PREPARDE D5.1 Report on requirements for data centre accreditation.

Project Information			
Project Identifier	*To be completed by JISC*		
Project Title	PREPARDE: Peer REview for Publication & Accreditation of Research Data in the Earth sciences		
Project Hashtag	#preparde		
Start Date	1 July 2012	**End Date**	30 June 2013
Lead Institution	University of Leicester		
Project Director	Dr Jonathan Tedds		
Project Manager	Dr Sarah Callaghan		
Contact email	sarah.callaghan@stfc.ac.uk		
Partner Institutions	University of Leicester British Atmospheric Data Centre (BADC) US National Centre for Atmospheric Research (NCAR) California Digital Library (CDL) Digital Curation Centre (DCC) University of Reading Wiley Faculty of 1000 Ltd		
Project Webpage URL	http://proj.badc.rl.ac.uk/preparde/wiki		
Programme Name	*Managing Research Data*		
Programme Manager	Simon Hodson		
Document Information			
Author(s)	Sarah Callaghan		
Project Role(s)	Project Manager		
Date		**Filename**	
URL	*If this report is on your project web site*		
Access	☐ Project and JISC internal	☒ General dissemination	

Document History		
Version	**Date**	**Comments**
8	15 July 2013	Final draft from the PREPARDE project. Note that these

		guidelines will likely be continually updated, though it is hoped that the main principles will remain the same.

Guidelines on recommending data repositories as partners in data publication

Draft guidelines, version 8 (15 July, 2013)

This document outlines the requirements for data repositories intent on providing a dataset as part of the scientific record. This may be either as a cited dataset (linked from and supporting a journal article or data paper), or as a published entity in its own right (published formally by the hosting repository).

This document is primarily intended as a resource for journal editors and publishers who wish to determine quickly and easily whether a repository is suitable to host data which is the basis of a research publication. It may also be of interest to researchers looking for a suitable repository for their data and those wishing to start a new data repository, as well as other parties with an interest in data publication and repository management.

These guidelines are intended to cover all the all the data associated with a scientific publication, from the small subsets that form the "data behind the graph" to the whole dataset underlying the research article.

For data publication, a repository must be actively managed in order to:

1. **Enable access to the dataset**
    a. Ensure that data will be accessible (either as open data, or provide information on conditions of access and a clear point of contact).
    b. Have a policy in place allowing appropriate access for peer reviewers, as required as part of support for the data peer-review process.
        i. In the context of data, peer reviewers are experienced researchers who produce or use data in the same field as the data being published.

2. **Ensure dataset persistence**
    a. Have a clear and public assertion of responsibility to preserve the data and provide access to the data over the long term.
    b. Have an appropriate, formal succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope.
    c. Repositories must develop and implement suitable quality control measures to ensure the metadata is correct and the data themselves are maintained and curated to avoid degradation.
        i. User feedback can and should be used to strengthen and correct the metadata as needed.
    d. Assign globally unique persistent IDs to the published datasets and maintain a repository-managed URI associated with each of those IDs. These URIs should also be associated with versions of the datasets.
    e. Permanent IDs for the dataset must resolve to a publicly accessible landing page which must:
        i. be open and human readable (and it would be preferred that they should also be provided in a format which is machine readable)
        ii. describe the data object and include appropriate metadata and the permanent identifier (used to identify the page in the first place)
        iii. be maintained, even if the data has been retracted.

3. **Ensure dataset stability**
    a. Stability means that the exact same version of the dataset that was cited can be returned to when the citation is resolved.
    b. If dataset versioning is supported, new versions should be permanently identified and linked from the original, published dataset landing page, without overwriting the original version linked from the article). The database should provide time stamped versions of archival data.

4. **Enable searching and retrieval of datasets**
    a. Allow users to easily determine whether a dataset has been peer reviewed or been subject to an equivalent level of scientific quality assurance.
    b. Provide appropriate metadata about the dataset in human readable form on the landing page (see point 2.e), and when possible standardized machine readable formats e.g. DataCite metadata schema http://schema.datacite.org
    c. Provide access to allow metadata for the datasets to be searched and retrieved through interfaces designed for both humans and computers.

5. **Collect information about repository statistics**
    a. Publish statistics on the level of access to any deposited item that is publicly accessible, to contribute to metrics of the item's publication impact.
    b. Publish information to enable journals and depositors to assess its take-up in the community it aims to serve, e.g. about any operational agreement with a well-established journal, learned society or equivalent body.

The following appendix lists ways by which a repository can demonstrate that it meets these mandatory criteria. It is split into generic schemes and subject-specific schemes.

Appendix: Repository Accreditation Initiatives

The following pages list ways by which a repository can demonstrate that it meets the recommendations given in the previous section. It is split into generic schemes and subject-specific schemes.

1. Generic resources

The following resources may be of value when identifying which repositories are suitable for use. Note that only the first two headings are actual accreditation schemes, but the remaining listed resources may be of use when determining if a data repository is suitable for data publication.

Trusted Digital Repository
(http://www.trusteddigitalrepository.eu/Site/Trusted%20Digital%20Repository.html)

- Any of the three certification levels outlined by TrustedDigitalRepository
    - Basic Certification is granted to repositories which obtain Data Seal of Approval (http://www.datasealofapproval.org/) certification;
    - Extended Certification is granted to Basic Certification repositories which in addition perform a structured, externally reviewed and publicly available self-audit based on ISO 16363 or DIN 31644;
    - Formal Certification is granted to repositories which in addition to Basic Certification obtain full external audit and certification based on ISO 16363 or equivalent DIN 31644.
- For more information, see the Alliance for Permanent Access to the Records of Science Network (APARSEN) report on peer review of digital repositories: http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D33_1B-01-1_0.pdf

ICSU World Data System
(http://www.icsu-wds.org/)

- Regular or network membership of the ICSU World Data System
    - Details of the evaluation criteria for membership can be found at http://icsu-wds.org/images/files/WDS_Certification_Summary_11_June_2012.pdf

DataCite (http://www.datacite.org/)

- Contractual arrangement with a DataCite managing agent for the purposes of minting DOIs.

Data repository directories

- Inclusion in a data repository directory that identifies repositories with standing in the scholarly community and publishes its selection criteria. Current examples include Re3data (http://www.re3data.org/) and Databib (http://databib.org/)

2. Subject-specific resources

The following list of subject specific resources may be of use, but are not intended to be an exhaustive list of the resources or subject areas with an interest in data publication.
It is hoped that different subject areas will update these guidelines with information about resources in their fields.

a. Geosciences

MEDIN Marine Environmental Data and Information Network (http://www.oceannet.org/)

- Data centre accreditation via MEDIN
    - Details of the accreditation process available at http://www.oceannet.org/data_submission/documents/medin_dac_accred_proc_v1.1_sept10.doc and list of accredited repositories is at http://www.oceannet.org/data_submission/index.html

IODE International Oceanographic Data and Information Exchange (http://www.iode.org/)

b. Life Sciences

BioSharing (http://www.biosharing.org)

- Registry of data and metadata reporting standards for different types of life science data (http://biosharing.org/standards).
- Catalogue of databases in the life sciences described according to the community-defined, uniform, generic description of the core attributes (**bioDBcore** (http://biosharing.org/biodbcore).