The University Of Sheffield.

University Library.

Report to the University of Sheffield Research Data Management Service Delivery Group

Research Data Management Technical Infrastructure: A Review of Options for Development at the University of Sheffield

Date: Version 2. 12/10/2014
Version 1. 12/05/2014

Author: John A. Lewis

Executive Summary

This report reviews the options available for the development of a technical infrastructure, the software and hardware systems, to support Research Data Management (RDM) at the University of Sheffield. The appropriate management of research data throughout the data lifecycle, during and after the research project, is considered good research practice. This involves **data management planning** during the research proposal stage; **looking after active data**, its creation, processing, storage and access during the project; and **data stewardship**, long-term curation, publishing and reuse of archive data after the end of the project.

Good RDM practice benefits all stakeholders in the research process: Researchers, will secure their data against loss or unauthorized access, and may increase research impact through publishing data; Research institutions may consider research data as 'special collections' and will need to minimise risk to data and damage to reputation; Research Funders wish to maximise the impact of the research they fund by enabling reuse; Publishers may wish to add value to research papers by publishing the underlying data.

Many research funders now mandate RDM procedures, particularly Data Management Planning (DMP), and the UK research councils policies have contributed to the RCUK common principles on data policy. Notice must be taken of the EPSRC Expectations of organisations receiving EPSRC funding. These include the requirements that the organisation will:

- Publish appropriately structured metadata describing the research data they hold - therefore the institution must create a public data catalogue.
- Ensure that EPSRC-funded data is securely preserved for a minimum of ten years – therefore the institution must create a data archive.
- Ensure that effective data curation is provided throughout the full data lifecycle – therefore the institution must provide the necessary human and technical infrastructure required.

Institutions in receipt of EPSRC funding are expected to be compliant with these expectations by 1st May 2015. The University of Sheffield Research Data Management Policy was developed in response to the RCUK principles and EPSRC expectations. This states that the University will develop infrastructure and services to support research data management in consultation with researchers.

The local infrastructure does not exist in a vacuum and interacts with, and is dependent upon a range of other services and processes in an information ecosystem. At one end of the continuum of research data curation is the local storage of data and metadata (data identification, description and documentation), usually accessible to the project team only. At the other end are international discipline-based data repositories or national data centres that publish research data, facilitating its discovery and access. Research institutions lie in the middle of this continuum and provide the means to move research data and metadata from their local, unpublished state to an international published state. Some institutions now publish research data, either by modifying the institutional repository (IR) to accommodate datasets in addition to research papers or through a data repository, being a new instance of repository system running alongside the IR. However, discipline-based repositories are considered the most appropriate facility for data publishing, due to their

configuration for the data types and metadata formats associated with the research community they serve.

The repository is here defined as a software system composed of three layers – a user interface, a database holding metadata records, and a storage layer holding the actual research data bitstreams. In some implementations, often known as data registries, data catalogues or metadata stores, the repository holds only the metadata records and links to the data stored elsewhere.

This report focuses on the outcomes of projects at UK HEIs funded by the JISC 'Managing Research Data' programmes 2009-11 and 2011-2013. Generally the infrastructure architectures examined have been developed in response to the functional requirements derived from researcher workflows. The major functional components of the RDM technical infrastructure for the institution are:

- **Metadata capture system** – in order to identify, describe and document the research data as they are created, captured and processed and record the context, conditions, variables and instrument settings. This may be accomplished manually, by the researcher filling in forms, or automatically, concurrent with data capture, by using appropriate equipment.
- **Active research data management system** – Active research data needs to be accessed rapidly, may require large computational resources and may require stringent security and access arrangements. A number of collaborative systems and virtual research environments have been developed to fulfil these requirements. These can be considered to comprise of a filestore and a **data registry** (sometimes known as a metadata store or asset registry).
- **Research Data Repository** – will be an appropriate place for preservation and publishing of archive research data for which there is no discipline-based repository or data centre available. The catalogue and archive functions of the repository may be separated.
- **Research Data Catalogue** – holds the metadata records of published (but not necessarily open access) research data. The data themselves may be held in a discipline-based data repository outside the institution or in an institutional data archive.
- **Research Data Archive** – preserves data not, or not yet, submitted to discipline-based data repositories. The associated metadata records will be held in the research data catalogue.
- **Current Research Information System (CRIS)** – manages the metadata associated with researcher identity, project information, research costing, grant applications and awards.

These components may overlap in function, but need to be interoperable to provide seamless RDM. Alternative approaches to an infrastructure composed of diverse components, where ensuring interoperability may be problematic, are provided by data grids and micro-services. The technical infrastructure must fit into the researcher workflow, making RDM processes automatic and virtually invisible to the researcher as far as possible. This is so as not to burden the researcher with additional work or changes to their practice. Products, processes and practices that have been developed by a researcher community should be adopted, adapted and developed for the needs of other researchers, rather than new solutions developed.

The choice of technical infrastructure components and the approach of implementation will need to be considered with regard to the infrastructure and expertise already present. Integrating and modifying existing components may be as expensive, in terms of development work, as implementing new infrastructure. Installing and configuring free open-source software may prove

expensive in terms of development, compared with proprietary systems. At the University of Sheffield there is currently no system that adequately supports collaborative active data management – a virtual research environment, or 'academic dropbox'. There is little information available regarding the use of data and metadata capture systems, although systems such as laboratory information management systems may be used by some research groups at the institution. It is feasible that the Symplectic CRIS may be configured for use as a research data registry. The ePrints institutional repository, WRRO, may be configured for use as a research data catalogue, but this relies on agreement between the WRUC member institutions. In order to implement an independent institutional research data repository, expertise will be needed to do the necessary development work.

A shared approach to RDM services is being investigated by the White Rose and N8 consortia (of which the institution is a member); a WR Research Data Catalogue and N8 shared data archiving service having been proposed. The development of RDM services delivered through the White Rose Grid and N8 HPC grid infrastructure need to be explored. Attention should be paid to the national research data service is being piloted through the DCC and JISC. The great benefits of the shared approach demand that support for collaborative projects establishing shared RDM services should be a priority.

This report briefly describes the eighty most commonly used components of RDM technical infrastructure at UK HEIs. The report describes evaluations, reviews and comparisons of these components, gives examples of established RDM services and highlights the recent projects at UK HEIs which were involved in developing these services.

By way of conclusion, a number of recommendations are made regarding the choice of infrastructure components to be made and implementation strategy to be considered. These recommendations take into account the current situation of technical infrastructure at the institution and the constraints on time and cost. Attention is drawn to the development of shared RDM services with collaborating institutions. Finally the proposal is put forward that a number of technical infrastructure components of an integrated RDM service are first piloted with EPSRC funded research projects to ensure compliance with EPSRC expectations by 1st May 2015.

Contents

1. Introduction

The purpose of this report is to indicate options available for the development of a technical infrastructure to support research data management (RDM) at the University of Sheffield. RDM, its situation within academic research and recent drivers towards change are defined. The processes involved in RDM and the elements of the supporting technical infrastructure are examined. The local context, of RDM technical infrastructure at the University of Sheffield and collaborating institutions, is explored. The range of technical infrastructure components available and evaluations of these components are reviewed. Instances of fully-functioning RDM technical infrastructure and many of the recent research projects that developed and piloted RDM technical infrastructure components are briefly described. Finally, recommendations for suitable technical infrastructure components are proposed.

1.1. Research Data Management

The research data collected to test a research assertion must be managed in an appropriate manner to be considered good research practice. Good RDM practice is now required of researchers by many research institutions and by most research funders. Increasingly research funders are demanding long-term curation of some of the data resulting from the research they fund, so that they may be available for re-use. The value of those data and the impact of the original research are increased by re-use. RDM may be considered to involve three broad areas of activity:-

- Data management planning, during the research proposal and grant application stage.
- Looking after 'live' or 'active' data as they are collected, processed, shared and stored.
- Data Stewardship - Long-term curation of research data and data publishing, making data discoverable and reusable.

1.2. Research data management drivers

JISC[1] have supported the development of RDM practice over the last fourteen years by funding projects involving HEIs through a number of programmes, in particular the Managing Research Data Programmes 2009-11[2] and 2011-13[3]. The DCC[4] was established in 2004 with JISC funding, to support expertise and practice in RDM. Since 2011 the DCC have offered tailored support in the development of policy, services and infrastructure for Higher Education Institutions, and are the foremost source for information and advice in the development of RDM infrastructure (Jones et al. 2013). Infrastructure refers to the hardware, software and human resources necessary to support the RDM services and processes. This report focuses on the technical infrastructure, the software and hardware components available.

[1] Joint Information Services Council (JISC) http://www.jisc.ac.uk/
[2] JISC MRD 09-11 http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx
[3] JISC MRD 11-13 http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata.aspx
[4] Digital Curation Centre (DCC) http://www.dcc.ac.uk/

The need for good RDM practice is recognised by all stakeholders involved in the research process. These include:

- **Researchers**, who may be part of a research project team, which may include members of many different institutions. Researchers need to secure their data against loss or unauthorised access. Making data available to reuse allows verification, promotes integrity and increases research impact.
- **The research institution** (may be a HEI) or body employing the researcher and providing the facilities. Research data may be considered part of an institution's special collections. Institutions will also wish to minimise risk to the data and damage to their reputation.
- **The research funder** (usually a research council, charity or a HEI) who may mandate RDM procedures such as the creation of a Data Management Plan (DMP) and the deposit of data to repositories. Research Funders may support facilities for data curation such as data centres. Funders wish to increase the return on their funding, through the reuse of data.
- **Governments**, who fund research councils and other funding bodies, are concerned to derive as much value as possible from publicly funded research.
- **Publishers of research papers**, who may publish the underlying data, seeking to add value to the publication process.

The EPSRC policy framework on research data[5], published in May 2011, puts forward the EPSRC expectations[6] of organisations receiving EPSRC funding, concerning the management and provision of access to EPSRC funded research data. These nine expectations were developed from seven guiding principles[7] which are aligned with the RCUK common principles on data policy[8]. Institutitions in receipt of EPSRC funding are expected to be fully compliant with these expectations by 1st May 2015. In terms of RDM Infrastructure, the pertinent expectations (EPSRC, 2013) are:

"Research organisations will ensure that appropriately structured metadata describing the research data they hold is published (normally within 12 months of the data being generated) and made freely accessible on the internet; in each case the metadata must be sufficient to allow others to understand what research data exists, why, when and how it was generated, and how to access it"

"Research organisations will ensure that EPSRC-funded research data is securely preserved for a minimum of 10-years from the date that any researcher 'privileged access' period expires or, if others have accessed the data, from last date on which access to the data was requested by a third party;"

"Research organisations will ensure that effective data curation is provided throughout the full data lifecycle... The full range of responsibilities associated with data curation over the data lifecycle will be clearly allocated within the research organisation, and where research data is subject to restricted access the research organisation will implement and manage appropriate security controls;"

[5] Engineering and Physical Sciences Research Council (EPSRC) policy framework on research data http://www.epsrc.ac.uk/about/standards/researchdata/Pages/policyframework.aspx
[6] EPSRC expectations http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx
[7] EPSRC principles http://www.epsrc.ac.uk/about/standards/researchdata/Pages/principles.aspx
[8] Research Councils UK (RCUK) common principles on data policy http://www.rcuk.ac.uk/research/datapolicy/

The University of Sheffield Research Data Management Policy[9] was developed in response to the RCUK principles and EPSRC expectations. Of the eight points of policy, the following (The University of Sheffield, Research and Innovation Services, 2014) are particularly relevant for RDM Infrastructure:

"The primary responsibility for effective research data management during the course of research projects lies with lead researchers. However, all researchers, including postgraduate and undergraduate students undertaking research, have a personal responsibility to manage effectively the data they create."

"Unless the terms of research grants or contracts provide otherwise, data generated by research projects are the property of the University of Sheffield. Researchers should exercise care in assigning rights in data to publishers or other external agencies."

"The University will provide support for research data management, including… ...Additional infrastructure and services for research data management, to be developed in consultation with researchers."

The research institution is the body responsible for providing the researcher with facilities for research and is therefore responsible for providing the researcher with the necessary infrastructure and services to support RDM. Design of this infrastructure must be based upon the researcher workflow, so as not to burden the researcher with additional work or by changing their practices, and where possible, making RDM processes virtually automatic and invisible to the researcher.

1.3. Research data lifecycle

Data collected during a research project will include the research data themselves; experimental, observational, modelled data etc. together with the metadata that describes these data in detail and documentation describing the context of the research, details of the research project and the processes involved. Data here will be defined as the numerical and textual information collected by analysis or measurement from the research samples or objects – but not the samples or objects themselves. For example, a collection of images or of tissue samples will not be considered data until there is textual or numerical information, such as identifiers (names or ID numbers), descriptions and relationships, associated with them. In this document, data refers to digital data, although the same management principles apply to analogue data formats. However, it is best practice to digitise research data, making its discovery and reuse easier.

From the initial drawing-up of a research proposal and grant application, data will be collected and managed. This includes documentation about the project, people and bodies involved, grant application, data management plans, experimental protocols, possibly test data or data collected from previous projects for re-use and a literature review or bibliography.

Once underway, a research project will collect or create raw data, which, during the project, will usually be processed to create derived or processed data. There may be many different iterations of processing, resulting in many sets of derived data. Eventually a set of 'results' data will be selected as the basis of the research publication(s) output by the project. All these sets of data can be

[9] The University of Sheffield Research Data Management Policy http://www.shef.ac.uk/ris/other/gov-ethics/grippolicy/practices/all/rdmpolicy

considered **active data**, which will need to be quickly accessible and easily shared between collaborators. All these sets of active data will need appropriate documentation to describe the processes involved in their creation and modification.

After the project has finished, researchers will need to select data for curation on a long-term basis. This may have been decided in agreement with the research funder during the initial planning stage of the project. Curation in the context of RDM, refers to archiving, preservation and adding value through transformation and reuse. These **archive data** selected for curation, may need further processing (validation, cleaning, anonymisation or redaction) before submitting to an appropriate repository. The associated metadata will be needed to provide the necessary information for citation and re-use. There may be the facility to add new metadata or documentation, generated by data reuse, to the curated dataset. Data not required for curation needs to be disposed of in an appropriate manner.

1.4. Data documentation, metadata and data collections

Data need to be documented to be understood and managed. **Data documentation** indicates the conditions and processes involved in the creation or collection of the data, the processing of the data and the context of the research. Detailed documentation is essential for verification and reuse. Adequate data documentation is necessary to determine provenance, licensing and access arrangements and preservation requirements. Research data need to be documented at three levels:

- Project level – providing an overview of the research context and design.
- File level – describe the relationships between files or database tables.
- Item level – describing, for example, the meaning of a variable in a table.

(Research Data Mantra, 2014, UKDA, 2014)

Metadata are a highly structured subset of core data documentation. Metadata are structured so that they may be indexed and stored within a database, thereby facilitating data organisation and discovery, and machine to machine interoperability. By considering its function, metadata may be divided into three layers:

- Core metadata (Datacite, 2011, p. 8) or Catalogue metadata - creator name, publisher, title and an identifier are required for discovery and correct citation of the data. This could possibly include some subject description or classification details.
- Detail metadata or Administrative metadata – provides generic dataset description. This includes access, preservation and technical metadata, and is required for the long-term curation of the data. This will include more detailed classification / subject description.
- Discipline specific metadata (also known as Reuse metadata) - This documents aspects of the dataset that will be of interest to researchers wishing to validate the research process or re-use the data. This will consist of experimental protocols, instrument settings, and relationships with other elements of the dataset, other files within a data collection or other data collections. This will provide very detailed classification / subject description, providing the fine-grained attributes of data necessary for accurate discovery and location of elements within a dataset. Discipline specific information is frequently held in unstructured formats, so could be considered data documentation rather than metadata.

(Ensom, 2013 and IDMB, 2011)

Data collections are typically organised by reference to a particular survey or research topic and may cover a specific geographic area and time period. The UKDA defines a data collection as typically comprised of three components: data, documentation and metadata. Code is occasionally considered a fourth component (Ensom and Corti, 2012, p. 3).

1.5. Data repository or Data registry?

A repository is a content management system, or digital asset management system, which may be considered to consist of three elements – a user interface front-end, and a database layer and storage layer back-end. The database holds records of entities, consisting of metadata elements as a series of fields. The storage layer, containing the actual data bitstreams, may be a file system on the repository server, or a file system on a local server or a remote server that is independent of the repository system. This may include cloud storage or a hybrid storage system. Usually in a repository system, only metadata records are held within the database, not the data objects themselves. This is due to the larger size of data objects which results in slower indexing / access speeds. Storage designation is handled by a storage controller or storage resource broker.

Although repositories were originally designed to manage digital documents, most can be modified to manage any data type. Different repository systems may be configured for different organisational structures. Repositories may range in the granularity of data described: The entity described by a repository record, the data object, may be a single row within a database or spreadsheet, a single file, or a collection of interrelated files constituting a dataset. In the Essex ePrints [6.1.4][10] context, the 'eprint', the key entity, is the 'data collection' which consists of a set of metadata and files (Ensom and Wolton, 2012). Datasets may be grouped as collections or groups, the 'User' entities may be grouped as a 'Community'.

A 'Data registry', or 'Data catalogue' or 'Metadata store', is a repository system that holds only metadata records. The data themselves are held on a local or remote file storage system, so the metadata record points to the data store filepath or URL. By using the appropriate metadata standards, metadata records may be exchanged between registries, giving rise to the possibility of national and international registries. Repository systems were originally designed to curate textual digital objects, but are now being modified to curate digital objects of all formats and sizes, with the view of extending their purpose to curate research data (Gutteridge, 2010). As they have been designed to manage, curate and publish research outputs, they may perhaps provide the ideal platform for a catalogue for the data underlying the research outputs.

The ANDS[11] strategy has been to separate the data storage function from the cataloguing function, the dissemination function and access control are provided by the metadata store (ANDS, 2011b).

1.6. Research data ecology

In considering the implementation of an Institutional RDM technical infrastructure, it can be useful to consider the ecological approach (Robertson et al. 2008), to gain a better understanding of the

[10] Essex Research Data http://researchdata.essex.ac.uk/ [see section 6.1.4 for information]
[11] Australian National Data Service (ANDS) http://ands.org.au

interactions between repositories and services. The local infrastructure does not exist in a vacuum and must interact with, and is dependent on, a diverse range of entities and processes in the information ecosystem.

In determining the place of a repository, registry or other infrastructure component in the overall information ecosystem, it may be helpful to identify a range of components by considering their data storage coverage and specialisation (ANDS, 2011a), which may be:

- Local – personal, project or departmental server for active data storage.
- Storage associated with an instrument or facility.
- Institutional storage for active data – networked filestores.
- Institutional repository for archive data.
- Multi-institutional project data storage (CARMEN [6.2.5] for example).
- Research council data centre – for archive data including longitudinal data.
- Discipline based repository – for active and archive data. National or International coverage.

Metadata storage will have a similar range (ANDS, 2011b), which may be:

- Local, project and instrument based metadata store – spreadsheets or databases associated with the data.
- Institutional data registry or repository.
- National data registry (ANDS).
- Discipline based metadata store – may be international, national or multi-institution based.

A number of practitioners have suggested that, regarding research data curation, "***the Institutional repository is the repository of last resort***" (Haywood, 2013), since discipline based repositories are better configured for the types of data and specialised metadata formats associated with the research community they serve. However the importance of the institutional research data services (tier 3) in the hierarchy of rising value and permanence (**Figure 1.** below) is emphasised in the Royal Society report 'Science as an open enterprise' (2012).

This is reiterated by Simon Hodson (2012), who maintains that Institutional research data services are essential because:

- The institution is where the data are created and can be captured. Institutions implementing RDM infrastructure will make data discovery and curation possible.
- Joining the gulf between curated data in national / international data services and uncurated and inaccessible data in individual or project collections.
- Elevating data to national / international data services from temporary and inaccessible individual collections. Important data collections may emerge as they become discoverable.

The data catalogue component of an institutional infrastructure would ideally adhere to the formats and schemas used by a national data registry under development. The existence of tools for interoperability and for deposit to the major data centres should also be a consideration in the selection of a repository system.

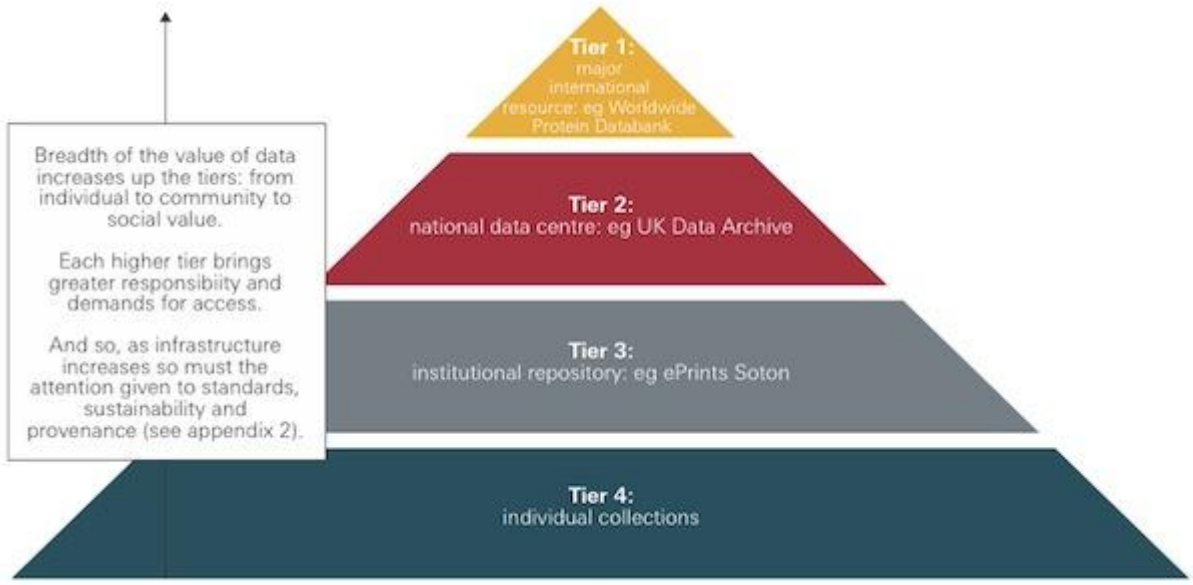

Figure 1. The data pyramid - a hierarchy of rising value and permanence (Royal Society, 2012)

1.7. Development of RDM services

The DCC have created a guide to developing RDM services at HEIs (Jones et al. 2013), which breaks down the development, process into a number of components, as visualised in **Figure 2**.

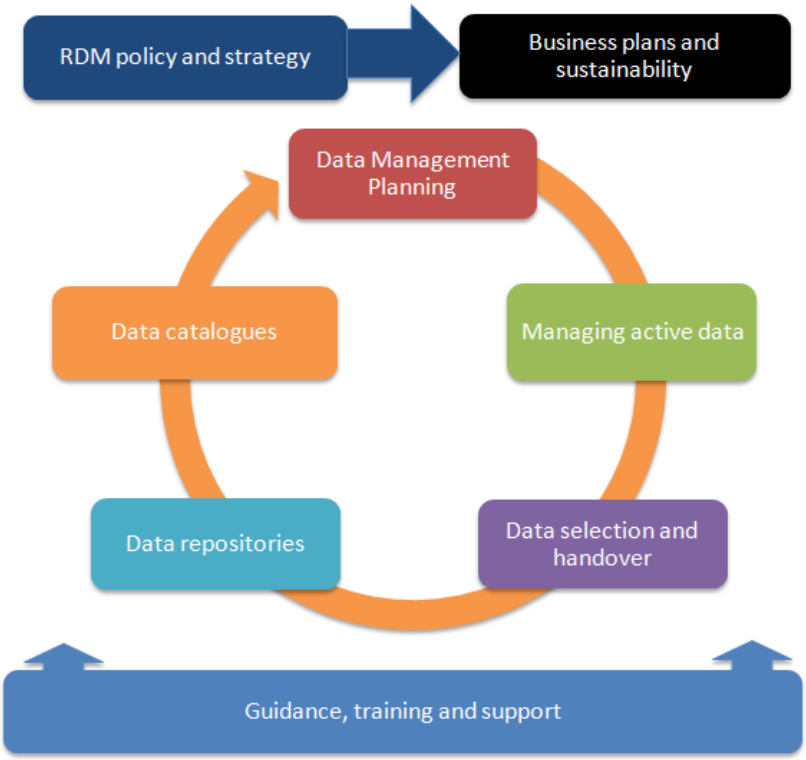

Figure 2. The components of an RDM service as envisaged by the DCC (Jones et al. 2013, p. 5)

The approach to the development of RDM Services for HEIs, recommended by the DCC involves:

- Assembling a steering group composed of senior representatives of stakeholder group – senior researchers and institutional support service managers.
- Appointing an RDM service development group to undertake the work.
- Carrying out a gap analysis, to determine gaps between current position and the aimed-at future position, and requirements gathering surveys to determine stakeholders needs.
- Development of RDM policy and strategy. Developing a policy first may be useful as a motivating factor, but may lead to problems if the proposed infrastructure and services cannot be realised. Alternatively, the development of policy may be subsequent to the defining of strategy.
- Designing services to meet local and external needs – putting in place the infrastructure required to support these services.
- Piloting these services to test that they are fit for purpose.

The DCC have published a case study detailing the development and implementation of an RDM strategy (Rans & Jones, 2013). Design of the technical infrastructure required to support the RDM Service is considered below.

2. RDM Technical Infrastructure Architecture

2.1. Technical infrastructure components

Generally, the infrastructure architecture cases examined below have been designed around functional requirements derived from researcher workflows. Some innovative infrastructure platforms have been architected to manage research data throughout the lifecycle, but most infrastructure projects have had to take into account current systems and have engineered modifications to facilitate interoperability (Hitchcock, 2012).

All components of the RDM technical infrastructure need to be interoperable. This is achieved through adherence to data and metadata formats and standards allowing data and metadata exchange between interoperable systems.

2.1.1. Major functional components of the RDM infrastructure (see **figure 3**):

- **Current Research Information System (CRIS)** [4.7] - manages the metadata associated with researcher identity, project information, research costing, grant applications and awards. The CRIS provides a register (inward facing catalogue) of the researcher's published outputs with associated citation metrics and may act as a means to deposit publications into the Institutional Repository if interoperable. The CRIS may feasibly be used to push metadata only records of research data to the Institutional Repository, or in itself, function as a catalogue of the institution's published research data outputs.
- **Data Management Planning Tools** [4.8] - These are facilities for aiding researchers compile the DMP required by some funding organisations to be submitted as part of the grant application process. Such tools provide templates for the different major funders and may be customised for the institution. The tools may be accessed via institutional login and the DMPs created, stored in the institutional CRIS.
- **Data and Metadata Capture** [4.10] - Metadata capture (data cataloguing) may be accomplished simply by providing an interface for researchers to fill out online forms. It is best for this process to be automated where possible to reduce the amount of manual annotation required of researchers. As well as reducing 'double-keying', which is frustrating for researchers, the number of errors introduced (inevitable through manual input) is reduced. Automatic metadata capture, concurrent with data capture, may be facilitated by using appropriate instruments and equipment and save data to the laboratory, departmental or facility file store or the institutional network. Electronic lab books, electron microscopes and other imaging instruments, genetic sequencing and analysis instruments may feed data to a project based Laboratory Information Management Systems (LIMS) [4.10.1]. Ideally, data and metadata need to be transferred to central active data storage.
- **Active Data Storage (including HPC Grid storage)** [4.5] - Active research data need to be rapidly accessed, easily shared between collaborators with access being controlled through stringent security arrangements. Working datasets, which change constantly (as they are being created, added to, processed and edited), will require Read-Write access, frequent back-up and may require large computational resources. However, much 'active data' will be immutable and only require Read-only access, so therefore may possibly be moved to 'Archive data' storage. This needs consideration in order to provide cost-effective storage.

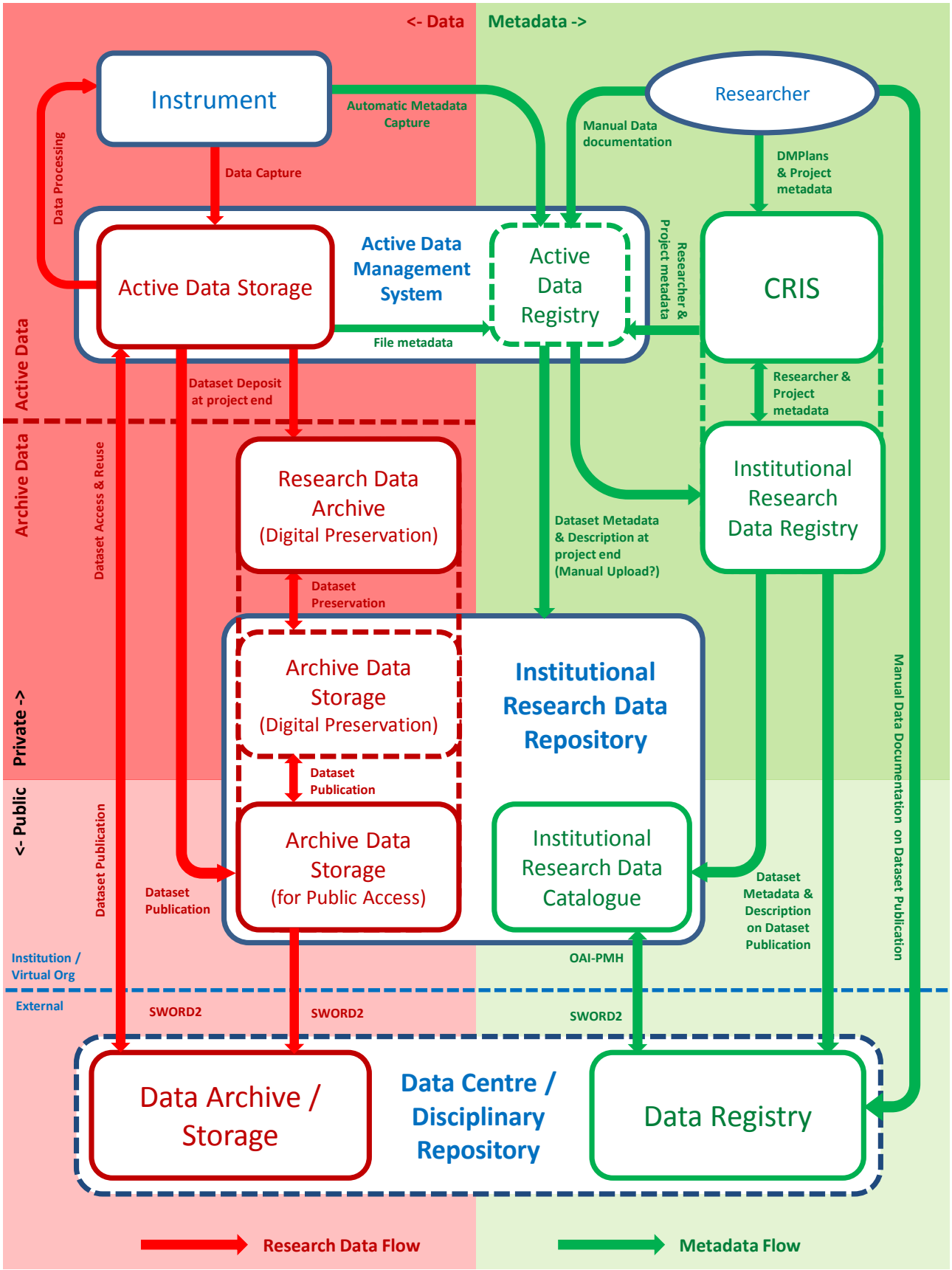

Figure 3. RDM technical infrastructure data flow – utilises the concept of the four quadrants of research data curation systems (Lewis, 2013).

- **Active Data & Collaboration Management** [4.5] - Collaborative computing systems, such as HPC Grids, Dropbox-like cloud storage services, Virtual Research Environments (VRE) and Laboratory Information Management Systems (LIMS), have been developed to accommodate the need for secure, read-write access to shared storage. Active data management systems may be considered to be comprised of three functional components: a storage layer, a data registry (or metadata store or asset registry) database layer and a User interface layer. In some cases these components will be integrated into a single system, in other cases, the metadata may be handled by the CRIS and storage by the institutional network.
- **Research Data Selection / Deposit Facility** - The institution may provide a service to help researchers appraise their data, assess the preservation requirements, help with submission to the institutional repository and help with submission to external repositories.
- **Archive Data Storage and Digital Preservation** [4.4 & 4.3] - Many archive data storage arrangements distinguish between permanent archival storage, maybe through an external service, and operational storage on a local server, holding ingested files for processing and access copies of the data. This distinction is due to the slow access speed and higher cost of retrieval from archival storage. Control of the system will be mediated by a Storage resource broker. Data selected for long-term preservation will require storage that ensures the file remains immutable. This 'Bit preservation' requires constant management and regular back-up to a variety of media including tape storage and off-site or cloud storage. A number of vendors offer a digital archiving service, but because of the high costs involved in retrieval, such archival storage services are most appropriate for back-up copies not access copies of data. Digital preservation refers to the active management of the processes that ensure 'bit preservation' data immutability and ensure that the material will remain accessible in perpetuity. A Research Data Archive will be required to preserve data not, or not yet, submitted to discipline-based data repositories. The associated metadata records are held in a research data registry or catalogue.
- **Research Data Registry** - The Registry is defined here as an inward-facing catalogue that holds the metadata records of unpublished research data. The data themselves will be held in an institutional data archive. The data and metadata may be eventually published by ingest into a discipline-based data repository outside the institution or into the Institutional Data Repository. The CRIS may function as a research data registry, providing researchers with an interface to record metadata in order to register a dataset.
- **Research Data Repository** [4.2 & 4.3] - Archive data is possibly best managed in a discipline based repository or data centre, whilst the Institutional repository is the 'repository of last resort', as previously discussed. The institutional data repository (or the institutional repository, if it has been modified to accommodate data) is an appropriate home for datasets for which there is no discipline based repository or data centre, or for temporary storage before being submitted to a data centre. A wide range of repository systems are available, from proprietary, externally managed services to open source software systems. Many have been designed to manage specific media, or designed to provide digital preservation or Research data management specific functions (a Digital Preservation System is a repository that manages the active preservation of the content). The Institutional Research Data Repository may provide a public catalogue for all published data created at the institution – the data being held by data centres / discipline based repositories as well as that held by the institution. The catalogue and

archive storage functions of the repository may be separated. In such an arrangement, the archive storage function may be achieved using an external service or in an institutional data archive, but access and deposit managed seamlessly through the repository platform. Where the repository holds only metadata records, it may be considered a Research Data Catalogue.

- **Research Data Catalogue** [4.6 & 4.2] - The Data Catalogue is defined here as a publicly-accessible catalogue, holding the metadata records of published research data. The data themselves may be held in a discipline-based data repository outside the institution or in an institutional data archive. The Research Data catalogue may be provided by a number of repository platforms. The selection of the underlying metadata schema is fundamental and consideration must be given to the schema used by the proposed National Data Registry[12]. Many institutions favour the Datacite metadata schema[13], subscription to which provides the means to mint DOIs and assurance of a standard level of preservation.

Many implementations involve an overlap in these functional components; for example the CRIS may provide a research data registry (inward facing); Laboratory instruments may be part of a LIMS, so that automated data capture is an integral part of the collaborative active data management system; Data storage and data catalogue (outward facing) may be separate systems or may be combined in a data repository. The storage / archive function of the data repository may be achieved using an external archive service, but access and deposit managed seamlessly through the repository platform.

2.1.2. Data grids and Micro-services

The data grid is a form of RDMI architecture in which middleware applications allow researchers to manage data across grid infrastructures. Grid computing involves a distributed infrastructure served by interoperable software services, 'middleware', allowing resource sharing; the resulting 'Grid' may be considered a 'virtual organisation' (Foster et al. 2003). Data grids permit the sharing of computational resources, storage resources, network resources, code repositories and catalogues. Access to Grid resources are controlled by a Resource management system, or Storage resource broker (SRB). Several middleware toolkits are available, including open source options Globus [4.13.1] and MyGrid [4.10.10].

The micro-services architecture approach considers the repository as a 'set of services' rather than a 'place' (Abrams et al. 2009). Each function in a workflow is embodied in a self-contained micro-service, which is joined with other micro-services in a 'pipeline' to produce complex processes. The micro-services approach has been developed by the California Curation Centre[14] and put into practice at the University of California Digital Library (CDL) Merritt repository [6.3.5]. At the University of Oxford, micro-services are built on an underlying Fedora repository platform, creating the Databank repository system [4.2.4], used as the platform for Oxford ORA-Data [6.1.15].

The iRODS [4.1.3] software system allows the management of a distributed workflow through the chaining of micro-services. iRODS software is termed 'adaptive middleware' and allows for a more flexible customisation of data management functions than can be achieved using a SRB system.

[12] DCC Research Data Registry Pilot http://www.dcc.ac.uk/projects/research-data-registry-pilot
[13] Datacite metadata schema 3.0 http://schema.datacite.org/meta/kernel-3/ [See section 4.12.1.]
[14] CDL Microservices https://wiki.ucop.edu/display/Curation/Microservices

These functions or micro-services are coded as 'rules', which may be compiled together to produce larger macro-level functionality.

Hydra [4.2.5] is multi-purpose repository framework based on a micro-services architecture. The main components are a Fedora repository platform [4.2.3], SOLR indexing software [4.13.2], Blacklight discovery interface [4.6.6] and the Hydra plugin, a 'Ruby on rails' library, which facilitates workflow in digital object management (Awre, 2012). Hydra has been implemented at the University of Hull [6.1.10] and the University of Virginia [6.3.10] with provisions made for curating research datasets. At Hull micro-systems implement workflows that allow deposit of materials via the CRIS, Converis [4.7.3], the Sakai VLE [4.5.2] and Sharepoint [4.5.3].

2.2. Functional requirements

The functional requirements of the RDM Infrastructure may be derived from analysis of stakeholder activities, particularly researcher workflows. Many of the JISC RDMI projects have carried out data audits and investigated researcher workflows and use case scenarios in order to specify infrastructure requirements. The following list is derived from the findings of several of the JISC RDMI projects: ADMIRe (Sero Consulting, 2012; Parsons and Berry, 2012), CKAN for RDM (Winn et al. 2013), KAPTUR (Garrett et al. 2012), Orbital (Stainthorp, 2012) and RoaDMaP (2013) [see section 7.1. for more information about these JISC RDMI projects].

Researcher requirements:

a) For active data

- Direct capture of data (and metadata) from instrument.
- As much automated metadata annotation as possible, such as project level metadata (researcher identity and grant information) imported from the CRIS.
- Network that provides adequate storage (personal and project) which is regularly backed up with speedy access to large data volumes.
- Secure, authenticated access mechanisms are required, especially for sharing sensitive data; usually involves institutional authentication mechanisms (Shibboleth [4.12.10]).
- Ability to share data with collaborators inside and outside the institution ('Academic Dropbox').
- Mechanisms for secure data destruction.
- Mechanisms for data transformation as required for data curation (such as anonymisation, aggregation and format transformation).

b) Depositing archive data

- User friendly data upload facility (like Dropbox [4.5.6]).
- Customisable workflows for creating or importing metadata and uploading file.
- Simple process for ingest of large data collections (multiple files) and association of collections with single metadata record (dataset record).
- Controlled lists for some metadata fields.
- Support for versioning of datasets.
- Clear choice of license options.

- Specify granular access rights to files at data object and collection level.
- Embargo options for metadata and files.
- Mechanisms for secure data destruction.

c) For data discovery and reuse

- Effective search and discovery mechanisms, using subject-specific terminology. Controlled vocabularies of keywords with auto-complete function.
- Enable immediate access to datasets.
- Access to datasets held outside the repository.
- Support access to very large datasets.
- Means of access to restricted data, where the metadata is visible; a 'contact owner' button.
- Linking dataset to context / reuse metadata or data documentation – describing the process of data generation.
- Related data and research publications indicated and linked to.
- Support for granular access to data and associated metadata.
- Visualisation and data analysis tools to give summary data or overview of data. Support query and processing of data on the repository server rather than after download.
- Support for free tagging - adding discipline specific tags or metadata to datasets.
- Federated catalogues allowing searching across multiple institutions.
- Advice on data citation.
- Citation data produced, demonstrating impact.

Additional RDM service requirements:

- Customisable metadata schema.
- Support multiple ingest protocols.
- Staged deposit workflow – allows administrative area for quality check / validation.
- Enable selective metadata harvesting.
- Enable extraction of metadata and data in open format.
- Support open standards and exposure of metadata.
- Support multiple content licensing – exposed clearly.
- Support technical metadata.
- Support generation of persistent unique identifiers.
- Support open methods of authentication.
- Ability to remove data to access controlled area, a dark archive for embargoed data
- Ability to delete data, generating a tombstone reference.
- Access to metadata through library catalogue – OAI-PMH [4.11.3] endpoint required.
- Support reporting – analysis of repository content, download and view metrics.
- Enable creation and retrieval of an audit trail, reporting management actions.

2.3. Institutional considerations

Expediency may perhaps determine the development of the Institutional RDM technical infrastructure. In the current climate of budget constraint and with the need to demonstrate value for money, there should be a focus on appraising the systems currently in place, and determining whether these may be modified to fit the proposed infrastructure. Modification of existing components will require local expertise or employment of developers, often the more expensive aspect of system implementation. Thus, work will be needed in costing the options available: building upon and integrating existing components, or otherwise implementing a new fully-integrated system, possibly a proprietary system, replacing existing components where necessary.

The least the institution needs to do for the development of the RDM technical infrastructure:

- Implement institutional policy – ***'Additional infrastructure and services for research data management, to be developed in consultation with researchers.'*** Therefore a research data audit is recommended to determine researcher practices.
- Fulfil Funder requirements – ***'Research organisations will ensure that appropriately structured metadata describing the research data they hold is published...'*** Therefore a data catalogue is required by 1st May 2015.
- Promote and facilitate good RDM practice. Training and guidance resources need to be developed.
- Select sustainable, inexpensive, open options (open for interoperability and sustainability). Business cases will need developing for the various options available.
- Take into account projected future requirements. This involves a consideration of risks to services through the removal of funding (for example the AHDS[15] data centre no longer received funding after 2008, so stopped functioning).

2.4. Requirements gathering methods

In developing the institutional RDM strategy, the DCC recommends using both requirements-gathering and gap analysis methods (Jones et al. 2013). The DCC provide a number of tools for the purpose and have published a case study detailing the use of these tools (Rans and Jones, 2013).

A number of UK institutions have used the Data Audit Framework (DAF)[16] developed by JISC and HATII for requirements gathering. The DAF provides a set of survey methods, questionnaire and interview frameworks in order to identify, locate and describe research data assets and determine how they are being managed. The AIDA Toolkit[17] has also been developed for institutional self-assessment of the readiness and capabilities for management of digital assets and digital preservation.

The Collaborative Assessment of Research Data Infrastructure and Objectives (CARDIO)[18] is a benchmarking tool for RDM strategy development developed from key aspects of DAF and AIDA and

[15] Arts and Humanities Data Service (AHDS) http://www.ahds.ac.uk/
[16] Data Audit Framework (DAF) http://www.data-audit.eu/index.html
[17] AIDA Toolkit http://aida.da.ulcc.ac.uk/wiki/index.php/Main_Page
[18] Collaborative Assessment of Research Data Infrastructure and Objectives (CARDIO) http://cardio.dcc.ac.uk/

other tools. The DCC recommend using CARDIO in conjunction with the other tools, the emphasis being on strategic planning and identifying gaps between the current situation and best practice.

3. The University of Sheffield RDM Technical Infrastructure Considerations

3.1. Local infrastructure components

Some components of an integrated infrastructure already exist at the University of Sheffield:

- The Institutional repository WRRO[19] is a shared service hosted at Leeds University and based on the ePrints platform [4.2.1]. There is administrative expertise at the University of Sheffield Library but software development expertise is located in Leeds.
- The Research and Innovations Service administer an institutional (inward facing) catalogue of research publications, MyPublications[20], which was used to facilitate the REF2014 process. This is built on the Symplectics Elements 4.4 [4.7.1] Research Information Management System (RIMS) and hosted by the Corporate Information and Computing Services (CICS). The Symplectics RIMS is a complete and integrated system, but as yet, its full functionality is perhaps not exploited at Sheffield.
- MyPublications is linked to the SAP platform [4.13.4] Corporate Information System (CIS) to import researcher identity and Academic unit information.
- MyPublications is linked to WRRO by a connector so that metadata may be pushed into, and files uploaded to, WRRO and outputs published easily.
- The University Research Management System (URMS)[21] is a web based tool used for costing and pricing research, obtaining approvals, managing awards and post award administration. It is a module of the CIS and is provided and developed by CICS.
- A Digital Asset Management System (DAMS) for library special collections and National Fairground Archive[22] based on the ContentDM [4.2.10] repository platform. Metadata is harvested by the library catalogue system Primo.
- Library management system, Alma, and Resource discovery tool or OPAC, Primo [4.6.7].
- CICS Network provides network filestore for individuals and departments. Files are available on and off campus, through Shibboleth [4.12.10] authentication. Off-campus access is via the internet or via VPN. Filestores are regularly backed up.
- CICS provides cloud storage and collaboration resources using Google Services [4.5.7].
- CICS provides a Linux based high performance computing (HPC) cluster, Iceberg[23], which is the Sheffield node of the White Rose Grid (WRG)[24]. Iceberg provides backed-up storage, space for using and storing very large amounts of data and facilities for project group level collaborative work. Several research groups at Sheffield are, or have been involved in WRG facilitated collaborations, such as the CARMEN Portal [6.2.5], which uses the iRODS grid management system and Pegasus[25] which used the SRB management system.
- Sheffield is a member of the N8 HPC Grid, currently operating Polaris[26], an SGI HPC cluster.
- Sheffield is a member of other Grid computing projects, GridPP[27] and WUN[28].

[19] WRRO – White Rose Research Online http://eprints.whiterose.ac.uk/
[20] MyPublications https://www.shef.ac.uk/ris/post-project/mypublications
[21] URMS - University Research Management System http://www.sheffield.ac.uk/ris/application/pricing/urms
[22] University of Sheffield Library Digital Collections http://cdm15847.contentdm.oclc.org/cdm/
[23] WRGrid Iceberg http://www.shef.ac.uk/wrgrid/iceberg
[24] White Rose Grid (WRG) http://www.wrgrid.org.uk/
[25] Pegasus http://hridigital.shef.ac.uk/pegasus
[26] Sheffield N8 HPC http://www.shef.ac.uk/wrgrid/n8

3.2. Consortia options

The University of Sheffield is currently involved in two consortia established to share services and foster collaboration. Groups from both consortia are investigating the feasibility of sharing RDM resources – infrastructure components, training materials and expertise.

3.2.1. White Rose Universities Consortium (WRUC)[29]

The WRUC collaboration between the universities of Leeds, Sheffield and York, established in 1997, supports research, knowledge exchange, and teaching and learning through a number of projects. The White Rose Grid (WRG), established in 2002, offers grid infrastructure and HPC for the WRUC. The consortium manages the shared institutional repository for research publications, WRRO, established in 2004, and for theses WREO[30]. Work done recently by the White Rose Libraries Systems Architecture Group (2013) has identified the various components of each institution's infrastructure in order to establish areas of interoperability and shared experience.

RDM Infrastructure components in operation at Leeds:

- The shared Institutional Repositories WRRO and WREO, based on the EPrints platform and housed on a Leeds server.
- A Symplectics CRIS component, for managing publications.
- EPrints-Symplectics connector.
- 'Kristal'[31], an in-house built grants management component of the CRIS.
- SAP based Corporate Information System.
- Shibboleth identity authentication.
- 'LUDOS'[32], a Digitool [4.2.11] Digital asset management system. This is to be replaced by an EPrints platform data repository.
- VRE / Research portal on SunGard Luminis platform [4.5.8] – to be replaced with Microsoft Dynamics [4.5.9] product suite.
- III Sierra [4.6.8] based library management system.
- HPC as part of WRG and N8 Polaris HPC.

At Leeds, the RoaDMaP project [7.1.15] is continuing due to interim funding by the University, to investigate service scoping and development. Proposed developments in the RDM infrastructure at Leeds include the choice of EPrints for the data registry or ***'discovery metadata store which points to the data'*** (Proudfoot, 2013a), the testing of Arkivum [4.4.1] hardware for archive data storage capabilities and its interaction with EPrints, and the integration of DMPonline [4.8.1] with Kristal (Proudfoot, 2013b).

---

[27] GridPP http://www.gridpp.ac.uk/
[28] Worldwide Universities Network http://www.wun.ac.uk/
[29] White Rose Universities Consortium (WRUC) http://www.whiterose.ac.uk/
[30] White Rose ETheses Online WREO http://etheses.whiterose.ac.uk/
[31] Knowledge Research Innovation System at Leeds (KRISTAL) http://www.leeds.ac.uk/forstaff/news/article/3826/get_started_with_kristal
[32] Leeds University Digital Objects (LUDOS) http://ludos.leeds.ac.uk/ludos/

RDM Infrastructure components in operation at York:

- The WRRO EPrints repository.
- A PURE [4.7.2] CRIS.
- EPrints-PURE connector.
- 'York Research Database'[33] – a registry of research publications and project information; a component of the PURE CRIS (equivalent to MyPublications at Sheffield).
- Shibboleth identity authentication.
- 'YODL'[34], the DAMS based on Fedora Commons [4.2.3] platform. Currently this holds small volumes of Humanities research data.
- 'Youshare' [4.5.10] an in-house built VRE / Portal system for sharing data and software.
- 'Alfresco' [4.5.11] an open source content management system, developed by the department of Biology as a research collaboration system.
- ExLibris Alma Library management system, with a Primo discovery interface 'Yorsearch'[35].
- Google Drive – University supported cloud storage and collaboration space.
- 'York Identity Manager' (IDM)[36] local authentication system (Shibboleth).
- HPC as part of N8 Polaris HPC and WRG.

Various RDM infrastructure developments at York are being considered include integrating DMPonline with the PURE CRIS and using Rosetta [4.3.3] for digital preservation. DataStage [4.5.1], Dropbox and Amazon Glacier [4.4.2] with AWS [4.5.12] are under consideration for active research data management; CKAN [4.2.2], DataBank [4.2.4] and Figshare [4.3.1] are being considered for data storage and discovery and Arkivum [4.4.1] is being investigated for archive data storage (Allinson, 2013).

The White Rose Research Data Working Group (2013) undertook a high level assessment of the options for a shared research data repository service, resulting in three options:

- 'The Bakery' – a regional service hosted and managed by one institution which the others pay to use.
- 'The Cake Mix' – a recommended best practice service to be deployed and managed in the individual institutions.
- 'The Recipe' – Individually developed solutions to agreed standards, which may be modified to suit local requirements.

It is considered possible for any of these options to share elements of the infrastructure, such as published API, data catalogue, storage management system or raw storage, and that these shared services could be organised on a regional or national basis.

[33] York Research Database https://pure.york.ac.uk/portal/en/
[34] York Digital Library (YODL) https://dlib.york.ac.uk/
[35] Yorsearch http://yorsearch.york.ac.uk/
[36] York Identity Manager (IDM) https://idm.york.ac.uk/idm/user/login.jsp

3.2.2. N8 Research Partnership[37]

The University of Sheffield is a member of the N8 Research Partnership, a collaboration of eight research intensive universities in northern England (with Durham, Lancaster, Leeds, Liverpool, Manchester, Newcastle and York). The N8 provides the shared HPC facility and an equipment sharing initiative n8equipment.org.uk[38]. The N8 RDM Architecture Working Group has drafted a reference systems architecture model for RDM across the N8 institutions. This model is useful for visualising the components of the RDM infrastructure of each member institution; to date, three of the eight institutions, including Sheffield and York, have yet to submit maps. The N8 RDM Archiving and Curation Working Group is investigating the feasibility of a shared storage service and developing data appraisal and curation processes and policies.

The N8 institutions' experience regarding RDM infrastructure implementation is limited to Leeds, Manchester and Newcastle and results from the JISC RDM projects RoaDMaP [7.1.15] at Leeds, MiSS [7.1.10] at Manchester, and Iridium [7.1.7] at Newcastle. The RDM infrastructure is most highly developed at Newcastle [6.1.14], where a suite of tools[39] are offered including a CKAN data portal[40], currently under development. Newcastle have developed their own research information systems as part of this suite, which comprises 'MyImpact' (a researcher profile and publication information system), 'MyProject' (a project and awards management system), VRE and eScience Central (Collaboration and workflow tools) and a Research Data Catalogue (linking data, projects and publications), although this is currently a proof of concept system.

Common infrastructure components at the five N8 member institutions outside WRUC:

- EPrints based Institutional Repositories at Durham, Lancaster, Liverpool and Newcastle.
- Fedora Commons used for Manchester IR 'eScholar' and Durham University Library special collections.
- Agresso / pFACT [4.13.5] financial management tool used at Durham, Lancaster, Liverpool and Manchester.
- Oracle Financials [4.13.6] used at Durham and Manchester.
- Oracle based CIS and CRIS 'ISIS' developed at Liverpool.
- CKAN used for Data Repository at Newcastle.
- PURE used at Lancaster.
- DMPonline used at Lancaster, whilst Manchester developed their own DMP tool.

[37] N8 Research Partnership http://www.n8research.org.uk/
[38] n8equipment http://www.n8equipment.org.uk/
[39] Research Data Management Tools, University of Newcastle https://research.ncl.ac.uk/rdm/tools/
[40] CKAN, Research data management, University of Newcastle https://research.ncl.ac.uk/rdm/tools/ckan/

3.3. Recent reviews of RDM service development at Sheffield

3.3.1. Research & Innovation Services RDM project - Case studies

During 2011-12, the University's Research and Innovation Committee commissioned a Research Data management Scoping Project to:

- Establish funders' current and potential RDM requirements.
- Identify gaps between funders' requirements and university practices.
- Identify and characterise support needs from pathfinder projects.
- Explore the university's capabilities for meeting RDM requirements and to propose sustainable, viable extensions to support services.

The pathfinder interviews provided insights into the researchers' perspectives, many of which had a bearing on RDM. Many find the idea of making research data publicly available, contentious and there is much confusion between open access to research articles and access to data. Many researchers need guidance in choosing between a range of storage options, and do not distinguish between active data storage, back-up, mirroring and archiving. Researchers will support initiatives that are researcher led and involve little bureaucracy. They will tend to adopt RDM practices that fit in with their work flows, the type of data they work with and that are aligned with their culture. There is a need for clarity on the issue of data ownership.

A list of actions that will potentially lead to the establishment of a University RDM support infrastructure was drawn up through a SWOT analysis of the current position. The resulting project report (Kane et al. 2012) recommends the following actions with implications for technical infrastructure:

- **Data organisation** – Update University's RDM Policy with metadata standards developed across the HE sector. Develop guidance on metadata for researchers. Establish a network of RDM expertise across the university.
- **Data management and planning** – Clarification of roles and responsibilities with regard to RDM support. Development of Data management plan (DMP) templates.
- **Data storage and back-up** - Clearer guidance on available storage options, the costs involved and advantages / disadvantages of these options is required. Guidance on data retention is needed. Departments should account for RDM in their business continuity plans.
- **Data sharing** – Create incentives to encourage data sharing (and RDM generally). Develop case studies that demonstrate the benefits to researchers of data sharing. Expand the University's RDM policy to cover data sharing and open access to data.
- **Data Repository** - Develop a repository for research data that are unsuitable for external archiving, possibly in collaboration with White Rose or N8 partners.
- **Data Catalogue** - Develop a system to catalogue, document and log all datasets held by the University, datasets held elsewhere and third party datasets acquired by the University.
- **Data ownership –** Clear guidance is needed regarding IP and ownership of data.

The report concludes that the R & I committee must decide which of these actions to support and prioritise and recommends that the University should investigate cost estimation and recovery models that enable research projects' RDM costs to be covered and investigate incentivising good RDM practices.

3.3.2. White Rose Services Repositories review

A review of the consortium's repositories and related research information infrastructure was commissioned by the White Rose Library directors in order to establish a five year roadmap for the development of the repositories (Kay and Stevens, 2012). The review highlights the emerging need for research data storage and the broader future role for repositories in the developing research data ecosystem.

Key recommendations with a bearing on the technical infrastructure include:

- The WR consortium should be maintained and considered the default channel for delivery of shared research services.
- Continuing ePrints development for the current service in parallel with future service design.
- Prioritise WRRO connector developments and ePrints upgrades to enhance deposit workflows.
- Working together to design and validate a research information infrastructure framework and to determine the role of WR services in this framework.
- Testing the feasibility of micro-services architecture for the future research information infrastructure.
- Assessing the potential of ePrints platform to meet priority requirements of the framework.
- Taking an incremental approach to development of infrastructure rather than a 'big bang' implementation.

The report discusses the challenges involved with using a single platform, ePrints, to provide a wide range of functionality. Integration with other institutional systems, such as the CRIS (Symplectics at Sheffield and Leeds, PURE at York) is considered difficult and the slow development of the connector was considered to have a negative impact on the WRRO service. Future core ePrints developments may not be in line with WR requirements, but by investing in local development work or commissioning developments through 'ePrints Services' this may be achieved.

The report strongly recommends testing the feasibility and business case for a micro-services architecture [as described in section 2.1.2.], particularly the Hydra system [4.2.5]. It is suggested that using micro-services to provide the required functions of the infrastructure will avoid duplication of functions provided by several components of the alternative ePrints-based architecture, and will mitigate the problems of integration with external components. It is also noted that although Hydra is integrated with Fedora in current implementations, there is a possibility of integration with ePrints providing the same functions as Fedora. However, implementation of the micro-services approach will require substantial investment and as yet there is no wide community of such repositories.

The report recommends that adoption of the micro-services approach offers the best long-term strategy and proposes a micro-services system trial during 2013-2014. In the short to medium term, the report proposes continued investment in the current repository infrastructure, focussed on mechanisms of deposit via Symplectics and PURE. This will mean a commitment to maintaining the relevant connectors after ePrints upgrades. Service enhancements will need to be tested on the current ePrints platform.

3.3.3. White Rose Consortium shared research data management services

This feasibility study by the DCC, commissioned by the WR library directors, was carried out in during 2013. RDM Service components were grouped into human and technical infrastructure components and their feasibility was assessed using six criteria: institutional benefits, service costs, staff resources, suitability of component, drivers for development and level of risk.

The resulting report (Rans et al. 2013) makes the following recommendations regarding technical infrastructure:

- **Investigate joint purchase of active data storage hardware**. Each organisation is investigating storage options including that of the WRG HPC facilities and virtualised storage from external providers. Investigation of current practice indicates a preference for direct control over data held locally – with many researchers preferring hard drives on desktop PCs to institutional Filestores.
- **Scope projected data storage requirements for the next 3-5 years by engaging with researchers**. The survey found a wide variation in storage requirements and that the determination of future storage requirements will prove difficult.
- **Investigate federated back-up storage using partner sites**. The current practice amongst researchers was found to vary widely, with inconsistent procedures reported at Sheffield, where no formal arrangement is in place to ensure good practice.
- **Investigate setting up collaborative spaces**. The survey indicated a wide range of collaborations in operation with some holding of third party data. York host 'YouShare', a portal for sharing data and software, and GoogleDrive is widely used at York (as well as Sheffield and Leeds). The report suggests a good case for developing a shared service if all partners require a DropBox like facility. Such a facility, it was suggested, will potentially strengthen collaborations within the consortium.
- **Share technical experience in the development of repository architecture**. The EPSRC deadline of 2015 is likely to drive institutions toward individual solutions. To deliver a shared data repository infrastructure, collaboration needs to be established before individual efforts have developed beyond the point where they can be easily abandoned or integrated. There has been some discussion about extending the WRRO to accommodate datasets, but Leeds and York are investigating alternative options.
- **Development of a formal WRC functional requirements specification**. The report recognises that long-term curation will probably be achieved through a blend of local and external services and that the institutional repository will be positioned as the 'repository of last resort' for research data.
- **Collect disciplinary requirements for tools supporting data ingest, metadata creation and preservation**. Research data shows a wider variation in ingest and deposit requirements than found with publications. The three institutions have experience in managing publications ingest to the WRRO and there is experience of the deposit of other materials at YODL in York and with the Timescapes project[41] in Leeds (using DigiTool). A collaborative approach will avoid duplication of effort. The report recommends the development of ingest and preservation tools if necessary.

[41] Timescapes: An ESRC Qualitative Longitudinal Initiative http://www.timescapes.leeds.ac.uk/

- **Liaise with key data centres to develop plugin deposit tools**, facilitating easy upload to external repositories. The report notes that there is little information about the use of external data repositories, or how much research data is managed in this way.
- **Share deliberations and decisions made about the use of DataCite metadata schema v.3.0 and DataCite DOIs**. Use of the same identifier service will facilitate the creation of a centralised data catalogue or portal.
- **Harmonise explorations of options for data catalogue development**. There is an opportunity for the development of a shared service, but the delivery deadline of 2015 impacts on the time available for successful collaboration. Local options are being explored - York Research Database (using the PURE CRIS front-end) and the use of ePrints as a data catalogue at Leeds.

4. Infrastructure Components

Components of the RDM Infrastructures established by higher education institutions are briefly considered below. The component function, the software / platform underlying the component and component interoperability are described, any evaluations identified, and institutions employing the component, particularly in an RDMI context, are noted. The list of components is not exhaustive, the most relevant and popular are reviewed, and the components are loosely categorised by function so there may be considerable overlap.

4.1. Integrated systems and integrating components

4.1.1. Dataflow http://www.dataflow.ox.ac.uk/

A two stage RDMI consisting of an active data management system, DataStage [4.5.1], with a data repository, DataBank [4.2.4] built on a Fedora Commons platform. The system uses 'Bagit' specifications [4.11.2] to transfer files to a SWORD2 [4.11.1] compliant archive. This system was developed by the Admiral project [7.2.1] and is being piloted at the University of Oxford [6.15], having been implemented during the Damaro project [7.1.3]. Dataflow is being evaluated by the Universities of Leeds and the Yorkshire & Humberside Metropolitan Area Network. [See section 5.5 for Newcastle University's Iridium evaluation of DataStage and DataBank].

4.1.2. Orbital Bridge https://github.com/lncd/Orbital-Bridge

The pilot RDMI built at the University of Lincoln centres on the 'Orbital Bridge' application (Jackson, 2012), which integrates an institutional facing data registry built on CKAN, a public data catalogue built on EPrints and the Research Management system 'Nucleus' (Stainthorp, 2012). The Orbital Bridge provides an interface, the 'Researcher Dashboard' (Winn, 2013b) allowing researchers to access and add information about projects, funding, outputs and datasets [6.1.12].

4.1.3. iRODS https://www.irods.org/index.php/Introduction_to_iRODS

Integrated Rule-Oriented Data-management System (iRODS) is a software system that allows the management of a distributed workflow through the chaining of micro-services. See section [2.1.2.] for more information. The iREAD project [5.14] evaluated iRODS for use in the CARMEN portal.

4.1.4. ONEIS http://www.oneis.co.uk/research

ONEIS for research provides three connected modules to support the whole university research lifecycle. 'Research Manager' is a CRIS component, providing management of grant applications, project costing, ethics and project approval and data management plans. 'PhD Manager' provides administration of PhD registration, supervision, progress review and completion. 'Data Curator' provides a means of long-term secure storage, description of data with rich metadata, data discovery and access, and administration of licensing and access rights. These ONESIS components were developed in collaboration with the University of Westminster.

4.2. Repository platforms

4.2.1. EPrints http://www.eprints.org/

The most widely used platform for institutional repositories (used for WRRO), EPrints is open source and free to use. Bespoke design, hosting and maintenance services are available. EPrints is built from Apache web server, MySQL [4.13.3] and Perl components and recommended to run on a UNIX-like operating system.

A number of plugins have been developed by the EPrints user community, some of which modify EPrints to handle datasets: The ReCollect plugin[42] has been developed by UK Data Archive and the University of Essex to implement a dataset metadata profile; Datacite DOI registration plugin[43]; SWORD 2.0 broker plugin[44]; Arkivum A-Stor storage backend plugin[45]. In addition to these, a number of projects have developed integration of EPrints with other components: KAPTUR [7.1.8] developed integration with Datastage and Figshare; The Orbital Project [7.1.12] developed the Orbital Bridge, which integrates EPrints with CKAN and other components [see above 4.1.2].

A number of H.E. Institutions use EPrints for their institutional data repository - Universities of Essex [6.1.4], Southampton [6.1.18] and West of England (UWE) [6.1.21]. The eCrystals repository [6.2.3], also at Southampton, runs on an EPrints platform. University of Leeds have chosen EPrints for their proposed data repostitory, as reported from the EPrints User Group workshop (Proudfoot, 2013a) at Leeds, October 2013.

4.2.2. CKAN http://ckan.org/

CKAN is an open source data management system developed by the OKF[46] to provide access to open data. Technologies used include PostgreSQL database engine [4.13.7], SOLR search [4.13.2], Python backend and Javascript frontend. It has a modular architecture with optional extensions - APIs surrounding a core system. CKAN is part of the RDMI implemented by Bristol [6.1.1] and Lincoln [6.1.12] and being trialled by the Kaptur project [evaluation at 5.6], and at Newcastle [Iridium CKAN use case 5.5]. The DART project[47] at Leeds uses CKAN for their data portal [6.2.2].

4.2.3. Fedora Commons http://www.fedora-commons.org/

(Flexible Extensible Digital Object Repository Architecture) - originally developed by Cornell University for managing digital content (DAMS). Fedora is RDBMS-independent and has been tested with MySQL, Oracle [4.13.6], PostgreSQL, Microsoft SQL and Derby (it is provided with Derby embedded). The Fedora Commons distribution includes Apache Tomcat, Derby SQL and Java components. Many workflow and service components and plug-ins have been developed to integrate Fedora within an RDM infrastructure. UK HEIs using the Fedora Commons platform include University of York Digital Library (YODL) and the Archaeological Data Service (ADS). Data repositories based on Fedora include 3TU Datacentrum [6.3.3], DANS Easy [6.3.4] and RUresearch [6.3.9].

[42] ReCollect Plugin http://bazaar.eprints.org/280/
[43] DataCite DOI Registration Plugin http://bazaar.eprints.org/307/
[44] RJ Broker for SWORD 2.0 http://bazaar.eprints.org/335/
[45] Arkivum A-Stor Storage Backend Plugin http://bazaar.eprints.org/313/
[46] Open Knowledge Foundation (OKF) http://okfn.org/
[47] DART project http://dartproject.info/WPBlog/

4.2.4. **Databank** (Fedora) http://www.dataflow.ox.ac.uk/index.php/databank

A data repository based on the Fedora Commons platform, designed by the Admiral project at the University of Oxford. [See section 4.1.1 for a description of DataFlow components and 5.5 for the Iridium evaluation of DataBank].

4.2.5. **Hydra** (Fedora) http://projecthydra.org/

Hydra is a multi-purpose repository framework based on a micro-services architecture. The main components are a Fedora repository platform, SOLR indexing software [4.13.2], Blacklight discovery interface [4.6.6] and the Hydra plugin, a 'Ruby on rails' library, which facilitates workflow in digital object management [as described in section 2.1.2]. Hydra is the platform for the University of Hull Digital Repository, Hydra [6.1.10], University of Virginia Libra [6.3.10] and the LSE digital Library[48].

4.2.6. **VITAL** (Fedora) http://www.vtls.com/products/vital

A Fedora based repository system developed through the Arrow project. This is used at Arrow, Monash University's research repository [6.3.1].

4.2.7. **Islandora** (Fedora) http://islandora.ca/about

Islandora is an open source Content Management System developed by the University of Prince Edward Island, built on a base of Fedora, Drupal [4.13.9] and Solr. This platform is used for the University of St Andrews Digital Collections Portal[49].

4.2.8. **DSpace** http://www.dspace.org/

DSpace is an open source repository system based on Apache server, PostgreSQL or Oracle and Perl. DSpace is the platform used for University of Edinburgh Datashare [6.1.3] and EDINA ShareGeo repository [6.2.1], Open Research Exeter [6.1.5], the University of Hertfordshire Research Archive (UHRA) [6.1.9], the Queen Mary University of London, Centre for Digital Music Research Data Repository (C4DM-RDR) [6.1.16] and DSpace at Cambridge [6.1.2].

4.2.9. **Datastar** https://sites.google.com/site/datastarsite/

Datastar is an open source repository system developed by Cornell University and Washington University. Designed to support collaboration and data sharing among researchers during the research process, and to promote publishing or archiving data and high-quality metadata to discipline-specific data centers, and/or to the institution's own digital repository.

4.2.10. **ContentDM** http://www.contentdm.org/

ContentDM is a proprietary repository software from OCLC, in use at University of Sheffield for digital collections at the Library Special Collections and National Fairground Archive [see 3.1].

4.2.11. **DigiTool** http://www.exlibrisgroup.com/category/DigiToolOverview

This is proprietary repository software from Exlibris, in use at University of Leeds for LUDOS.

4.2.12. **Equella** http://www.equella.com/

This is proprietary repository software in use for the institutional repository at Royal Holloway Research Online[50] and Oxford Brookes RADAR[51]. At Nottingham[52] it is integrated into Moodle and

[48] LSE digital Library http://digital.library.lse.ac.uk/
[49] University of St Andrews Digital Collections Portal https://arts.st-andrews.ac.uk/digitalhumanities/
[50] Royal Holloway Research Online (RHRO) http://digirep.rhul.ac.uk/

used to house and share digital teaching resources (except audio and video files, which are recommended to be uploaded to Kaltura). Equella is the platform used for the research data repository at Griffiths University [6.3.2] and was also being considered by Nottingham ADMIRe for the data repository / metadata store (Berry and Parsons, 2012a)[see 5.2].

4.2.13. Archimede http://www.bibl.ulaval.ca/archimede/index.en.html

An open source repository system designed at Laval University Library based on the DSpace model. The system was designed with internationalisation in mind, so it has an easily modified multilingual interface. The system is not platform dependent, and based on open source components – Java, Apache Ant, MySQL (recommended) and Lucerne.

4.2.14. ARNO https://www.h-net.org/announce/show.cgi?ID=127076

Repository software developed by the ARNO Project (Academic Research in the Netherlands Online): partners were the universities of Amsterdam, Twente and Tilburg. ARNO is based on Apache, Oracle and Perl architecture.

4.3. Digital preservation (repository) systems and services

4.3.1. Figshare http://figshare.com/

This is a cloud based, open access repository for research outputs. Data is persistently stored under CC license. Unlimited storage is offered for publicly accessible data, whereas private data is provided with 1Gb free storage. The service is supported by Digital Science[53] , the providers of Symplectics Elements [4.7.1] and Projects [4.5.4]. The company now offers 'Figshare for institutions', providing a cloud based data repository service. 'Figshare for Institutions' provides a 'project area' for collaborative working (includes collaborators from outside the institution). The cloud storage facility is provided by Amazon Web Services (S3 and Glacier), though Figshare can hold metadata only 'Stubb' records for data held elsewhere. Figshare is integrated with Symplectics [4.7.1], ORCID [4.12.4] and Altmetric [4.12.6], uses Shibboleth authentication [4.12.10] and mints DOIs from Datacite [4.12.1]. Figshare is OpenAIRE[54] compliant and a CLOCKSS[55] collaborator. This service is used by Imperial College London and University of Oxford in the UK.

4.3.2. Dataverse http://thedata.org/

Dataverse Network is an open source web application for publishing, citing, analysing and preserving research data, which may be installed by any institution. The architecture is based on PostgreSQL, Lucerne (SOLR) and Java. Rossetta is storage agnostic: any storage service may be integrated via the Storage Abstraction Layer, using plugins. The preservation workflow is configurable, and the metadata schema is highly customisable. The system architecture involves an Oracle database, uses Java APIs and a SOLR search engine. The preservation process incorporates the PRONOM global digital format registry[56] for file characterisation. This application supports the data repositories at Harvard University [6.3.6] and John Hopkins University [6.3.7].

---

[51] RADAR, Oxford Brookes University Resource Bank https://radar.brookes.ac.uk/radar/access/home.do
[52] University of Nottingham, Equella https://equella.nottingham.ac.uk/institutions.do
[53] Digital Science, Macmillan http://digital-science.com/
[54] OpenAIRE https://www.openaire.eu/
[55] CLOCKSS (Controlled LOCKSS) http://www.clockss.org/clockss/Home
[56] PRONOM global digital format registry http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

4.3.3. Rosetta http://www.exlibrisgroup.com/category/RosettaOverview

Rosetta is a proprietary digital preservation system from Exlibris, and the successor to DigiTool. The system is based on a distributed architecture which is scalable and flexible and provides continual preservation actions for long-term curation. The system is based on the OAIS[57] model and conforms to the TDR[58] requirements. The system integrates easily with Exlibris Primo for the discovery function.

4.3.8. Archivematica https://www.archivematica.org/

Archivematica is a free open source CMS / repository system for managing preservation of digital content, designed by Artefactual. The system has a micro-services architecture which provides a suite of tools for processing digital objects from ingest to access in compliance with the OAIS reference model. Archivematica uses METS, PREMIS and Dublin Core standards, uses the Bagit format [4.11.2] and is incorporates PRONOM. The system provides a 'dashboard' for public search and access, and for administrative / archiving functions. Files are stored in local or remote storage services, managed through multiple configurable pipelines. Once processed, the DIP (Dissemination Information Package) may be uploaded to a CMS for web access – AtoM [4.6.3] is the (bundled) default access system, but ContentDM [4.2.10] may also be used. Implementations include Columbia University Libraries, Harvard Business School, MoMA, and Yale University Library.

4.3.9. RODA http://www.roda-community.org

The Repository of Authentic Digital Objects is a free open source digital preservation system built in Java. This is based on the Fedora Commons platform [4.2.3] which provides the 'RODA Data Services' backend, above which is the 'RODA Core Services' layer, based on secure web services, which handles the ingest workflow, preservation planning, repository querying and administrative functions. On top of this lays the RODA web user interface (based on the Google web toolkit), providing access through download and embedded web viewers. The preservation actions (compliant with the OAIS model) and migration services are handled by a separate component. RODA is interoperable with other systems, such as AtoM. Used by the National Library of Wales, King's College London and University of Dundee.

4.3.10. Preservica http://preservica.com/

Rreservica is a proprietary digital preservation system developed by Tessella. The software as a service system is built from Java and provides flexible storage options – cloud storage using Amazon Glacier, S3 and RDS, local file servers or a hybrid system. The web interface includes embedded rendering / viewers. Interoperability with other systems through APIs (Java, REST and SOAP) and OAI-PMH [4.11.3], and integrates with Sharepoint [4.5.3] and discovery systems [4.6]. Preservation workflows are configurable, file formats are characterised by PRONOM and technical metadata extracted using JHOVE[59]. Customers include National Libraries of Australia, Estonia, Finland and Latvia; UK National Archives; Met Office; UK Parliament; Wellcome Library.

---

[57] Open Archival Information System (OAIS)
http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=57284

[58] Trusted Digital Repository (TDR)
http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf?urlm=161690

[59] JHOVE (JSTOR / Harvard Object Validation Environment http://jhove.sourceforge.net/

4.4. 'Archive Data' storage

4.4.1. Arkivum http://www.arkivum.com/

Arkivum provides a digital archiving service, certified to ISO27001. Three copies of the data are kept, two at geographically separate data centres and one at an escrow service. Data is uploaded from the institutional network using a local gateway appliance, the A-Stor (a file server), to the Arkivum data centre. This may be achieved within the institutional firewall. Entered a data archiving framework agreement with JANET[60].

4.4.2. Amazon Glacier http://aws.amazon.com/glacier/

A proprietary cloud storage and backup service, optimised for data that are infrequently accessed and for which short retrieval time is not critical, thus a low cost option for long-term data storage. Geographical location of data storage may be chosen to meet regulatory requirements.

4.4.3. Amazon S3 http://aws.amazon.com/s3/

The Amazon Simple Storage Service provides a web-services interface for secure cloud storage of unlimited volumes of data. This service offers fast retrieval compared to Amazon Glacier.

4.4.4. DuraCloud http://duracloud.org/

DuraSpace offers this commercial hosted service providing cloud infrastructure for data preservation and access. This is used for the ICPSR repository [6.3.11].

4.4.5. ARCserve http://www.arcserve.com/gb/default.aspx

A proprietary data backup service that offers a range of options including backup to cloud, disc or tape and high data availability with continuous data protection.

4.5. 'Active data' management and collaboration platforms

4.5.1. Datastage http://www.dataflow.ox.ac.uk/index.php/datastage

The active data management component of Dataflow, appears as a mapped drive on the researcher's computer (an 'Academic Dropbox') and provides metadata annotation and repository submission functions. Datastage is being tested by the Universities of Essex, Hertfordshire and QML Centre for Digital Music [See section 4.1.1 for a description of DataFlow components and 5.5 for the Iridium evaluation of DataStage].

4.5.2. Sakai CLE https://sakaiproject.org/

Sakai CLE provides a suite of resources for collaboration and project management. Resources include the means to store, organise and share files, facilities for blog, chat and managing forums, and a glossary providing contextual definitions. Sakai CLE provides the VLE Part of the Hydra infrastructure at Hull [6.1.10], VRE at Newcastle [6.1.14], Bath[61], Lancaster[62] and Monash [6.3.1]. The software has been evaluated by the Research360 [5.8] and Iridium [5.5] projects.

[60] JANET Joint Academic Network https://www.ja.net/
[61] iSusLab http://www.bath.ac.uk/csct/isuslab/
[62] Lancaster Centre for e-Science (LCeS) https://sakai.lancs.ac.uk/portal

4.5.3. Microsoft Sharepoint http://en.wikipedia.org/wiki/Microsoft_SharePoint

SharePoint is an established Web application platform introduced by Microsoft in 2001. The platform provides a range of Web tools, including intranet portals, document and file management, collaboration, social networks, extranets, websites, enterprise search, and business intelligence. Sharepoint is part of the infrastructure at the University of Southampton [evaluation at 5.4].

4.5.4. Digital Science Projects http://www.digital-science.com/products/projects

'Projects' is a research project management desktop application for Mac. It allows researchers to manage research activity, track changes to files, manage backup and restore previous versions of files, and to annotate and organise files and folders easily. This application integrates seamlessly with Figshare, though there is no institutional form to date.

4.5.5. D4Science http://www.d4science.eu/

D4Science is a European e-infrastructure project which provides a mechanism for of data e-infrastructure interoperability. The mechanism is based on the gCube software framework, which allows distributed virtual organisations to collaborate and share resources by managing the cloud / grid middleware thus configuring their own VREs.

4.5.6. Dropbox https://www.dropbox.com/

Dropbox is a commonly-used collaboration and cloud storage service, free to individuals (for volumes up to 2Gb) with added components for organisation subscription, such as file recovery, version tracking and phone support. Data is protected by 256-bit AES and SSL encryption and Two-step verification & mobile passcodes. Dropbox may store clients' data on servers in another country.

4.5.7. Google Drive http://www.google.co.uk/enterprise/apps/education/products.html

Also known as Google Documents, Google provides collaborative and cloud storage services for educational institutions, offering 30Gb storage per user and integration with its email service and text, voice and video chat service. Security features include two-step authentication and encrypted connection to servers. A vault service is offered for secure archiving of content. Google provides these services for the Universities of Sheffield and York.

4.5.8. Luminis http://www.ellucian.co.uk/Solutions/Ellucian-Luminis-Platform/

Luminis ia a collaboration portal platform, originally provided by SunGard, now Ellucian. Used for the VLE at the University of Leeds.

4.5.9. Microsoft Dynamics http://www.microsoft.com/en-gb/dynamics/

Dynamics is a collaboration platform for customer relationship management and enterprise resource planning. This will be used for the VRE at University of Leeds, replacing Luminis.

4.5.10. YouShare http://www.youshare.ac.uk/

A web-based portal for sharing data and software, developed at the University of York and funded by HEFCE. The portal allows the sharing of data and services in a secure online environment, the execution of analysis code and analysis of data, and the curation of data, analysis code and experimental protocols.

4.5.11. Alfresco http://www.alfresco.com/

Alfresco is an open source content management system, which interfaces with Google mail and drive (forthcoming) and the institutional filestore. This is being developed at the University of York

for use as a 'Research Lab management system' and at St. Andrews for data archiving involving a Fedora Commons repository (Allinson, 2013).

4.5.12. Amazon Web Services (AWS) http://aws.amazon.com/

AWS provides a wide range of cloud-based services for organisations, including: cloud computing and applications, storage (Glacier [4.4.2] and S3 [4.4.3]), databases, networking & virtual private cloud (VPC), analytics and deployment, identity and access management.

4.5.13. Huddle http://www.huddle.com/

Huddle is a collaboration platform that is designed for content sharing, document management, project and workflow management, secure intranet and extranet service. This is marketed as a 'Sharepoint alternative'.

4.5.14. Kaltura http://corp.kaltura.com/Video-Solutions/Education

Kaltura provides an open source video management platform with a focus on universities deploying videos within their organisation. This platform includes collaborative video editing and publishing components.

4.5.15. THREDDS Data Server (TDS) http://www.unidata.ucar.edu/software/thredds/current/tds/

THREDDS (Thematic Real-time Environmental Distributed Data Servcies) Data server provides catalogue, metadata and data access for scientific datasets. TDS is open source Java middleware, and is used for part of the 3TU Datacentrum infrastructure [6.3.3].

4.5.16. HUBzero http://hubzero.org/

HUBzero is an open source content management system designed for collaborative working and data sharing for scientific research and education. It is the platform for Purdue University Research Repository [6.3.8].

4.6. Catalogue software / Access platforms

4.6.1. DataFinder https://github.com/bhavanaananda/datafinder

Open source software, developed at the University of Oxford to provide a catalogue of research data. The metadata schema has been developed for full description of data, people responsible, how they were generated, access arrangements, links to publications etc. Datafinder integrates with Databank software and is designed to be used, with minimal modification, by other HEIs as part of their RDM infrastructure.

4.6.2. ReDBox (Fedora) http://www.redboxresearchdata.com.au/

Redbox has been designed as a metadata store / catalogue for research data. This provides workflows and interfaces for metadata creation. ReDBox is a research data registry, so the research data is assumed to be stored elsewhere, but data and related documentation / files may be uploaded to the system. ReDBox has been developed with, and is therefore closely integrated with Mint [4.12.8], a name authority and vocabulary system. Development was supported by the ANDS.

4.6.3. ICA AtoM https://www.ica-atom.org/

Designed by Artefactual , 'AtoM' (acronymn for 'Access to Memory') is a free web-based archival description software that is based on International Council on Archives ('ICA') standards. This system

allows organisations to create standards-based descriptions of their archival holdings and subsequently publish them to the World Wide Web.

4.6.4. XMC Cat http://d2i.indiana.edu/xmccat

This is a metadata catalogue storing rich metadata describing data objects stored in files, repositories or on the web. Metadata schemas are composed of concepts that describe data. In XMC Cat, the XML metadata schemas are partitioned into concepts, which act as the unit of metadata storage. This allows for a dynamically adaptable query interface.

4.6.5. OpenLink Virtuoso http://virtuoso.openlinksw.com/

A hybrid, multi-model data server architecture allows Virtuoso to offer Relational, XML and RDF data management, full text indexing, linked data, web application and document web server function and web service deployment (SOAP or REST).

4.6.6. Blacklight discovery interface http://projectblacklight.org/

Blacklight is an open source discovery interface for any SOLR index. Blacklight is a Ruby on Rails gem which accommodates heterogeneous data. This is part of the infrastructure for Hydra, the IR at the University of Hull [6.1.10].

4.6.7. ExLibris PRIMO http://www.exlibrisgroup.com/category/PrimoOverview

Primo is the discovery interface that offers a single search box for the whole range of a library's collections, be they locally managed or remote electronic content. This provides the discovery interface for the libraries of the Universities of Sheffield and York.

4.6.8. III Sierra http://sierra.iii.com/

The Sierra platform provides a suite of library services, including a resource Discovery interface. With similar functionality to Ex Libris Primo, this is the Resource Discovery interface used by the University of Leeds library.

4.6.9. Greenstone http://www.greenstone.org/

Open source multilingual digital library software, able to handle a wide variety of media formats.

4.7. Current Research Information Systems (CRIS)

4.7.1. Symplectic Elements http://www.symplectic.co.uk/product-tour/

Elements allows the Research Office to manage their researchers' published outputs by importing records from external sources such as WoS, Scopus, CrossRef and Figshare, and by allowing researchers to import details from Google Scholar, Mendeley, Endnote, Refman and Bibtex. Research information including HR, finance and grants administration data is managed and may be imported from legacy databases. Faculty information and academic profiles are managed and reported. Elements integrates with Eprints, Fedora and DSpace repositories through community developed plugins. Elements is the RIS in use at the Universities of Sheffield and Leeds.

4.7.2. PURE http://info.scival.com/pure

Pure provides comprehensive research information management. Pure aggregates data from awards management, HR, finance, student administration and other institutional sources. Publication data is retrieved from external sources such as Scopus, WoS, PubMed, Worldcat and Mendeley to populate Pure with information about researcher outputs. Integration with Dspace, ePrints, Fedora and

Equella supports automatic population of the institutional repository. Pure is the RIS in use at the Universities of York, Lancaster and Edinburgh.

4.7.3. Converis http://www.converis5.com/

Developed by Avedas, a Thomson-Reuters company, Converis is in use at Hull and integrated into the Hydra infrastructure. Converis appears to have the same functionality as Elements and Pure. This CRIS adheres to Research information standards – CERIF, CASRAI, VIVO and ORCID (see below).

4.8. Data management planning (DMP) tools

4.8.1. DMPonline https://dmponline.dcc.ac.uk/

The DMPonline tool has been developed by the DCC to help researchers create data management plans. The tool contains templates that represent the requirements of various funders and institutions. Guidance is provided during the process and the DMP may be exported in a variety of formats. The tool is used by the University of Lancaster.

4.8.2. DMPTool https://dmptool.org/

Released in 2011, this tool was designed by collaboration of number of U.S. universities, the DCC and DataONE.

4.9. Metadata Generators

4.9.1. Datacite metadata generator http://www.datacite.org/node/102

A single HTML form which can be used to generate DataCite Metadata Kernel 3.0 XML. Metadata is generated by populating text boxes and selecting values from drop-downs. The results can then be saved to a file. This tool has been created by Marcin Paluch and is available from https://github.com/mpaluch/datacite-metadata-generator.

4.9.2. Dublin Core Metadata Generator http://www.dublincoregenerator.com/

creates a Dublin Core standard metadata for your research data.

4.10. Data capture and workflow management systems

4.10.1. LIMS

Laboratory information management systems manage all aspects of laboratory processes, from data capture, sample management and instrument control to workflow, document and personnel management. **LIMSfinder** http://www.limsfinder.com/ provides information about the numerous LIMS available. An **Open Source LIMS** is available at http://www.institutelabauto.org/ProductList/informatics/PL-OS-LIMS.html.

4.10.2. Digital Lab books

ELN (Electronic laboratory notebooks) are computer applications designed to document experiments as an alternative to paper laboratory notebooks. Listings of open source and commercial ELNs may be found at **Atrium Research** http://www.atriumresearch.com/html/eln.htm ,
Laboratory Automation Products and Services (Open Source)
http://www.institutelabauto.org/ProductList/informatics/PL-OS-ELN.html (& Commercial)
http://www.institutelabauto.org/ProductList/informatics/PL-ELNs.html

Other examples include:

- **Open Atrium** is a team collaboration tool http://openatrium.com/
- **Laboratory Logbook** http://lablog.sourceforge.net/
- **Open Source ELN** http://sourceforge.net/projects/eln/
- **Indigo ELN**: The Open-Source Chemistry Electronic Lab Notebook http://ggasoftware.com/opensource/indigo/eln
- **MyLabbook** - Drupal based OS ELN http://mylabbook.org/
- **OpenWetWare Lab Notebook** http://openwetware.org/wiki/Lab_Notebook
- **Quartzy** http://www.quartzy.com/
- **LabAssistant** http://labassistant.en.softonic.com/mac
- **My Lab** http://www.mylab.fi/en/
- **Docollab (Sparklix)** https://www.docollab.com/#/
- **eCAT** http://www.researchspace.com/electronic-lab-notebook/index.html
- **E-Notebook for Chemistry (Wingu)** http://www.cambridgesoft.com/Ensemble_for_Chemistry/ENotebookforChemistry/Default.aspx
- **LabArchives** http://www.labarchives.com/ in partnership with BioMed Central, acts as the default storage system for supplementary data published with articles in BMC journals.

4.10.3. Bioconductor http://www.bioconductor.org/

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. More data management software resources for Biomolecular research are provided by **BBMRI** http://www.bbmri-wp4.eu/node/45 and **Biocompare** http://www.biocompare.com/Software/.

4.10.4. COLWIZ http://www.colwiz.com/

COLWIZ provides web, desktop and mobile apps to facilitate individual and collaborative research, saving precious time for researchers at every stage: from an initial idea, through collaboration, to publication of results.

4.10.5. gCube http://www.gcube-system.org/

A large software framework designed to abstract over a variety of technologies including data, process and resource management on top of Grid/Cloud enabled middleware.

4.10.6. Kepler https://kepler-project.org/

An open source java based application to help create, execute and share analyses and data to create scientific workflows.

4.10.7. Labguru http://www.labguru.com/

This is a project management system for Life Sciences. Facilitates project management and collaboration, linking research data, protocols, results, published papers and integrating external data.

4.10.8. Labtrove http://www.labtrove.org/

Labtrove is a data-centric digital infrastructure for supporting research. The software was developed at the University of Southampton as a result of experience gained through eScience research

projects such as CombeChem[63], eBank[64], eCrystals [6.2.3], R4L[65], Smart Tea[66] and oreChem[67]. Research infrastructure components – repository, LIMS, pervasive computing and RDF, are integrated into a blogging / social network paradigm. In Labtrove, the data is associated with the project metadata, at the point of, or prior to, creation. Therefore researchers can recreate and adapt experiments, using automated procedures and instrument settings. The system provides the necessary framework for good data management and curation.

4.10.9. Labview http://www.ni.com/labview/products/

Labview is a graphical development environment providing a range of tools for data acquisition, instrument control, data management and reporting.

4.10.10. MyGrid http://www.mygrid.org.uk/

A suite of tools designed to "help e-Scientists get on with science and get on with scientists". The tools support the creation of e-laboratories and have been used in domains as diverse as systems biology, social science, music, astronomy, multimedia and chemistry. Includes MyExperiment [4.10.11] and Taverna [4.10.17].

4.10.11. MyExperiment http://www.myexperiment.org/

This is a social networking site and Virtual Research Environment (VRE) designed for people to share, discover and reuse workflows and build communities. MyExperiment was developed using the MyGrid software suite [4.10.10] by a team from the universities of Southampton, Manchester and Oxford.

4.10.12. MyTea VLab http://mytea.org.uk/vlab

VLab enables the formation of a Smart Research Framework, helping the creation and preservation of the record. VLab extends the model of a digital infrastructure for supporting research (repositories, LIMS, pervasive computing & basic RDF underpinning) to incorporate the online blog paradigm, where a data centric system with control over visibility and sharing are essential.

4.10.13. MIEN http://neurosys.msu.montana.edu/applications/mien/

Model Interaction Environment for Neuroscience - a package of interface and library code intended to make a number of scientific modeling, data markup, and data storage tasks easier. Many of the extension functions of MIEN are devoted to neuroscience tasks, but the core MIEN package is a general purpose scientific modeling and data visualization tool with a flexible extension system.

4.10.14. Omero http://www.openmicroscopy.org/site/products/omero

OMERO manages images from the microscope to publication using a central repository. Data can be viewed, organized, analyzed and shared from anywhere via the internet, from a desktop app (Windows, Mac or Linux), from the web or from 3rd party software.

4.10.15. OBiBa http://www.obiba.org/

OBiBa provides open source software components for biobanks and biomolecular research.

---

[63] CombeChem http://www.combechem.org/
[64] eBank http://www.ukoln.ac.uk/projects/ebank-uk/
[65] R4L - Repository for the Laboratory http://r4l.eprints.org/
[66] Smart Tea Project http://www.smarttea.org/
[67] oreChem – Cyberinfrastructure for Chemistry Project http://www.openarchives.org/oreChem/

4.10.16. SCAPE http://www.scape-project.eu/

SCAPE is an open source infrastructure platform that executes institutional digital preservation strategies, for very large, complex and heterogenous collections of digital objects, by extending repository functionality with semi-automated workflows. The system integrates Fedora Commons, Taverna and Hadoop.

4.10.17. Taverna http://www.taverna.org.uk/

Taverna is an open source and domain-independent Workflow Management System – a suite of tools used to design and execute scientific workflows and aid *in silico* experimentation. The Taverna suite was written in Java by the MyGrid team [4.10.10] and includes the Taverna Engine (enacting workflows), Taverna Workbench (desktop client application) and Taverna Server (allows remote execution of workflows). Taverna has been widely deployed, particularly in Bioinformatics and Chemistry, is hosted by the University of Manchester and supported by JISC, EPSRC, BBSRC[68], ESRC[69] and FP7[70].

4.10.18. Yogo http://neurosys.msu.montana.edu/

The Yogo Data Management System is a set of software tools created to enhance the process of data annotation, analysis and web publication. The system provides a set of easy to use software tools for data sharing by the scientific community. It enables researchers to build their own custom designed data management systems. Another branch of the system provides tools for viewing anatomical and physiological data.

4.11. Data transfer protocols

4.11.1. SWORD http://swordapp.org/about/

Simple Web-service Offering Repository Deposit (SWORD) is a lightweight protocol for depositing content from one location to another. It is a profile of the Atom Publishing Protocol (APP) and designed to 'lower the barriers to deposit' any content into repositories.

4.11.2. BagIt http://tools.ietf.org/html/draft-kunze-bagit-10

The BagIt file packaging format is a hierarchical file packaging format for storage and transfer of digital content. A 'bag' is a structure to enclose a 'payload' and descriptive 'tags', and does not require knowledge of the payload's internal semantics. Also at: https://github.com/LibraryOfCongress/bagit-java

4.11.3. OAI-PMH http://www.openarchives.org/pmh/

The Open Archives Initiative Protocol for Metadata Harvesting is a low-barrier mechanism for repository interoperability. OAI-PMH is a set of six verbs or services that allow data providers to expose structured metadata and make them available for harvesting by service providers' requests.

[68] Biotechnology and Biological Sciences Research Council (BBSRC) http://www.bbsrc.ac.uk/
[69] Economic and Social Research Council (ESRC) http://www.esrc.ac.uk/
[70] Seventh Framework Programme http://cordis.europa.eu/fp7/home_en.html

4.12. Identifier services and identity components

4.12.1. DataCite http://www.datacite.org/

Datacite is an international organisation which supports research data archiving, access and citation by assigning persistent identifiers to datsets. An institution may join DataCite in order to have DOIs minted for its datasets.

4.12.2. DOI http://www.doi.org/

The Digital Object Identifier is an character string used to uniquely identify any object. The DOI provides an actionable (clickable), interoperable, persistent link to metadata about the object, including the URL where the object is located. The DOI for an object is permanent, whereas the location and other metadata may change, therefore the DOI may be used for persistent citation.

4.12.3. CERIF http://www.eurocris.org/Index.php?page=featuresCERIF&t=1

The Common European Research Information Format (CERIF) is a standard for managing and exchanging research information. It provides a data model that describes the research domain, defining research entities; researchers, projects, organisations, outputs and funding, and the relationships between these entities. CERIF has been developed by EuroCRIS[71].

4.12.4. ORCID http://orcid.org/

ORCID provides a persistent identifier for individual researchers, so that their identity is unambiguous. Automatic links to research outputs, publishing activities and grant applications are supported. ORCID is now integrated with Symplectic Elements and Figshare, (Hamnel, 2012).

4.12.5. CASRAI http://casrai.org/

The Consortia Advancing Standards in Research Administration Information are developing a common data dictionary and advance best practice for research information exchange and reuse.

4.12.6. Altmetric http://www.altmetric.com/

Altmetrics, supported by Digital Science (as is Figshare, Projects and Symplectics), collects article level metrics by tracking mention of schoralry articles. Three products are offered – 'Altmetrics explorer', for viewing, analysing and reporting metrics for articles in the database; 'Embeddable badges' which add metrics to a journal platform or application; Altmetric API, giving flexible metric display.

4.12.7. VIVO http://vivoweb.org/about

VIVO is an open source semantic web platform and ontology for representing researchers and their associated training, background, activities, organizations, and outputs including publications and research resources. VIVO has been developed and implemented at Cornell University in association with other projects including CASRAI, ORCID and EuroCRIS.

4.12.8. Mint http://www.redboxresearchdata.com.au/

Mint is an open source name authority and vocabulary system that provides services to web applications. Mint was developed with ReDBox on the Fascinator platform with support by the ANDS.

[71] EuroCRIS – European Organisation for International Research Information http://www.eurocris.org/

4.12.9. BRII Registry http://brii.medsci.ox.ac.uk/

The Building the Research Information Infrastructure project (BRII) at the University of Oxford, aimed at developing infrastructure, built on semantic web technologies, enabling efficient sharing of research information. The registry implemented at Oxford, integrated into the Fedora infrastructure, forms a part of the Oxford DAMS and as such, benefits from data preservation. This system is not yet available for other institutions.

4.12.10. Shibboleth https://shibboleth.net/

Shibboleth is a very widely deployed federated identity authentication system. It is an open source, free software system that provides single sign-on capabilities for individual access to protected online resources within and between organisations. Shibboleth is employed for user authentication at Sheffield, Leeds, York and many other HEIs.

4.13. Other software systems and platforms of interest

4.13.1. Globus Toolkit https://www.globus.org/toolkit

The Globus Toolkit is an open source set of software components enabling the sharing of services and resources across the 'grid'. The toolkit includes software for security, information infrastructure, resource and data management, monitoring and discovery. Services and resources may be shared across institutional and geographical boundaries whilst retaining local autonomy.

4.13.2. Apache SOLR https://lucene.apache.org/solr/

SOLR is an open source search platform developed by the Apache Lucerne project. Features include full-text search, hit highlighting, faceted search, near real-time indexing, database integration, rich document handling and geospatial search. SOLR is written in Java, has REST- like HTTP/XML and JSON APIs and runs as a standalone full-text search server.

4.13.3. MySQL http://www.mysql.com/

MySQL is the world's most popular, free open source relational database application. MySQL is the database component of EPrints and an optional database for Fedora commons.

4.13.4. SAP https://www.sap.com/uk/solution/industry/higher-education-research.html

SAP provides a suite of software tools for university management processes.

4.13.5. Agresso / pFACT http://www.unit4software.co.uk/products/agresso

Agresso is a range of Enterprise Resource Planning (ERP) software tools.

4.13.6. Oracle http://www.oracle.com/index.html

Oracle provides a range of database systems and Enterprise management resources.

4.13.7. PostgreSQL http://www.postgresql.org/

PostgreSQL is a free open source object-relational database system. PostgreSQL is one of the databases that may be incorporated in CKAN, Fedora Commons and DSpace.

4.13.8. Open Journals System https://pkp.sfu.ca/ojs/

A platform for hosting your own OA journals.

4.13.9. Drupal https://drupal.org/

Drupal is a free open source content management system, which may be used to provide a web-based user interface for many applications (such as catalogue databases).

4.13.10. Moodle https://moodle.org/

Moodle is a free open source Learning Management System / Virtual Learning Environment (VLE).

4.13.11. Hadoop http://hadoop.apache.org/

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.

5. Reviews, Evaluations and Comparisons of Infrastructure Elements

Several of the JISC research data management infrastructure projects published the results of reviews, evaluations and comparisons they carried out to select the infrastructure components and RDM tools they were to trial. A number of projects published the results of user requirements surveys, conducted to choose from the components and tools available. These are indicated below, where the findings may be briefly described.

5.1. CKAN for Research Data Management in an Academic Setting

A workshop was held on 18th February 2013, facilitated by the JISC MRD programme, to investigate the use of CKAN for RDM. The workshop featured presentations from Bristol and Lincoln and discussions fed into a user requirements gathering exercise. CKAN capabilities in fulfilling these requirements are expressed in the output of this work, which is available at: http://lncn.eu/mxz2 (Winn et al. 2013).

5.2. Admire [7.1.1]

The project issued document outlining considerations in developing a Research Data Management repository strategy, which included a review of repository software: http://admire.jiscinvolve.org/wp/files/2013/05/ADMIRe-RDM-Repository-Strategy-Requirements.pdf (Berry and Parsons, 2012b).
The EQUELLA digital repository system, in use as a DAMS at Nottingham, was piloted for use as a data repository. The evaluation of the data repository pilot was reported at: http://admire.jiscinvolve.org/wp/files/2013/05/ADMIRe-EQUELLA-Research-Data-Repository-Pilot.pdf (Berry and Parsons, 2012a). Key issues revealed include the need for manual validation of the metadata entered through a wizard, the workflow requirements of obtaining a DOI and storing non-open datasets.

5.3. Data.bris [7.1.4]

The CKAN data portal platform was investigated by the data.bris project to provide a public read-only catalogue of research data publications (data discovery) and also to manage controlled-access active data (collaborative sharing). The team gave a presentation on their evaluation of CKAN at the JISC 'CKAN for research data management in an academic setting' workshop - reported on the blog at: http://data.bris.ac.uk/2012/12/18/ckan-and-data-bris/ (Price, 2013a), and the slides are available at: http://data.bris.ac.uk/files/2013/02/databris-ckan.pdf (Price, 2013b). CKAN has now been adopted as part of the implemented infrastructure at Bristol [6.1.1].

5.4. Datapool [7.1.5]

The JISC Datapool project at the University of Southampton was concerned with modifying Microsoft Sharepoint to create data deposit interfaces, consisting of project and dataset forms that can collect metadata to feed through to the EPrints repository. The metadata profile of the EPrints institutional repository was extended using the ReCollect plugin, adapting the repository for research data. A report was published, describing the integration, available at: http://EPrints.soton.ac.uk/352813/3/EPrints-sharepoint-report-final10.pdf (Hitchcock and White 2013).

5.5. Iridium [7.1.7]

The Iridium project involved the assessment and testing of a number of systems:

- A categorisation and brief review of 67 RDM tools and infrastructure components was carried out and reported at: http://research.ncl.ac.uk/media/sites/researchwebsites/iridium/iridium_external_tools_assessment_17_5_2013_v1_PGR_LW.pdf (Iridium Support Team, 2012).
- An evaluation of DataStage and DataBank, reported at: http://iridiummrd.wordpress.com/2013/02/14/iridium-evaluation-of-datastage-and-databank-research-data-management-tools-from-dataflow-project/ (Wood, 2013).
- CKAN use case report at: http://research.ncl.ac.uk/media/sites/researchwebsites/iridium/iridium_CKAN_case_study_12_6_2013_v1_BA.pdf (Allen, 2012)
- Sakai integration into RDMI http://iridiummrd.wordpress.com/2011/11/22/research-data-management-at-euro-sakai-2011/ (Martin, 2011).

CKAN was adopted as the platform for the research data portal at the University of Newcastle. Information on the CKAN data portal is available at: http://research.ncl.ac.uk/rdm/tools/ckan/

5.6. Kaptur [7.1.8]

The Kaptur project carried out an evaluation of technical systems in May 2012, to judge their suitability for the management of visual arts research data. A set of user requirements was created with which to evaluate the technical system capabilities, based on: software type and cost, storage requirements, interface requirements, system requirements and institutional requirements. Seventeen software systems were chosen to evaluate and five were short-listed by their high scores: Dataflow, DSpace, EPrints, Fedora and Figshare. These were then measured against a more detailed set of requirements, and EPrints was deemed the most viable option, particularly since it was already in use at the partner institutions. However, Figshare and Dataflow were strong contenders and fulfilled some of the requirements that EPrints did not. Therefore two pilots were implemented, an integration of EPrints with Figshare and an integration of EPrints with Datastage. The findings were reported in the 'Kaptur Technical analysis report' at: http://www.research.ucreative.ac.uk/1239//1/Kaptur_technical_analysis.pdf (Garrett et al. 2012).

In November 2012, it was agreed that neither of the two pilots were viable and that an integration of EPrints and CKAN, not available for the earlier technical analysis, would be piloted. It was determined that the EPrints-CKAN instance, although integration was not fully possible at the time, was a stronger, sustainable model and worth continuing to develop in the future (Gramstadt, 2013).

5.7. Orbital [7.1.12]

The concept of a 'minimum viable product for RDM' was developed for the Orbital project, and its feature set considered to be authentication, storage, hosting/publishing, licensing, persistent URI and analytics. CKAN was chosen as the platform for the data repository, as it was found to meet these requirements 'out of the box', and for many other reasons as reported in the blog post at: https://orbital.blogs.lincoln.ac.uk/2012/09/06/choosing-ckan-for-research-data-management/ (Winn, 2012). An evaluation of the use of CKAN for RDM, presented at two conferences, is available at: http://eprints.lincoln.ac.uk/9778/ (Winn, 2013a).

5.8. Research360 [7.1.14]

The Research360 project at the University of Bath, carried out a survey of user requirements for a research data repository, published at: http://opus.bath.ac.uk/34082/ (Cope, 2013). Development of the technical infrastructure involved integration of ePrints and the HCP file storage system, described at: http://opus.bath.ac.uk/35532/3/Research360_EPrints_HCP_Report_FINAL.docx.pdf (Research360, 2013). Modification of Sakai to enable deposit of material into a SWORD2 compliant repository is described at: http://opus.bath.ac.uk/35540/3/Research360_Sakai_Development_Report_FINAL.pdf (Research360 Project, 2012).

5.9. Roadmap [7.1.15]

Research data repository functional requirements were compiled by the RoaDMap repository working group. The criteria were based on the Kaptur project review and enlarged to account for the local context. The draft repository functional requirements are available at: http://library.leeds.ac.uk/downloads/file/389/data_repository_platform_functional_requirements (RoaDMaP repository working group, 2013), and discussed in a blogpost at: http://blog.library.leeds.ac.uk/blog/roadmap/post/163 (Proudfoot, 2013c).

Initially it was considered expedient to build upon the existing EPrints infrastructure, although Dataflow offered a better fit with project needs, so was considered. Dataflow however, revealed technical issues (the link between DataStage and DataBank), so other platforms were considered. Using three case studies, the functional requirements were tested against the three main candidates for repository platform: EPrints, CKAN and DataFlow. EPrints was eventually chosen for a pilot service, given the short timescale given for EPSRC compliance (Proudfoot et al. 2013).

5.10. SMDMRD[72]

User requirements were gathered by questionnaire and interview, using DAF methodology. The main user requirements were:

- Seamless access, or command line access with batch import/export support
- easy to use web interface for searching published datasets
- advanced metadata-based search function
- customisable metadata and RDF support
- dataset version control
- multi-level access control
- linking data to published papers (DOI or handle.net)

In order to choose a platform for a prototype data management system, the project compared installations of Fedora Commons (using an Islabdora Drupal module), Dspace, DataVerse and DataFlow in fulfilling the following criteria:

- Meeting user requirements out of the box
- Ease of install, getting it running and maintenance
- Ease of customisation
- How many standards supported
- How well developed, supported and widely used

[72] Sustainable Management Of Digital Music Research Data
http://rdm.c4dm.eecs.qmul.ac.uk/category/project/smdmrd

The project team favoured DataFlow because of DataStage functions, but it was still under development. DSpace was found to be easiest to install and run, much online help being available. Queen Mary University also has a DSpace institutional repository already. The team found Fedora difficult to install and run, and Dataverse limited in its functionality, particularly metadata customisation. Eventually, DSpace was chosen for the pilot data management system, with the intention to combine it with DataStage to integrate researcher workflows, using the SWORD protocol to transfer datasets. The report on platform choice is available at: http://rdm.c4dm.eecs.qmul.ac.uk/platform_choice (Fabiani, 2012).

5.11. UWE Managing Research Data [7.1.9]

The project team chose EPrints for the research data repository at UWE because it is already in use for the Institutional repository, therefore no further funding was necessary and they had the skills necessary to repurpose the system. They were in communication and sharing knowledge with other institutions that had already used EPrints for data publication. These factors are discussed in the document available at: http://www2.uwe.ac.uk/services/library/using_the_library/Services%20for%20researchers/eprints-data-repository-uwe.pdf (Holliday, 2012).

5.12. Loughborough University UK HE Research Data Management Survey[73]

A survey of UK HEIs was conducted to determine their plans for future RDM services and received responses from 38 institutions. Regarding technical infrastructure components and tools, the results revealed that 6 (16%) institutions had an operational Research Data Service, with 25 (66%) developing one. Most institutions were storing or aimed to store both data and metadata, 2 planned to hold only metadata and 2 planned to hold just the data. Regarding the software system the service used or was intending to use:

- EPrints – 11 institutions
- DSpace – 4
- PURE – 4
- Symplectics – 2
- Converis – 1
- Figshare – 1
- iRODS – 1
- Other systems - 13 (included DataFlow, Fedora/Hydra, Equella and in-house developed)

A report on the survey results, and links to the survey results, are available at: http://blog.martinh.net/2013/10/metadata-is-love-note-to-future-uk.html (Hamilton, 2013).

5.13. St Andrews [6.1.17]

CKAN was investigated as the platform for a pilot RDM system at the University of St Andrews, as part of the JISC funded C4D project [7.1.2]. A list of user requirements was composed, with which to measure CKAN suitability, which contributed to the work done at the 'CKAN for Research Data Management Workshop' [5.1], as published in the blogpost at: https://research-computing.wp.st-andrews.ac.uk/2013/03/15/ckan-for-research-data-management/ (Plietzsch, 2013a).

[73] Loughborough University UK Higher Education Research Data Management (RDM) Survey http://blog.lboro.ac.uk/rdm/tag/uk-hei-rdm-survey/

CKAN was chosen for the pilot, evaluation process being described in the blogpost at: http://research-computing.wp.st-andrews.ac.uk/2013/11/27/using-ckan-for-research-data-management/ (Plietzsch, 2013b).

5.14. iREAD [7.3.1]

The iRODS evaluation and demonstrator project provided an evaluation and demonstration of the iRODS system, assessing the capabilities of a demonstrator system against use-case requirements from the CARMEN project. The evaluation is available at: http://www.wrg.york.ac.uk/iread

5.15. DANS Easy [6.3.4]

The process of deciding between Fedora, ePrints and DSpace, for the DANS Easy data repository service, is described in the paper at: http://www.ais.up.ac.za/digi/docs/bogaards_paper.pdf (Bogaards, 2009)

5.16. DCC

The DCC provide a catalogue of RDM tools and services at: http://www.dcc.ac.uk/resources/external/tools-services

5.17. ANDS

Provides information on technical resources at: http://www.ands.org.au/resource/techdocs.html ; and on metadata stores solutions at: http://ands.org.au/guides/metadata-stores-resources.html .

5.18. JISC Digital Media

Advice on various aspects of managing digital media collections may be found at: http://www.jiscdigitalmedia.ac.uk/managing

6. Active Institutional Infrastructure examples

Several institutions have now established institutional research data repositories, or host discipline-based or multi-institutional project-based research data repositories. The UK based institutional data repositories open to external view, are listed below. Some discipline-based data repositories based at UK HEIs are also listed, together with a number from institutions outside the UK.

6.1. UK institutional data repositories and catalogues

6.1.1. Bristol Data.bris http://data.bris.ac.uk/data/

The Open data repository is still under development. CKAN has been selected for the data repository and functions as a catalogue of research data. This integrates with the PURE RIS, which functions as a catalogue of research outputs (Price, 2013a).

6.1.2. DSpace at Cambridge https://www.repository.cam.ac.uk/

The institutional repository is now able to preserve and publish research data.

6.1.3. Edinburgh Datashare http://datashare.is.ed.ac.uk/

This repository is based on DSpace. The technical infrastructure at Edinburgh involves integration with PURE, active data infrastructure and the DMPonline tool.

6.1.4. Essex Research Data http://researchdata.essex.ac.uk/

This data repository is built on the EPrints platform, modified using the ReCollect plugin to accept datasets. The service includes allocation of Datacite DOIs.

6.1.5. Open Research Exeter https://ore.exeter.ac.uk/repository/

Based on the DSpace platform, material may be deposited via Symplectic. ORE's content includes journal articles, conference papers, working papers, reports, book chapters, videos, audio, images, multimedia research project outputs, raw data and analysed data. Exeter's three former repositories (The Exeter Research and Institutional Content Archive (ERIC), Digital Collections Online (DCO) and the Exeter Data Archive (EDA)) were merged into ORE and all previous content is still available via the same permanent link. The merger took place in March 2013.

6.1.6. GSoA RADAR http://radar.gsa.ac.uk/

This ePrints based Glasgow School of Art institutional repository accepts a wide range of objects including research data. This repository was the subject of a case study for the KAPTUR project.

6.1.7. Glyndwr University Research Data Catalogue

http://glynfo.glyndwr.ac.uk/course/view.php?id=41&section=11

A Data Catalogue facility of the bespoke CRIS.

6.1.8. Goldsmiths Research Data Catalogue http://eprints-data.gold.ac.uk/

Goldsmiths research data catalogue is built on the ePrints platform and results from the work done for the KAPTUR project.

6.1.9. Hertfordshire UH Research Archive http://rdm.herts.ac.uk/rdm/uh-research-archive.html

This is a DSpace institutional repository that is being expanded to include a data catalogue and a research data archive.

6.1.10. Hull Hydra https://hydra.hull.ac.uk/

The digital repository at Hull is built on the Hydra micro-services architecture [see 2.1.2 and 4.2.5]. The repository is designed to hold a wide range of digital resources including research datasets.

6.1.11. ICL http://spiral.imperial.ac.uk/

Researchers are advised that the preferred repository will be the discipline specific repository, failing that they should use the Imperial College repository: the DSpace based institutional repository, Spiral. Library becoming licensed to mint DOIs and will be developing a data catalogue.

6.1.12. University of Lincoln Researcher Dashboard https://orbital.lincoln.ac.uk/

The Researcher Dashboard is the interface for the Data deposit workflow, facilitated by the 'Orbital Bridge' application (Stainthorp, 2013). This links the various components of the RDMI: an EPrints IR for published research papers, network storage, Lincoln's Awards Management System and a CKAN based data registry (Stainthorp, 2012).

6.1.13. LSE http://eprints.lse.ac.uk/

Researchers are recommended to deposit data in 'LSE Research Online', the EPrints institutional repository (may now contain datasets). The LSE Digital Library http://digital.library.lse.ac.uk uses the Hydra platform for preservation and access of digital collections (using Fedora and SOLR but not the Blacklight discovery layer); interoperability between the Digital Library and the ePrints repository is planned.

6.1.14. University of Newcastle https://research.ncl.ac.uk/rdm/tools/

A Research Data infrastructure has been implemented at Newcastle which includes a CKAN data portal (for archiving and publishing data) together with a number of in-house built systems – a MyProject (a project and awards management system), MyImpact (a researcher profile and publication information system), a Research Data Catalogue (linking data, projects and publications), a VRE and e-Science Central (research collaboration tools).

6.1.15. Oxford ORA-Data https://databank.ora.ox.ac.uk/

Provided by the Bodleian Libraries, ORA-Data replaces Databank and Datafinder (but still Fedora based) as the archival store for digital data produced resulting from research by Oxford academics. ORA-Data is designed to hold records of datasets, irrespective of their location. If the actual data are stored elsewhere, a metadata record for datasets can be created and made available in ORA-Data and that includes a link to the location of the dataset. This is a component of the DataFlow infrastructure at Oxford, alongside DataStage which provides local management of active research data, including metadata annotation and a collaborative workflow. DataStage facilitates the deposit of data (with accompanying metadata) into ORA-Data and other data repositories. The RDM infrastructure also includes the Online Research Database Service (ORDS)[74] and the institutional repository, Oxford University Research Archive (ORA)[75].

[74] Online Research Database Service (ORDS) http://ords.ox.ac.uk/
[75] Oxford University Research Archive (ORA) http://ora.ox.ac.uk/

6.1.16. QMUL C4DM-RDR http://c4dm.eecs.qmul.ac.uk/rdr/

The Research Data Repository at Queen Mary University of London, Centre for Digital Music is a Dspace based repository [Discussion of the process of selection at 5.10]. This repository was specifically configured for long-term preservation and sharing of multimedia file formats.

6.1.17. St. Andrews https://risweb.st-andrews.ac.uk/portal/en/

The Pure CRIS holds records of Datasets and other research outputs and makes them accessible through the 'PURE Research Portal'. The DSpace Institutional Repository, 'Research@StAndrews:FullText' http://research-repository.st-andrews.ac.uk/, holds full text papers submitted through Pure.

6.1.18. ePrints Soton http://eprints.soton.ac.uk

The EPrints institutional repository at the University of Southampton has extended the existing the list of data types accepted to include datasets and experiments, using the ReCollect plugin. The EPrints Soton now holds research data underlying published research (papers) outputs. Another strand of work, using Sharepoint to catalogue and share active data, has yet to be implemented. The University of Southampton has a federated approach to repository management and so there are a number of instances of ePrints being used by departments to curate their research outputs.

6.1.19. University of the Arts London Data Repository http://www.researchdata.arts.ac.uk/

This repository for research data is built on an ePrints platform.

6.1.20. UCA Research Online http://www.research.ucreative.ac.uk/

UCARO is the institutional repository and accepts a wide range of research outputs including research data. This is built on the ePrints platform.

6.1.21. UWE Research Data Repository http://researchdata.uwe.ac.uk/

An instance of EPrints was modified for use as the data repository at UWE. The project developed its own metadata profile for research data, having decided against subscribing to the Datacite scheme and before the Recollect plugin became available.

6.1.22. Warwick http://wrap.warwick.ac.uk/

Researchers may register their datasets with the institution through WRAP (the EPrints based institutional repository) in the same way as they would give information about publications. Datasets may be deposited in WRAP.

6.2. Discipline-based research data repositories hosted by UK HEIs

6.2.1. Edina ShareGeo http://edina.ac.uk/projects/sharegeo/

Not an institutional repository, but based at The University of Edinburgh, here, DSpace has been customised to offer a repository that eases both the deposit and discovery of geospatial data.

6.2.2. Leeds DART Data Portal http://dartportal.leeds.ac.uk/

The Detection of Archaeological Residues using Remote-sensing Techniques (DART) research project maintains a CKAN data portal for the open data outputs from the project.

6.2.3. eCrystals at the University of Southampton http://ecrystals.chem.soton.ac.uk/

The University of Southampton department of Chemistry holds data from X-ray diffraction experiments in an ePrints repository. Each ePrint instance consists of Bibliographic data, data collection parameters and files; the files include raw data (.hkl), visualisations (.jpg), experimental conditions (.htm), structure determination outputs, final structural result (.cif and .cml) and a validation report.

6.2.4. UKDA http://www.data-archive.ac.uk/

Not an institutional, but a national social and economic research data repository based at the University of Essex. The UKDA provides the UK Data Service[76], which curates key quantitative and qualitative data, UK Data Service ReShare[77], curating data from ESRC funded research and the HDS[78] (successor to the AHDS). These are housed on a modified ePrints repository platform.

6.2.5. CARMEN Portal http://www.carmen.org.uk/portal

The CARMEN Portal is a VRE to support e-Neuroscience, providing storage and processing services over a Grid infrastructure. The CARMEN system is a three-tier web architecture consisting of a web portal, an application layer and a storage layer, developed by a collaboration of researchers from 11 UK universities. The Java portal allows the user to access data and to create and run analysis tool on remote servers. The storage layer is shared between MySQL databases and a SRB (Storage Resource Broker) system. The application layer consists of Java servlets, providing a middleware layer that bridges storage and portal.

6.3. Institutional and discipline-based research data repositories outside the UK

6.3.1. Monash University http://arrow.monash.edu.au/vital/access/manager/Index

Arrow, the research repository at Monash provides a place for researchers to store and manage research data and related publications. The university provides LaRDS (Large Research Data Store) for research datasets storage, which is used for collaboration using the Confluence wiki and Sakai VRE, and publishing data via the research repository Arrow. Monash also hosts a number of project based research data repositories. Research datasets are catalogued through the various current RDM platforms. This metadata may be harvested by the Research Data Australia (RDA) service, which provides a national research data catalogue. Monash does not have an institutional research metadata repository (catalogue) since this service is provided at the national level (Jones, 2013). The software system employed is 'VITAL'.

6.3.2. Griffith University http://equella.rcs.griffith.edu.au/research/logon.do

The Research Data Repository is based on Equella, and participates in the RDA catalogue. Some research data collections may be discovered using the Research Hub service at: http://research-hub.griffith.edu.au/collections.

6.3.3. 3TU Datacentrum http://datacentrum.3tu.nl/en/home/

3TU.Datacentrum, a collaboration of TU Delft, TU Eindhoven and University of Twente Libraries, provide a data repository, storing datasets from technical and scientific research in the Netherlands,

[76] UK Data Service http://www.ukdataservice.ac.uk/
[77] UK Data Service Reshare Repository http://reshare.ukdataservice.ac.uk/
[78] History Data Service (HDS) http://hds.essex.ac.uk/

and data processing services. Datacentrum is built on Fedora Commons and THREDDS [4.5.15] dataserver architecture.

6.3.4. DANS EASY https://easy.dans.knaw.nl/ui/home

Easy is the online archiving system provided by the Data Archiving and Networked Services (DANS), an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organisation for Scientific Research (NOW). The repository is built on Fedora Commons architecture.

6.3.5. California Digital Library Merritt Repository https://merritt.cdlib.org/

Merritt is built on a micro-services architecture providing digital curation through a series of devolved, independent but interoperable services. By devolving functions to a set of small self-contained services, they are easier to deploy, maintain and develop, leading to a flexible system able to respond to diverse needs and an ever changing technical environment. One of the central services is the Curation Storage micro-service, which supports a set of behaviors for manipulating and retrieving entities and their properties. Interaction with the Storage service is provided via a Java procedural API, a command line API, and a RESTful web API. The micro-services available are listed at: https://confluence.ucop.edu/display/Curation/Microservices.

6.3.6. Harvard Dataverse Network http://thedata.harvard.edu/dvn/

The Harvard Dataverse Network is a repository for sharing, citing and preserving research data; open to all scientific data from all disciplines worldwide. This is built on the Dataverse repository application and is part of the Dataverse Network.

6.3.7. Johns Hopkins Data Archive https://archive.data.jhu.edu/dvn/

The JHU Data Archive runs on the Dataverse repository software platform and is part of the Dataverse network.

6.3.8. Purdue University Research Repository https://purr.purdue.edu/

PURR provides an online, collaborative working space and data-sharing facility, based on the HUBzero platform.

6.3.9. Rutgers University Research Data Portal https://rucore.libraries.rutgers.edu/research/

RUresearch makes research data available to the scholarly community and provides a collaborative workspace for data processing and reuse. The system also provides access to supplementary resources, codebooks, lab books and publications to give context to the data. RUresearch is built on Fedora Commons architecture.

6.3.10. University of Virginia Libra http://libra.virginia.edu/

The University of Virginia institutional repository, Libra, is built on the Hydra micro-services platform and now accepts research datasets.

6.3.11. ICPSR http://www.icpsr.umich.edu/icpsrweb/ICPSR/index.jsp

The Inter-university Consortium for Political and Social Research provides a discipline-based data repository, located at the University of Michigan, Ann Arbor. This repository is built on the DuraCloud platform. ICPSR provides a range of other data curation tools and services.

A list of research data repositories, including discipline-based, national and institutional research data repositories can be found at **Databib**[79] http://databib.org/ and at **re3data.org**[80] http://www.re3data.org.

[79] Databib http://databib.org/
[80] re3data.org Registry of Research Data Repositories http://www.re3data.org

7. RDMI Project outputs

7.1. Outputs from the JISC RDMI 2011-2013 projects

7.1.1. ADMIRe http://admire.jiscinvolve.org/wp/

This JISC RDM Technical Infrastructure strand project involved creation of a blueprint for RDM infrastructure at the University of Nottingham. Initially a two layer RDMI service was envisaged – an active data management layer featuring collaborative tools and an archive layer featuring preservation and publishing. A revised model was proposed, the ADMIRe RDAS Research Data Archiving System (Sero Consulting, 2012), the focus of the development model being on opportunities afforded by current university infrastructure and systems. The RDAS was piloted using the Equella repository platform for a metadata store / data catalogue (Berry and Parsons, 2012a).

7.1.2. C4D http://cerif4datasets.wordpress.com/

The aim of the C4D project was developing a framework for incorporating metadata into CERIF such that research organisations and researchers can better discover and make use of existing and future research datasets, wherever they may be held. C4D built upon the IRIOS project work, which developed a platform for managing research information exchange using CERIF - supports the import of grant data (inputs) from Research Councils and publication data (outputs) from HEIs. The system allows inputs to be linked to outputs, which can then be exported in CERIF and used in other CERIF-compliant information systems (e.g. Pure, ePrints and, in the near future, RMAS).

7.1.3. DaMaRO http://damaro.oucs.ox.ac.uk/

This project, based at the University of Oxford,pulled together previous developments and elements to create a federated institutional infrastructure based upon a locally developed platform 'Dataflow'. WP6 was concerned with the development of software to facilitate the capture of metadata from existing research database systems 'DaaS', 'DataStage' and 'Colwiz'. The project also investigated automatic metadata capture from institutional and cloud-hosted storage, from 'LabTrove' and from research using 'Sharepoint'. WP7 was concerned with Data storage – developing an ingest service for SWORD-compliant datasets, integrating local data stores with 'DataBank', the Bodleian Libraries' archival standard research data repository; WP8 was concerned with data discovery and access, with the creation of 'Datafinder' a semantically aware data catalogue.

7.1.4. Data.bris http://data.bris.ac.uk/jisc-project/

The data.bris project developed a service providing a front end for the institutional research data storage facility (RDSF) and an institutional data repository. RDFS provides space for sharing data and curating data. Originally designed around security and restricted access, RDFS therefore needed modification to allow sharing and publication of data. Two types of access to research data are provided by data.bris: Research data publication, defined as read-only access to published data; Research-active data sharing, defined as read-only access to unpublished data (sharing), or read/write access (collaborative sharing). The CKAN data portal platform was adopted for the public point of access to published datasets (published dataset catalogue). Metadata is harvested from data.bris for the CKAN portal. The metadata store is a SPARQL 1.1 service. The new RIS (PURE) will serve as the Institutional repository and be fully integrated into the RDM infrastructure (Steer, 2012).

7.1.5. Datapool http://datapool.soton.ac.uk/

The Datapool project was carried out at the University of Southampton. A primary requirement was to find a mechanism for data upload and description, with a drop-box like sharing service, together with parallel deposit with external data services using SWORD protocol. Datpool developed services based on existing platforms at Southampton - ePrints and Sharepoint. Eprints provides a native interface for data capture and description, but Sharepoint does not, therefore these needed development. EPrints has been modified for research data using the ReCollect plug-in. The Sharepoint platform was considered appropriate as it is a versatile platform, providing multiple tools, and the university IT service has a long-term commitment to supporting it. Sharepoint provides a deposit interface facilitating metadata capture but does not provide storage for data – instead, it captures pointers to the data storage location (Hitchcock and White, 2013).

7.1.6. Research Data @ Essex http://www.data-archive.ac.uk/create-manage/projects/rd-essex

This project piloted the use of ePrints for a comprehensive research data repository. Resulting outputs were a metadata profile for describing research data and 'ReCollect', an ePrints plug-in to implement the profile. The metadata profile was developed from the 3 layer model produced by the IDMB project and mapped to the Datashare, DDI 2.1 and INSPIRE schemas. The pilot repository was tested with sample datasets from Essex departments and also ingested from the ESRC Data Store. The project highlighted gaps in the system: a lack of a facility for tagging multiple files with metadata; Limits to the size of a file that can be uploaded, so ePrints could hold metadata only records for large files that had to be stored elsewhere. Solutions to these problems included improvements to SWORD 2 provision for data transfer or using 'BitTorrent' like peer to peer sharing (Ensom, 2013).

7.1.7. Iridium http://research.ncl.ac.uk/iridium/

The Iridium project produced a policy and a pilot infrastructure for RDM at Newcastle University, to enable research data curation throughout the data lifecycle. For infrastructure development, Iridium undertook evaluation of many tools and infrastructure components including CKAN, Dataflow and SWORD [5.5]. The pilot infrastructure includes a public Research Data catalogue based on the CKAN platform, integrated with a bespoke internal facing research data registry, which is a component of the bespoke CRIS (MyImpact & MyProjects - all internal access) and the external facing ePrints Institutional Repository[6.1.14].

7.1.8. KAPTUR http://www.vads.ac.uk/kaptur/

The Kaptur project, led by the VADS[81] aimed to develop a model of best practice in the management of research data in the visual arts. The project investigated the nature of arts research data, and the diverse methods of managing the curation of data outputs. Amongst the project outputs was a technical analysis report (Garrett et al. 2012) [5.6]. The project has also led to the piloting of data repositories at Goldsmiths [6.1.8] and UAL [6.1.19], and the extension of the institutional repositories at GSoA [6.1.6] and UCA [6.1.20] to hold research data.

7.1.9. MRD at UWE http://www1.uwe.ac.uk/library/sc/mrd-project.aspx

The MRD project, which ran at University of the West of England (UWE), developed a pilot data management system for two research centres in the Faculty of Health and Life Sciences. The project aimed to develop processes and infrastructure which integrated with existing research infrastructure

[81] Visual Arts Data Service (VADS), University of the Creative Arts http://www.vads.ac.uk/

and which were in keeping with the current culture and practices at the university. The project outputs included a seven stage roadmap, development of online RDM guidance and training materials and the development of an RDM service and institutional policy. The project developed a metadata profile for research data and implemented an instance of ePrints, modified for data as a pilot repository for data.

7.1.10. MiSS http://www.miss.manchester.ac.uk/

The MiSS (MaDAM into Sustainable Service) project at the University of Manchester, aimed to deliver a RDM infrastructure by building on the experience of the MaDAM (Manchester Data Management) project. The project outputs included an institutional RDM policy which underpins the RDM service, that provides a point of contact for RDM enquiries, a website for RDM nformation and resources and support for data management planning (for which a local DMP tool was developed). A system allowing researchers to annotate, store, publish and preserve their data is under development.

7.1.11. Open Exeter http://as.exeter.ac.uk/library/resources/openaccess/openexeter/

This project, at the University of Exeter, investigated how researchers create, manage and use research data through case studies. This was followed on with the development of an advocacy and governance framework to embed RDM policy across the university. A third work strand, involved with development of technical infrastructure, resulted in the fully functioning research data repository.

7.1.12. Orbital http://orbital.blogs.lincoln.ac.uk/

The Orbital project was initiated to develop and implement a research data management infrastructure at the University of Lincoln. The technical infrastructure was piloted with the School of Engineering, taking into account the challenging requirements of engineering researchers. This has been extended to provide data-driven services across the whole university. The project developed institutional policy, which was approved, and a programme of support and training for research staff and students. The project has resulted in the implementation of the 'Researcher Dashboard', a single interface for the RDM infrastructure that links to projects, publications and data.

7.1.13. PIMMS http://proj.badc.rl.ac.uk/pimms

PIMMS (Portable Infrastructure for the Metafor Metadata System) developed tools to capture information about the workflow of running simulations, from the design of experiments to the implementation of experiments via simulations running models. PIMMS provides a local portal for research groups to view and search their own content, and publish their metadata content to institutional, national and international services. PIMMS also includes data node software so that data documented with PIMMS can also be published to the web.

7.1.14. Research360 http://www.ukoln.ac.uk/projects/research360/

This project aimed to develop RDM policy, training resources and pilot infrastructure at the University of Bath. A survey of researcher RDM practices was carried out, that established the need for development of data management policy and coordination, and problems with data storage and sharing needed addressing. A data management policy was developed, guidance provided through a data management website, and training resources and workshops designed. The project investigated development of existing infrastructure – integrating Sakai VRE with SWORD2 deposit protocol and

piloting ePrints for a research data repository. The project concluded that considerable resource will be needed to sustain the pilot RDM service and that further areas of RDM need to be investigated.

7.1.15. RoaDMaP http://library.leeds.ac.uk/roadmap-project

This project investigated RDM requirements using three case studies, from different subject disciplines and being at different stages of the award process (pre-award, live-award and post-award), and through a survey of researchers at the University of Leeds. The various workpackages covered all aspects of RDM. The outputs included the University of Leeds RDM Policy, development of the RDM website, DMPonline development, and the development of training resources. Technical infrastructure was investigated through a pilot vitualised storage system and the development of a set of functional requirements for a research data repository (Proudfoot et al. 2013).

7.1.16. RDTK http://research-data-toolkit.herts.ac.uk/

The project delivered a toolkit for researchers at the University of Hertfordshire, which provides advice on all aspects of research data management. The project also developed technology demonstrators for RDM infrastructure, including a cloud services (active data management) pilot, a document management pilot and a pilot data repository system. An instance of DSpace was installed, tested with data deposit via SWORD2 protocols, and tested for integration with PURE and DataStage. In due course, the DSpace institutional repository will be expanded to include data deposit.

7.1.17. SWORD-ARM http://archaeologydataservice.ac.uk/research/swordarm

The Archaeological Data Service (ADS) developed an automated deposit facility, by building upon the SWORD2 protocol and integrating other sources of research and project metadata. The resulting process helps and encourages researchers to deposit their data with the ADS.

7.2. Outputs from the JISC RDMI 2009-2011 projects

7.2.1. Admiral http://imageweb.zoo.ox.ac.uk/wiki/index.php/ADMIRAL

Admiral Project at the University of Oxford, created a two-tier federated data management infrastructure for life science researchers. The first tier provides locally managed data storage and staging facility (an ADMIRAL server, which became DataStage) to meet their local data management needs for the collection, digital organization, metadata annotation and controlled sharing of biological datasets. The second tier provides a data preservation and publication platform created and managed by the Bodleian Library Service as part of the Oxford Research Archive initiative (the Databank server), providing an easy and secure route for archiving annotated datasets to an institutional repository, The Oxford University Data Store, for long-term preservation and access, complete with assigned DOIs and Creative Commons open access licences. Two principles were involved (1) Sheer curation - making data management and deposit a seamless part of normal day-to-day research activity, irrespective of the instrument, software or OS used to capture data. (2) Curation by addition - researchers are not required to have everything in place all at once, or in any particular sequence. Data is accepted in any state and then allowed to be improved until ready for publication. Research is often an incremental process, and may change through time, so an incremental, unconstrained approach to data gathering was taken by this project.

7.2.2. FISHnet http://www.fishnetonline.org/home

This project involved investigating the data management practices of freshwater biology researchers and their concerns about data protection, access control and copyright. This information was used to develop a web-based data management and sharing tool, which was piloted and then fully implemented. The tool is built into the Freshwater*Life* website[82].

7.2.3. I2S2 http://www.ukoln.ac.uk/projects/I2S2/

Infrastructure for Integration in Structural Sciences (I2S2) aimed to identify requirements for a data-driven research infrastructure in Structural Science, with a focus on Chemistry. Two pilot were established to investigate the business processes of research and the benefits of an integrative approach, particularly to issues of scale (laboratory to national) and issues of boundaries between institutions.

7.2.4. IDMB http://www.southamptondata.org/

At the University of Southampton, the Institutional Data Management Blueprint Project, aimed to create a framework for RDM suitable for the whole institutional and that facilitates e-research practice. The project analysed requirements from a range of disciplines using a diverse range of data. The project developed an institutional RDM policy and investigated the use of ePrints and Sharepoint for RDM infrastructure. The project was followed by Datapool.

7.2.5. Incremental http://www.lib.cam.ac.uk/preservation/incremental/

Through a series of semi-structured interviews, the project team identified the current practices of researchers (PIs, post-docs and PhD students) at the universities of Glasgow and Cambridge. From these findings, services and resources were developed to help researchers manage their research data. Support services including RDM advice websites were developed for each institution and the institutional repository, DSpace at Cambridge [6.1.2], enhanced to accept research data.

7.2.6. MaDAM http://www.library.manchester.ac.uk/aboutus/projects/madam/

The Manchester Data Management project aimed to develop a pilot RDM infrastructure by focussing on the practices and requirements of biomedical researchers. The project had an iterative, user-driven, bottom-up developmental approach and aimed to produce a technical and governance solution flexible enough to meet the needs of all disciplines across the institution. The pilot infrastructure was successfully developed to manage data throughout the data curation lifecycle, from data capture to storage to publication and reuse. The project also developed DMP service for researchers at the university. MaDAM was followed by MiSS [7.1.10] to investigate the pilot infrastructure sustainability and rollout.

7.2.7. PEG-Board http://www.paleo.bris.ac.uk/projects/peg-board/

This project explored the data management needs of the Palaeoclimate research community, by means of DAF survey of members of the BRIDGE[83] research group, drawing up a number of use case scenarios. The project developed tools to facilitate data manipulation, publishing, discovery and reuse, climate data management policies and guidelines, and metadata schemas for palaeoclimate data.

---

[82] Freshwater*Life* website http://new.freshwaterlife.org/

[83] BRIDGE (Bristol Research Initiative for the Dynamic Global Environment) http://www.bristol.ac.uk/geography/research/bridge/

7.2.8. SUDAMIH http://sudamih.oucs.ox.ac.uk/

The Supporting Data Management for the Humanities project aimed to develop services for the management of data addressing the requirements of researchers from the Humanities at the University of Oxford. The project determined user requirements by means of DAF survey and developed pilot services to test. The project outputs included training materials and workshops, which take into account for the 'life's work' nature of Humanities research, and a pilot 'DaaS' (Database as a service) system, providing online databases designed for typical humanities datatypes. The pilot DaaS was further developed into a full service through the VIDaaS project.

7.2.9. VIDaaS http://vidaas.oucs.ox.ac.uk/

The VIDaaS (Virtual Infrastructure with Database as a Service) project developed the pilot DaaS service as part of the Online Research Database Service (ORDS) at the University of Oxford. ORDS allows researchers to create, edit, search and share relational databases, online through the university's private cloud environment. Although funded by another JISC programme[84], this project developed a service that is an integral part of the current RDM infrastructure at Oxford [6.1.15].

7.3. Outputs from other relevant projects

7.3.1. iREAD http://www.wrg.york.ac.uk/iread

The iRODS evaluation and demonstrator project provided an evaluation and demonstration of the iRODS system. The project, based at the University of York, implemented the iRODS system, deployed via the White Rose Grid (WRG). The capabilities of the demonstrator system were assessed against use-case requirements from the CARMEN project. The WRG hosts one of the nodes of the CARMEN e-science infrastructure, providing collaborative workspace and data archiving facilities. The project demonstrated the benefits of using the iRODS system in place of the CARMEN SRB system.

7.3.2. CARMEN http://www.carmen.org.uk/about

CARMEN is an e-Science pilot project funded by the EPSRC, which started in 2006. The CARMEN consortium initially involved 19 researchers from 11 universities (including Sheffield, York, Manchester and Newcastle). The project developed a virtual laboratory for neurophysiology enabling storage, sharing and processing of data (neural activity time and image series), analysis code and expertise in a distributed environment. The current system architecture [described above in section 6.2.5] utilises the UK e-Science Grid infrastructure, within which CARMEN Active Information Repository Nodes (CAIRNs) have been built, which are the functional units of the system.

[84] JISC UMF Shared Services and the Cloud programme
http://www.jisc.ac.uk/whatwedo/programmes/umf.aspx

8. Conclusions and Recommendations

The EPSRC expectations of organisations receiving EPSRC funding requires the research data created as a result of that funding, be effectively curated and securely preserved and for metadata describing this research data to be created and published by 1st May 2015. The University of Sheffield Data Management Policy was developed in response to the EPSRC expectations. This policy states that the university will provide support for research data management, including infrastructure and services to be developed in consultation with researchers.

In addition to surveying researchers to determine their RDM practices and attitudes, it is appropriate to develop these services to fit seamlessly in with the researcher workflow and not burden the researcher with significant change to their work practices. Rather than a researcher filling out forms, metadata may be exchanged between systems (CRIS, VRE and repository for example) thus reducing re-keying. Products, processes and practices that have been developed by a research community should be adopted, adapted and developed for the needs of other researchers, rather than entirely new solutions developed. As indicated by Jones et al. (2013: 14):

> ***"…close engagement with researchers is critical when designing RDM systems to ensure their applicability and uptake."***

8.1. Infrastructure components and implementation strategy

In choosing the components of the technical infrastructure and strategy for instituting it, the following options must be considered:

- A 'Big Bang' implementation or an incremental roll-out, allowing the service to be adopted slowly.
- A system wide generic infrastructure or a 'Bottom-up' project based approach, pilot infrastructure components being tested possibly throughout the whole lifecycle of a research project.
- Utilising existing infrastructure upon which new services are developed, or implementing new infrastructure. Integrating existing components may require investment in terms of development work, but implementing new infrastructure will also be costly.
- Choosing components where there is local experience or choosing components where there is as yet, little community of practice.
- Choosing open source components, which require local expertise and development, or proprietary components, supported but expensive.

Perhaps the greatest factor in considering these options is that some RDM services, data catalogue and data archiving, need to be implemented by May 2015. Therefore, in providing a 'bare compliance' option, it may be expedient to utilise the facilities that are already in place, if they provide the necessary functions with little modification or development to achieve integration.

8.2. Published research data catalogue or repository

To some extent, choice of the infrastructure component providing the public research data catalogue is contingent on the development of the WRRO as a catalogue of all published research outputs, research data as well as research papers. A White Rose Research Data Catalogue, perhaps implemented as a separate instance of ePrints as with WREO, would need to contain only catalogue records, not the datasets themselves, which would be held in a research data archive. The institutional ownership of research data will likely require the local control, if not location of the preservation and storage functions in the research data archive.

EPrints is already integrated with Symplectic at Sheffield and Leeds, and with PURE at York, so dataset metadata may be automatically imported into an ePrints research data catalogue. EPrints would need to be modified to handle a research dataset metadata profile with the ReCollect plugin. This dataset metadata will be automatically recorded by the Symplectics' system, some will be harvested from external repositories (now harvests from Figshare), and researchers manually input all other necessary metadata fields. At Open Research Exeter, research data is deposited via Symplectic into the DSpace repository.

If the decision is made not to go ahead with a shared WR research data catalogue, then a local based system may be implemented that combines the catalogue and archive functions together as an institutional research data repository. A local instance of ePrints could be considered, as there is much local experience in the use of ePrints, Symplectics and the connector, and a willingness to share knowledge within the RDM practitioner community. Alternatively, other repository and cataloguing systems should be considered. The open source systems Dspace, Fedora Commons, Datafinder, Hydra and CKAN have a growing community of users, willing to share expertise. A number of proprietary systems, such as ContentDM for which there is local expertise, must also be considered.

8.3. Research data archive

A facility for preservation of research data that has not been submitted to an external repository, needs to be provided by the institution to comply with EPSRC expectations. Such a data archive may offer a preservation service for unpublished research data also. Repository systems, that have been designed for or modified for research data, provide the archival storage function and the catalogue function. Alternatively, specialist long-term archiving and preservation systems exist. Possible candidates here include Rosetta (as the library has experience with ExLibris systems), Figshare for Institutions (as it is supported by the providers of, and integrated with Symplectic), Arkivum (as it is involved in the JANET data archiving framework agreement) and Dataverse (one of the open source data preservation platforms available).

8.4. Active data management

Currently at Sheffield, collaborative functionality is provided by Google Drive and the HPC facilities, but there is no institutional Virtual Research Environment (VRE) as such. A number of JISC infrastructure projects investigated the use of collaboration tools such as DataStage, Sharepoint and

Sakai, to provide a VRE which is integrated with the CRIS, file servers, data archive and / or a data repository. Surveys of researchers have shown a requirement for 'Academic Dropbox' facilities, which allow the sharing of data and for 'Social network' style annotation (Garrett et al. 2012).

For RDM, one function required of the VRE is that of data registry, defined here as an inward-facing data catalogue. This is built into DataStage (at Oxford) and Sharepoint (at Southampton), though a number of institutions incorporate a separate data registry component in their active data management facility, for example the use of CKAN at Lincoln and PURE at Bristol. Attention perhaps should be paid to YouShare developed at the University of York. The capabilities of the WRG infrastructure for collaborative active data management should also be investigated.

8.5. Data and metadata capture

There is currently a lack of information about the data and metadata capture tools being used and developed at the University of Sheffield, although systems such as laboratory information management systems may be used by some research groups at the institution. Experience in the use of such tools needs investigation to inform the choice of and development of an active data management infrastructure capable of integrating these tools.

8.6. Final remarks

The great benefits of a shared approach, in terms of saving money and time, should mean that engaging in collaborative efforts to establish shared services is a priority concern. Opportunities to collaborate in the development of a WR Research Data Catalogue and the proposed N8 shared data archiving service must be exploited. The development of RDM services delivered through the White Rose Grid and N8 HPC grid infrastructure need to be explored. Attention should be paid to the national research data service being piloted by the DCC and JISC.

With consideration of the time constraint of compliance to the EPSRC expectations, it may be appropriate to pilot a range of RDM services, implementing different components of the RDM technical infrastructure, with a number of EPSRC research projects. If the pilot component proves sustainable, then it will contribute to the incremental roll-out of a fully integrated RDM service, whilst fulfilling the requirements of the EPSRC expectations.

9. References

9.1. Works cited in the text [all URLs accessed on 29/04/14 unless stated]

Abrams, S., Cruse, P. and Kunze, J. (2009) Preservation is not a place. *International Journal of Digital Curation*, *4*(1) pp. 8-21. http://www.ijdc.net/index.php/ijdc/article/view/98

Allen, B. (2012) *Iridium project CKAN use case*. Newcastle: Iridium Project, University of Newcastle. http://research.ncl.ac.uk/media/sites/researchwebsites/iridium/iridium_CKAN_case_study_12_6_2013_v1_BA.pdf

Allinson, J. (2013) *Infrastructure and Systems Audit: Report to the White Rose Libraries Systems Architecture Group*. York: University of York Library.

ANDS (2011) *ANDS and Data Storage*. ANDS. http://ands.org.au/guides/storage.pdf

ANDS (2011) *Metadata Stores Solutions Guide*. ANDS. http://ands.org.au/guides/metadata-stores-solutions.pdf

Awre, C. (2012) Hydra UK: Flexible Repository Solutions to Meet Varied Needs. *Ariadne* Issue 70 http://www.ariadne.ac.uk/issue70/hydra-2012-11-rpt

Berry, M. and Parsons, T. (2012a) *EQUELLA data repository pilot*. Nottingham: ADMIRe Project, University of Nottingham Information Services. http://admire.jiscinvolve.org/wp/files/2013/05/ADMIRe-EQUELLA-Research-Data-Repository-Pilot.pdf

Berry, M. and Parsons, T. (2012b) *Research Data Management Strategy Requirements: Support for creating, archiving and sharing research data*. Nottingham: ADMIRe Project, University of Nottingham Information Services. http://admire.jiscinvolve.org/wp/files/2013/05/ADMIRe-RDM-Repository-Strategy-Requirements.pdf

Bogaards, L. (2009) Easy on Fedora: Building on the latest generation repository infrastructure. In: *The African Digital Scholarship and Curation Conference 2009*. 12-14 May 2009 CSIR Conference Centre, Pretoria, South Africa. Pretoria: University of Pretoria. http://www.ais.up.ac.za/digi/docs/bogaards_paper.pdf

Brown, M., Parchment, O. and White, W. (2011) *Institutional data management blueprint*. Southampton: IDMB Project, University of Southampton. http://eprints.soton.ac.uk/196241/

Cope, J. (2013) *Institutional Data Repository User Stories*. Bath: University of Bath. http://opus.bath.ac.uk/34082/

Datacite (2011) *DataCite Metadata Schema for the Publication and Citation of Research Data.* Datacite http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf

EDINA and Data Library, University of Edinburgh (2012) *Research Data Mantra*. University of Edinburgh. http://datalib.edina.ac.uk/mantra

Ensom,T. (2013) *Research Data Essex repository development and ingest*. Colchester: UKDA, University of Essex. http://www.data-archive.ac.uk/media/402401/rde_repositorydevelopmentingestreport.pdf

Ensom, T. and Corti, L. (2012) *Research Data Management at the University of Essex*. Colchester: UKDA, University of Essex. http://www.data-archive.ac.uk/media/375377/rde_researchdatamanagementassessment.pdf

Ensom, T. and Wolton, A. (2012) *Opening up research data at Essex: experiments with EPrints*. Colchester: UKDA, University of Essex. http://www.data-archive.ac.uk/media/368772/rde_or2012_notes.pdf

EPSRC (2013) *EPSRC expectations of organisations in receipt of EPSRC research funding*. Swindon: EPSRC http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx

Fabiani, M. (2012) And the winning platform is… *Research Data Management at the Centre for Digital Music Blog*, 5th January. London: Queen Mary University of London, Centre for Digital Music. http://rdm.c4dm.eecs.qmul.ac.uk/platform_choice

Foster, I., Kesselman, C. and Tuecke, S. (2003) The Anatomy of the Grid: Enabling Scalable Virtual Organizations In: Berman, F., Fox, G. and Hey, A.J.G. (eds) *Grid Computing: Making the Global Infrastructure a Reality.* New York : J. Wiley. http://toolkit.globus.org/alliance/publications/papers/anatomy.pdf

Garrett, L., Silva, C. and Gramstadt, M. (2012) *Kaptur Technical analysis report*. London: Visual Arts Data Service (VADS), University for the Creative Arts. http://www.vads.ac.uk/kaptur/outputs/Kaptur_technical_analysis.pdf

Golding, D. (2014) *Towards a reference architecture for Research Data Management*. N8 RDM Working Group. University of Leeds [Unpublished]

Gramstadt, M-T. (2013) *KAPTUR Final Report*. London: Visual Arts Data Service (VADS), University for the Creative Arts. http://www.vads.ac.uk/kaptur/outputs/KAPTUR_final_report.pdf

Gutteridge, Christopher (2010) Using the Institutional Repository to publish research data. In: *Open Knowledge Conference (OKCon) 2010, London, UK.* http://eprints.soton.ac.uk/270885/

Hamilton, M. (2013) Metadata is a Love Note to the Future - UK Higher Education Research Data Management (RDM) Survey. *Martin Hamilton's blog [Digital Futures]*, 9th October. Loughborough: Loughborough University. http://blog.martinh.net/2013/10/metadata-is-love-note-to-future-uk.html

Hamnel, M. (2012) Figshare and Symplectic – ORCID launch partners. *Digital Science Blog* 16th October. Digital Science. http://www.digital-science.com/blog/posts/figshare-and-symplectic-orcid-launch-partners

Haywood, J. (2013) Caretakers of the Present, Guardians of the Future: The digital Research Data Challenge. In: *RDMF: Funding Research Data Management,* Aston 25th April 2013. DCC. http://www.dcc.ac.uk/events/research-data-management-forum-rdmf/rdmf-special-event-funding-research-data-management#sthash.WoDs86Oy.dpuf

Hitchcock, S. (2012) To architect or engineer research data repositories. *JISC Datapool project Blog*, 17th December 2012. Southampton: University of Southampton. http://datapool.soton.ac.uk/2012/12/17/to-architect-or-engineer-research-data-repositories/

Hitchcock, S. and White, W. (2013) *Towards research data cataloguing at Southampton using Microsoft SharePoint and EPrints: a progress report*. Southampton: University of Southampton. http://eprints.soton.ac.uk/352813/

Hodson, S. (2012) Institutional Data Repositories and the Curation Hierarchy: reflections on the DCC-ICPSR workshop at OR2012 and the Royal Society's Science as an Open Enterprise report. *JISC MRD Blog,* 6th August 2012. London: JISC. http://researchdata.jiscinvolve.org/wp/2012/08/06/institutional-data-repositories-and-the-curation-hierarchy-reflections-on-the-dcc-icpsr-workshop-at-or2012-and-the-royal-societys-science-as-an-open-enterprise-report/

Holliday, L. (2012*) EPrints as a data repository at UWE*. Bristol: Managing Research Data Project, University of the West of England. http://www2.uwe.ac.uk/services/library/using_the_library/Services%20for%20researchers/eprints-data-repository-uwe.pdf

Iridium Support Team (2012) *Iridium external RDM tools assessment*. Newcastle: Iridium Project, University of Newcastle. http://research.ncl.ac.uk/media/sites/researchwebsites/iridium/iridium_external_tools_assessment_17_5_2013_v1_PGR_LW.pdf

Jackson, N. (2012) A Bridge to the Skies. *Orbital Blog*, 4th September. Lincoln: Orbital Project, University of Lincoln. http://orbital.blogs.lincoln.ac.uk/2012/09/04/a-bridge-to-the-skies/

Jones, S. (2013). *Bringing it all together: a case study on the improvement of research data management at Monash University*. DCC RDM Services case studies. Edinburgh: Digital Curation Centre. http://www.dcc.ac.uk/sites/default/files/documents/publications/case-studies/Monash-case-study.pdf

Jones, S., Pryor, G. & Whyte, A. (2013) *How to Develop Research Data Management Services - a guide for HEIs.* DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides/how-develop-rdm-services

Jones, S., Pryor, G. & Whyte, A. (2013) Components of research data management support services [Diagram]. In: Jones, S., Pryor, G. & Whyte, A. *How to Develop Research Data Management Services - a guide for HEIs*. (p 8). DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/sites/default/files/documents/publications/RDMcomponents.PNG

Kay, D. and Stevens, O. (2012) *White Rose Repositories Service Review*. Sheffield: Sero Consulting Ltd.

Lewis, S. (2013) The four quadrants of research data curation. *Edinburgh Research Data Blog,* 6th December. Edinburgh: University of Edinburgh Research Data Management Action Group. http://datablog.is.ed.ac.uk/2013/12/06/the-four-quadrants-of-research-data-curation-systems/ [Accessed 13/10/14].

Martin, A. (2011) Research Data Management at Euro Sakai 2011. *Iridium Blog*, 22nd November. Newcastle: Iridium Project, University of Newcastle. http://iridiummrd.wordpress.com/2011/11/22/research-data-management-at-euro-sakai-2011/

Parsons, T. and Berry, M. (2012) *Research Data Management Technical Requirements*. University of Nottingham & JISC ADMIRe Project. http://admire.jiscinvolve.org/wp/files/2013/05/ADMIRe-RDM-Technical-Requirements-Report.pdf

Plietzsch, B. (2013a) CKAN for Research Data Management. *Research Computing Blog*, 15th March. St Andrews: University of St Andrews. https://research-computing.wp.st-andrews.ac.uk/2013/03/15/ckan-for-research-data-management/

Plietzsch, B. (2013b) Using CKAN for Research Data Management. *Research Computing Blog*, 27th November. St Andrews: University of St Andrews. http://research-computing.wp.st-andrews.ac.uk/2013/11/27/using-ckan-for-research-data-management/

Price, S. (2013a) data.bris at CKAN for Research Data Management workshop. *Data.bris Blog,* 19th February. Bristol: data.bris project, University of Bristol. http://data.bris.ac.uk/2013/02/19/data-bris-at-ckan-workshop/

Price, S. (2013b) data.bris Use case, role and functionality for CKAN adoption. In: *CKAN for Research Data Management in an Academic Setting*. JISC workshop. 18th February 2013, HEFCE Offices, London. Bristol: data.bris project, University of Bristol. http://data.bris.ac.uk/files/2013/02/databris-ckan.pdf

Proudfoot, R. (2013a) EPrints and Research Data, 15th October. *RoaDMaP Blog,* 23rd October. Leeds: RoaDMaP Project, University of Leeds. http://blog.library.leeds.ac.uk/blog/roadmap/post/184

Proudfoot, R. (2013b*) RDM at the University of Leeds: update report for White Rose Service Development Group*. Leeds: University of Leeds.

Proudfoot, R. (2013c) Research data repository requirements. *RoaDMaP Blog*, 5th June. Leeds: University of Leeds. http://blog.library.leeds.ac.uk/blog/roadmap/post/163

Proudfoot, R., Phillips, B., Banks, T. and Blyth, G. (2013) *RoaDMaP Final Repot (draft)*. Leeds: RoaDMaP Project, University of Leeds. http://library.leeds.ac.uk/downloads/file/536/roadmap_final_report

Rans, J. and Jones, S. (2013) *RDM strategy: moving from plans to action*. DCC RDM Services case studies. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/developing-rdm-services

Rans, J., Lyon, L. and Duke, M. (2013) *White Rose Consortium Shared Research Data Management Services: A Feasibility Analysis*. Bath: DCC, University of Bath.

Research360 (2012) *Sakai-SWORD2 Integration Development Report*. Bath: Research360 Project, The University of Bath.
http://opus.bath.ac.uk/35540/3/Research360_Sakai_Development_Report_FINAL.pdf

Research360 (2013) *EPrints Integration with the Hitachi Content Platform*. Bath: Research360 Project, University of Bath.
http://opus.bath.ac.uk/35532/3/Research360_EPrints_HCP_Report_FINAL.docx.pdf

RoaDMaP repository working group (2013) *Data repository platform functional requirements*. Leeds: RoaDMaP Project, University of Leeds.
http://library.leeds.ac.uk/downloads/file/389/data_repository_platform_functional_requirements

Robertson, R.J., Mahey, M. and Allison, J. (2008) *An ecological approach to repository and service interactions*. Bath: UKOLN.
http://www.ukoln.ac.uk/repositories/digirep/images/a/a5/Introductoryecology.pdf

The Royal Society (2012) *Science as an open enterprise: The Royal Society Science Policy Centre report*. London: The Royal Society. http://royalsociety.org/policy/projects/science-public-enterprise/report/

The Royal Society (2012) The Data Pyramid – a hierarchy of rising value and permanence [Diagram]. In: The Royal Society (2012) *Science as an open enterprise: The Royal Society Science Policy Centre report*. (p. 60). London: The Royal Society.
http://researchdata.jiscinvolve.org/wp/files/2012/08/Data-Pyramid2.jpg

Sero Consulting (2012) *University of Nottingham Research Data Archiving System: Process model and use case analysis.* Nottingham: JISC ADMIRe Project, University of Nottingham.
http://admire.jiscinvolve.org/wp/files/2013/05/ADMIRe-RDM-Process-Model-and-Use-Case-Analysis.pdf

Stainthorp, P. (2012) Orbital deposit of dataset records to the Lincoln Repository: workflow. *Orbital Blog,* 6th December. Lincoln: Orbital Project, University of Lincoln.
http://orbital.blogs.lincoln.ac.uk/2012/12/06/orbital-deposit-of-dataset-records-to-the-lincoln-repository-workflow/

Stainthorp, P. (2013) Throw down the SWORD. *Orbital Blog,* 7th May. Lincoln: Orbital Project, University of Lincoln. http://orbital.blogs.lincoln.ac.uk/2013/05/07/throw-down-the-sword/

Steer, D. (2012) data.bris architecture. *Data.bris Blog* 3rd February. Bristol: data.bris project, University of Bristol. http://data.bris.ac.uk/2012/02/03/data-bris-architecture/

The University of Sheffield, Research and Innovation Services (2014) *Research Data Management Policy.* Sheffield: The University of Sheffield. http://www.shef.ac.uk/ris/other/gov-ethics/grippolicy/practices/all/rdmpolicy

UK Data Archive (2014) *Documenting your data.* UKDA. http://www.data-archive.ac.uk/create-manage/document

White Rose Libraries Systems Architecture Group (2013) *WR Systems Matrix*. (unpublished).

White Rose Research Data Working Group (2013) *Research Data Repository – Options paper*. (unpublished).

Winn, J. (2012) Choosing CKAN for research data management. *Orbital Blog*, 6th September. Lincoln: Orbital Project, University of Lincoln. https://orbital.blogs.lincoln.ac.uk/2012/09/06/choosing-ckan-for-research-data-management/

Winn, J. (2013a) Open data and the academy: an evaluation of CKAN for research data management. In: *IASSIST 2013*, 28-31 May 2013, Cologne. http://eprints.lincoln.ac.uk/9778/

Winn, J. (2013b) The Researcher Dashboard. *Orbital Blog*, 3rd May. Lincoln: Orbital Project, University of Lincoln. http://orbital.blogs.lincoln.ac.uk/2012/09/04/a-bridge-to-the-skies/

Winn, J. et al. (2013) CKAN RDM Requirements. In: *CKAN for Research Data Management in an Academic Setting*. JISC workshop. 18th February 2013, HEFCE Offices, London. Lincoln: University of Lincoln. http://lncn.eu/mxz2

Wood, L. (2013) iridium – evaluation of DataStage and DataBank research data management tools from DataFlow project. *Iridium Blog* 14th February. Newcastle: Iridium Project, Newcastle University. http://iridiummrd.wordpress.com/2013/02/14/iridium-evaluation-of-datastage-and-databank-research-data-management-tools-from-dataflow-project/

9.2. Alphabetical index of entities noted in the text

Mint http://www.redboxresearchdata.com.au/	38
MiSS http://www.miss.manchester.ac.uk/	54
Monash University repository http://arrow.monash.edu.au/vital/access/manager/Index	49
Moodle https://moodle.org/	40
MRD at UWE http://www1.uwe.ac.uk/library/sc/mrd-project.aspx	53
My Lab http://www.mylab.fi/en/	35
MyLabbook http://mylabbook.org/	36
MyExperiment http://www.myexperiment.org/	36
MyGrid http://www.mygrid.org.uk/	35
MyPublications https://www.shef.ac.uk/ris/post-project/mypublications	17
MySQL http://www.mysql.com/	39
MyTea VLab http://mytea.org.uk/vlab	36
N8 Research Partnership http://www.n8research.org.uk/	20
n8equipment http://www.n8equipment.org.uk/	20
OAI-PMH http://www.openarchives.org/pmh/	37
OBiBa http://www.obiba.org/	36
Omero http://www.openmicroscopy.org/site/products/omero	36
ONEIS http://www.oneis.co.uk/research	25
Online Research Database Service (ORDS) http://ords.ox.ac.uk/	47
Open Archival Information System (OAIS) http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=57284	29
Open Atrium http://openatrium.com/	35
Open Exeter http://as.exeter.ac.uk/library/resources/openaccess/openexeter/	54
Open Journal System https://pkp.sfu.ca/ojs/	39