Information

Services

The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Research Data Management Strategy Requirements

Support for creating, archiving and sharing research data

Authors: Mark Berry, Thomas Parsons

Date: 27/11/12

Abstract

The following document outlines the key areas that should be considered when creating a Research Data Management (RDM) repository strategy. The report considers external repositories, standards and available options within Nottingham. It is suggested that this information is reviewed before the development of a high-level Information Services (IS) repository strategy.

1. Introduction

The following document identifies key areas that should be considered when creating The University of Nottingham (Nottingham) repository strategy. It is not intended to be the strategy, but instead should provide pointers and be considered as a landscape assessment from the perspective of RDM.

2. Research Data Repository Strategies

The JISC funded DISC-UK DataShare project (2007-2009) produced comprehensive guidelines for the development of a repository strategy that includes research data. It is recommended that this document be read in conjunction with the guidance provided from:

- The JISC funded resource keeping research data safe.
- The Open Access Scholarly Information Sourcebook - Business Aspects of Institutional Repositories

3. Data Repositories

3.1. Existing Nottingham Repository Software

Nottingham already has existing capability in the area of data repositories and talks are in progress to consider new software. These are the known examples within IS at present, although other repositories may be hosted by researchers using Nottingham infrastructure.

3.1.1. Equella

EQUELLA is digital repository software that is managed by IS and supported by Pearson. Further work is scheduled to assess whether it could function as either a data repository in its own right or as a metadata catalogue.

Equella Documentation - Online documentation for Equella

UoN Equella Test Servers

3.1.2. Ex-Libris DigiTool and Rosetta

Rosetta is the successor to DigiTool and may play a future role in any repository strategy. Nottingham DigiTool currently houses digital copies of Historic Collections.

Rosetta supports the entire OAIS model high-level functions: ingest, storage, management, administration, preservation and delivery - but its key strength is continuous preservation actions to support long-term curation of digital artefacts.

3.1.3. ePrints

EPrints has been used at Nottingham since 2001 and is the data repository of choice for UWE, Southampton and Exeter.

UWE: EPrints Repository Press Release - Press Release for UWE's project to use EPrints for research data

3.1.4. External Data Repository Software

The following software is established and utilised by other HEIs, so should be considered within any repository strategy.

JISC RDM DataFlow Project (University of Oxford):

DataFlow implements a two-stage data management infrastructure: DataStage allows to locally work with, annotate, publish, and permanently store research data, while valuable datasets are preserved and published via the institutional DataBank platform.

LabArchives:

LabArchives is a web-based commercial laboratory notebook that allows research groups to manage, securely store and publish their research data. LabArchives has recently signed a partnership with BioMed Central open access publisher to make this platform the default storage system for supplementary data to be published with research papers submitted to BMC journals.

DSpace

DSpace Introductory Video

CKAN

CKAN: Open Data Surfacing and Visualisation Software - Software solution for making open data accessible, with tools enabling the publishing, sharing, finding and using of datasets. Aimed at data publishers (mainly governmental and corporate) wanting to make their data open and available.

Dataflow

Dataflow (Oxford Labs) - Two-stage data management infrastructure: DataStage allows research groups to work with and annotate research data; DataBank allows the institution to publish and preserve the data. Published datasets are assigned DOIs to make them citable. The software is open source, written in Python, and uses the SWORD protocol for packaging research data and associated metadata for deposit into the repository.

3.2. HEI Data Repositories

The following repositories serve as useful examples of other HEIs work in this area.

Edinburgh DataShare:

Edinburgh DataShare is an online digital repository of multi-disciplinary research datasets produced at the University of Edinburgh. Researchers who have produced research data associated with an existing or forthcoming publication, or which has potential use for other researchers, are invited to upload their dataset for sharing and safekeeping. A persistent identifier and suggested citation will be provided.

UWE: EPrints Repository for Data

UWE Research Data Repository is under development, as a project to adapt EPrints to include datasets as well as publications

3.3. External Data Repositories

The following external repositories provide an indication of currently accessible data repositories, should these wish to be factored into any strategy.

Dryad:

Dryad is an international repository of data underlying peer-reviewed articles in the basic and applied biosciences.

figshare:

This unaffiliated platform allows researchers to store (and eventually to share and publish open access) all their research output, including figures, datasets, tables, videos or any other type of materials. Emphasis is made on its potential use for dissemination of unpublished and/or negative results.

arXiv service:

arXiv is primarily an archive and distribution service for research articles. arXiv provides support for data sets and other supplementary materials only in direct connection with research articles submitted, either storing them as ancillary files on arxiv or as linked datasets in the DataConservancy subject-based data repository.

The Australian Repositories for Diffraction Images (TARDIS)

TARDIS/MyTARDIS is a data repository system for high-end instrumentation data capture from facilities such as the Australian Synchrotron. The program records the data generated by the experiment, catalogues it and transfers it back to the home institution (where the researcher can analyse the data).

UK Data Archive

National collection of research data for Social Science and the Humanities

Data must be submitted by first completing a Data Review Form at the Economic and Social Data Service

How to Deposit (UK Data Archive)

3.4. Data Repository Indexes

Further information and examples of data repositories can be found under the following resources.

3.4.1. OpenDOAR

OpenDOAR Search - Directory of Open Access Repositories: Advanced Search for Repositories by: Subject Area, Content Type, Repository Type, Country, Language and Software. There is also an API to search the database, allowing query results to be returned as XML data.

3.4.2. Databib

Databib - Collaborative, annotated bibliography of primary research data repositories. Can be searched alphabetically or by subject area, and through Advanced Search on Title, Description, Authority, Subject and Access.

3.4.3. Key Data Repositories list

Description of Key Data Repositories from: Data Deposit Scenarios (Sword).

3.4.4. DataCite Repositories list

The British Library DataCite service maintains a comprehensive list of data repositories and their subject areas.

3.5. Digital Repositories, Preservation, and File Formats

A repository strategy for data retention should include details of how files will be curated in the long-term. This then implies that care and guidance should be provided to researchers on deposit of data files, so as to identify file formats that are at risk of obsolescence.

File formats - Chapter from Stephen Abrams for the DCC discussing aspects of format description, validation, and characterisation that may assist with long-term curation and usability of data.

Characterising and Preserving Digital Repositories: File Format Profiles - ARIADNE Article exploring the past and present file format profiles of repositories, showing how digital and institutional repositories are changing and rapidly grosing to target new types of digital content, including data and teaching materials as well as research papers.

3.6. The OAIS Standard

When devising a strategy it is useful to review best practice for standards, which in this case is the Open Archival Information System (OAIS) Reference Model is ISO Standard 14721:2003. It defines the standards for an archive that has accepted the responsibility for long-term preservation of information and for making it available to a designated community.

The high-level functions of an OAIS are:

- Ingest of digital objects
- Storage of digital objects
- Data management
- Administration
- Preservation planning
- Provision of access

The requirements of an OAIS are:

- Negotiate for and accept appropriate information from information producers
- Obtain sufficient control of the information provided to the level needed to ensure long term preservation
- Determine, either by itself or in conjunction with other parties, which communities should become the designated community and, therefore, should be able to understand the information provided
- Ensure that the information to be preserved is independently understandable to the designated community, without needing the assistance of the experts who produced the information
- Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original
- Make the preserved information available to the designated community

3.7. Repository Support

Finally, it is worth noting that a data repository will require long-term maintenance, funding and support to continue. The high storage requirements for a data repository (>1PB) will place a growing burden on any institution that houses research data, therefore it is worth reviewing the reflections from those who support repositories internationally.

Repositories Support Project (RSP) - JISC-funded initiative providing guidance and advice on institutional repositories to UK HEIs. The principle aim of the project is "to increase the pace of institutional adoption by providing practical assistance and advice based on available solutions, with an emphasis on operational issues to do with the installation, implementation and deployment of institutional repositories."

Repository software survey - Nov 2010 survey from the RSP, comparing repository software solutions

4. Conclusion

The review has highlighted areas that will be of interest to those formulating a repository strategy and highlighted current practice. The area of data repositories is growing rapidly and initiatives like figshare and Dryad will no doubt be matched by similar offerings in the near future. Although the recent changes in RCUK research data policy[1], make it appear mandatory that Nottingham create or partner with other institutions to provide at least a basic level of data repository functionality.

[1] http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx