Jisc RDM Shared Service Pilot Initial Statement of Requirements

Introduction

The aim of the Research Data Management Shared Services project is to procure a pilot service that provides a platform for ingesting, publishing, safely storing, archiving and preserving research data. This service will enable researchers to publish and share research data and for Higher Education Institutions to meet policy expectations of research funders as well as adhering to best practice. The pilot will be funded, procured and managed by Jisc.

Purpose

This document is intended to inform Jisc staff, project co-design partners, future users of the service and potential suppliers and of the requirements are for the elements of a Research Data Management Shared Service and to inform the procurement, development and management of the service. This document is an initial scope of requirements and these requirements will be subject to review and refinement by Jisc, the community and interested parties.

Funder Requirements

The service must enable institutions to meet funder requirements and expectations, such as the EPSRC research data policy[[1]](#footnote-1). The service should also enable institutions to efficiently provide funders with reporting information for Research Excellence Framework[[2]](#footnote-2) and EC’s OpenAIRE[[3]](#footnote-3).

Scope of the Service

The first phase of the service will not include provision for active ‘live’ data storage for objects that are being created, transformed and worked on by researchers using their own workflows and will focus on a system that will allow the ingest, publication, long term storage and preservation of finalised data objects for publication or archiving.

Top Level Requirements

The Broad top-level requirements for the Pilot Shared Service are:

* The service must enable researchers and institutions to meet funder requirements and expectations, such as the EPSRC research data policy[[4]](#footnote-4). The service should also enable institutions to efficiently provide funders with reporting information for Research Excellence Framework[[5]](#footnote-5) and EC’s OpenAIRE[[6]](#footnote-6).
* The service must meet and comply with relevant information security, legal and ethical standards.
* The service must provide researchers with user-friendly methods to deposit, publish, share, discover, access and cite research data objects.
* The service must be able to store, archive, preserve and provide access to a range of research data objects from multiple disciplines and with differing file formats, sizes and volumes.
* The service must work towards interoperability and exchange of metadata with institutional systems, (CRIS, publication repositories etc.) and external systems and services (national data centres, scholarly communications services, UK Research Data Discovery Service, ResearchFish, ORCID).
* The service must use existing identity management, authorisation and access protocols and allow for different access policies for different types of data.

Draft Requirements

|  |  |
| --- | --- |
| Ref | DRAFT Requirement |
| **1. Ingest/Deposit** |
| 1.1 | Ingest/Transfer of multiple Submission Information Packages (SIP) through a user friendly web interface with institutional branding. The SIP should include the data objects, any related supporting material, software and discovery, administration and preservation metadata. |
| 1.2 | Offer customisable metadata profiles at the dataset level.g. ReCollect ‘generic’ RDM specification, disciplinary specifications and RIOXX and REF requirements for research outputs. Metadata must be able to map to DataCite 3.0 and meet OpenAIRE requirements |
| 1.3 | Support for mandatory metadata fields and controlled vocabulary drop-downs to support global systems e.g. DataCite and institutionally defined metadata. |
| 1.4 | Support auto-completion of metadata fields by harvesting information and institutional and other systems linked to project and person identities e.g. CRIS, ORCID |
| 1.5 | Support auto-generation of preservation metadata from dataset object files and file systems (e.g. maintaining file structure and hierarchies and associated metadata where applicable) |
| 1.6 | Support for existing disciplinary and other metadata that accompanies a deposit? |
| 1.7 | It will be possible to select and configure the required level of automation within the ingest workflow. |
| 1.8 | Support association of single or multiple files with one metadata record and allow granularity (e.g. hierarchies associated with the files) to be shown in the metadata |
| 1.9 | Support for web uploads and extensibility to support additional ingest methods and protocols as required e.g. SWORD, FTP and a simple process for ingesting large datasets |
| 1.10 | Datasets will be assigned DataCite digital object identifiers (DOI’s) with institution specific prefixes |
| 1.11 | Support for integrity checks during the transfer of large datasets through checksums |
| 1.12 | Support for ingest for large datasets from physical storage media at a specific location  |
| 1.13 | Support ingest and validation of multiple file types including multimedia  |
| 1.14 | Allow assignment of publication embargo periods for both metadata, data and individual files within data |
| 1.15 | Enable relationships with other web resources and other digital objects |
| 1.16 | Data will be checked for viruses on ingest |
| 1.17 | File format identification to be performed prior to validation. Including support for inspecting common container formats (ZIP, TAR, possibly ARC/WARC). Also identifying encrypted files that are unaccompanied with passwords, |
| 1.18 | Support for institutional quality check of data and metadata prior to selection, appraisal, publication and preservation. |
| 1.19 | Include a means for recording appraisal decisions relating to ingested objects and individual elements within and associated with them. |
| 1.20 | Authors will be able to add rights/access/licence information at the point of upload/ingest |
| **2. Publication** |
| 2.1 | Creation of an institutional branded landing page with data download (or information on data access if download option is not available) , metadata and contextual information for each record |
| 2.2 | DataCite DOI must resolve to landing page |
| 2.3 | Support for dataset preview within landing page |
| 2.4 | Support and show links to scholarly communication systems e.g. is related to, is cited by etc. |
| 2.5 | Support for usage statistics and metrics on landing page e.g. citations, downloads, altmetrics |
| 2.6 | Allow metadata-only records for research data that cannot be openly shared (for ethical, commercial and legal reasons) and provide support in the metadata to enable this |
| 2.7 | Support versioning of datasets and associated DOIs including information about the latest and past versions |
| 2.8 | Display file size, format, versioning and software application and version used to create the dataset, as well as software required to view the dataset. |
| 2.9 | Display all available metadata associated with a dataset on request |
| 2.10 | Generate a suggested data citation format and allow depositors to have the option of specifying a data citation format. |
| 2.11 | Show dataset license, rights and checksum information |
|  **3. Registry and Discovery** |
| 3.1 | Create institutionally branded data registry that allows search of the institution’s data holdings |
| 3.2 | Allow search across all of an institution’s holdings. |
| 3.3 | Offer effective search and discovery mechanism including a variety of browse options and keyword searching |
| 3.3 | Support for faceted search e.g. by discipline, author, date, location etc.  |
| 3.4 | Support browsing of institutional dataset records |
| 3.5 | Contain a private user area ‘workspace’ showing users’ deposited datasets including those under embargo.  |
| 3.6 | User workspace to allow data records to be bookmarked for future reference |
| 3.7 | Contain metadata for institutional data held elsewhere e.g. in national data centres |
| **4. Data and Metadata Exchange and Integration** |
| 4.1 | Offer RESTful API functionality to data and metadata users, conforming to open standards, with outputs in JSON/XML |
| 4.2 | Create a simple metadata structure that would enable extraction and publication of metadata and data in an open format e.g. to enable migration to another repository platform. |
| 4.3 | Import or harvest metadata and associated files from a variety of sources using established open standards such as OAI-PMH, OAI-ORE. |
| 4.4 | Harvest / ingest / import metadata from other institutional and external systems e.g. funder, grant number, equipment, author identifier, project code. |
| 4.5 | Ability to support externally managed unique entities e.g. identifiers such as ORCID, |
| 4.6 | Ability to transfer and harvest metadata from object identifier systems e.g. DataCite, CrossRef |
| 4.7 | Link metadata and data from other research and administrative systems using open standards. |
| 4.8 | Link to associated research outputs, potentially in other systems e.g. Published papers. |
| 4.9 | Harvest data from and supply data to research information management systems, content management systems and internal and external research systems and services |
| 4.10 | Data to be harvested into UK Research Data Discovery Service (UKRDDS) |
| **5. Access** |
| 5.1 | Enable institution to specify system administrator and other user roles with differing levels of access to different groups of archived data (data owner, research group, institution, etc) with appropriate controlled permissions and access rights to perform different functions (e.g. undertake preservation actions) |
| 5.2 | Provide appropriate access and download of data as defined by the researcher/institution  |
| 5.3 | Enable depositor and repository manager to specify granular access rights and controlled permissions settings to the deposited data where necessary  |
| 5.4 | Support embargo options for metadata and/or associated files |
| 5.5 | Support open methods of authentication e.g. Shibboleth, OAuth, OpenID, SAML |
| 5.6 | Support Microsoft Active Directory authentication  |
| 5.7 | Support links to data sets, or metadata, which may be stored on different storage platforms and file systems |
| 5.8 | Support links to data sets, or metadata ,which may be stored in other national or international repositories |
| 5.9 | Ability to remove data from public view to an access controlled area, a dark archive or full deletion and amend the dataset landing page to create a tombstone page which includes institutional takedown policies and contact details. |
| 5.9 | Support permanent data destruction to meet any legal or ethical requirements. |
| 5.10 | Support safeguards against accidental or malicious deletion. |
| 5.11 | Support retention of metadata records for provenance where data has been deleted. |
| 5.12 | Allow admin staff to upload data on behalf of other staff. |
| 5.13 | Multiple users must be able to use the system at any one time |
| **6. Storage (Access and Archival)** |
| 6.1 | To provide guaranteed secure, corruption free, expandable data storage in numerous formats, for a minimum of ten years per data item, from the last time the item is requested/accessed, possibly indefinitely.  |
| 6.2 | Allow data to be stored on different types of storage platforms e.g. cloud, local (institutional and Jisc), Jisc Infinity Data Centre, Data Archiving Systems, HSM, offline. |
| 6.3 | Storage to be peered to the JANET network |
| 6.4 | Allow use of appliances and storage located in the Infinity Data Centre |
| 6.5 | Enable secure storage of sensitive data |
| 6.6 | Storage to be archived and preserved using appropriate policies |
| 6.7 | Data objects to be archived and preserved for 10 years since last access, with the capability to do this automatically as per institution’s policy |
| 6.8 | Integrate with archive file appliance or archival system |
| 6.9 | Data to be backed up and archived to avoid any loss of data. |
| **7. Archival Management** |
| 7.1 | Comply with an archival management policy that will meet funder mandate and institutional data management policies |
| 7.2 | Creation of an Archival Information Package to be stored for a configurable period (e.g. minimum of 10 years since date of last access for EPSRC data outputs) |
| 7.3 | Backup and replicate datasets with a minimum of 3 copies of data stored in different locations |
| 7.5 | Put in place procedures to deal with failover and recovery of data from archive.  |
| 7.6 | Ability to connect to a range of storage systems, including the front end storage appliance to an off-site data archive and offline storage. |
| 7.5 | Provide an appropriate method of ensuring the integrity of datasets, such as checksums |
| 7.6 | Provide a schedule and method of media migration |
| 7.7 | Where problems of data loss or corruption occur, a reporting/notification system will prompt appropriate action. |
| 7.8 | Support for a retention and review workflow defined by institutional policies |
| **8. Data Preservation** |
| 8.1 | Data is to be readable by available software within the timeframe it is archived, taking advantage of newer/improved preservation tools over time |
| 8.2 | Preservation description information (PDI) is persistently associated with the relevant content information. The relationship between a file and its metadata/documentation must be permanent. |
| 8.3 | Accurately record and maintain relationships between different representations of a file (for example, from submitted originals to dissemination and preservation versions that will be created over time). |
| 8.4 | Support the PREMIS metadata schema and use relevant parts of the schema to store preservation metadata e.g.  |
| 8.5 | Store technical metadata extracted from files (for example that which is created as part of the ingest process). |
| 8.6 | Allow preservation plans (such as file emulation/virtualisation, migration/normalisation) to be enacted on individual or groups of files.  |
| 8.7 | Allow customisable workflows for migrated content – ability to keep, delete, record or store data so can re-generate original |
| 8.8 | Automated checking of files will be carried out post-migration to ensure the quality of preservation actions |
| 8.9 | Record actions, migrations and administrative processes that occur with a full audit trail. Any actions that move or change a file and/or its metadata, should be captured as a PREMIS compliant event |
| 8.10 | Allow for disposal of data where appropriate and record such actions |
| 8.11 | Allow emulation tools to be used on the data if necessary |
| 8.12 | Support the incorporation of new digital preservation tools (for migration, file validation, characterisation etc) as they become available. |
| **9. Reporting** |
| 9.1 | The solution must be able to provide access and usage statistics from both the user side and the public side, and present them in a usable way. |
| 9.2 | Generate a variety of reports (or allows access to the data for local reporting tools) analysing repository content, including records associated with an academic unit, a funder, grant number, data creator and analysis of data by format, file size, location and embargo and access conditions. |
| 9.3 | Report on preservation actions undertaken |
| 9.4 | Hold funder and grant metadata associated with datasets |
| 9.5 | Be able to report on REF compliance for relevant research objects and meet reporting requirements for REF  |
| 9.6 | Be OpenAire compliant for both text and data publications |
| 9.7 | Support reporting by funder to confirm a particular funder's open access data / deposit requirement has been met. |
| 9.9 | Generate reports on collection content growth pattern over time. |
| 9.10 | Generate reports on the number and origins of repository end users, including by domain type and geographical location. |
| 9.11 | Report and respond to changing requirements of designated communities. |
| 9.12 | Report on data that needs to be disposed of (based on 10 years from date of last access mandate for example) |
| 9.13 | Report on number and speed of data downloads and log number of data requests. Reports should be possible over specified periods of time and against specified subsets of the repository (by Funder, by Creator, by academic unit). |
| **10. General** |
| 10.1 | User interface branding to be defined by institutions and compliment the look and feel of the institutions’ web presence and to be able to be updated as appropriate. |
| 10.2 | Compliance with the Open Archival Information System (OAIS) reference model and audit the system against ISO16363. |
| 10.3 | Technical support and help should be available to institutional staff and researchers. Training must be provided |
| 10.4 | Pilot HEI’s and researchers will be supported in the adoption and testing of the system |
| 10.5 | Any software chosen for elements of the system will be under active development |
| 10.6 | A community of users from Pilot HEI’s and the wider community will exist around chosen software or systems to enable sharing of use cases, workflows and to promote developments |
| 10.7 | User interface intuitive to users, as evidenced by user testing from pilot instituions. |
| 10.8 | User interface meets W3C WCAG 1.0 accessibility requirements. |
| 10.8 | User Interface to integrate with modern web browsers on multiple platformsAny web-based front-end should operate using the HTML5 framework and work without extra software on both the PC Windows 7/10 platform and Apple Mac Platform, specifically with Safari v8 or newer. |
| 10.9 | Ability to support rapid prototyping, so proof of concept work can be carried out quickly and easily |
| 10.10 | Detailed documentation available covering the technical architecture, system APIs and processes of the software and systems |

Method

Top level requirements and priorities were gathered from the community at Jisc Research at Risk consultation events and workshops. These events have been crucial in shaping the structure of the proposed system. Detailed requirements have been gathered from Jisc Managing Research Data projects and Research Data Spring projects. Useful information sources that have fed into these requirements are:

* Leeds Research Data Management Pilot Roadmap[[7]](#footnote-7)
* ADMIRe Project[[8]](#footnote-8)
* Kaptur[[9]](#footnote-9)
* SWORD[[10]](#footnote-10)
* Filling the Data Preservation Gap[[11]](#footnote-11)
* A consortial approach to integrated RDMS[[12]](#footnote-12)
* DataVault[[13]](#footnote-13)
* Jisc Learner Analytics Operational Requirements

This document also used from the Manchester Metropolitan University Draft Requirements Specification

These requirements are an initial draft and will be subject to community consultation and refinement, including prioritisation.

1. <https://www.epsrc.ac.uk/about/standards/researchdata/expectations/> [↑](#footnote-ref-1)
2. <http://www.ref.ac.uk/> [↑](#footnote-ref-2)
3. https://guidelines.openaire.eu/en/latest/ [↑](#footnote-ref-3)
4. [htsps://www.epsrc.ac.uk/about/standards/researchdata/expectations/](https://www.epsrc.ac.uk/about/standards/researchdata/expectations/) [↑](#footnote-ref-4)
5. <http://www.ref.ac.uk/> [↑](#footnote-ref-5)
6. https://guidelines.openaire.eu/en/latest/ [↑](#footnote-ref-6)
7. <https://library.leeds.ac.uk/roadmap-project-outputs> [↑](#footnote-ref-7)
8. <http://admire.jiscinvolve.org/wp/files/2013/05/ADMIRe-RDM-Technical-Requirements-Report.pdf> [↑](#footnote-ref-8)
9. <http://www.vads.ac.uk/kaptur/outputs/Kaptur_technical_analysis.pdf> [↑](#footnote-ref-9)
10. <http://swordapp.org/2012/07/data-deposit-scenarios/> [↑](#footnote-ref-10)
11. <https://www.york.ac.uk/borthwick/projects/archivematica/> [↑](#footnote-ref-11)
12. <http://figshare.com/articles/Project_Report_A_consortial_approach_to_integrated_RDMS/1480451>

 [↑](#footnote-ref-12)
13. <http://libraryblogs.is.ed.ac.uk/jiscdatavault/reading/> [↑](#footnote-ref-13)