A Digital Curation Centre
'working level' guide

Where to keep research data

DCC Checklist for Evaluating Data Repositories

Angus Whyte (DCC)

Please cite as: Whyte, A. (2015). 'Where to keep research data:
DCC checklist for evaluating data repositories' v.1 Edinburgh: Digital Curation Centre.
Available online:
www.dcc.ac.uk/resources/how-guides

Preface

What is the Digital Curation Centre?

The Digital Curation Centre (DCC) is a world-leading centre of expertise in digital information curation.

Our primary aims are:

- Build research organisations' capacity, capability and skills in managing data
- Support knowledge exchange and development of good practice in digital curation

What guidance publications does DCC produce, and for what audiences?

We aim to help a broad audience including research support, library and IT service staff, as well as data producers. Normally the guidance is not specific to disciplines. It includes:

Briefings outline current topics, explaining their importance, relevant roles and responsibilities, current issues and upcoming challenges.

How-to Guides provide working-level knowledge, background concepts and practical steps towards implementing capabilities in your organisation, and how to align these with best practice.

Checklists also provide working-level knowledge, but cover less of the background. They aim to ensure that practitioners have addressed the full scope of a challenging curation topic, and provide further sources.

Fold-outs are quick start guides for researchers, summarising good practice in relation to a specific data management topic.

Case studies describe how some organisations develop and deliver curation. They aim to complement a How-to Guide or Checklist, and illustrate practical challenges and lessons learnt.

Examples are shorter and offer 'who, what, where, how, when' summaries of approaches to RDM service delivery.

Catalogues offer half page entries profiling key resources in RDM. These include the Tools and Services Catalogue, and Disciplinary Metadata Directory.

How can I reuse the content of this publication?

Introduction

Who is this checklist aimed at?

This checklist aims to assist research support staff in UK Higher Education Institutions whose task is to help researchers make informed choices about where to deposit data. It is also relevant to managers with responsibility for defining policy on Research Data Management (RDM).

The checklist complements '*Five steps to decide what data to keep - DCC Checklist for Appraising Research Data.* ([1]) In common with the '*Five steps*' checklist, the guidance is meant to provide a template. It can be adapted to complement the services the Institution provides or recommends to researchers, and policies that govern how those services operate. The checklist reflects current UK funders and institutional policies, which can be tracked on the DCC website.[2]

What does this checklist cover and what does it exclude?

Choosing a long-term[3] service to look after data means asking questions similar to those you ask when choosing a publisher; 'if I hand this over, will they review it, safeguard the content, and make sure it is accessible for as long as it is of value?' This checklist relates these questions to the following key considerations:

1. **is a reputable repository available?**
2. **will it take the data you want to deposit?**
3. **will it be safe in legal terms?**
4. **will the repository sustain the data value?**
5. **will it support analysis and track data usage?**

A repository can refer to any storage service that offers a mechanism for managing and storing digital content.[4] To help you identify the best options for you, we do the following:

- describe different types of service and suggest some pros and cons of each.
- list sources of information on where to find candidates of the various types.
- introduce three levels of service that can help match a repository to the depositor's requirements for their data collection.

The checklist is mainly concerned with external third-party repositories that offer a managed service to the research community. Guidance on more generic cloud-based storage services is covered but only to the extent of recommending sources of further guidance. This checklist does not provide criteria for choosing the platforms your institution might use to *provide* a data repository or catalogue service. Another DCC guide covers that: *How to Evaluate Data Repository and Catalogue Platforms* (2016).

When should the checklist be applied?

Long-term storage choices should be considered at an early stage in research. Before deciding on the location(s) there should be a shared understanding of how the points below will be dealt with. They may be already documented in a pre-award Data Management Plan (DMP).

- What types of data will be produced, using what methods and formats
- What documentation and structured information (metadata) will be produced
- Legal or ethical factors affecting data sharing e.g. consent, privacy, copyright or commercial considerations
- Storage and security needs, considering access by collaborators and the volumes likely to be produced
- How long the data needs to be retained
- How decisions on what data to keep will be made
- How data that is suitable for sharing will be shared
- What costs you will need to meet, including time to prepare data for long-term use

If the research funder requires a DMP, they will typically expect a preferred deposit location to be named in it.[5] However, unless the funder specifies where data should be deposited, the choice will often be made later in the research. Even with a DMP in place, the choice of repository can be revised later to reflect the nature of the data that has actually been collected.

---

[1]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
[2]DCC Policy resources are available at www.dcc.ac.uk/resources/policy-and-legal
[3]The phrase long-term is used here to mean 'beyond the end of the research project'. 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 -5 years.
[4]Repositories Support Project (n.d.) ' What is a repository'. Available at: http://www.rsp.ac.uk/start/before-you-start/what-is-a-repository/
[5]DCC guidance on preparing a DMP is available here: www.dcc.ac.uk/resources/data-management-plans

How do we define data and data collections?

This guide follows 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".*[6] Data could be any of the following: [7]

- **source data** - all data collected, created and used by the research, including data held elsewhere
- **assembled datasets** - data extracted or derived from the above
- **referenced data** - any subset of the above that has been used in analysis or to draw conclusions. Consistent with whatever is considered 'supplementary material' to research findings in your domain.

We use the term data collection for any combination of the above with other information or digital objects that would be needed to enable access and interpretation.[8] This could include, for example, a notebook, protocol, software or set of instrument calibrations. What you include in a data collection should take account of:

- **reasons for keeping it** - these are the most important factor in determining what metadata, documentation, dependent software or other objects will be needed to interpret and use the data. A single research project could easily produce a number of data collections, each matching a different potential use. For example different outcomes from research may be written up in a number of articles for different audiences and publishers, and each article could have a related data collection deposited in a different repository;
- **research domain the data was collected for** - as a repository service will have expectations about the evidence needed to reproduce findings in the domain;
- **the user community** – the depositor may expect a repository to give added value by making the data visible to a particular community, who will also have expectations about how the data is presented or packaged with metadata and other information;
- **repository terms and conditions** - some repositories allow different access permissions and/or licence conditions to apply to the various files in a data collection. Others may be less flexible, meaning that if different conditions apply to subsets of the data or to software, or to documentation, then these would need to be split into separate collections or even deposited in different repositories.

What are the options?

There are hundreds of repositories worldwide. Some cater for a specific research domain, while others are general-purpose repositories. They may be called something other than a repository, for example a data centre or archive.

The choice of storage location will mainly depend on whether the data can be shared publicly or not. Some pros and cons of the available choices are shown in Table 1.

Where should I share data for public access?

As a general rule, a domain repository is likely to offer the best home for data that can be publicly shared. Currently the preferred way of sharing is as Journal article supplementary material, according to one survey.[9] That may change as more data repositories become available, and more Journals recommend depositing in them.

Some may prefer to share data through a departmental, project, or personal web site. In that case the institution should be provided with a record of the data collection, and preferably the option of holding a copy. This checklist can be used for reference in discussion with researchers, to consider whether small local facilities are enough to meet the requirements, and identify alternative solutions that may be available centrally from the institution, or externally.

Small data collections and websites can grow into domain repositories. They are not necessarily an inferior choice if the host institution has the underlying support and infrastructure available to sustain their long-term growth.

---

[6] *C.L. Borgman (2015). Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press*

[7] *Adapted from: Peter Burnhill, Muriel Mewissen & Adam Rusbridge (2014) ' Where data and journal content collide: what does it mean to 'publish your data" Presented at 'Dealing with Data Conference', 26 August 2014, University of Edinburgh Library. Avaliable at: https://www.era.lib.ed.ac.uk/handle/1842/9394*

[8] *The Research Data Alliance defines a data collection as 'an aggregation that contains digital objects and digital entities. The collection is identified by a PID and described by metadata' (see 'Core Terms & Model v1.6' Data Foundation & Terminology Working Group, available at: dx.doi.org/10.15497/06825049-8CA4-40BD-BCAF-DE9F0EA2FADF) For other definitions of the term 'data collection' see, e.g., the Dryad data repository ('What is a data collection' at http://datadryad.org/pages/faq) and Open Knowledge Labs 'Data collections' (at: dataprotocols.org/data-packages/index.html). The term 'information package' is used in the OAIS standard, see section 4.2.2.1 in CCSDS (2012) Reference Model for an Open Archival Information System (OAIS) Available at: public.ccsds.org/publications/archive/650x0m2.pdf*

[9] *According to Wiley's 'Researcher Data Insights' survey two-thirds of the survey respondents who say they share data do so as supplementary materials; see Wiley Exchanges blog, 3rd Nov. 2014 available at: http://exchanges.wiley.com/blog/2014/11/03/how-and-why-researchers-share-data-and-why-they-dont/*

Sharing data for public access

Domain/discipline-specific data repository, data centre or 'scientific database'
- Pros - most likely to offer both the specialist domain knowledge and data management expertise needed to ensure your data collection is properly kept and used
- Cons - most likely to be selective, requiring advance planning of the effort needed to meet high standards for metadata and documentation

General-purpose data repository: e.g. Dryad, Figshare, Zenodo
- Pros - most likely to offer useful search, navigation and visualisation functionality
- Cons - requires scrutiny of terms and conditions to ensure consistency with your funder, journal or institution's policies on cost recovery, copyright/IP, long-term preservation

Institutional data repository
- Pros - most likely to accept any data of value, especially if no suitable home can be found for it elsewhere, and to ensure that policy requirements for long-term access are met
- Cons - unlikely to be as well-resourced as either general-purpose or domain repositories

Journal supplementary material service
- Pros - most likely to comply with the journal or publisher's requirements
- Cons - may be costly, unlikely to offer a data repository's functionality or long-term solution

Departmental, project or personal web page
- Pros - might provide functionality tailored to your data collection and/or your existing data users and peer network
- Cons - least likely to make your data collection visible to new users and contacts, or to sustain long-term access to your data collection

Keeping data secure for the long-term

Institutional data archive or vault
- Pros - most likely to have considered the total costs of long-term storage, and to ensure that policy requirements for long-term access are met
- Cons - may be less likely to offer the same ease of use as third-party storage or archiving services

Safe centres or havens
- Pros - most likely to meet stringent security requirements for handling sensitive data, and to ensure that legal requirements for data protection are met
- Cons - may be less likely to offer similar levels of digital preservation as a data archiving third-party service or institutional data archive

Cloud storage third-party services
- Pros - most likely to offer easy to use file store and share functionality
- Cons - long-term reliability and costs of data retrieval may be unpredictable; terms and conditions need careful scrutiny to ensure it complies with policy requirements for long-term access and other legal requirements, e.g. a data centre location within the European Union

Data archiving third-party services
- Pros - likely to offer cost-effective long-term storage with guarantee of accessibility, including data that may not be shareable for confidentiality reasons
- Cons - less likely to offer administrative interface to manage access and preservation policies (although some services offer repository integration)

Table 1. Pros and cons of typical options

Where should I keep confidential data for the long-term?

The main focus of the guide is on identifying repositories for data sharing, rather than on managing confidential data. If your data is classed as confidential you should weigh up the convenience and cost of a third-party external storage versus your institution's offering, taking into account your trust in third parties to manage the risks of data loss or disclosure over the long-term.

Confidential data should generally not be deposited with a repository designed to maximise visibility and accessibility. However an open access repository can meet some needs for data that cannot be fully open by:

- temporarily embargoing data before publication
- publishing a metadata record for data that is held elsewhere under controlled access

Your institution's IT support or research ethics contact may be able to advise on a local safe centre for data management.

There are shared services for confidential data, for example:

- Jisc works with the UK's Administrative Data Research Network (ADRN) to develop network links between centres that deal with sensitive data, and to provide higher assured access management for users
- for researchers based at any UK academic institution or ESRC-funded research centre, the UK Data Service (UKDS) has developed protocols for handling sensitive and confidential data. They provide depositors who have offered data that meets their selection and appraisal criteria with three levels of data access, see Table 2.

	Open data	Safeguarded data	Controlled data
Security requirement	Suitable for fully anonymised data or data with agreement to publish personal details	Partially anonymised data or data with agreement to publish personal details, and where owner wishes to track usage	Too detailed, confidential or sensitive to be downloaded
Level of access	Accessible without user registration	Accessible to authenticated users	Accessible to authenticated users, using secure remote access or secure onsite room
Legal conditions	Under open licence, either Open Government Licence (OGL) for Crown Copyright data or Creative Commons for other data	Requiring an End User Licence and, where appropriate, special conditions agreed to, or data owner approval	Requires user accreditation and registration through training and approval by a data access committee

Table 2. Three levels of data access requirement and conditions (source UK Data Service[10])

A number of third-party cloud storage and data archiving services are certified to the ISO 27001 standard for information security (see best practice recommendations and relevant standards below). Costs aside, two main considerations here will be the acceptability of the provider's terms and conditions and what the provider offers to guard against risks such as data loss and unauthorised access.

[10]*See UKDS at: http://ukdataservice.ac.uk/deposit-data/how-to/regular-depositors/negotiate*

What about software code?

It is important to treat software code with similar levels of care as data and make arrangements for its storage and preservation. This particularly so when specific software is needed in order to use or interpret a research data collection. In some cases the software may be more essential to reproducing a piece of research than the data. The Software Sustainability Institute provides further guidance on benefits and methods of software reservation, including guidance on code repositories at: http://software.ac.uk/resources/guides/choosing-repository-your-software-project

What about the Jisc shared RDM service?

Jisc is aiming to provide a shared RDM service, including a data repository. This is scheduled to be available from late 2017, subject to a successful pilot phase. The service will offer institutions the ability to provide many of the capabilities described later in this guide, including policy and standards compliance, and will provide ease of use and interoperability. Institutions will be able to brand the service, so that to the depositor it is presented as an integral part of their institution's RDM service. Further information is available on the Jisc website.

Where can I find a data repository?

Funders

Some UK Research Councils and other major funders stipulate that data produced in a project they fund is offered to a speific data centre or repository identified in their policy (see Table 3).[11]

Others funders may mandate deposit to one of a number of recommended repositories. For example the Wellcome Trust provides a list of data repositories and database resources in its *Guidance for Researchers* on data sharing.[12] These are organised by life sciences study type, for example microarray, or nucleotide databases.

Repository registries

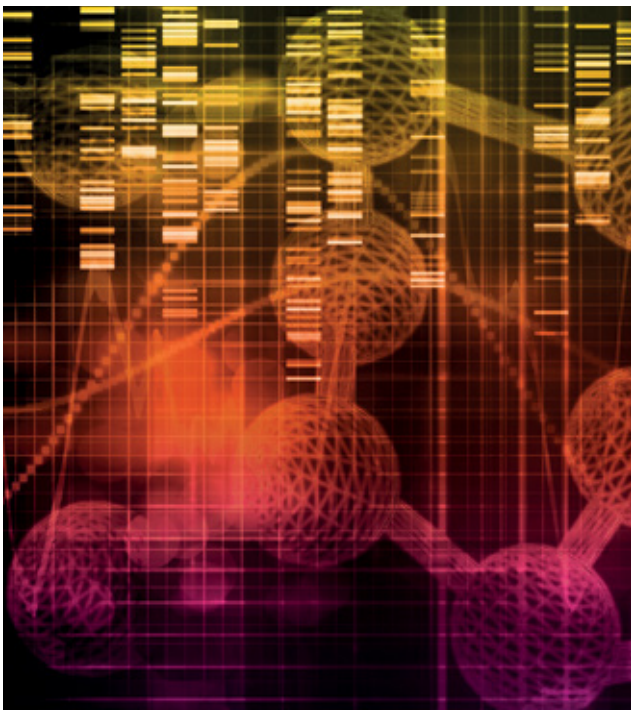
The sources below offer facilities to search and browse descriptions of research data repositories, and are also good reference points for advice about data sharing.

Re3data

The re3data.org registry covers research data repositories internationally and across disciplines. It is managed under the auspices of DataCite, an international organisation that issues Digital Object Identifiers (DOI's) for data. Listed repositories are reviewed before inclusion, and you can search them by discipline, host country, and various content types. Re3data merged in 2015 with the registry Databib.org, hosted by Purdue University Libraries and also endorsed by DataCite.

Biosharing

The biosharing.org catalogue includes bioscience databases described according to domain guidelines and standards. It is partly compiled with the support of Oxford University Press and Re3Data.org.



[11] *DCC (n.d.) Funder policies overview, available at: www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies and the SHERPA/JULIET directory of Open Access policies at: www.sherpa.ac.uk/juliet/index.php*

[12] *Guidance for Researchers available at: www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Guidance-for-researchers/WTX060360.htm*

Funder	Data deposit policy
AHRC	Significant electronic resources or datasets must be made available in an accessible depository for at least three years after the end of the grant. In Archaeology, the Archaeology Data Service (ADS) must be consulted within three months of the start of the proposed research and data must be offered for deposit within three months of project completion. Source: Arts & Humanities Research Council (2015) Research Funding Guide, available at: **www.ahrc.ac.uk/funding/research/researchfundingguide/annexes/gc23acknowledgementofsupport/**
BBSRC	The BBSRC does not run its own data centre but provides examples of existing databases and public repositories that it supports in its data policy. Source: Biotechnology and Biological Sciences Research Council (2010) Data Sharing Policy, available at: **www.bbsrc.ac.uk/about/policies/position/policy/data-sharing-policy/**
ESRC	Grant holders must ensure that data created as a result of research projects are made available for reuse. Data can be deposited with the ESRC data service provider (currently UK Data Service), or an appropriate responsible digital repository, provided certain criteria (FAIR principles])are met. In all cases the grant holder must provide metadata for resource discovery via the UK Data Service to maximise the discoverability of ESRC data assets. Source: Economic and Social Sciences Research Council (2015) Research Data Policy FAQs, available at: **www.esrc.ac.uk/about-esrc/information/data-policy.aspx**
EPSRC	Research organisations must ensure that data is securely preserved for a minimum of 10 years beyond the end of any researcher privileged access period. This may be in a domain or publisher-specific repository that commits to the continued accessibility of retained data, with 'legal safeguards'. This means 'legislation governing access to, or otherwise affecting, the security of information held in digital or electronic form'. Source: Engineering and Physical Sciences Research Council (2014) Clarifications of EPSRC expectations on research data management, available at: **www.epsrc.ac.uk/about/standards/researchdata/expectations/**
MRC	It is the responsibility of the study director or unit director to meet the requirements for his/her studies. Studies may share their data by archiving their data collection (or a subset) at a discipline-based repository, for instance the UK Data Archive, or at an institutional repository that can preserve the data and make them available to users. This may be particularly suitable for legacy data collections and for studies that no longer actively collect data or receive funding. Source: Medical Research Council (2011) MRC policy on sharing of research data from population and patient studies, available at: **www.mrc.ac.uk/research/research-policy-ethics/data-sharing/**
NERC	Researchers are expected to offer data for deposit in the NERC-funded data centre relevant to their subject or discipline. These include: British Oceanographic Data Centre, National Geoscience Data Centre, British Atmospheric Data Centre, NERC Earth Observation Data Centre, Polar Data Centre, Environmental Information Data Centre, UK Solar System Data Centre, and the Archaeology Data Service. Information on all data held within the data centres will be made available through the NERC Data Discovery Service. Source: Natural and Environmental Sciences Research Council (2011) Guidance notes for the NERC data policy, available at: **www.nerc.ac.uk/research/sites/data/policy/**
STFC	STFC would normally expect data to be managed through an institutional repository, , a university, a laboratory or an independently managed subject specific database. Several data centres, services and research portals are in place, such as the UK Solar System Data Centre, the Chemical Database Service and the Diamond Data Portal. These are generally organised on a subject basis rather than serving the outputs of the whole council, and deposit is not mandated. Source: STFC Scientific Data Policy, at: **www.stfc.ac.uk/about-us/freedom-of-information/scientific-data-policy/**

Table 3. UK Research Council policies on depositing in data repositories

Data journals

Data journals consist of data articles that describe how, why and when a dataset was collected and any derived data product. Rather than presenting any analysis or conclusions, a data article may present arguments about the value of the data for future analysis. A data journal will not normally host data itself but recommend where it should be deposited, and then link to it. This tends to make them useful sources of advice about repositories. Examples include:

- *Scientific Data* (Nature Publishing Group) which requires authors to submit their dataset to "community-recognised data repositories", listed by discipline atwww.nature.com/sdata/data-policies/repositories

- *Ubiquity Press* series of open access metajournals have dedicated open access repository on the Dataverse network.[13] See: www.ubiquitypress.com/site/research-integrity/

Journal policies

If you are targeting particular journals to publish your research, you should check for any policies on data. Journals are increasingly requiring authors to deposit the data underlying their articles in a recognised repository, to complement or replace any in-house facility for supplementary materials. Examples include:

- Public Library of Science (PLOS) recommends repositories it recognises as "trusted within their respective communities" and also points to re3data as a more general source. See journals.plos.org/plosone/s/data-availability

- *Science's* standing policy is that when a paper is published, archival data relevant to its results or methods must be deposited in a publicly accessible database

- many journals in the ecology domain have adopted the *Joint Data Archiving Policy* [14] coordinated by the Dryad data repository, which recommends deposit in Dryad or other repositories that meet certain criteria like those described in this guide

Learned & professional societies

A society relevant to the research domain may offer advice on data sharing that includes recommendations about where to deposit data. Few societies offer a data repository at the moment but this may change. For example the Royal Society of Chemistry launched the open access *Chemical Sciences Article Repository* in 2013, with the intention of expanding it to include data.

Jisc purchasing frameworks for cloud storage and data archiving

Jisc offers all Janet connected organisations access to cloud storage, data centre, and data archiving services that have been procured for the sector. This helps your institution avoid procurement costs, and ensures that providers offer terms and conditions that meet the common needs of UK institutions, e.g. a high level of assurance on legal compliance, meeting information security and quality assurance standards. Further information and advice on choosing cloud providers is available from Jisc at: www.ja.net/products-services/janet-framework-agreements .

Safe centres

Your institution's central IT service should be able to offer guidance on locally available centres for holding sensitive data. The Administrative Data Network maintains a list of safe centres where researchers may access, but not download, confidential and sensitive data. Some of these enable secure data pooling with collaborators, rather than archiving of the results, but may contain useful pointers towards suitable facilities. Further information is available at adrn.ac.uk/protecting-privacy/secure-environment/safe-centres .

[13] *See Ubiquity Press at http://www.ubiquitypress.com/site/research-integrity/ and Dataverse at: https://dataverse.harvard.edu/dataverse/ubiquity-press?q=&types=dataverses*
[14] *Joint Data Archiving Policy. Available at: datadryad.org/pages/jdap*

What is the right service level for the data?

Data repositories are commonly thought of as offering different levels of curation and preservation. For example the Royal Society report *Science as an Open Enterprise* talks about 'tiers of repository service as shown in Figure 1.

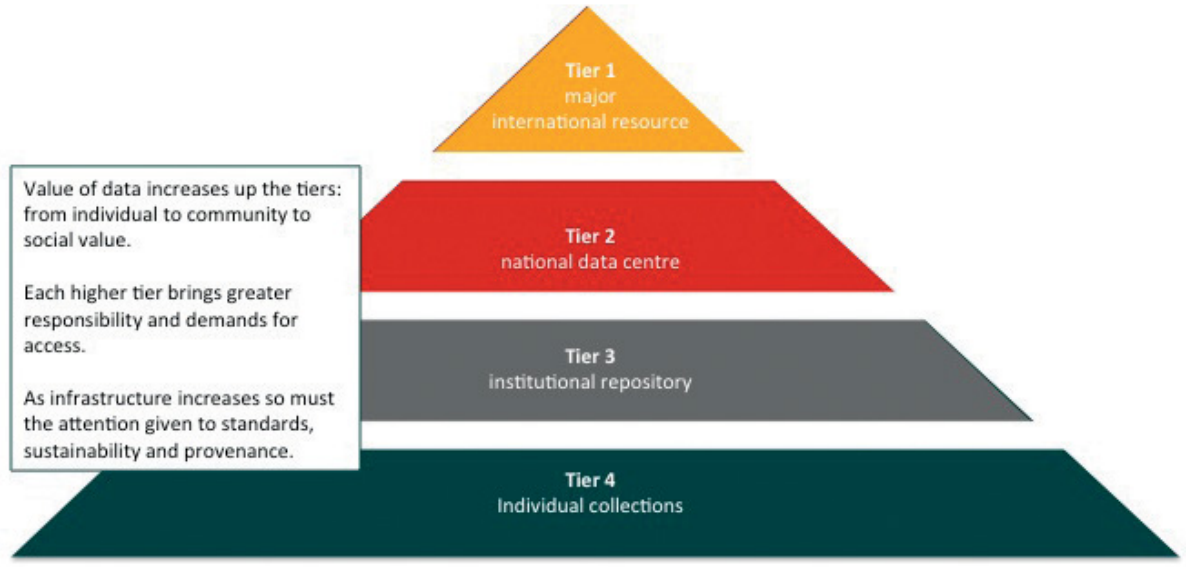


Figure 1. Hierarchical view of repositories, adapted from Royal Society report Science as an Open Enterprise,p.60.[15]

The approach in this guide is slightly different; rather than classify a repository into a single tier we assume that different aspects of its service may be important for particular data types and may offer the following three levels of capability:

Service Capability Levels		
Level 1	**Level 2**	**Level 3**
☐ Consistent with funding body mandates and with minimum levels of community standards, including the mandatory elements of Data Seal of Approval, and US National Digital Stewardship Alliance (NDSA) preservation level 1.	☐ Aligns with best practice standards and community guidelines, e.g. those from Force11 Data Citation Implementation Group (DCIG), NDSA preservation level 2, and SCAPE Preservation Guidelines	☐ Exceeds sector norms, e.g. through certification on the higher levels of international standards for Trusted Digital Repositories or relevant community guidelines e.g. NDSA preservation levels 3-4.

Table 4. Three levels of service capability

[15] *Royal Society (2012) Science as an Open Enterprise. Available online: http://royalsociety.org/policy/projects/science-public-enterprise/report/*

Checklist: is it the right repository for your data?

The checklist that follows addresses the five key questions posed in this guide:

1. is the repository reputable?
2. will it take the data you want to deposit?
3. will it be safe in legal terms?
4. will the repository sustain the data value?
5. will it support analysis and track data usage?

By going through each question and assessing the level that best fits the needs of the data to be deposited, you should be able to make a balanced decision on how well the repositories (or other solutions) on your shortlist meet the data's needs.

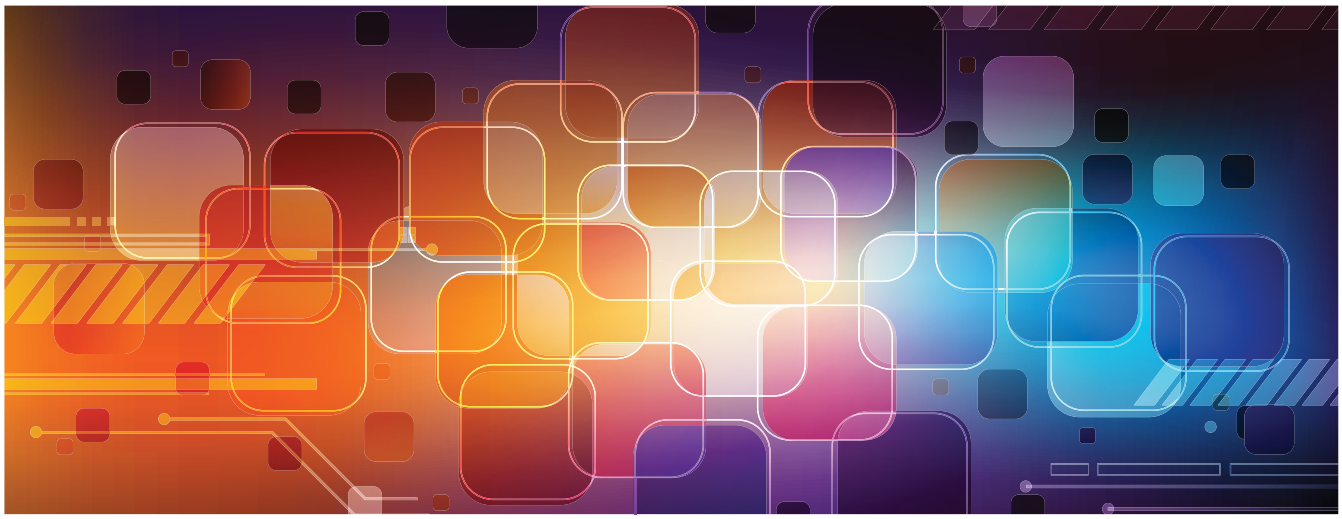
Please note: the checklist offers a shortcut through relevant policies and standards. It offers no guarantee a repository complies with any specific standard or funding body policy. If you need to check compliance please refer to the relevant source of further guidance at the end of the guide.

1. Is the repository reputable?

Reputation		
Level 1	**Level 2**	**Level 3**
☐ Listed in the re3data or Biosharing registries, or broadly recognised in the research domain	☐ Endorsed by a relevant funder, journal, or learned/ professional society	☐ Certified to an appropriate international standard

Points to consider

Certification is relatively new - using Re3data.org you can search for certified repositories. Certification standards are new however. There are many good repositories that are not compliant, and that is likely to continue for some time. See ' Where can I find a data repository?' for pointers to repositories endorsed by funders, journals and learned societies.



2. Will the repository take the data you want to deposit?

Collection policy

Level 1	Level 2	Level 3
☐ Accepts research data regardless of the data type or domain	☐ Focuses on data types or domains similar to that which you have to deposit	☐ Has an international reputation in the domain, or for publishing data similar to that which you have to deposit

Points to consider

Factors likely to affect the repository collection policy. This may be referred to as a Collections Development Policy. Otherwise if there is nothing by that name visible on the site, the 'deposit guidelines' should be checked. With the obvious exception of general-purpose repositories, most accept data that relates to a particular institution, funding body, research topic or domain. Data may need to meet certain criteria such as:

data or study type - a focus on outputs of particular data collection methods or instruments. e.g. questionnaire-based surveys, clinical trials, microarrays, space telescope observations, qualitative social research, meteorological observations, medical imaging, crystallographic images, or audio samples.

file format - repositories commonly prefer open rather than proprietary formats; usually this is stated as a preference rather than a hard rule, but any format that carries licence conditions with cost implications for the repository is likely to be excluded (e.g. certain moving image formats).

Repositories will also be governed by terms and conditions that may exclude data on various grounds such as copyright or data protection (see Question 3 in the checklist).

3. Will the data be safe in legal terms?

For this criterion we first consider the basic legal terms and conditions to check. Here only one capability level is given, as a repository either will or will not comply. Then we consider licensing, disclosure risk and access control, where a repository may offer different levels of capability you can match to your needs. Note that Re3data provides relevant details for the repositories it lists, under 'terms'.

Legal terms and conditions

Level 1

- ☐ **Personal data** or data which may identify individuals when linked to other data should not be stored outside the European Economic Area, unless in a legal jurisdiction that ensures personal data is adequately protected
- ☐ By agreeing to the terms and conditions the depositor will not be breaching other **Data Protection** principles, or the terms of any confidentiality agreement with data subjects or owners (e.g. consent form, consortium agreement)
- ☐ By agreeing to the terms and conditions the depositor will not be in breach of **copyright**, or any contract terms covering **Intellectual Property** in the research, (e.g. the grant conditions or a consortium agreement)
- ☐ Anything deposited that is not publicly accessible can be retrieved by the institution in response to a valid **Freedom of Information** request

Depositor responsibilities

☐	**Licensing:** You can assign rights information to a collection, applying standard terms and a limited range of open access license options	☐	You can apply a standard end user licence and, where appropriate, select from a range of special conditions or data owner approval for sharing	☐	You can apply bespoke access conditions or your own licence where appropriate
☐	**Disclosure risk:** Depositors must confirm that data was collected or created in accordance with legal and ethical criteria prevailing in the data producer's geographical location or discipline	☐	Repository limits access according to conditions appropriate to the level of disclosure risk you inform them about	☐	Repository reviews and manages disclosure risks in the data you deposit with them
☐	**Access control:** Enables depositors to restrict file/ object access and edit permissions to approved users for specific periods	☐	Enables depositors to restrict which users can access, edit, move or delete files/objects	☐	Enables depositors to define roles and restrict file/object actions accordingly

Points to consider

Check prior commitments - the person responsible for the data (e.g. Principal Investigator or data steward) must consider any contractual commitments made to research funders, suppliers, partners, participants or subjects. These include consortium agreements, and informed consent forms, as the right to archive the data must be obtained. Guidance on getting informed consent for archiving is available from the UK Data Service.[16]

Data protection - Your institution will provide guidance on personal data and how to comply with the Data Protection Act 1998, normally through a research ethics committee, Data Protection Officer or Records Manager. Further information is available from the *Information Commissioner's Office* (ICO) [17] and the *UK Data Service*.[18] It is important to note that EU Data Protection regulations are likely to be strengthened in 2016, and will include significant changes in service providers' obligations to citizens whose personal data they control.[19] This will affect both institutions and the service providers they use.

Copyright: If your research has made use of **third-party** data under licence, you should check whether the licence permits you to archive it for reuse, seek permission if the effort in doing so makes a significant difference to the long-term value of your data, or exclude it from the data you archive. Copyright law determines whether copyright exists in data produced in your research, for example, whether it meets the legal definitions of either a database or literary work. In law these are very broadly defined, and it is safer to assume copyright does exist than that it does not. So you also need to determine who owns the data. If the data was **produced by a researcher** employed by your institution, by default you should assume the institution owns the copyright, unless a research contract or other valid agreement stipulates otherwise. If a **student** produced it, they will be the owner of copyright, unless there is a valid agreement to the contrary.

There is also further guidance on legal aspects of RDM in the DCC publication *Five Things You Need to Know About RDM and the Law: DCC Checklist on Legal Aspects of RDM*, available at: www.dcc.ac.uk/resources/how-guides/rdm-law .

If in doubt, you should get advice from your institution's legal team. Further information is also available from *Jisc Customer Services* (customerservices@jisc.ac.uk) and the archived *Jisc Legal site*.[20]

---

[16] *UK Data Service (n.d) Consent for data sharing. Available at: ukdataservice.ac.uk/manage-data/legal-ethical/consent-data-sharing.aspx*
[17] *ICO (n.d.) Guide to data protection. Available at: ico.org.uk/for-organisations/guide-to-data-protection/*
[18] *UK Data Service (n.d.) Data protection. Available at: ukdataservice.ac.uk/manage-data/legal-ethical/obligations/data-protection.aspx*
[19] *European Commission (2015) 'Data Protection Day 2015: Concluding the EU Data Protection Reform essential for the Digital Single Market' Fact Sheet. Available at: europa.eu/rapid/press-release_MEMO-15-3802_en.htm*
[20] *Jisc Legal (2011) 'Who owns copyright in works created in universities and colleges' Available at: www.jisclegal.ac.uk/ManageContent/ViewDetail/ID/1893/Who-owns-copyright-in-works-created-in-universities-and-colleges-6-June-2011.aspx*

4. Will the repository sustain the data value?

There are many ways that data producers, their institutions, the research community and wider society can obtain value from data that is published and archived for the long-term (further guidance on reasons to keep data is available[21]). The main ways that a repository can add value are by making the data FAIR, i.e. making it *findable, accessible, interoperable*, and as *reusable* as possible for as long as required.

The repository must provide a landing page, with a stable identifier for the data you deposit. It must also enable the data collection to be cited, and preferably should publish further metadata to make the collection more widely discoverable.

Connections to other research management systems are becoming essential. The repository must enable links from the data to other research outputs. Preferably it should also allow metadata to be automatically harvested by other institutional systems and by external systems, such as discovery services, to maximise exposure. Exposing this metadata can exploit any links to the IDs of the data producer (e.g. ORCID), and related outputs. Ideally Linked Open Data schemas should be used to expose the metadata, enabling relationships between producers, facilities and outputs to be comprehensively analysed, in order to inform institutional research strategy.

Findable, accessible and interoperable

Level 1	Level 2	Level 3
☐ **Metadata publishing:** Data collections are catalogued in a repository according to funder expectations so that they are discoverable by title, creator, and date of deposition	☐ Repository publishes other pertinent information as metadata fields to enhance cross-disciplinary discovery	☐ Metadata is catalogued to enhance reuse according to sector-leading standards, or to fulfil domain-specific purposes
☐ **Stable identifiers:** Enables a DOI or other open standard identifier to be assigned to a landing page for each ingested dataset/ collection	☐ Supports assignment of related persistent IDs per dataset/ collection	☐ Supports assignment of multiple persistent IDs at different levels of granularity within dataset/ collection
☐ **Discovery metadata:** Provides Datacite mandatory metadata and exposes it according to open access repository protocols	☐ Provides metadata elements to enable broader discovery (e.g. geo-spatial) to reflect best practice changes and local needs	☐ Exposes discovery metadata as Linked Open Data to optimise automatic discovery
☐ **Metadata harvesting:** Sufficient information can be harvested about data deposited with third-party repositories, to meet funders' needs for metadata on locally produced data and its relation to other outputs	☐ Metadata can be routinely harvested with links to data producer IDs (e.g. ORCID), any grant information and related outputs, enabling it to meet the institution's research administration needs	☐ Metadata on the externally held research data is sufficiently structured and organised to enable it to inform institutional strategy

[21] *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*

The repository should commit to open standards for file formats, and support linking of the data collection to contextual information that helps interpret it. For domain repositories, that could include support for domain-specific metadata terms or vocabularies.

In terms of preservation, the repository must have clear guidelines on what would happen to the data in the event that it ceases operation or changes its scope. That must include preserving the integrity of the data deposited, version control information, and preferably all aspects of the digital content that are significant for reuse.

Data reusability

Level 1	Level 2	Level 3
☐ **File checking and preservation planning:** Supports virus check and file format validation at ingest, and provides format inventory	☐ Supports file normalisation to selected preservation formats	☐ Supports format migration planning and implementation at dataset/ collection level
☐ **Data integrity & fixity:** Enables a digital fingerprint (e.g. checksum) to be assigned to all files in collection when ingested	☐ Provides routine integrity checks, and documents the results	☐ Supports replace/repair of corrupted data, and provides an audit trail on request of all checks performed and their results
☐ **Domain metadata & context information:** Supports linking of dataset to related records according to open repository protocols and standard identifiers	☐ Supports limited domain-specific metadata at repository level, e.g. geospatial terms	☐ Supports domain or content-specific metadata or vocabulary terms at dataset/ data collection level
☐ **Continuity strategy:** Enables data to be held in storage that's backed-up to a different location	☐ Enables data to be stored in storage that's backed-up to two separate locations, one of which is located off-site	☐ Enables data and metadata distribution across multiple locations according to defined rules/ policies
☐ **Version control:** Ingested objects and latest edits are time-stamped	☐ Change history for object and metadata is available on request	☐ Integrates with software versioning repository (e.g. Github) to link archived collections with dynamic data or software

Points to consider

Don't be afraid to ask - some of the details mentioned in the tables may be difficult to find. The Re3data catalogue can help here. Its listings include the repository's persistent identifier, metadata standards used, interlinking between data and publications (enhanced publication), and whether it supports versioning. Other details may be available on the repository's website, or by contacting them. If you cannot easily find details about a repository that matter to you contact the repository management team directly, as it helps drive up standards for everyone.

5. Will the repository support analysis and track data usage ?

A repository should provide its depositors with an indication of how many times the data has been viewed and downloaded. The repository should do more than supply basic metadata to allow a data collection to be cited; it is essential that users can search across the repository contents on basic fields e.g. creator, title, date of production. A repository should preferably offer functionality that amplifies whatever characteristics make the data valuable. This may simply be that it provides evidence underpinning another published output, in which case the only essential characteristic is to be able to follow the link to a publication. Preferably the repository will also encourage re-use by helping users to analyse the content, for example by offering domain-specific search features, extracting key terms from the data, or supporting visualisation of features extracted from it.

Analysis and usage tracking

Level 1	Level 2	Level 3
☐ **Citation and usage tracking:** Data collections are catalogued in a repository according to funder expectations so that they are discoverable by title, creator, and date of deposition	☐ Provides usage statistics for access (e.g. page views) and for downloads of the data collection if it is public	☐ Metadata about the deposited data is harvested by a national, international or domain-based data registry
☐ **Content search:** Enables a DOI or other open standard identifier to be assigned to a landing page for each ingested dataset/ collection	☐ Some content types may be searched on best practice metadata and retrieved content can be viewed	☐ Search and browse terms can vary per collection, according to collection-specific facets
☐ **Data mining & visualisation:** n/a	☐ Supports the extraction and visualisation of patterns in structured datasets, or of domain- relevant entities or terms	☐ Supports automatic discovery of semantically related entities

Points to consider

Weigh-up costs against benefits: The basic-level criteria are essential if you need to meet the expectations of UK Research Councils or other public research funders. However funders recognise that data management costs need to be offset against benefits, and it is up to you to judge how far the benefits a repository offers outweigh any costs, whether those are charges levied or costs of preparing the data for deposit.

Bibliography: policy, good practice and standards

Funding body policy expectations

Research funders typically use data policy to communicate their expectations at a general level, i.e. stipulating what they want done to safeguard data, rather than how a repository should do that. Usually funder policies apply to the Principal Investigator grant recipient. However some policies require, recommend, or imply that specific repository capabilities are provided. The checklist draws on the following examples to help define a minimum level of policy expectations. Please note that other funders may have different minimal compliance requirements.

EPSRC Research Data Policy Expectations

The UK Engineering and Physical Sciences Research Council mandates that institutions in receipt of research funding meet its expectations towards data management. It states nine expectations, which align with the broader Research Councils UK Common Principles on Data Policy. Expectations (v) to (viii) state the need for institutions to ensure researchers have capabilities available for data retention, persistent identification, metadata publication, and curation. While they apply specifically to EPSRC-funded research, many UK universities have used them as a basis for institutional RDM policy that applies more widely. Source: EPSRC (2011) Policy Framework on Research Data Expectations, available at: https://www.epsrc.ac.uk/about/standards/researchdata/expectations/

European Commission Guidelines on Data Management in Horizon 2020

The Guidelines are applicable to all projects participating in the Pilot on Open Research Data in Horizon 2020. While their main focus in on Data Management Plans, the Guidelines are accompanied by a Model Grant Agreement that requires projects to deposit data and associated metadata in repositories. Further details are available at: European Commission (2013) ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Concordat on Open Research Data

The (UK) Concordat on Open Research Data was produced by a multi-stakeholder working group, which includes RCUK, Jisc, the Wellcome Trust and Universities UK. The concordat "…aims to establish a set of expectations of good practice with the intention of establishing open research data as the desired position for publicly-funded research over the long-term". It provides generic guidance on the different responsibilities of researchers, their employers, and funders of research. Available at www.rcuk.ac.uk/research/opendata/

Journal policy criteria

Scientific Data - Repository Evaluation Criteria

Nature Publishing Group data journal Scientific Data requires datasets to be made available to editors and referees when manuscripts are submitted. Datasets must be shared with the scientific community as a condition of publication, and authors need to identify a data repository suitable for their datasets. As well as recommending specific repositories the journal identifies criteria it uses to select appropriate repositories based on earlier work in the PREPARDE project.[22] The criteria are available at: www.nature.com/sdata/data-policies#repo-suggest

Public Library of Science (PLOS)

Data Availability Policy states that authors "…are encouraged to select repositories that meet accepted criteria as trustworthy digital repositories". The journal policy lists repositories it regards as trusted and state that "If no specialized community-endorsed open repository exists, institutional repositories that use open licenses permitting free and unrestricted use or public domain, and that adhere to best practices pertaining to responsible data sharing, sustainable digital preservation, proper citation, and openness are also suitable for data deposition.". Further details are available at: journals.plos.org/plosone/s/data-availability

[22] *Callaghan, S. et al (2014) 'Guidelines on Recommending Data Repositories as Partners in Publishing Research Data' Vol. 9, No. 1, pp. 152-163 doi:10.2218/ijdc.v9i1.309*

Good practice recommendations

There are many sources of advice on best practices for storing, depositing and preserving data. The following have been consulted for this guide:

FAIR Guidelines

FORCE11 (Future of Research Communication and E-scholarship) proposes criteria for FAIR (Findable, Accessible, Interoperable and Reusable) principles for data objects. FORCE11 is "a community of scholars, librarians, archivists, publishers and research funders that has arisen organically to help facilitate the change toward improved knowledge creation and sharing".[23] The principals are available at: www.force11.org/group/fairgroup/fairprinciples

Data Citation Implementation Group (DCIG) Recommendations

In 2014 the FORCE11 Data Citation Implementation Group (DCIG) produced a set of common guidelines to operationalise compliance with the widely endorsed Joint Declaration of Data Citation Principles (JDDCP). The DCIG recommendations provide further detail in the areas of data citation, archiving, and machine accessibility. They are intended to be consistent with the FAIR Guidelines (above) and are described in: Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R. R., Duerr, R. et al. (2015). 'Achieving human and machine accessibility of cited data in scholarly publications', PeerJ PrePrints. Available at: peerj.com/preprints/697/

OpenAire Guidelines for Data Archive Managers

OpenAire2020 is a European Commission funded initiative to realise the vision of an open access infrastructure for scholarly communication and research information, primarily relating to research outputs of European funding streams. This includes mechanisms to harvest research data metadata from data archives, with the aim of supporting open science and tracking research impact. In 2013 OpenAire issued guidelines for data archives on applying the DataCite Metadata Schema to ensure compatibility with the OpenAIRE infrastructure. These make certain recommended DataCite properties mandatory. The guidelines are available at: https://guidelines.openaire.eu/en/stable/data/index.html

NDSA Guidelines on Preservation Levels.

The US National Data Stewardship Alliance (NDSA) Levels of Digital Preservation are described as "… a tiered set of recommendations for how organizations should begin to build or enhance their digital preservation activities. A work in progress by the NDSA, it is intended to be a relatively easy-to-use set of guidelines useful not only for those just beginning to think about preserving their digital assets, but also for institutions planning the next steps in enhancing their existing digital preservation systems and workflows." NDSA (2014) Available at: www.digitalpreservation.gov/ndsa/activities/levels.html

SCAPE Preservation Guidelines

The SCAPE project aimed to enhance the state of the art in digital preservation with a particular emphasis on scalability that is capacity to handle numerous digital objects that may be individually very large, heterogeneous or complex. The project collated preservation guidelines contained in the report 'D20.6 Final best practice guidelines and recommendations', available at: www.scape-project.eu/deliverable/d20-6-final-best-practice-guidelines-and-recommendations

Relevant international standards

These include standards concerning Quality Assurance, Information Security, and status as a Trusted Digital Repository. The latter group of standards assesses repositories on a broad range of operational and management criteria.

Trusted digital repository: ISO 16363 and related standards

The European Commission has endorsed a three-level framework that defines progressively more rigorous assessment of trustworthiness for digital repositories. Trustworthiness encompasses the organisation, technology and resourcing needed to operate an open access data repository. The three levels follow below.

1. Basic certification - based on 16 quality guidelines in the Data Seal of Approval (DSA).[24] This covers the repository's relationship with data producers and consumers, and its creation, storage and re-use of digital data. Repositories self-assess their compliance, and can get certification when their assessment is successfully peer-reviewed through the DSA organisation.International Council for Science World Data System (ICSU-WDS)[25] membership: repositories are assessed on similar criteria to the DSA when joining. A WDS-DSA Partnership group has proposed the DSA-WDS Catalogue of Common Requirements to better align the two standards. See further details below.

2. Extended Certification is granted to Basic Certification repositories which in addition perform a structured, externally reviewed and publicly available self-audit based on the ISO16363 standard or DIN 31644 standards.[26]

[23] *'About FORCE11', available at: www.force11.org/about*
[24] *Data Seal of Approval. Available at: www.datasealofapproval.org/en/*

3. Formal Certification is granted to repositories that in addition to Basic Certification go through a successful external audit based on ISO 16363 or the equivalent DIN 31644, also known as the nestor seal.[27]

ISO 16363 certification is relatively new, costly and time-consuming, so in the short term is only likely to be found in large data centres or domain repositories that are already well known to set high standards, such as the UK Data Archive.

Some Trusted Digital Repository standards have more take-up in certain disciplines than others, e.g. *Data Seal of Approva*l in social sciences, or ICSU-WDS in earth sciences.

DSA-WDS Catalogue of Common Requirements

This catalogue of 18 criteria was developed by the DSA–WDS Partnership Working Group on Repository Audit and Certification, a Working Group (WG) of the Research Data Alliance. It aims for a set of harmonized common requirements for certification of repositories at the basic level, drawing from criteria already put in place by the Data Seal of Approval (DSA) and the ICSU World Data System (ICSUWDS). An additional goal of the project was to develop common procedures to be implemented by both DSA and ICSUWDS. Ultimately, the DSA and ICSU WDS plan to collaborate on a global framework for repository certification that moves from the basic to the extended level. Further details are available at: https://rd-alliance.org/group/rdawds-certification-digital-repositories-ig/wiki/request-comments-dsa–wds-common

Information security: ISO 27001

The ISO 27000 group of standards concern management of information security, including systems specification, implementation, evaluation, auditing and risk management. ISO 27001 certification is likely to matter if your data collection includes any confidential items, such as un-anonymised data relating to human subjects, that require closed or restricted access.

Quality assurance: ISO 9000

The ISO 9000 group of standards identifies aspects of quality management - what it means, how to do it effectively, and how to get it audited. ISO 9000 certification does not provide guarantees about research data quality, but it likely to indicate that a repository has rigorous QA processes and management procedures.

MIT Libraries Data Repository Comparison Template

MIT Libraries Data Management Services developed this template to help researchers understand differences in features among data repositories under consideration. The template was designed to be filled out by data management consulting staff to inform researchers in selecting a repository for deposit of their data. Each repository evaluated can be described in the template according to types of characteristics, including: administration, scope, metadata, policies, workflows, preservation, and other features. See MIT Libraries (2016). 'Data Repository Comparison Template' Cambridge, MA: MIT Libraries. Available online: libraries.mit.edu/repository-comparison

Leiden University Information Sheets

Leiden University Libraries developed a catalogue in which they evaluate local and (inter)national services relevant for Leiden researchers. The catalogue provides information sheets comparing these services on functionalities, certifications, organizational features, conditions of use etc. They are also evaluated for consistency with the University's policy, and thus to relevant funder requirements. Researchers can browse the catalogue by discipline, or by well-known lifecycle models. The catalogue is available at: https://vre.leidenuniv.nl/vre/lrd/Pages/About.aspx

Acknowledgements

This publication was produced with the support of Jisc. The checklist was informed by the PREPARDE project 'Guidelines on recommending data repositories' (Callaghan et al 2014). We much appreciate comments on earlier drafts from Mathew Addis (Arkivum), Libby Bishop (UK Data Service) Dorothy Byatt (University of Southampton), Chris Gibson and colleagues (University of Manchester), Catherine Jones (STFC), Varsha Khodiyar (F1000 Research), Gareth Knight (London School of Hygiene & Tropical Medicine), Mathew Mayernick (NCAR), Susan Manuel (Loughborough University), Jenny Mitcham (University of York), Joan Starr (California Digital Library), Jonathan Tedds (University of Leicester), Joachim Wackerow (GESIS) and Chris Brown (Jisc).

References

Callaghan, S., Tedds, J., Kunze, J., Khodiyar, V., Mayernick, M. Lawrence, R., Murphy, F., Roberts, T. and Whyte, A. (2014) 'Guidelines on recommending data repositories as partners in publishing research data' Proceedings IDCC14, San Francisco, Feb. 25, 2014

PREPARDE project website (n.d.) available at: www2.le.ac.uk/projects/preparde

UK Data Archive *Collections Development Policy* (2012) Available at: www.data-archive.ac.uk/media/54773/ukda067-rms-collectionsdevelopmentpolicy.pdf

[26]ISO 16363 available from the CCSDS website at: public.ccsds.org/publications/archive/Forms/DispForm.aspx?ID=298

[27]Dobratz et al, 'The nestor Catalogue of Criteria for Trusted Digital Repository Evaluation and Certification' Texas Digital Library 8(2), 2007. Available at: https://journals.tdl.org/jodi/index.php/jodi/article/view/199/180

Follow the DCC on
Twitter: @digitalcuration, #ukdcc