as a starting point for other repositories.

2. The Collections Development Policy overarching selection and appraisal priorities and criteria

The UK Data Service's Collection Development Policy (CDP) states that it acquires data to meet three central purposes:

- *Potential secondary use and analysis for research*: to enable researcher access to materials originally created to inform and support research, but asking new questions or undertaking a restudy or follow-up study;
- *Teaching and learning use*: to provide access to data that have been created or repurposed specifically for teaching purposes to enhance use of the collection and support skills development;
- *Replication and validation of research*: to provide researchers with access to original research that allows checks for consistency, validation of sources and validation and replication of research outputs, both published and unpublished.

The CDP further lays out four discrete 'Curation Categories' for which it treats all data collections.

- *CURCAT1: Data collections selected for long-term curation.* These data collections are made available for download, or accessible via online access tools.
- *CURCAT2: Data collections selected for "short-term" management.* These data collections will not (initially) be retained for long-term preservation, rather they will be backed-up (i.e., bit-level preservation only), made accessible and discoverable through online access tools (including Nesstar, InFuse, UKDS.Stat, etc.) or via in-house repository software (ReShare).
- *CURCAT3: Data collections selected for 'delivery' only*, e.g., where data from third parties are accessed via APIs/web services and delivered to end users via a UK Data Service interface. Issues such as level of trust in owner, what documentation/metadata are required, and how rights/registration are handled need to be agreed.

- *CURCAT4: Data collections selected for "discovery" only*. These collections will not be brought formally into the holdings of the UK Data Service, they will exist only in other (institutional) repositories, but the UK Data Service will harvest (or in exceptional circumstances, create) metadata records to allow these data collections to be found more easily.

- A fifth category (CURCAT5) relates to preservation-only which falls outside the scope of the UK Data Service Appraisal, and which is handled by the UK Data Archive. Data collections may be moved into higher or lower categories over time if the need arises.

Alongside these purposes sits the UK Data Forum's UK Strategy for Data Resources for Social and Economic Research (UK Data Forum, 2013-2018) which has identified priority data resources, driven by research and policy needs as:

- Longitudinal data on people and families;
- Addressing multidisciplinary challenges (e.g. biosocial data, links with environmental data);
- Data about organisations;
- Customer databases;
- Social media as research data;
- Cross-sectional data on people and families;
- International data access;
- Administrative data
- New and novel forms of data

The CDP must be supported by selection and appraisal criteria that are both transparent and accountable to enable monitoring of processes, planning and justification for UK Data Service activities and spending. Materials will be rigorously reviewed to determine if they can be viably managed, preserved and distributed.

The high-level criteria used to review materials considers matters relating to the value of the data:

- *Relevance to the remit*: The content of the materials meet the ESRC's current strategies and are within the 'Scope of Collections'. This includes being high-quality, authoritative, reliable and up-to-date with 'good temporal and spatial coverage', with long and consistent time series (as appropriate), and strong opportunities for comparable research;
- *Scientific or historical value*: The materials are considered important resources either for research purposes and/or meet current demands in research trends and/or there is research and/or teaching activity in the subject area covered by the data;
- *New sources or types of data*: The materials fill a gap in the collection and/or have been requested by users;
- *International value*. The materials are considered to be important for international research;
- *Uniqueness/risk of loss*: The materials form a rare and complete resource of value to the research community and/or are at risk of loss if not acquired; data collections which are costly to reproduce or to which continued access to data is threatened; data are not generally available in any other form e.g. paper;
- *Usability/redistribution/operational benefits*: The materials have been well prepared according to data format/documentation and ethical/legal guidelines and fall within Scope; accession to the Service makes the resources more accessible. Where data are made completely open by the data owners, that is publically accessible without registration, the UK Data Service may continue to make

accessible versions of the data because it fits well with the portfolio of other data[1];

- *Replication data and resources*: Materials and related outputs required to replicate published or unpublished research, or undertake a restudy.

Where materials fall outside of scope, efforts are made to determine whether another archive, data centre or institutional repository is more appropriate for dissemination, curation or preservation of these data.

There are also certain criteria for not accepting data even though the value is deemed to be high:

- Legal and ethical issues: there are insurmountable rights management issues e.g. consent, IPR, copyright and Data Protection issues which are unable to be satisfactorily resolved and where full use of the data would not be possible without infringing legislation;
- Lack of sufficient contextual materials to enable re-use;
- Old/unreadable formats that are difficult to convert or make usable, or impossible to recover.

3. Appraisal criteria for submissions and data discovery

A core set of criteria is used for judging *strategic high-level value*, a second set for assessing *user need and analytic value* and a third set relating to *usability and accessibility*. These are classified into three stages. Value criteria are used as a first stage for appraising data for selection. If data do not pass the first stage, that is they are out-of-scope and do meet the core Collections Development remit, they are unlikely to pass to Stage 3 to be assessed for their accessibility and/or usability. ESRC data (CURCAT 2) are not appraised using value criteria but are subject to Stage 3 checks only.

STAGE 1: Criteria relating to strategic high-level value

- UK Data Service remit as set out in the CDP list above (3 options: Potential secondary use and analysis for research; Teaching and learning use; or Replication and validation of research). At least one must apply
- data relevant to the UK Data Forum's UK Strategy for Data Resources for Social and Economic Research and/or identified as strategic priorities by the ESRC. (These subject areas may change over time)
- collections with a national mandate to share (e.g. ESRC data policy, ESRC-funded longitudinal studies)
- data identified as gaps in the Collection or selected for acquisition in a given period by the DISAC (identified by Producer Type or Data Type)
- data from key long-term series already held by the UK Data Service with existing high usage (mostly produced by government)

STAGE 2: Criteria relating to user need and analytic value

- data useful for supporting decision-making and policy formation
- data useful for international, longitudinal or comparative research (e.g. time series or harmonised)
- authoritative source of high quality data that are/have been widely used and cited
- data directly linked to a peer-reviewed published paper
- microdata otherwise not available to the research community (e.g. disclosive data)
- data directly requested by a critical number of users
- new and novel forms of data
- unique, unrepeatable data and those that are costly to reproduce
- data which may assist in the development of data curation skills within the UK Data Service

STAGE 3: Criteria relating to potential usability and accessibility

- data that have definable conditions within the access spectrum

[1] Anticipated access method will not affect the appraisal of data

- legal and ethical agreements and protocols used do not prohibit re-use, e.g. consent to share/reuse data, IPR, copyright, DPA
- data have appropriate supporting documentation/contextual information
- data are produced in, or can be converted at minimal or no cost to, a suitable format for reuse through a UK Data Service delivery platform (without need for specialised software)

- data are more openly accessible or where a national licensing model will offer considerable savings
- third party host is a reliable and reputable supplier of data
- third party host can confirm that data supplied are likely to be stable and not change over time
- third party supplier/host can confirm that suitable preservation processes are used to manage and curate data
- third party host/supplier can confirm that a persistent identifier can be made available

Where a collection is deemed to be of high or commercial value but cannot be made accessible via the UK Data Service it can be referred to the UK Data Archive for negotiation about a possible costed preservation-only route (CURCAT 5).

Stage 3 criteria could be formalised for external use to apply a UK Data Service 'Kitemark' to data that pass the checks. This could be used to highlight data of excellence housed in other repositories.

4. Using the Appraisal Grid for presenting submissions

The Appraisal Grid lists the three sets of criteria/stages in blocks. A single member of the UK data Service Data Appraisal Group (DAG) is assigned a 'proposer' role charged with taking forward a data submission through to a DAG discussion and decision. The proposer completes the header information and tick all of the relevant boxes that apply, or write N/A. At least one of the UK Data Service high-level value criteria must be checked in order for the offer to move forward to the Data Appraisal Group for discussion. The proposer also must include a brief statement about potential use/re-use value in the Grid. Key URLs to third party hosted websites, data catalogues, data access points or documentation must be added. Recommendations for access, delivery and adding value are listed by the proposer on the Grid.

The following table sets out the Curation category by the appraisal stages required:

CURATION CATEGORY	**Appraisal Criteria Stage 1**	**Appraisal Criteria Stage 2**	**Appraisal Criteria Stage 3**
CURCAT1 Curated	√ At least one high-level value must apply	√	√
CURCAT2 Short-term management: Online tools	√ At least one high-level value must apply	√	√
CURCAT2 Short-term management: ReShare*			√
CURCAT3 Third party hosted data for delivery only	√ At least one high-level value must apply	√	√ Attention paid to supplier authority and reliability; and data format and conversion issues

CURCAT4 Third party hosted data for discovery only	√ At least one high-level value must apply	√	√ Attention paid to supplier authority and reliability; data format and the degree of documentation
CURCAT5	√ At least one strategic high-level value must apply	√	√ Attention paid to supplier authority and reliability; and data format and conversion issues

* If a collection is deemed to be exceptionally valuable, in that it becomes well-used or heavily cited, or is related to collections already in the main collection within ReShare, it can be promoted to CURCAT1, via a formal appraisal.

5. Using the Appraisal Grid to make decisions for ingest and access routes

The proposer completes the grid (Appendix A) with a number of recommendations for ingest and access pathway. The Data Appraisal Group (DAG) uses a collection's completed Appraisal Grid, together with the presenter's case for support, to agree decisions about the collection's outcome. DAG makes decisions about

- whether to acquire the collection or not
- its curation category (CURCAT 1-5)
- its anticipated access and delivery mechanism; options for CURCAT 1 and 3 are
    - Access type: Open; Safeguarded (End user licence; special licence; special conditions); Controlled (approved/accredited researcher access)
    - Data delivery system: Download system; Nesstar; UKDS.Stat; Infuse, WICID; Casweb; Secure Lab remote access; Secure Lab safe room; Qualibank; ReShare; Data Services as a Platform; Open API; Other added value
    - Suitability for developing into a teaching dataset
- whether it might merit any high-level promotion, user-oriented activity such as a guide, workshop or any special communication activity
- a possible or expected release date; how urgent ingest might be, deciding low, medium or high priority for processing

Decisions are recorded in an acquisitions database (known as Mirage), which are available to the Ingest team when the study is handed over for ingest work.

6. References

[URLs checked 13 February 2018]

UK Data Forum (2013) UK Strategy for Data Resources for Social and Economic Research http://www.esrc.ac.uk/files/news-events-and-publications/news/2013/uk-strategy-for-data-resources-for-social-and-economic-research/.

NERC Data Value Checklist (2013) http://www.nerc.ac.uk/research/sites/data/policy/data-value-checklist/.

DCC Appraisal and Selection (2008) http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/appraisal-and-selection.

Appendix A: UK Data Service Collection Development Appraisal Grid

Acq No. [insert acquisition number]
Depositor:
Title:
Short description:

Strategic value	Yes	No	Comment
Potential secondary use and analysis for research			
Teaching and learning use			
Replication and validation of research			
Relevance to the UK Data Forum Data Strategy			
Longitudinal data on people and families			
Addressing multidisciplinary challenges (e.g. biosocial data)			
Data about organisations			
Consumer databases			
Social media as research data			
Cross-sectional data on people and families			
International data			
Administrative data			
Big data			
Does the collection have a mandate to share with UK Data Service?			
Have these data been recommended by Data Infrastructure Strategic Advisory Committee?			
User need and analytic value			
Data useful for supporting decision-making and policy formation			
Data useful for international, longitudinal or comparative research (e.g. time series or harmonised)			
Authoritative source of high quality data and / or widely cited			
Microdata otherwise not available to the research community (e.g. disclosive)			
Data directly requested by a critical number of users			
Unique, unrepeatable data or data that are costly to reproduce			
Are there related data already in the Collection?			
Statement on potential use:			

Usability and accessibility			
Data have definable conditions within the access spectrum			
Legal and ethical agreements do not prohibit re-use, e.g. consent to share/reuse data, IPR, copyright, DPA			
Data has adequate supporting documentation/contextual information			
Data are produced in, or can be converted at minimal or no cost to, a suitable format for reuse (without need for specialised software)			
Do data exceed 1GB in size?			If yes, inform Big Data Team and Data Security Manager
Third party host (e.g. API): is reliable and reputable			
Third party host: is the data likely to change/be updated over time?			
Third party host: suitable preservation processes are used to manage and curate data			
Third party host: a persistent identifier can be made available (robust access)			

Proposer Recommendations		
Proposer name (autocomplete)		
Recommended acceptance outcome?	□ Accept □ Unable to accept	□ Referred ReShare □ Referred external
Curation Category	□ CURCAT 1 (long-term curation) □ CURCAT 2 (short-term management)	□ CURCAT 3 (delivery only) □ CURCAT 4 (discovery only) □ CURCAT 5 (preservation only)
Anticipated access (CURCAT 1 and 3)	□ Open, without registration □ Safe-guarded (End user licence) □ Special conditions	□ Special licence □ Controlled access (Approved researcher)
Data Delivery	□ Download System □ Nesstar □ UKDS.Stat □ InFuse □ WICID □ Casweb □ Secure Lab remote access	□ Secure Lab safe room □ Qualibank □ ReShare □ Data Lake □ Open API □ Other added value
Suitable for high-level promotion/workshop/communication activity?		
Release date deadline		
Expected release date		
Recommended Priority Level	□ High □ Medium	□ Low