Research at Risk:

RDM Shared Service Project 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 pilot will not include provision for active ‘live’ data storage and will focus on a system that will allow the ingest, publication, long term storage and preservation of finished data objects for publication or archiving. However future enhancements to the service will include integration with platforms used to store active data.

Draft Requirements

|  |  |
| --- | --- |
| Ref | DRAFT Requirement |
| **1. Ingest/Deposit** | |
| 1.1 | Creation of a Submission Information Package (SIP) through a user friendly web interface with institutional branding |
| 1.2 | Offer customisable metadata profiles e.g. ReCollect ‘generic’ RDM specification, disciplinary specifications and RIOXX for text. 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 |
| 1.4 | Support auto-completion of metadata fields by harvesting information from institutional and other systems linked to project and person identities e.g. CRIS, ORCID |
| 1.5 | Support auto-generation of preservation metadata |
| 1.6 | Support association of single or multiple files with one metadata record and allow granularity |
| 1.7 | Support for web uploads and additional ingest methods and protocols e.g. SWORD and a simple process for ingesting large datasets |
| 1.8 | Datasets will be assigned DataCite digital object identifiers (DOI’s) |
| 1.9 | Support multiple file types including multimedia |
| 1.10 | Allow assignment of publication embargo periods for both metadata and data |
| 1.11 | Allow depositors to create relationships with physical objects (including location) |
| 1.12 | Data will be checked for viruses on ingest |
| 1.13 | Data will be validated using appropriate tools |
| 1.14 | Support for institutional quality check of data and metadata prior to publication |
| **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 |
| 2.6 | Allow metadata-only records for research data that cannot be openly shared (for ethical, commercial and legal reasons) |
| 2.7 | Support versioning of datasets including information about the latest and past versions |
| 2.8 | Display file size, format and software application and versioning |
| 2.9 | Generate a suggested data citation format and allow depositors to specify data citation format. |
| 2.10 | Show dataset license information |
| 2.11 | Display usage statistics and impact measures for the dataset |
| **3. Registry and Discovery** | |
| 3.1 | Create institutionally branded data registry that allows search of the institution’s data holdings |
| 3.2 | 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 etc. |
| 3.4 | Support registry browsing |
| 3.5 | Contain a user area ‘workspace’ showing users’ deposited datasets |
| 3.6 | User workspace to allow data to be ‘saved’ or bookmarked |
| 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 API functionality to data and metadata users, conforming to open standards. |
| 4.2 | Enable extraction of metadata and data in an open format e.g. to enable migration to another repository platform. |
| 4.3 | Support current open standards and support exposure of metadata in different ways. |
| 4.4 | Import or harvest metadata and associated files from a variety of sources using established standards such as OAI-PMH, OAI-ORE. |
| 4.5 | Harvest / ingest / import metadata from other institutional and external systems e.g. funder, grant number, equipment, author identifier, project code. |
| 4.6 | Ability to support externally managed unique entities e.g. identifiers such as ORCID, |
| 4.7 | Ability to transfer and harvest metadata from object identifier systems e.g. DataCite, CrossRef |
| 4.8 | Display linked data from other systems using open standards. |
| 4.9 | Link to associated research outputs, potentially in other systems e.g. Published papers. |
| 4.10 | Harvest data fro and supply data to research information management systems |
| **5. Access** | |
| 5.1 | Provide open access and download of data where possible |
| 5.2 | Enable depositor and repository manager to specify granular access rights to the deposited data where necessary |
| 5.3 | Support embargo options for metadata and/or associated files |
| 5.4 | Support open methods of authentication e.g. Shibboleth, OAuth, OpenID |
| 5.5 | Support Microsoft Active Directory authentication |
| 5.6 | Support access to data sets which may be stored on different storage platforms and file systems |
| 5.7 | Support access to data sets which may be stored in other national or international repositories |
| 5.8 | Ability to remove data from public view to an access controlled area, a dark archive or full deletion |
| 5.9 | Support permanent data destruction to meet any legal or ethical requirements. |
| **6. Storage (Access and Archival)** | |
| 6.1 | Allow data to be stored on different types of storage platforms e.g. cloud, local, HSM |
| 6.2 | Storage to be peered to the JANET network |
| 6.3 | Allow use of appliances and storage located in the Infinity Data Centre |
| 6.4 | Enable secure storage of sensitive data |
| 6.5 | Storage to be backed up and archived using appropriate policies |
| 6.6 | Data objects to be stored and archived for 10 years since last access |
| 6.7 | Integrate with archive file appliance or archival system |
| 6.8 | 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 minimum of 10 years since date of last access |
| 7.3 | Minimum of 3 copies of data stored in different locations |
| 7.4 | 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. |
| **8. Data Preservation** | |
| 8.1 | Data is to be readable by necessary software within the timeframe it is archived |
| 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 | support the PREMIS metadata schema and use it to store preservation metadata. |
| 8.4 | store technical metadata extracted from files (for example that which is created as part of the ingest process). |
| 8.5 | allow preservation plans (such as file migration/normalisation) to be enacted on individual or groups of files |
| 8.6 | Automated checking of significant properties of files will be carried out post-migration to ensure these properties are adequately preserved |
| 8.7 | record actions, migrations and administrative processes that occur |
| 8.8 | allow for disposal of data where appropriate and record such actions |
| 8.9 | Allow emulation tools to be used on the data if necessary |
| 8.10 | Support the incorporation of new digital preservation tools (for migration, file validation, characterisation etc) as they become available. |
| **9. Reporting** | |
| 9.1 | Generate a variety of reports 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. |
| 9.2 | Report on preservation actions undertaken |
| 9.3 | Hold funder and grant metadata associated with datasets |
| 9.4 | Be OpenAire compliant for both text and data publications |
| 9.5 | Support reporting by funder to confirm a particular funder's open access data / deposit requirement has been met. |
| 9.6 | Generate reports on collection content growth pattern over time. |
| 9.7 | Support preservation function by generating reports on repository content. |
| 9.8 | Generate reports on the number and origins of repository end users, including by domain type and geographical location. |
| 9.9 | 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 | Comliance with the Open Archival Information System (OAIS) reference model. |
| 10.2 | Technical support and help should be available. |
| 10.3 | Pilot HEI’s and researchers will be supported in the adoption and testing of the system |
| 10.4 | Any software chosen for elements of the system will be under active development |
| 10.5 | A community of users will exist around chosen software or systems to enable sharing of use cases, workflows and to promote developments |
| 10.6 | User interface intuitive to users |
| 10.7 | User interface meets W3C WCAG 1.0 accessibility requirements. |
| 10.8 | Ability to support rapid prototyping, so proof of concept work can be carried out quickly and easily |
| 10.9 | Detailed documentation available covering the technical architecture 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[[4]](#footnote-4)
* ADMIRe Project[[5]](#footnote-5)
* Kaptur[[6]](#footnote-6)
* SWORD[[7]](#footnote-7)
* Filling the Data Preservation Gap[[8]](#footnote-8)
* A consortial approach to integrated RDMS[[9]](#footnote-9)
* DataVault[[10]](#footnote-10)
* Jisc Learner Analytics Operational Requirements

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. <https://library.leeds.ac.uk/roadmap-project-outputs> [↑](#footnote-ref-4)
5. <http://admire.jiscinvolve.org/wp/files/2013/05/ADMIRe-RDM-Technical-Requirements-Report.pdf> [↑](#footnote-ref-5)
6. <http://www.vads.ac.uk/kaptur/outputs/Kaptur_technical_analysis.pdf> [↑](#footnote-ref-6)
7. <http://swordapp.org/2012/07/data-deposit-scenarios/> [↑](#footnote-ref-7)
8. <https://www.york.ac.uk/borthwick/projects/archivematica/> [↑](#footnote-ref-8)
9. <http://figshare.com/articles/Project_Report_A_consortial_approach_to_integrated_RDMS/1480451>

   [↑](#footnote-ref-9)
10. <http://libraryblogs.is.ed.ac.uk/jiscdatavault/reading/> [↑](#footnote-ref-10)