REPORT
Information Services

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

Research Data Management Technical Requirements

Author(s):	Thomas Parsons, Mark Berry
Audience:	ADMIRe and IS stakeholders
Published:	13/09/2012

Contents

1. Introduction

This document specifies the user requirements for a system to support the creation, archiving and sharing of research data throughout the research lifecycle. Consideration of these requirements requires a holistic view of the systems and workflows in which research data is present; and this document attempts to elucidate this and produce a high-level specification of the requirements for systems to support research data management.

There are two key requirements that must be satisfied in order to comply with funding councils' mandates:

1. Data must be stored securely and backed up throughout the lifecycle of the project
2. Research projects must publish a supporting research dataset alongside any publication (e.g. a journal paper), unless there is sound justification for not making the data publically available (i.e. commercial or sensitive data sets).

Research projects must demonstrate compliance with these mandates. Compliance inspections are implied by funding councils and this necessitates the provision of systems to support these aspects.

It is important to note that this specification only applies to data in a digital format and not to physical data (e.g. papers, slides, specimens etc.). Physical data can be treated and archived in many differing ways and hence this area requires further work to ensure compliance.

The following requirements are derived from interviews with Nottingham researchers, review of policies, JISC MRD partner projects (particularly research360[1] at the University of Bath) and work by the SWORD team[2].

2. Research data storage

The requirements for storage appear relatively simple: data must be kept securely and backed up, and access to research datasets must be controlled based on IPR and confidentiality of the data. However, there are a number of more detailed requirements that fall under these headings:

2.1. Generic storage requirements

- Researchers need to manage all research data securely throughout the research data lifecycle, protected from loss through electronic backups and secure storage for physical records.
- Researchers need secure storage for both working and archived data.
- Researchers need to archive multiple datasets from their working data, both during the course of the research and after completion.

[1] http://blogs.bath.ac.uk/research360/
[2] http://swordapp.org/2012/07/data-deposit-scenarios/

- Researchers often need more than 20GB of secure, networked, backed up storage, with archival, versioning and metadata management facilities.

2.2. File transfer

- Researchers need to be able to transfer arbitrarily large files into storage. The client/server pair must be able to deal with network scalability issues, such as slow network speed or high latency, in order to adequately support large files.
- Researchers sometimes need to bulk transfer data between systems, so the storage solution should support bulk transfer and data migration.
- Some researchers need to stream files into storage over a period of time – for example, when a piece of equipment is producing new measurements which need to be streamed.
- External collaborators in industry and other institutions need access to use, update and manage both live and archived research data.

2.3. Access

- Researchers need to provide access to external collaborators and non-human users to use, update and manage live and archived research data.
- Researchers need to provide for the authentication of local and external researchers (collaborators) and control authorisation to specific data sets, in order to restrict access (by default) to the researchers who created the data.
- Deposit of files into storage must be possible by the originator/custodian of the data, by a person or entity acting on their behalf, and by a non-human user, such as a piece of equipment, automated process or software intermediary.

2.4. Sensitive or highly-confidential research data

- Research data archives need to include metadata-only records for research data that cannot be openly shared (for ethical, commercial and legal reasons).
- Research data archives need to include 'dark archives' that are managed but not publicly accessible.
- Researchers need facilities to enable the secure collection, management, maintenance and destruction of data in line with legal requirements, research funders' terms and conditions, and University and School policies.
- The University must facilitate the secure collection, management, maintenance and destruction of data in line with legal requirements, research funders' terms and conditions, and University and School policies.

2.5. Metadata for working data

- Researchers require a searchable catalogue or index of their working and archived data.
- Researchers need to be able to transfer file content alone, content together with metadata, metadata alone, and collection metadata.
- Researchers need generic and subject-specific solutions for metadata management, integrated with their storage solution.
- Researchers should provide enough metadata to enable data re-use. Metadata therefore needs to explain the context of the data, how it was collected, what intellectual property rights apply, and any other information, software or tools necessary for understanding and re-using the data.
- Each dataset that is indexed must create a unique reference (i.e. DOI).

2.6. Archive and preservation

- Researchers need to archive Open Access data sets, metadata-only records for embargoed or non-shared data, and 'Dark Archives' for managed but confidential data sets.
- Research data archives need to include research data that has been approved for Open Access sharing and re-use (subject to appropriate attribution).
- Research archives may have complex virus scanning requirements: note that Computer Science research may require the storage of viruses as research data.

3. Publishing a dataset

The following requirements are designed to facilitate the publication of a dataset, either at the end of a project or in support a research publication. Some key points are:

- Some researchers need an institutional data repository in order to comply with funders' requirements. Many subject-specific repositories are available, but they do not cover all subjects for which open data access is a funding requirement.
- Researchers need to archive datasets, both with full open access and also with embargoes and indefinite access restrictions.
- Some datasets must be made available with full Open Access sharing and (attributed) re-use rights, where this has been approved.
- Some datasets must be embargoed until they can be made accessible.
- Most working data or ‘milestone’ datasets are restricted as a matter of course, but may be made public after project completion.

- Some datasets must be "access on request".
- Some datasets need metadata-only public records for data that cannot be openly shared for ethical, commercial and legal reasons.
- Some datasets need all access, including access to metadata, to be restricted indefinitely ('dark archives' which need to be managed but not publicly accessible).
- Some researchers wish to archive artefacts such as software, virtual operating systems, and even hardware; this is likely to require project-based solutions.

Throughout these high-level requirements, there is a common workflow involved in creating and publishing a dataset. The workflow has six key steps:

1. The researcher creates metadata that describes the dataset
2. The researcher selects or uploads a dataset
3. The researcher submits a dataset for approval
4. An administrator reviews and approves the record
5. The record and data set are released with the relevant access permissions (public or restricted) and a unique identifier assigned
6. The dataset identifier is then cited

The following requirements are designed to facilitate these six steps. It is assumed that a web interface will be used unless otherwise specified.

3.1. Metadata capture

3.1.1. Metadata about a dataset

- Researchers must be able to create and edit metadata for data sets, and must provide enough metadata to enable discovery and re-use.
- Researchers must provide a minimum amount of metadata when depositing a data set.
- Some metadata matches statements from research proposals
- Researchers should provide enough metadata to enable data re-use. Metadata therefore needs to explain the context of the data, how it was collected, what intellectual property rights apply, and any other information, software or tools necessary for understanding and re-using the data.
- Overall Classification Schema to match UoN library requirement (e.g. Dewey Decimal).
- Subject specific classification schemas for dataset discovery may be required.
- The descriptions of metadata fields must be clear

- Metadata records can relate to datasets held elsewhere and do not have to be accompanied by a UoN stored dataset
- To ensure quality of metadata, interfaces should obtain as much data as possible automatically (e.g. pFACT/AMM), and human intervention may be necessary to check the metadata derived from these sources.

3.1.2. Metadata for reuse and search

- Data sharing restrictions must be made clear. These will have been defined in the research proposal; along with a timescale for public release and any restrictions.
- Intellectual Property Rights (IPR): Copyright/IP/licensing statement must be clearly captured and displayed.
- Metadata describing restricted access datasets (e.g. commercial or medical, animal etc) will need to be restricted itself and not accessible without the appropriate permissions.
- Metadata fields for browsing and search include: Researcher(s), Funding Body, Subject Area, Keywords, Department/Research Group, Project Code.

3.1.3. Core metadata fields (from UoN and Research360 JISC MRD project at the University of Bath)

Metadata field	**Category**
Date submitted	Administrative
Funding body	Administrative
Funding body reference	Administrative
Security Rating	Administrative
Access rights ("open", "embargoed until (date)", "embargoed indefinitely")	Administrative
Date available	Administrative
Research start date	Administrative
Data collection start date	Administrative
Term of funding	Administrative
Last day of project	Administrative
Licence	Administrative
Date copyrighted	Administrative

Project ID	Administrative
Last date modified	Administrative
Research project Title	Descriptive
Alternative titles	Descriptive
Abstract	Descriptive
Language	Descriptive
Subject	Descriptive
Primary Investigator (PI)	Descriptive
Lead Researcher	Descriptive
Researchers	Descriptive
Acknowledgements	Descriptive
Publisher/Rights holder	Descriptive
Spatial	Descriptive
Temporal	Descriptive
Output References	Descriptive
Data Sources (new)	Descriptive
Related datasets	Descriptive
Research methods	Descriptive
Limitations	Descriptive
Source of data	Descriptive
Required Resources (e.g. software)	Descriptive
Audience for reuse	Descriptive
Data retention period	Preservation
Raw Data Location (DOI)	Technical
Software requirements	Technical
Extent (data size)	Technical

File formats (multiple values)	Technical
Type (always dataset)	Technical
Contents (list of contained files)	Technical
Replaces/replaced by	Descriptive
Redaction (multiple sources joined)	Descriptive

3.2. Deposit of data files

- Researchers need to transfer both individual files and groups of files into storage when publishing final data sets.
- Researchers need to be able to add links to related data sets when publishing a data set.
- Researchers must be able to bypass the file deposit interface, if the dataset is already held elsewhere (e.g. a national subject data repository).
- Researchers need to create new data sets and package them in an archive format - one file per data set.
- Packaging and archive software should reject submissions exceeding limits on max size and number of files. To open such files, sufficient memory for decompression, and disk space for both zip and extracts is needed, implying a limit on how large a zipped dataset can be if it is to be re-usable.
- Seamless access (virtual disk), SFTP and/or command line access with batch import/export support for deposit of large datasets (>2gb).

3.3. Approval

- When a metadata record and dataset have been successfully created and deposited by the researcher, control must pass to a registered approver (with appropriate access rights) for validation and release.
- An approver must check both the metadata record and data files before releasing the metadata record and archiving the dataset. Due to the serious nature of Data Protection and IPR issues, the system must require explicit confirmation from an approver for this stage to complete.

3.4. Creation of unique identifiers

- Researchers need to make research data accessible using persistent URLs which can be registered to an identifier (e.g. DOI), cited in papers, and guaranteed in the long term. This is a requirement of some funders, and long-term preservation of the dataset and a URL pointing to it is a requirement of DataCite DOIs.

- Unique IDs, alongside persistent URLs which resolve the research data object, must be created upon approval of the dataset.
- Metadata schemas must support DataCite's minimum metadata standard, and systems should be flexible to cater for future standards including the forthcoming CERIF profile for datasets.
- Once a dataset is approved, a DOI(s) must be minted using Datacite (requiring Creator, Publisher, Publication Year and Title).
- For datasets that are highly confidential, then a unique identifier must be generated via an internal system. Metadata records must not be sent to DataCite (all metadata is publically available via the DataCite Metadata Search interface[3] and API).
- DOIs (Digital Object Identifiers) for each published dataset must be made available in the metadata record for citation purposes.
- Researchers should NOT be able to delete data sets once a DOI has been issued (e.g. 'write once read many' (WORM) storage cannot be modified once archived).

3.5. Retention and Preservation

- Researchers need access to a guaranteed and managed repository in order to make research data accessible using persistent URLs with registered identifiers.
- The dataset and metadata record should be archived and retained for the appropriate period (minimum of 7 years).

3.6. Administration (researcher perspective)

- Researchers need to be able to update an aspect of an object (e.g. the dataset or the metadata) with information regarding a related part of the object or a related work (e.g. the dataset associated with some metadata, or the publication derived from the data).
- Researchers need to be able to update the open/closed status and embargo dates of data sets.
- Researchers need to be able to update access controls on closed data sets.
- Researchers should NOT be able to delete data sets once a dataset and metadata record have been approved and released.
- Users should be able to create a new version of a dataset with the original version remaining in place.
- Researchers and administrators need to be able to update the link that each DOI resolves to (for all DOIs that they manage).

[3] http://search.datacite.org/ui

3.7. Data citation

- Every metadata record must be displayed via a web 'landing page' that is accessible via a unique identifier (e.g. a DOI or internal identifier). Access to this webpage is dependent upon the relevant permissions (public or restricted) and any embargo period.
- 'Landing pages' must list all related datasets e.g. Dryad,[4]
- A researcher should cite the landing page identifier and not the identifier to the dataset.
- Metadata records displayed on a landing page should not be editable by the researcher
- Landing pages may include images or other text that can be edited by the researcher or administrator e.g. The Journal of Open Archaeology Data[5]
- Researchers need to demonstrate compliance to the relevant authority e.g. Funding Council, Institution, School Managers etc

3.8. Data repository requirements

- Interfaces should be web-based.
- There must be multi-level dataset access policies.
- Data repositories must provide public access and machine interfaces (APIs) to search and browse metadata and access datasets for public access datasets.
- Access to confidential datasets via any human interface or machine interfaces (APIs) must be based on the permissions of logged-in users and universal-access guests. No public access should be available for these datasets.
- Embargo periods for datasets must be respected:
    - A landing page must not provide a direct link to the data set unless the embargo period has passed.
    - An embargoed dataset must not be accessible until the embargo period has expired.
- Software exposing access to data set archives should provide similar user interface features and appearance across multiple accession routes.
- Version control for datasets.
- Customisable metadata and RDF support.
- Link data with publications (DOI, handle.net).
- Easy to use web interface for searching published datasets.
- Advanced metadata-based search functionalities.

[4] http://datadryad.org/resource/doi:10.5061/dryad.bg8h5
[5] http://openarchaeologydata.metajnl.com/

- Usage statistics (e.g. using Google Analytics) and email of usage statistics to dataset owners.
- Researchers need good end-user documentation for any unfamiliar processes, e.g. sponsoring of collaborators in order to grant them access.
- Documentation and help should be available at an appropriate level for the main interfaces.
- Interfaces between components should use well known protocols (e.g. SWORD) or use REST.

3.8.1. Data repository interoperability requirements

- Researchers need to create data sets, package them in an archive format, and deposit them into archives using the SWORD2 protocol.
- Researchers need to deposit datasets into archives using the SWORD2 protocol. SWORD2 uses the Bagit format: A manifest, metadata, and payload files (in a directory structure), usually stored as zip files.
- Researchers need to select appropriate deposit targets (collections within repositories), deposit data sets into multiple repositories, deliver separate parts of objects to multiple separate destinations, and update parts of objects with information regarding related parts of the object or related works.
- Researchers need to deposit data sets into multiple repositories, and stage the migration from one repository to another. In some cases the deposit repository will be a staging environment for the data pending migration. The client should be able to discover if and when a server has migrated data.
- Researchers sometimes need to deliver objects to multiple destinations. Sometimes one part of an object needs to be sent to one location, and another part to a different location (for example, data to one location and metadata to another). If different parts of an object go to different locations (e.g. data and metadata), it should be possible for each location to know about the other(s).
- Allowing linked data by using RDF[6].

3.9. Administration of a data repository (by an RDM support service)

- Administrators need documented and secure web interfaces to update the configuration of RDM components.
- Administrators need a means to update the configuration of the various components needed for RDM.
- Web interfaces for administration must be secure.

[6] http://linkeddata.org/

- Administration interfaces should be comprehensively documented.
- RDM Systems must be integrated with existing University services, in particular to reuse data from existing systems (i.e. pFACT).
- The system must provide reports on datasets submitted, approved and access requests. This data will be used for compliance purposes, so should include options to filter data via funder, project, School, researcher, date range and type of data.

3.10. Management of RDM inputs and outputs

Although the specification in this document allows compliance to various mandates, it does not provide a holistic management system. For this to occur, each dataset that is submitted and approved, must be linked to records of a research project's planned outputs, documentation and publications. Thereby providing proof that a project has complied and the two requirements of storing data and archiving data have been met. This role requires a management system that is commonly referred to as a CRIS (Current Research Information System[7]), this is beyond the scope of this document, but does require significant further work.

4. Conclusion

The proposed RDM service requires a technical infrastructure that meets the core functionality described in this document. Interviews with UoN academics and the RDM survey demonstrate that a data repository would be used and is, in some cases, essential to secure funding or comply with existing funding requirements.

This high-level specification and the following questions should help to highlight whether any potential system will be suitable or not:

- How many of the user requirements does the platform meet out of the box?
- How easy is it to install, run and maintain?
- How easy is it to customise looks and metadata management?
- How many standards does it support?
- How well developed, supported, and widely used is it?
- How big is 'large' data? Generally conceived as >2gb
- Planned file formats that will be accepted (with justification if types are omitted).
- Expected volume of data over time.
- Infrastructural support (hardware, backups, etc).

[7] http://www.atira.dk/en/pure/