Arkivum
Every bit archived

# RDM Workflows and Integrations for Higher Education Institutions Using Hosted Services

A report written by Matthew Addis as part of the Jisc supported project “Small and Specialist: A consortial approach to building an integrated RDM system”.

DOI: dx.doi.org/10.6084/m9.figshare.1476832

**Arkivum Limited**
R21 Langley Park Way
Chippenham
Wiltshire
SN15 1GE
UK

+44 1249 405060

info@arkivum.com
@Arkivum
Arkivum.com

**Title**
RDM Workflows and Integrations for Higher Education Institutions Using Hosted Services

**Part no.**
ARK/REPT/ALL/332

**Version**
1.0

**Document date**
July 2015

**Status**
Release

**Arkivum Limited**
R21 Langley Park Way
Chippenham
Wiltshire
SN15 1GE
UK

+44 1249 405060

info@arkivum.com
@Arkivum
Arkivum.com

**Title**
RDM Workflows and Integrations for Higher Education Institutions Using Hosted Services

**Part no.**
ARK/REPT/ALL/332

**Version**
1.0

**Document date**
July 2015

**Status**
Release

# Table of Contents

# 1 Summary

This report has been created as part of the Jisc Research Data Spring[1] supported project "Small and Specialist: A consortial approach to building an integrated RDM system"[2] (CREST RDMS). The report investigates some of the workflows and processes involved in Research Data Management (RDM).

We define a RDM workflow to be:

> *The sequence of repeatable processes (steps) through which Research Data passes during its lifecycle, including the steps involved in its creation, curation, preservation, access and eventual disposal.*

The aim of the report is to inform subsequent stages of the Jisc CREST RDMS project where a shared-service will be designed and developed for RDM. The shared service will implement a subset of these workflows and be made available to small and specialist institutions and their researchers. Due to the small-scale nature of the project, the report focuses on specific UK examples from consortium partners, or institutions they work with, and is not intended to be a comprehensive investigation into RDM workflows. However, given the general paucity of documented RDM workflows, the authors hope that this report will still be of interest to the wider community.

When we investigated RDM workflows in current use, it became clear that the ability to define and automate RDM processes is an important factor in ensuring the quality, consistency and repeatability of RDM. Workflows help ensure clarity for all those involved, including researchers and support staff, by giving a clear description of what to do, how to do it, and when. There is currently much activity in UK institutions around RDM policy, raising awareness of the need and benefits of RDM, and supporting researchers through guidelines and training. However, these activities can lack the detail needed on a practical day-to-day level, or, as one researcher we spoke to commented, "What's needed is something actionable rather than aspirational". Workflows and their implementation in a RDM infrastructure helps fill this gap. The benefits of a workflow based approach include: cost reduction particularly when scaling up RDM activities; lower barriers to use which helps ensure researcher participation; and higher levels of confidence for an institution when addressing funding body expectations or trying to get the maximum value and impact from its research outputs. Much of this is simply part of good research practice where workflows can help to ensure that RDM gets embedded within day-to-day operations of an institution and its researchers.

The next stage of the work anticipated in the Jisc CREST RDMS project (subject to further funding) is to extend consultation on this document and RDM workflows to a wider community; define the specific workflows that are most important to support in the shared-service; and to create a 'blueprint' that defines what the service will be in practice and how it will operate.

---

1. https://www.jisc.ac.uk/rd/projects/research-data-spring
2. http://crest.ac.uk/blog/

# 2 Objectives

Much effort has been made by many institutions in defining RDM policy, raising RDM awareness amongst researchers, and supporting these researchers through providing guidelines and training materials. These policies and guidelines are increasingly supported by infrastructure for researchers to use, for example when depositing research outputs into an institutional repository, or for the institution itself to use, for example when assessing the impact of its research or checking conformance to funding body policy. However, well-defined processes containing clear step-by-step instructions on what to do, who should do it, and when are currently less well developed. Policy says what should be done. Training materials often says why it is important and what might go wrong if good practice isn't followed. But sometimes a gap exists in terms of clear descriptions of how RDM should be executed in practice.

This report aims to look at what workflows are typically in place, where the gaps are, and how hosted RDM services could help fill these gaps or simplify the processes. In particular, this report aims to:

- Describe the workflows/processes involved when researchers and institutions use or operate RDM infrastructure (tools, services, platforms).
- Provide practical examples of how RDM workflows are implemented and supported at a range of UK Higher Education Institutions, including integration of systems/tools.
- Define/compare the strengths/weaknesses when using hosted RDM infrastructure, on site infrastructure or a combination.

# 3 What do we mean by workflows?

'Workflow' is one of those overloaded terms with multiple definitions depending on discipline or application. For example, the term 'workflow' might be used in the context of business process management, the automation of scientific data processing, or the procedures for paper document handling – all of which are very different things. The result can often be confusion or miscommunication when different parties talk to each other from different perspectives, as is often the case in a multi-disciplinary area such as RDM. This makes it doubly important to define what we mean by 'workflow' in this report.

We start with two general-purpose definitions of 'workflow':

> *"The sequence of industrial, administrative, or other processes through which a piece of work passes from initiation to completion; the passage of a piece of work through this sequence."* **Oxford English Dictionary.**

> *"A workflow consists of an orchestrated and repeatable pattern of business activity enabled by the systematic organization of resources into processes that transform materials, provide services, or process information. It can be depicted as a sequence of operations, declared as work of a person or group, an organization of staff, or one or more simple or complex mechanisms."* **Wikipedia.**

These definitions show that people perform workflows as well as automated systems, workflows can involve lots of different stakeholders, and workflows will typically involve processing or transforming something. In the case of RDM, that 'something' is the research data and the stakeholders include researchers, librarians, systems administrators and service providers.

In this context, we can define workflows for Research Data Management as:

> *The sequence of repeatable processes (steps) through which Research Data passes during its lifecycle, including the steps involved in its creation, curation, preservation, access and eventual disposal.*

This definition could also be extended to include activities that take place before the research data is created, for example research data management planning and the creation of Data Management Plans (DMPs) for grant applications.

Different people will be involved in different RDM workflows. This allows us to group workflows according to the main actors involved. For example, the workflows that a researcher goes through might include:

- Plan how to manage the data (e.g. DMP)
- Do the research and create/analyse/use data
- Decide what data to keep/make accessible, who for and why
- Think about, and resolve, issues around ethics and privacy
- Deposit into a discipline-specific repository or institutional repository
- Publish papers, reference the data, disseminate the findings, promote the data

The steps won't necessarily always be in this order and often the set of steps may be repeated as a research project is executed.

Staff an institution will also go through various workflows. For example, librarians or research support staff might do the following:

- Define metadata and data standards, templates and checks
- QA/QC of research data deposits
- Review/approve requests by people to access the data
- Gather and report statistics on data usage
- Handle requests for data removal
- Curate and manage datasets to ensure they remain useful and usable
- Decide how to preserve data and when to perform preservation actions
- Licensing, embargoes, access requests/restrictions
- Auditing/checking/showing compliance with funder expectations and requirements

Workflows will often require interaction with RDM systems, infrastructure and external services. For example, workflows that might happen using specific tools or applications in the RDM infrastructure can include:

- Using an institutional repository to make data publicly accessible
- Minting a Document Object Identifier (DOI)
- Synchronising metadata between systems, e.g.:
    - Current Research Information System (CRIS)
    - Institutional Repository (IR)
    - Dataset registry, e.g. Jisc Research Data Registry
    - Funding body reporting system, e.g. Research Fish
- Moving/copying data to access platforms, e.g. Figshare
- Moving/copying data to archive for long-term storage.
- Planning and executing preservation actions, e.g. using Archivematica
- Deletion/disposal/removal of data and records involving all of the above.

# 4 Benefits of workflows and automation

Much of the complexity of current RDM workflows comes from the need for people (Researchers, Library staff, Research Officers, IT administrators) to use multiple systems that are not fully integrated. For example, a Researcher might have to enter the same metadata into multiple locations such as an Institutional Repository, CRIS, funding body research tracking system, or journal.

Complexity and lack of integration can mean wasted effort, more opportunities for mistakes, or increased likelihood for steps to be forgotten or delayed. Automation of workflows so that systems communicate directly with each other can help to substantially reduce these problems whilst providing a simplified and more seamless experience to the user.

The benefit of a simple and seamless experience for the researcher is hard to overstate and has been a recurring theme when we have looked at current RDM workflows in a range of institutions. If a researcher is faced with the need to spend extra time and effort on what they might consider 'box ticking' for compliance with University policy and funder expectations, then they may simply chose not to do some of the required steps, for example making an entry for their data set in an institutional repository or CRIS. This is especially true if they feel that their data is already 'out there' and accessible, e.g. as part of supporting information for a journal publication or because they have deposited it in a discipline-specific repository.

On the one hand, there is a clear case for the benefits of RDM as part of good scientific and research practice, e.g. science as an open enterprise[1]. There is growing evidence of the direct benefits to researchers[2] [3], e.g. citations and grant funding, and there is emerging evidence of the macro-economic benefits[4] [5] to institutions and nationally. But, on the other hand, unless RDM is simple and practical on a day-to-day level then researchers may simply do the 'minimum possible' to 'comply' and these benefits are not realised.

Workflows that focus on making the researcher's life easier can make a big difference to adoption of RDM by researchers, especially in the transition period we are currently in where good RDM is not yet the 'norm' in all disciplines and institutions. The problem can be compounded in smaller or specialized institutions that are not research intensive and don't have the staff or budget to create and manage an integrated RDM infrastructure and support/encourage researchers to use it. The benefit of automated workflows in this scenario is that they allow shared services to be developed that individual institutions can then subscribe to at a relatively low start-up and on-going cost. This approach is one of the objectives of the CREST RDMS project.

In summary, the potential benefits of RDM workflows, especially when automated, include:

---

1. https://royalsociety.org/~/media/royal_society_content/policy/projects/sape/2012-06-20-saoe.pdf
2. http://data.bris.ac.uk/files/2013/06/data-bris-benefits-report-V2.pdf
3. http://www.dcc.ac.uk/resources/briefing-papers/making-case-rdm
4. http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report,_Jan14_v1-04.pdf
5. http://ands.org.au/resource/open-research-data-report.pdf

- Simpler and easier RDM processes from a Researcher perspective, which both encourages adoption and lowers the cost of institutional support to the research base.
- Clear and repeatable RDM processes that help ensure higher levels of quality and consistency in RDM across the research base.
- Ability to deploy RDM as community-driven shared service(s) so that smaller institutions can ‘join forces’ to benefit from having access to a common RDM infrastructure.
- Scaling RDM up across a large research base using automation and ‘factory’ type approaches to achieve ‘economies of scale’ and move away from RDM being a manual and labour intensive endeavour.

Many of the case studies in this report look at ways in which RDM workflows have been, or can be, automated and to bring one or more of the benefits above.

# 5 Case studies and examples

The flowcharts and diagrams used in the case studies below are available as a file set from Figshare:

DOI: dx.doi.org/10.6084/m9.figshare.1476831

## 5.1 Workflows for guiding a Researcher in how to deposit research data

Most UK institutions now have policies in place for their research data. A list of institutions and their policies [1] is provided by the Digital Curation Centre. These policies are broadly similar and address the desire of the institution to: retain and make use of its research holdings; address legal and ethical considerations; and meet expectations from the various funding bodies supporting the research. However, individual policies differ at the detailed level. For example, there can be specific requirements on whether certain activities are mandatory or not such as the deposit of research data in an institutional repository. The workflows in this section show how funder expectations and institutional policies can be presented to researchers as flowcharts. This provides both clarity on what the researcher is expected to do and provides something that is typically much shorter and easier to interpret than the institution's underpinning policy document. The more prescriptive nature of a flowchart and the ability to convey the institution and funding body requirements in a single diagram means that researchers are both more likely to follow the process and, moreover, remember that there is a process to follow. Three examples of research data deposit workflows are included in this section.

- **The University of Southampton**. The University of Southampton is currently developing its research data management infrastructure in support of its RDM policy [2]. This is in part being driven by the EPSRC expectations [3] and clarifications [4] of having RDM in place as of 1 May 2015. The infrastructure uses an EPrints [5] repository hosted by Southampton to provide a central point for both publications and research data. The workflow [6] shows Researchers how they are expected to use this repository in the context of meeting the EPSRC expectations.
- **Loughborough University**. Loughborough University is using a combination of Figshare [7] for data publication and access, Symplectic Elements [8] as their Current Research Information Systems (CRIS), DSpace as their institutional repository [9], and Arkivum [10] to provide data archiving. These together make up the RDM infrastructure and support Loughborough's draft RDM policy [11]. The workflow from a Researcher perspective is very similar to that of

**Arkivum Limited**
R21 Langley Park Way
Chippenham
Wiltshire
SN15 1GE
UK

+44 1249 405060

info@arkivum.com
@Arkivum
Arkivum.com

**Title**
RDM Workflows and Integrations for Higher Education Institutions Using Hosted Services

**Part no.**
ARK/REPT/ALL/332

**Version**
1.0

**Document date**
July 2015

**Status**
Release

---

[1] http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies
[2] http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html
[3] https://www.epsrc.ac.uk/about/standards/researchdata/expectations/
[4] https://www.epsrc.ac.uk/files/aboutus/standards/clarificationsofexpectationsresearchdatamanagement/
[5] http://eprints.soton.ac.uk/
[6] http://library.soton.ac.uk/ld.php?content_id=11468527
[7] https://lboro.figshare.com/
[8] http://symplectic.co.uk/products/elements/
[9] https://dspace.lboro.ac.uk/dspace-jspui/
[10] http://arkivum.com/
[11] http://www.lboro.ac.uk/service/research/offcampus/docs/ResearchDataManagementPolicy-Draft.pdf

Southampton in that Figshare provides a single point of interaction and focus, and the complexity of the rest of the system is 'hidden'.

- **Imperial College London**. Imperial College London has recently published its RDM policy [12] as part of its RDM activities [13]. In contrast to Southampton and Loughborough, Imperial College does not require the use of a specific repository by default and instead leaves this as a choice for the Researcher whilst still making it clear that they are ultimately responsible for the research data that they create.

### Key points

- There is no 'one size fits all' workflow from a Researcher perspective. Differences in institution policy and RDM infrastructure will result in different workflows even when a Researcher is trying to meet the expectations of the same funding body.
- Workflows that minimise the burden on a Researcher, for example by providing a very simple 'one stop shop' for data deposit irrespective of the type of data or discipline, are likely to be easier for a Researcher to adopt and easier for an institution to monitor and support.
- Workflows that help Researchers realise the benefits of RDM, for example by making it easier for them to promote and share their research outputs, to get better citation and download rates, or to increase the chances of further funding or research collaborations, will all help incentivise Researchers. This 'carrot' approach is in contrast to workflows geared towards the enforcement of policy or funding body expectations, which can sometimes appear as a 'stick' from a Researcher perspective.

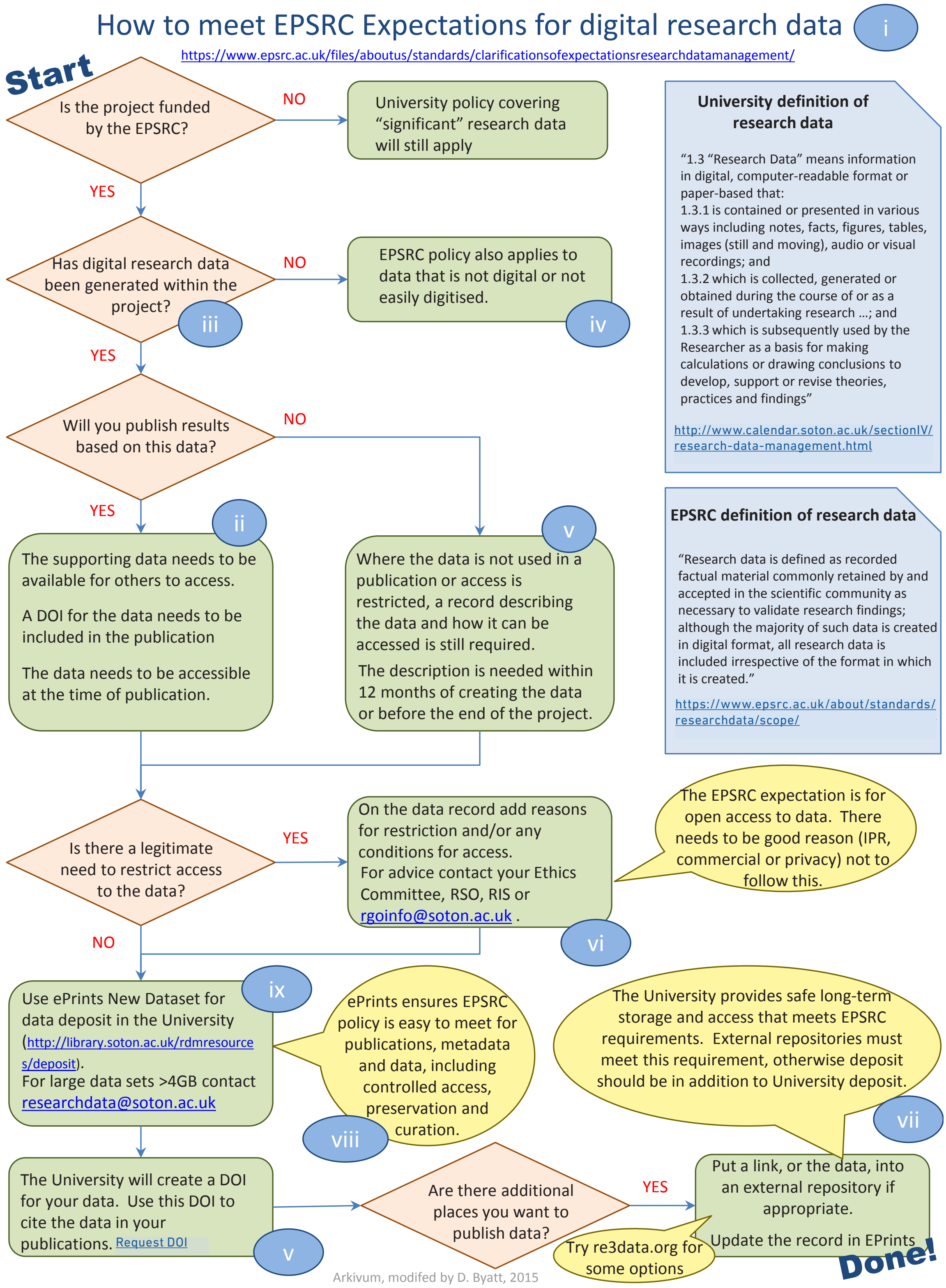
### Considerations for RDM as a service

- Workflows should be as simple as possible from a Researcher perspective, for example using a 'one stop shop' approach so a Researcher only has to interact with one system/service rather than many.
- Workflows should enable the Researcher to benefit directly from RDM, e.g. by automatic publication of data in locations that facilitate easy and online access by others.
- Uniformity of the workflow across disciplines and data types will make the workflow easier to implement, support and monitor from an institutional point of view.

---

[12] https://workspace.imperial.ac.uk/researchservices/Public/Imperial%20College%20RDM%20Policy.pdf
[13] http://www3.imperial.ac.uk/researchsupport/rdm

### 5.1.1 University of Southampton: Data deposit by Researchers to meet EPSRC expectations



How to meet EPSRC Expectations for digital research data (i)

https://www.epsrc.ac.uk/files/aboutus/standards/clarificationsofexpectationsresearchdatamanagement/

Start

Is the project funded by the EPSRC?

NO: University policy covering “significant” research data will still apply

YES

Has digital research data been generated within the project? (iii)

NO: EPSRC policy also applies to data that is not digital or not easily digitised. (iv)

YES

Will you publish results based on this data?

YES: (ii) The supporting data needs to be available for others to access.

A DOI for the data needs to be included in the publication

The data needs to be accessible at the time of publication.

NO: (v) Where the data is not used in a publication or access is restricted, a record describing the data and how it can be accessed is still required.

The description is needed within 12 months of creating the data or before the end of the project.

Is there a legitimate need to restrict access to the data?

YES: On the data record add reasons for restriction and/or any conditions for access.
For advice contact your Ethics Committee, RSO, RIS or rgoinfo@soton.ac.uk . (vi)

The EPSRC expectation is for open access to data. There needs to be good reason (IPR, commercial or privacy) not to follow this.

NO

Use ePrints New Dataset for data deposit in the University (http://library.soton.ac.uk/rdmresources/deposit).
For large data sets >4GB contact researchdata@soton.ac.uk (ix)

ePrints ensures EPSRC policy is easy to meet for publications, metadata and data, including controlled access, preservation and curation. (viii)

The University provides safe long-term storage and access that meets EPSRC requirements. External repositories must meet this requirement, otherwise deposit should be in addition to University deposit. (vii)

The University will create a DOI for your data. Use this DOI to cite the data in your publications. Request DOI (v)

Are there additional places you want to publish data?

YES: Put a link, or the data, into an external repository if appropriate.

Update the record in EPrints

Try re3data.org for some options

Done!

“1.3 “Research Data” means information in digital, computer-readable format or paper-based that:
1.3.1 is contained or presented in various ways including notes, facts, figures, tables, images (still and moving), audio or visual recordings; and
1.3.2 which is collected, generated or obtained during the course of or as a result of undertaking research ...; and
1.3.3 which is subsequently used by the Researcher as a basis for making calculations or drawing conclusions to develop, support or revise theories, practices and findings”

http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html

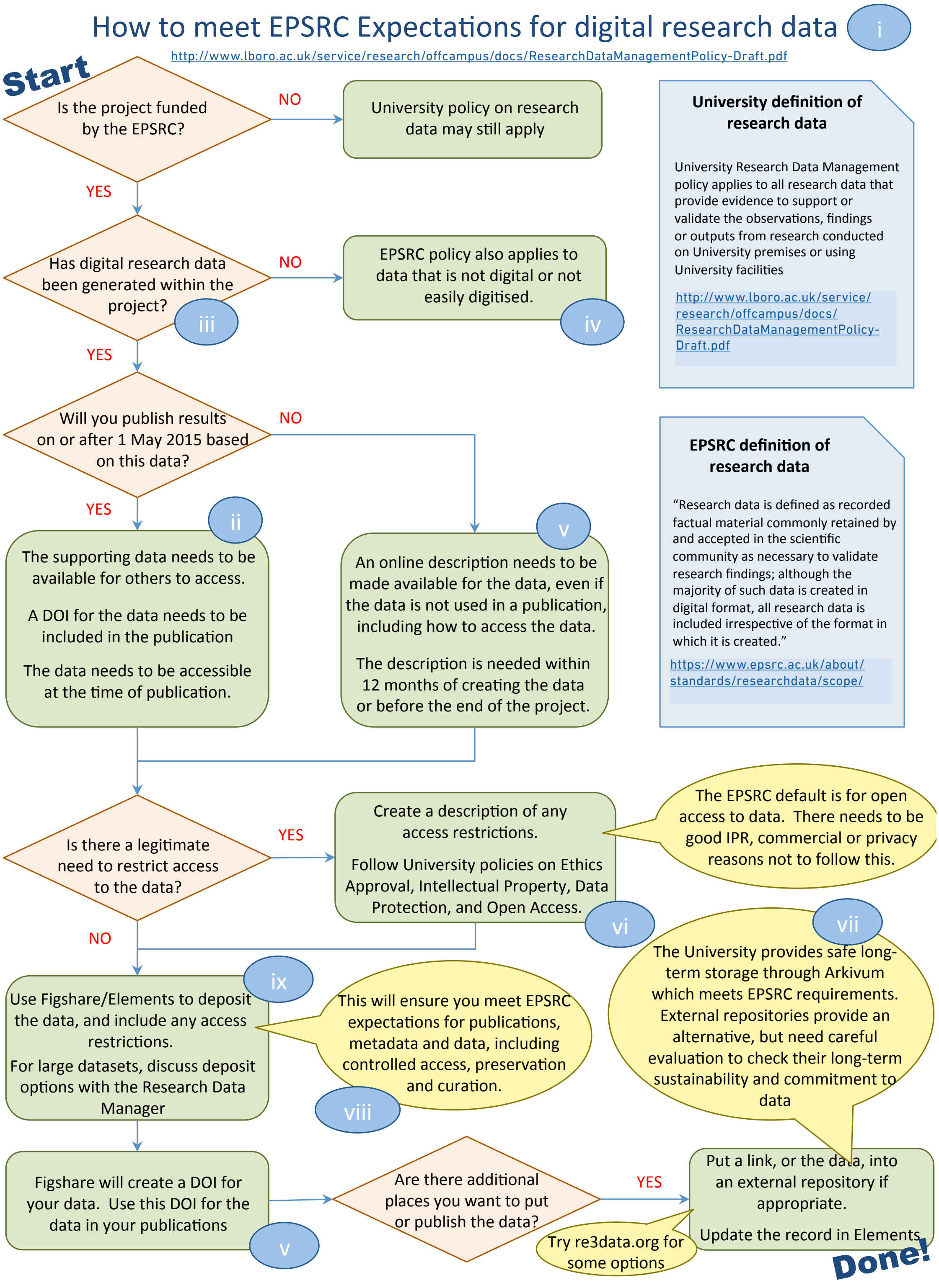
“Research data is defined as recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings; although the majority of such data is created in digital format, all research data is included irrespective of the format in which it is created.”

https://www.epsrc.ac.uk/about/standards/researchdata/scope/

Arkivum, modifed by D. Byatt, 2015

Figure 1. Workflow for the University of Southampton. Originally authored by Arkivum and subsequently modified by Southampton. Available from http://library.soton.ac.uk/ld.php?content_id=11468527. The Roman numerals refer to each of the specific EPSRC expectations. http://dx.doi.org/10.6084/m9.figshare.1477992

## 5.1.2 Loughborough University: Data deposit by Researchers to meet EPSRC expectations


How to meet EPSRC Expectations for digital research data
i
http://www.lboro.ac.uk/service/research/offcampus/docs/ResearchDataManagementPolicy-Draft.pdf
Start
Is the project funded by the EPSRC?
NO
University policy on research data may still apply
YES
Has digital research data been generated within the project?
iii
NO
EPSRC policy also applies to data that is not digital or not easily digitised.
iv
YES
Will you publish results on or after 1 May 2015 based on this data?
NO
YES
ii
v
Is there a legitimate need to restrict access to the data?
YES
NO
vi
vii
ix
viii
Are there additional places you want to put or publish the data?
YES
v
Try re3data.org for some options
Done!

**University definition of research data**

University Research Data Management policy applies to all research data that provide evidence to support or validate the observations, findings or outputs from research conducted on University premises or using University facilities

http://www.lboro.ac.uk/service/research/offcampus/docs/ResearchDataManagementPolicy-Draft.pdf

**EPSRC definition of research data**

"Research data is defined as recorded factual material commonly retained by and accepted in the scientific community as necessary to validate research findings; although the majority of such data is created in digital format, all research data is included irrespective of the format in which it is created."

https://www.epsrc.ac.uk/about/standards/researchdata/scope/

The supporting data needs to be available for others to access.

A DOI for the data needs to be included in the publication

The data needs to be accessible at the time of publication.

An online description needs to be made available for the data, even if the data is not used in a publication, including how to access the data.

The description is needed within 12 months of creating the data or before the end of the project.

Create a description of any access restrictions.

Follow University policies on Ethics Approval, Intellectual Property, Data Protection, and Open Access.

The EPSRC default is for open access to data. There needs to be good IPR, commercial or privacy reasons not to follow this.

Use Figshare/Elements to deposit the data, and include any access restrictions.

For large datasets, discuss deposit options with the Research Data Manager

This will ensure you meet EPSRC expectations for publications, metadata and data, including controlled access, preservation and curation.

The University provides safe long-term storage through Arkivum which meets EPSRC requirements. External repositories provide an alternative, but need careful evaluation to check their long-term sustainability and commitment to data

Figshare will create a DOI for your data. Use this DOI for the data in your publications

Put a link, or the data, into an external repository if appropriate.

Update the record in Elements

Figure 2. Workflow for Loughborough University. Originally authored by Arkivum and subsequently reviewed by Loughborough..The Roman numerals refer to each of the specific EPSRC expectations.

http://dx.doi.org/10.6084/m9.figshare.1477991

**Arkivum Limited**
R21 Langley Park Way
Chippenham
Wiltshire
SN15 1GE
UK

+44 1249 405060

info@arkivum.com
@Arkivum
Arkivum.com

**Title**
RDM Workflows and Integrations for Higher Education Institutions Using Hosted Services

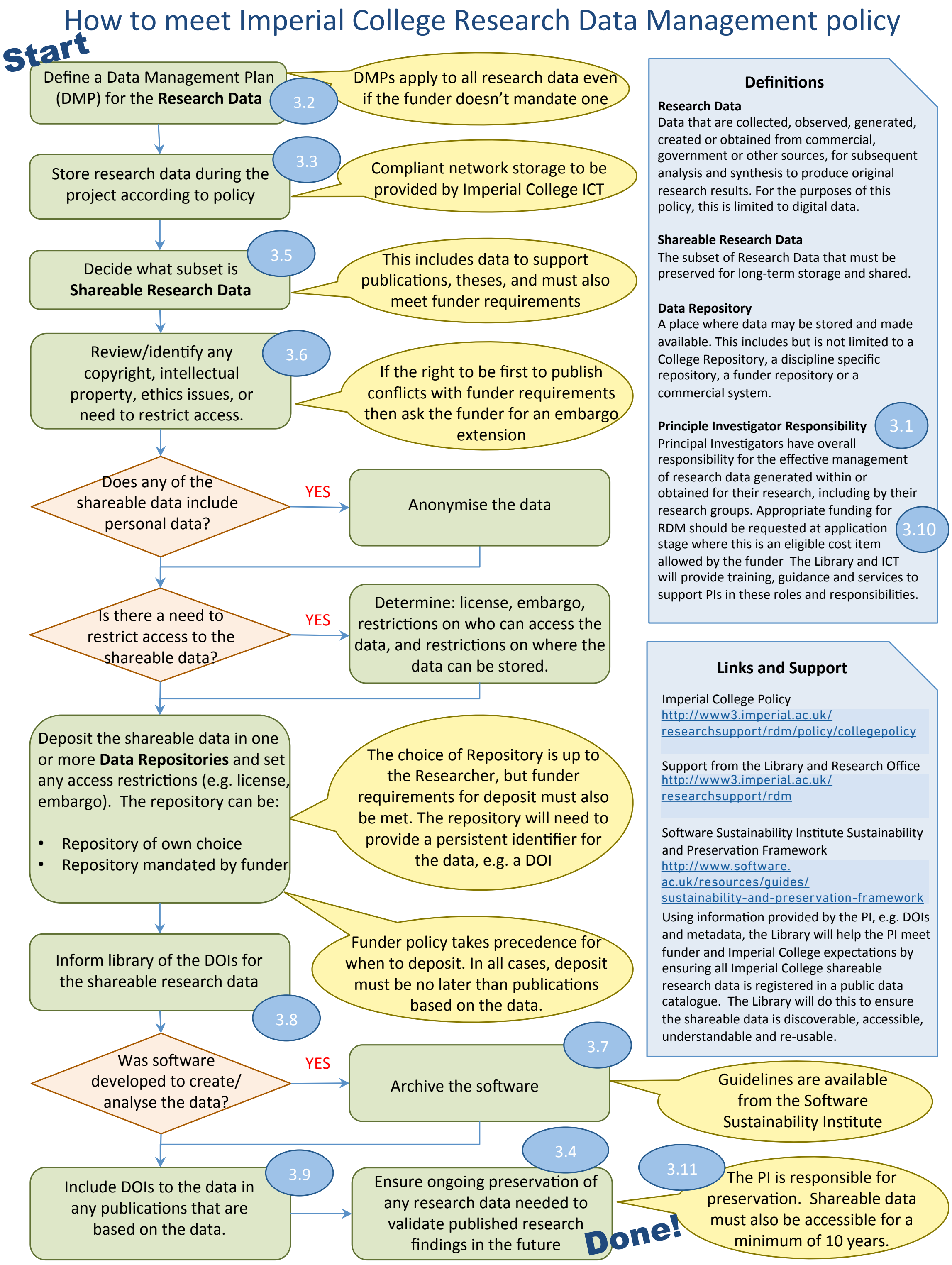
**Part no.**
ARK/REPT/ALL/332

**Version**
1.0

**Document date**
July 2015

**Status**
Release

## 5.1.3 Imperial College: RDM by Researchers to meet institutional policy



How to meet Imperial College Research Data Management policy

Define a Data Management Plan (DMP) for the **Research Data**

DMPs apply to all research data even if the funder doesn't mandate one

Store research data during the project according to policy

Compliant network storage to be provided by Imperial College ICT

Decide what subset is **Shareable Research Data**

This includes data to support publications, theses, and must also meet funder requirements

Review/identify any copyright, intellectual property, ethics issues, or need to restrict access.

If the right to be first to publish conflicts with funder requirements then ask the funder for an embargo extension

Does any of the shareable data include personal data?

Anonymise the data

Is there a need to restrict access to the shareable data?

Determine: license, embargo, restrictions on who can access the data, and restrictions on where the data can be stored.

Deposit the shareable data in one or more **Data Repositories** and set any access restrictions (e.g. license, embargo). The repository can be:

- Repository of own choice
- Repository mandated by funder

The choice of Repository is up to the Researcher, but funder requirements for deposit must also be met. The repository will need to provide a persistent identifier for the data, e.g. a DOI

Funder policy takes precedence for when to deposit. In all cases, deposit must be no later than publications based on the data.

Inform library of the DOIs for the shareable research data

Was software developed to create/ analyse the data?

Archive the software

Guidelines are available from the Software Sustainability Institute

Include DOIs to the data in any publications that are based on the data.

Ensure ongoing preservation of any research data needed to validate published research findings in the future

The PI is responsible for preservation. Shareable data must also be accessible for a minimum of 10 years.

**Definitions**

**Research Data**
Data that are collected, observed, generated, created or obtained from commercial, government or other sources, for subsequent analysis and synthesis to produce original research results. For the purposes of this policy, this is limited to digital data.

**Shareable Research Data**
The subset of Research Data that must be preserved for long-term storage and shared.

**Data Repository**
A place where data may be stored and made available. This includes but is not limited to a College Repository, a discipline specific repository, a funder repository or a commercial system.

**Principle Investigator Responsibility**
Principal Investigators have overall responsibility for the effective management of research data generated within or obtained for their research, including by their research groups. Appropriate funding for RDM should be requested at application stage where this is an eligible cost item allowed by the funder The Library and ICT will provide training, guidance and services to support PIs in these roles and responsibilities.

**Links and Support**

Imperial College Policy
http://www3.imperial.ac.uk/researchsupport/rdm/policy/collegepolicy

Support from the Library and Research Office
http://www3.imperial.ac.uk/researchsupport/rdm

Software Sustainability Institute Sustainability and Preservation Framework
http://www.software.ac.uk/resources/guides/sustainability-and-preservation-framework

Using information provided by the PI, e.g. DOIs and metadata, the Library will help the PI meet funder and Imperial College expectations by ensuring all Imperial College shareable research data is registered in a public data catalogue. The Library will do this to ensure the shareable data is discoverable, accessible, understandable and re-usable.

Figure 3. PI workflow for RDM at Imperial College London. The flowchart is Arkivum's understanding of Imperial's recently released RDM policy document and not an approved Imperial College representation. The numbers refer to specific policy requirements on the PI. http://dx.doi.org/10.6084/m9.figshare.1477993

## 5.2 Workflow for publishing data using DOIs for research data sets

### Overview

The availability of an institutional repository or other centrally provisioned RDM infrastructure is not always necessary in order for Researchers to go a long way in providing online access to their research results in a way that allows that data to be easily found and used by others and also aligns well with funding body expectations. RDM examples and guidelines by Henry Rzepa[1] in the Computational Chemistry group at Imperial College London[2] provide an informative case study. The aim here is for RDM is to follow F.A.I.R principles[3], namely that data should be Findable, Accessible, Interoperable and Reusable. FAIR has emerged from the eScience community and is independent of specific funding body requirements or institutional policies and instead sets out general principles for ensuring that scientific data is readily discoverable, accessible and usable. It should be noted that whilst emerging from the scientific community, there is no reason why FAIR principles cannot be applied more widely to other disciplines. The FAIR principles include the availability of both human and machine-readable descriptions of the data, the data itself being in a syntactically and semantically well defined form, and the data being easily locatable through persistent identifiers, for example DOIs[4]. Data can exist at multiple levels[5], be deposited in multiple systems, and be in different versions or formats including the Supporting Information[6] for a Journal paper, an online version in a repository such as Figshare, or an 'accepted manuscript' version with accompanying files in an institutional repository. All can have their own DOIs. One of the challenges is to simultaneously realise the benefits of online access and DOIs (FAIR) whilst ensuring the various versions of the data are connected together and it is clear to users of the data which version is the 'version of record'[7]. However, the benefit of making data accessible with DOIs still vastly outweighs any consequent issues of version reconciliation and cross-referencing.

It should be noted that data publication workflows and the use of DOIs is very much an evolving area, for example see the work of the Research Data Alliance (RDA) working group on publishing data workflows[8] and its recent survey results[9]. One area that the group is working on is how to cite evolving datasets[10] where, for example, a persistent identifier is used to refer to a specific subset of data within a data resource, e.g. online service, so it can be referred to and retrieved at a later date but in the form

---

[1] http://www.imperial.ac.uk/people/h.rzepa, ORCID: http://orcid.org/0000-0002-8635-8390

[2] http://www.imperial.ac.uk/chemistry/research/

[3] https://www.force11.org/group/fairgroup

[4] http://www.dcc.ac.uk/resources/how-guides/cite-datasets

[5] For example, in computational chemistry when simulating the structure or behaviour of a compound, an individual fileset might be from the execution of a model on an HPC facility. The fileset includes the input files, output files, parameters for the software tool, log files for the simulation run, and other associated files. Multiple filesets might be combined into a study, for example multiple simulation runs for variations of a compound's composition or structure. This constitutes a dataset. The results of one or more studies might then be reported and summarised in an article that presents more details of the method and software tools that were used to create and interpret the datasets.

[6] Examples of Supporting Information used by Journals include: PLoS http://journals.plos.org/plosone/s/supporting-information and Wiley https://authorservices.wiley.com/bauthor/suppinfo.asp

[7] http://www.ch.imperial.ac.uk/rzepa/blog/?p=14183

[8] https://www.rd-alliance.org/groups/rdawds-publishing-data-workflows-wg.html

[9] http://dx.doi.org/10.5281/zenodo.19107

[10] https://www.rd-alliance.org/filedepot?fid=667

that it existed at a given point in time. This has implications for repositories that hold growing or changing research datasets.

### Key Points

- FAIR provides a good set of general principles for making data accessible and reusable. Institutional RDM policy that aligns with FAIR principles is likely to help with both understanding and adoption in an institution's research base.
- FAIR provides principles not implementation details. FAIR data is not necessarily technically difficult to achieve, but it does require researchers to have discipline and specific examples and processes to follow so it is clear how to best implement FAIR principles in their particular research environment.
- Data sets can be multi-level, exist across multiple locations/repositories, and have different versions. DOIs form a solid basis for persistently referencing and linking together the various versions of data.

### Considerations for RDM as a service

- Researchers will want to publish data and articles in locations/services most appropriate to their research and discipline, e.g. particular journals or subject -specific repositories. This means there is a need for institutional RDM systems to support, where possible, the automated discovery, linking and copying of this data to create the institutional record. This ensures that the institutional record is consistent with external sources of data with minimal burden on the Researcher.
- The 'web of DOIs' that arises from multi-level datasets and descriptions means that there is a need for linking between systems. Benefits of DOI based linking, especially when bi-directional between institutional RDM systems and external data repositories/publications, includes a robust way for users of the data to both 'drill down' to the primary underpinning data and 'navigate back up' to contextual information and interpretation in article publications.

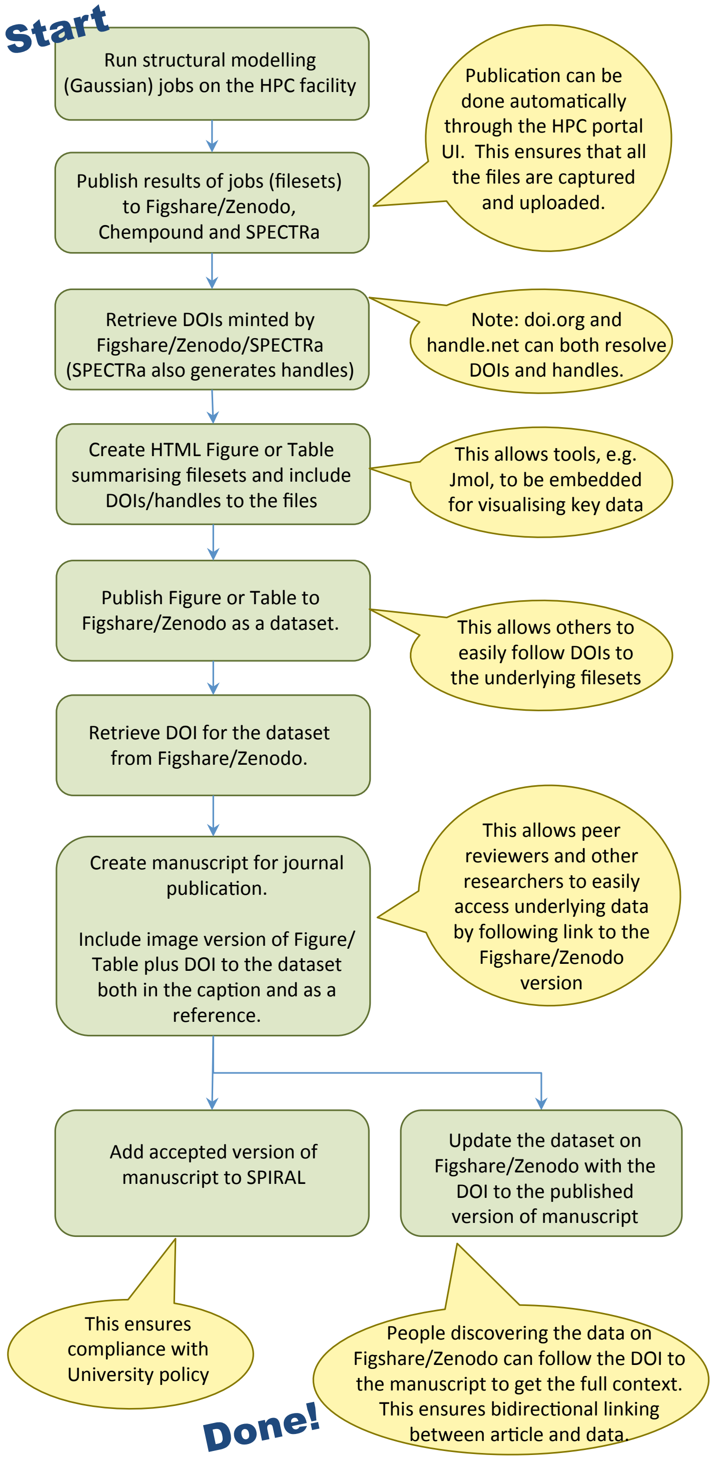
### Specific Workflow

The example below provides flowchart representation of the specific steps for computational chemistry research outputs generated by HPC simulation of chemical structure. Further details are available from a wiki page [11] developed by Henri Rzepa for Imperial Researchers.

---

11. https://wiki.ch.ic.ac.uk/wiki/index.php?title=Rdm:intro

# Use of DOIs for computational chemistry data



Start

Run structural modelling (Gaussian) jobs on the HPC facility

Publish results of jobs (filesets) to Figshare/Zenodo, Chempound and SPECTRa

Publication can be done automatically through the HPC portal UI. This ensures that all the files are captured and uploaded.

Retrieve DOIs minted by Figshare/Zenodo/SPECTRa (SPECTRa also generates handles)

Note: doi.org and handle.net can both resolve DOIs and handles.

Create HTML Figure or Table summarising filesets and include DOIs/handles to the files

This allows tools, e.g. Jmol, to be embedded for visualising key data

Publish Figure or Table to Figshare/Zenodo as a dataset.

This allows others to easily follow DOIs to the underlying filesets

Retrieve DOI for the dataset from Figshare/Zenodo.

Create manuscript for journal publication.

Include image version of Figure/Table plus DOI to the dataset both in the caption and as a reference.

This allows peer reviewers and other researchers to easily access underlying data by following link to the Figshare/Zenodo version

Add accepted version of manuscript to SPIRAL

This ensures compliance with University policy

Update the dataset on Figshare/Zenodo with the DOI to the published version of manuscript

People discovering the data on Figshare/Zenodo can follow the DOI to the manuscript to get the full context. This ensures bidirectional linking between article and data.

Done!

## Tools and Services

**HPC Portal**
Imperial College HPC facility
http://www.imperial.ac.uk/admin-services/ict/self-service/research-support/hpc/

**Gaussian**
Software tool for chemical structure simulations
http://www.gaussian.com

**Chempound**
Imperial college Chemical Database (Chempound)
http://www.chempound.net

**SPECTRa**
Imperial College Data Repository (DSpace)
https://spectradspace.lib.imperial.ac.uk:8443

**SPIRAL**
Imperial College Publication Repository (Dspace)
http://spiral.imperial.ac.uk

**Jmol**
3D Chemical Structure Viewer
http://jmol.sourceforge.net

**Figshare**
Public data sharing platform
http://www.figshare.com

**Zenodo**
Public data sharing platform
http://zenodo.org

## Example

**Journal Publication (Manuscript)**
The Houk–List transition states for organocatalytic mechanisms revisited
A. Armstrong, R. A. Boto, P. Dingwall, J. Contreras-García, M. J. Harvey, N. J. Mason and H. S. Rzepa, Chem. Sci., 2014, 5, 2057
DOI: 10.1039/C3SC53416B

**Table (Dataset)**
Rzepa, Henry S.; Harvey, M J; Mason, Nicholas; Dingwall, Paul; Armstrong, Alan; Contreras-García, Julia; Boto, Roberto (2013): Table 8. Houk-List Transition state analogues. figshare.
DOI: 10.6084/m9.figshare.832543

**Gaussian Simulation (Fileset)**
dc.title C 17 H 22 N 2 O 3
dc.type Gaussian job archive
dc.identifier.uri
http://hdl.handle.net/10042/25119/dc.identifier.uri
DOI: 10.14469/ch/19080

Figure 4. Flowchart representation of a data set publication process for computational chemistry data at Imperial College London. http://dx.doi.org/10.6084/m9.figshare.1477994

## 5.3 Workflows for checking and understanding copyright

### Overview

Funding bodies, for example RCUK [1], typically expect open access to both data and publications where possible, with minimal barriers to access including use of permissive licenses such as CC0 [2] or CC-BY [3]. However, there can be many cases where this is either not possible or where time and effort is needed to deal with rights issues before access can be provided. Issues include IPR such as copyright and patenting, handling of personal data, and the ability to support Freedom of Information (FOI) requests. Examples include research done in collaboration with a commercial sponsor, research involving human subjects, and research in creative disciplines where the rights issues can be complex. Researchers in these areas won't always be aware of the all the restrictions and implications of their research and need to work closely with their institution's Research Office, Ethics Committee or Enterprise Services. The result can be the need to restrict access [4], for example through embargoes, the need to restrict how data can be used, for example through specific licenses [5], the need for 'review/approve' workflows for dealing with access requests, the need to anonymise [6] or reduce the dataset into a form that can be made open, and the need to provide specialised and secure environments [7] within which users can access data.

The Jisc-funded Kultur project created a model of an institutional repository for use in the creative and applied arts. The project partners were the University of Southampton, University of the Arts London, University for the Creative Arts, and the Visual Arts Data Service, and Leiden University is an associate partner. Kultur has since been followed by Kaptur [8], which has completed a more comprehensive analysis of RDM in the creative arts. Work continues in the current Jisc CREST RDMS project on how to best build an RDM infrastructure for the visual arts [9] [10]. Kultur was notable for having a work package specifically dedicated to IPR issues [11] including an accompanying decision-making workflow [12], so provides an ideal case study for this report.

Scenarios in Kultur included: video of an artist's sculpture including work of others in the background; digitisation of analogue photographs of unknown origin; use of music within a video art production; and film-making involving actors and script/ideas by others. Kultur reported on how to create an IPR framework [13] to help researchers understand and resolve the IPR issues in scenarios such as these. The recommendation was for Researchers to be supported by an IP framework, which should be produced and enforced by some department, committee or structure within the University to ensure clear lines of responsibility and accountability. The

---

1. http://www.rcuk.ac.uk/research/openaccess/policy/
2. https://creativecommons.org/choose/zero/
3. https://creativecommons.org/licenses/by/3.0/
4. http://www.data-archive.ac.uk/create-manage/consent-ethics/access-control
5. http://www.dcc.ac.uk/resources/how-guides/license-research-data
6. http://www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation
7. http://adrn.ac.uk/protecting-privacy/secure-environment/safe-centres
8. http://www.vads.ac.uk/kaptur/
9. http://www.crest.ac.uk/an-update-from-the-rdiva-team-at-uca/
10. http://www.crest.ac.uk/gathering-rdiva-requirements/
11. http://kultur.eprints.org/documents.htm
12. http://kultur.eprints.org/docs/flowchart.pdf
13. http://kultur.eprints.org/docs/IP%20paper%203%20final.pdf

framework extends beyond the identification of IP issues to include disclaimer, take-down policy, terms of use, and acceptable use of content that is deposited into an Institutional Repository

## Key Points

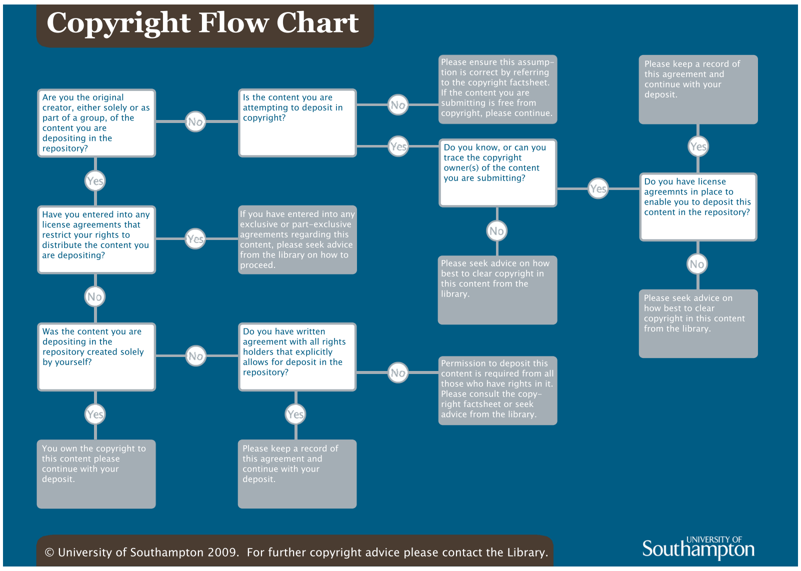
- Kultur recommended that Institutional Repository Guidelines should be: easy to follow and understand for non-IPR expert depositors; accurate and detailed but to not overly wordy or use convoluted legal language; visually appealing; and based around scenarios and content familiar to arts based academics.
- Kutur recommended that guidelines should include: (a) Copyright Fact Sheet; (b) A Scenario Set that is meaningful to arts-based institutions and academics; and (c) A Flow Chart that highlights common IP issues and leads the user to specific contacts and advice.

## Considerations for RDM as a Service

- Researchers should be guided through IP issues at the time of deposit, if not earlier, and the repository should support a workflow that allows institution staff to review and control what content in the repository is made accessible. Support should be included for end-users of the repository to report IP issues or concerns and request take-down.
- Different elements of a research data set may have different IP rights and hence require different licenses and access control.
- The IP position of items in a dataset can change over time and hence there needs to be a way to update the license terms and access control rules of individual items within a dataset.

## Specific Workflow

The flowchart developed in Kultur is shown below. This shows the workflow from a Researcher perspective. It is apparent from the flowchart that many of the ‘end points’ in the workflow involve the Researcher seeking further advice from the Library. This reflects the complexity of dealing with IPR issues in this discipline and how experts in the legal issues will often need to be involved.



Are you the original creator, either solely or as part of a group, of the content you are depositing in the repository?

Is the content you are attempting to deposit in copyright?

Please ensure this assumption is correct by referring to the copyright factsheet. If the content you are submitting is free from copyright, please continue.

Do you know, or can you trace the copyright owner(s) of the content you are submitting?

Please seek advice on how best to clear copyright in this content from the library.

Do you have license agreemnts in place to enable you to deposit this content in the repository?

Please keep a record of this agreement and continue with your deposit.

Please seek advice on how best to clear copyright in this content from the library.

Have you entered into any license agreements that restrict your rights to distribute the content you are depositing?

If you have entered into any exclusive or part-exclusive agreements regarding this content, please seek advice from the library on how to proceed.

Was the content you are depositing in the repository created solely by yourself?

You own the copyright to this content please continue with your deposit.

Do you have written agreement with all rights holders that explicitly allows for deposit in the repository?

Please keep a record of this agreement and continue with your deposit.

Permission to deposit this content is required from all those who have rights in it. Please consult the copyright factsheet or seek advice from the library.

© University of Southampton 2009. For further copyright advice please contact the Library.

Figure 5. Kultur copyright flowchart. The flowchart is copyright University of Southampton and is available from http://kultur.eprints.org/docs/flowchart.pdf

## 5.4 Workflows for research data archiving and access (EPrints and Arkivum)

### Overview

EPrints[1] is software that allows an institution to implement a repository for their research outputs. Currently, research outputs in an institutional repository typically consists of copies of publications in journals or at conferences. However, this is now increasingly starting to include the underpinning research data and publication repositories are moving to become publication and data repositories. Including research data in the repository brings with it the challenges of how to process and store this data so it can be retained and accessed over long timescales. EPrints is widely used in the UK[2]. Likewise, Arkivum provides data archiving as a service to many UK Universities for the long-term storage of research data[3]. Linking of Arkivum storage to EPrints provides a more integrated approach to long-term access to both publications and data. Integration is achieved through use of an EPrints plugin[4], which was developed to address use cases from the University of Leeds and the University of Southampton. The objective is to ensure research outputs are easily discoverable through EPrints, DOIs are allocated where necessary, and the underpinning research data is stored in a dedicated archive to ensure its long-term safety and to minimize costs. Research data is retrieved from the archive on-demand when access requests are received from users of EPrints.

### Key Points

- The institution controls what is physically archived and at what point in the process, not the researcher. The researcher is typically required to deposit with EPrints but then staff, e.g. in the library, determine if and when a given output is added to the archive. This allows selection policies and QC to be implemented before archiving takes place.
- Access to research data will sometimes require a mediated process whereby a request for access is made to the institution (via EPrints), which then decides whether to grant access and retrieve data from the archive. This provides a point of control where access policies can be implemented, e.g. for sensitive or confidential data, along with conformance to funding body expectations.
- EPrints, as with most other repository software, is not designed for large amounts of research data either in terms of data volume (e.g. multiple TBs of genomics data) or number of objects (e.g. millions of files of sensor data). An alternative workflows is needed to allow direct deposit to the archive and subsequent linking, description and QC through EPrints.

### Considerations for RDM as a service

- Multiple workflow options need to be supported for both deposit and access. These workflows need to provide an institution with control over what data to

---

1. http://www.eprints.org/uk/
2. For example, see the IRUS-UK members who have EPrints http://www.irus.mimas.ac.uk/participants/
3. http://arkivum.com/he/
4. http://wiki.eprints.org/w/Files/Configuration_and_User_Guide_for_version_2.1_of_the_EPrints/Arkivum_storage_plugin

archive, when, and how it can be accessed.

- Integration between metadata repositories and data archives needs to support a range of deployment models where parts of the solution might be hosting by an institution, by one or more third-parties, or a combination of the two.
- Data and publications are created at different points in the research lifecycle and will often be deposited at different times. The subsequent linking together of publications and archived data needs to be a simple process and automated where possible.

### Specific workflows

### Upload/download via EPrints

The deposit workflow supported by the current Arkivum EPrints plugin is shown below in Figure 6. This workflow could easily be adapted for other repository software and other archiving solutions.

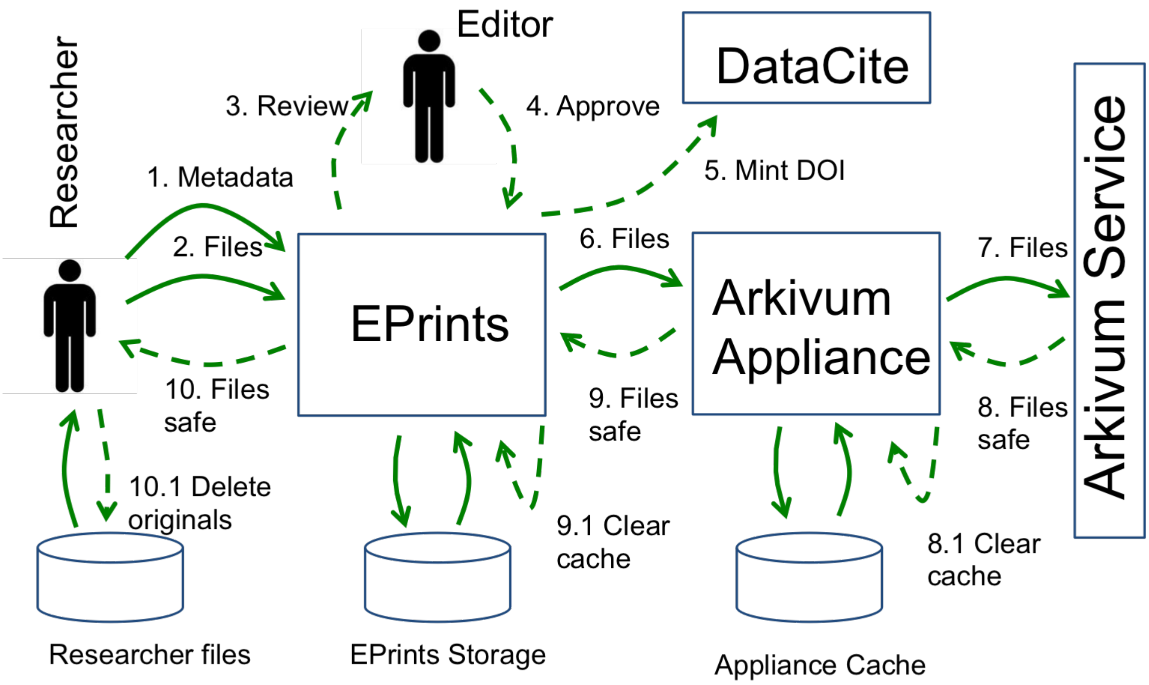


Figure 6. Deposit and archiving of research data using EPrints and Arkivum. http://dx.doi.org/10.6084/m9.figshare.1476831

1. The researcher creates a record for the research data in EPrints and adds descriptive metadata, e.g. conforming to a given schema such as recollect [ref].
2. The researcher adds files containing the research data to EPrints. EPrints stores these files locally on its server (EPrints Storage).
3. An Editor reviews the submission and works with the Researcher if necessary to ensure the deposit meets University policy, e.g. minimum metadata requirements or file formats.
4. The submission is approved.
5. DataCite is used to create a DOI for the dataset. The DOI resolves to the EPrints

repository.

6. The files are transferred to the Arkivum appliance (the gateway to the Arkivum archive service). The files are stored in the Appliance Cache. The files are checksummed and compared with the EPrints Storage copy to ensure transfer has been successful.
7. The files are encrypted (optional) and sent to the Arkivum Service for long-term retention.
8. The Arkivum Service confirms to the Arkivum Appliance that the files have been successfully archived (multiple copies in multiple locations). The appliance removes its local copy if needed, e.g. to free up space (8.1).
9. The EPrints server is notified that the files are now safely archived and it can remove its copy of the files if needed (9.1).
10. The researcher can use EPrints to see when their files have been successfully archived and can then chose to delete their local copy.

The workflow is designed so that there is a chain of custody between EPrints and the archive, which means it is possible to verify that files have been transferred successfully from one to the other. This chain of custody extends into the Arkivum Service. The use of checksums in EPrints means that the chain of custody can also be extended to the Researcher as well in order to achieve an end-to-end chain between the source of the data and its long-term storage locations. The original data and temporary copies in EPrints storage are not deleted until the data has been confirmed to be successfully archived and the chain-of-custody has been completed. This ensures that the original data is not removed or put at risk during the deposit workflow.

Subsequent access to data after it has been archived is shown in Figure 7.

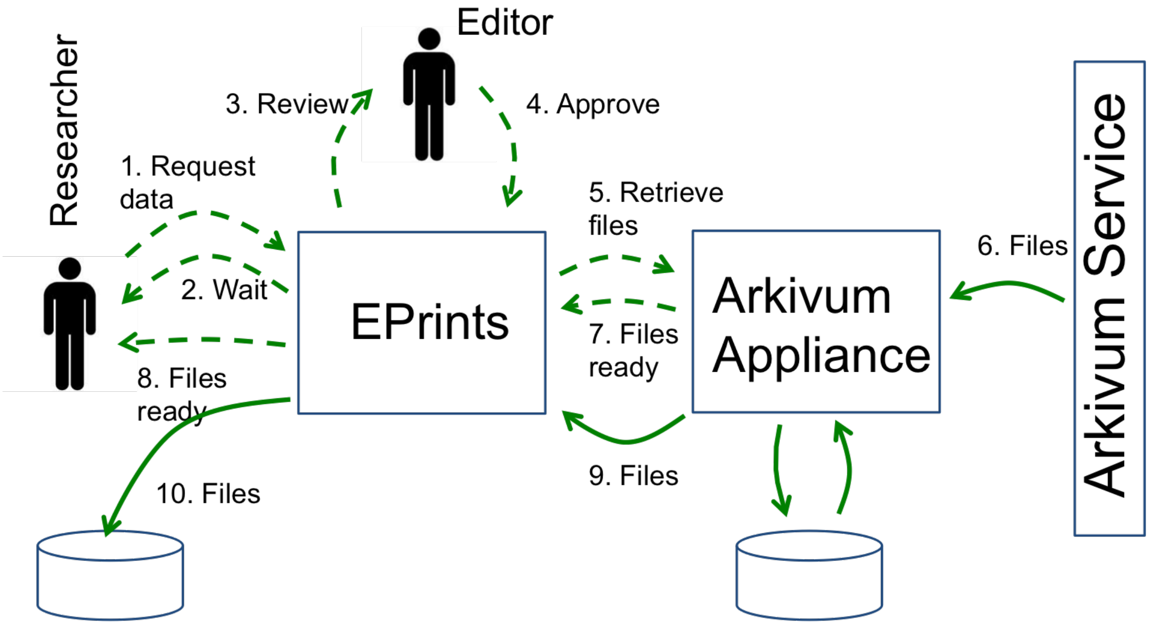


Figure 7. Access to archived research data through EPrints using a review/approve workflow.
http://dx.doi.org/10.6084/m9.figshare.1476831

1. A Researcher tries access to some research data through EPrints (e.g. they have followed a DOI in a publication that resolves to an EPrints landing page for the data). The researcher is told that the data is not immediately available and they need to submit an access request. The Researcher includes a simple justification of why they want to access the data.
2. The Researcher is informed that they need to wait for access to be approved, e.g. because the data is restricted or because the dataset is large and it will take time to restore and be made accessible.
3. The Editor reviews the request and works with the Researcher if necessary to clarify why they need the data and what they will use it for.
4. The request is approved by the Editor.
5. The EPrints server makes a request to the Arkivum appliance for the data to be retrieved from the Arkivum Service.
6. The files are retrieved and held in the Appliance cache.
7. EPrints is notified so it knows that the data is now available from the appliance.
8. The Researcher is told that the files are now ready to be accessed.
9. The Researcher accesses the data in the same way that they tried in Step 1 but this time they have direct access to the files. The files are streamed from the Arkivum appliance through the EPrints server and on to the Researcher.
10. The Researcher saves the files on their local storage so they can be used.

Note that some of the steps above may not always be necessary, e.g. the data is open access and there are no restrictions requiring access justification, or the data is small and is already in either EPrints local storage or on the Arkivum appliance and hence can be served immediately.

### Deposit via an uploader tool

In the case of large datasets, deposit and access through EPrints is not necessarily the most appropriate workflow. An alternative workflow is shown in Figure 8. This is not yet implemented in the current EPrints/Arkivum plugin. The workflow is at the design stage and is the result of discussions between ULCC, Arkivum and a range of EPrints users.

1. The Researcher uses an ‘Uploader’ tool to prepare a dataset for deposit. The dataset consists of files and metadata. Files are on the Researcher’s local storage. Metadata is entered by the Researcher to describe these files. When the Researcher has identified the files to be archived and has added the metadata then they hit an ‘Upload’ button.
2. The metadata is submitted to EPrints by the Uploader as part of creating an EPrints record for the dataset.
3. The files are copied from the Researcher’s local storage and are transferred directly to the Arkivum Appliance for archiving. The files do not get archived at this point and are instead held in a ‘pending’ state waiting for sign-off.
4. The appliance returns a URL to the files. This is a persistent URL that allows future access to the files.
5. The Uploader adds the URL to the record in EPrints so that EPrints knows

the location of the files in the Arkivum Appliance. This completes the EPrints submission.

6. The Editor is notified that a submission has been made. The Editor reviews the metadata.
7. The Editor uses the link to the data in the Arkivum Appliance to check that the data files are present and conform to University policy.
8. The Editor approves the submission.
9. EPrints notifies the Arkivum Appliance that the files have been approved. This releases the files from the 'pending' state.
10. The files are copied to the Arkivum Service and the files complete the archiving process as described in previous workflows.

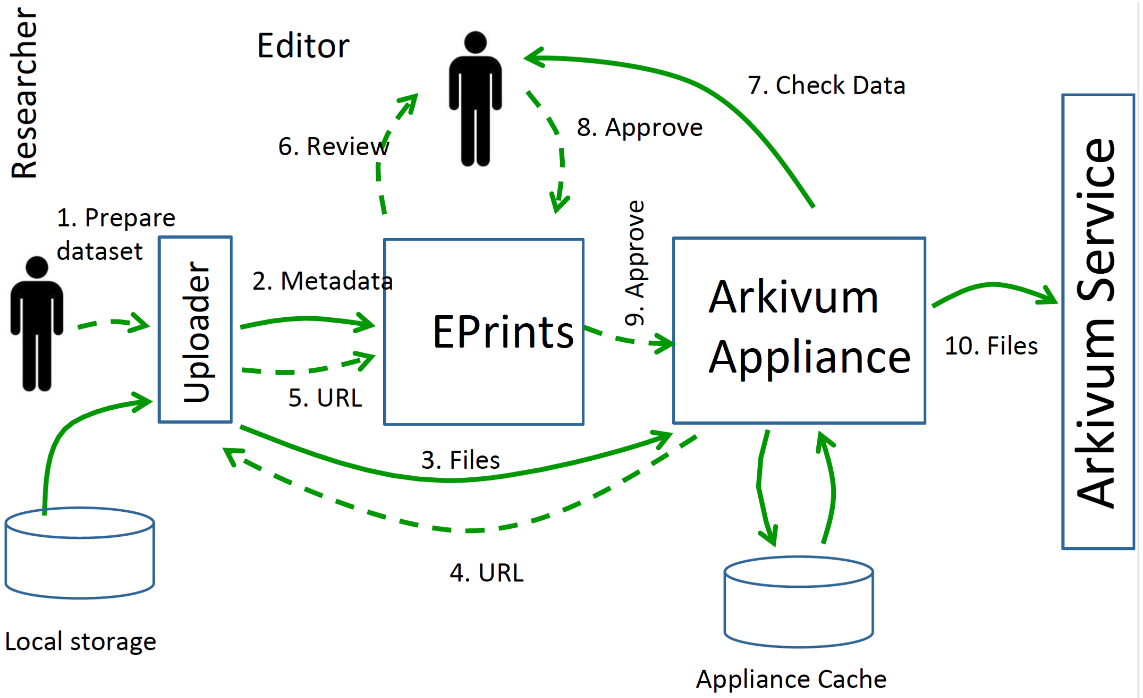


Figure 8. Direct deposit of large data sets to Arkivum using an uploader and linking to EPrints. http://dx.doi.org/10.6084/m9.figshare.1476831

The use of persistent URLs to the data in the archive and their embedding into the EPrints record allows the archive and the EPrints server to be in different physical locations and to be operated by different entities. For example, the Arkivum Appliance might be implemented at an institution because of the data volumes involved or because of security aspects of the data. The EPrints repository might be hosted by a third-party (e.g. as ULCC do for many of their customers). The model above allows the two to be joined together.

### Direct deposit and access

Conceptually, it is easy to imagine depositing files into a repository and moving them to an archive, or depositing files into an archive and then linking them to a repository record, but in practice there can be limitations on how easily this is achieved. Institutional Repositories and Research Information Systems are typically not designed to handle large volumes of data or large numbers of files. They don't tend to support data transfer mechanisms that are optimised for moving files between systems (for example, secure ftp, network file systems with domain integration, and Dropbox style private data stores such as OwnCloud [1]). 'Uploader' tools can help, but can also be a barrier because the Researcher needs to install and understand how to use this software.

The workflow below shows an approach where the Researcher is given a range of file upload and download options to data in the archive but with the Repository still remaining 'in charge' of access control and data visibility (e.g. embargos). The researcher creates the catalogue entry first in the Repository before uploading any data files. The repository will then present a number of options for getting data associated with that record into archival storage, and the researcher can choose the one which best suits their needs. This promotes the early creation of data catalogue entries and is also an opportunity to provide advice and instructions to the researcher for uploading their data files.

The workflow and description has been developed in collaboration with Tim Miles-Board and Rory McNicholl at ULCC. Implementation is planned for the next step of the Jisc CREST RDMS project.

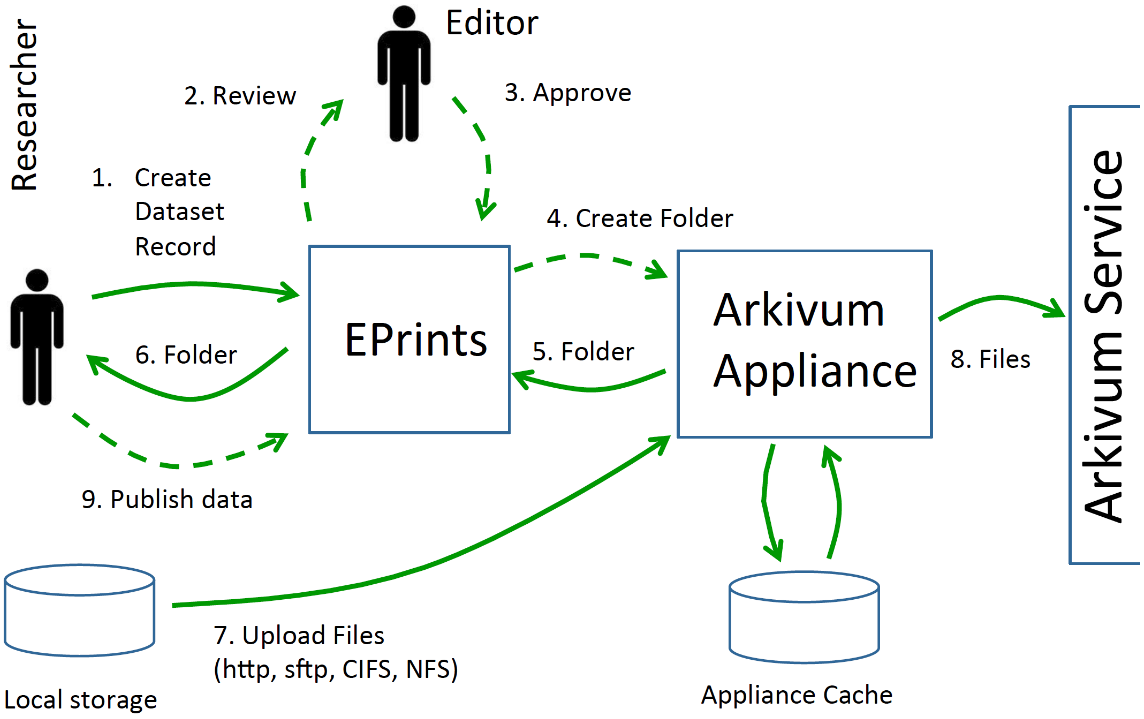


Figure 9. Direct upload to archive using data deposit folders. http://dx.doi.org/10.6084/m9.figshare.1476831

1. Researcher logs into EPrints (eg. using institutional single sign-on). The Researcher creates a new dataset record and enters metadata.
2. Researcher submits the dataset record, which may go into a review buffer or be made immediately live depending on institutional policy.
3. The Editor approves the dataset record.

---

[1] https://owncloud.org/

4. EPrints requests that a folder is created in the archive as the destination for the Researcher's files. The Arkivum appliance creates a permanent location for the files.
5. A unique and temporary folder name for the dataset location is returned to EPrints for use by the Researcher. This is a temporary handle that can be used to deposit data. The temporary folder name is time limited, e.g. it might expire after 30 days.
6. The folder name is returned to the user and they are given a range of options on how to upload their data. Options might include:
   a. Direct upload via EPrints. This option is best for uploading a small number of small files. EPrints uses the Arkivum API to transfer the file directly to the upload folder.
   b. CIFS. This option is best for uploading a large number of files or large files, but will likely only be available within the institution. The researcher copies their files to the upload folder using an smb:// URL provided by EPrints.
   c. Secure FTP. This option is best for uploading a large number of files or medium-to-large files. The researcher uses FTP client software to copy their files to the upload folder using an ftp:// URL provided by EPrints.
   d. ownCloud. This option is best for uploading a small number of medium-to-large files. The researcher uses the ownCloud web client to upload their files to the upload folder using a URL provided by EPrints.

Note that in each case the Researcher can (and should) provide a md5sum checksum of each file - this can be used to verify that the copy received is identical to the copy stored on the researcher's local storage.

7. EPrints emails a summary of these options, and a reminder of the upload expiry date, to the Researcher.
   a. The Researcher can add files to the temporary upload folder using any of the above options at any time before the expiry date. Note that temporary upload folders are only available for a fixed, configurable, length of time e.g. 30 days. The Arkivum appliance periodically checks and removes any expired upload folders. This does not affect the underlying persistent location for the data that is always visible to EPrints.
   b. EPrints sends regular email reminders to researcher to "complete" the dataset record
8. Files are replicated to the Arkivum service for safe storage.
9. After uploading is completed, researcher returns to EPrints to complete the dataset record:
   a. Researcher specifies how each file in the dataset can be accessed, for example:
      i. Publicly available (perhaps after an embargo period)
      ii. Available within UK federation only (where available)
      iii. Available within institution only

b. Researcher verifies files are all present and correct, e.g. using checksums where checksums were not provided at point of upload.

c. Researcher marks dataset record as completed.

d. EPrints makes the record/data accessible. This could involve a final review/approve step by the Editor (not shown).

The diagram below shows how the upload process might look from a Researcher perspective:

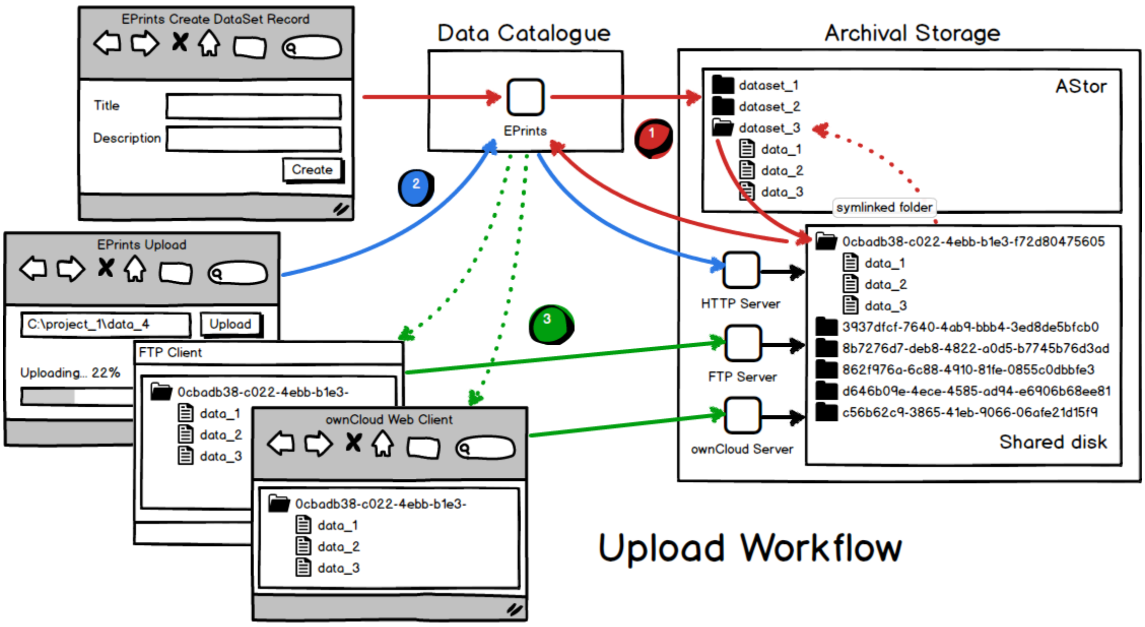


Figure 10. Researcher upload user interface mock-up (Created by Tim Miles-Board at ULCC). http://dx.doi.org/10.6084/m9.figshare.1477826

1. A researcher creates a new dataset record in EPrints - behind the scenes EPrints instructs the Arkivum appliance (the Arkivum Service) to create a new folder to receive the data files. After creating the dataset folder, the appliance creates a temporary upload folder (symlinked to actual dataset folder) and returns the name of the temporary folder to EPrints.
2. Since EPrints knows the name of the temporary folder, the researcher can upload data files directly to EPrints - EPrints will transfer the file to the upload folder using the Arkivum Service HTTP upload API.
3. Alternatively the researcher can get a URL from EPrints that will allow them to upload files using, for example, an FTP client or the ownCloud Web client.

If data set needs to be accessed, then the process is very similar to upload and works through temporary download folders and links (in a very similar way to how this is supported in Dropbox or other data sharing platforms). This allows a 'view' to be created over the underlying data in the archive so that just the subset that needs to be accessed is visible and downloadable.

The download process is as follows:

1. External user discovers dataset via Web search engine, or by searching /

browsing the data catalogue directly

2. If the dataset is “complete” (see upload workflow above), then EPrints shows metadata and lists all files in dataset (with checksums and associated access restrictions/embargoes)
3. External user clicks “Select files to download” button
4. External user prompted to login (if not already authenticated) or continue without authentication.
5. EPrints lists files that external user is authorised to download.
6. External user selects desired files.
7. EPrints informs the Arkivum Service that selected files can be downloaded.
8. The Arkivum Service creates a temporary folder for the download, containing links to the actual files which the user selected, and returns the temporary folder name to EPrints. Note that temporary download folders are only available for a fixed, configurable, length of time e.g. 30 days. The Arkivum appliance periodically checks and removes any expired download folders.
9. EPrints presents available download options to external user - the options presented will depend on the institutional setup but could include:
   a. HTTP. This option is best for downloading a small number of small files. The external user downloads the files individually from the download folder.
   b. Secure FTP. This option is best for downloading a large number of files or medium-to-large files. The external user uses FTP client software to copy files from the download folder using an ftp:// URL provided by EPrints.
   c. ownCloud. This option is best for downloading any number of medium-to-large files. The external user uses the ownCloud web client to download files from the download folder using a URL provided by EPrints. OwnCloud also provides the option to download all files in the download folder as a zip file.

The diagram in Figure 11 below shows how the download process might look from a Researcher perspective:

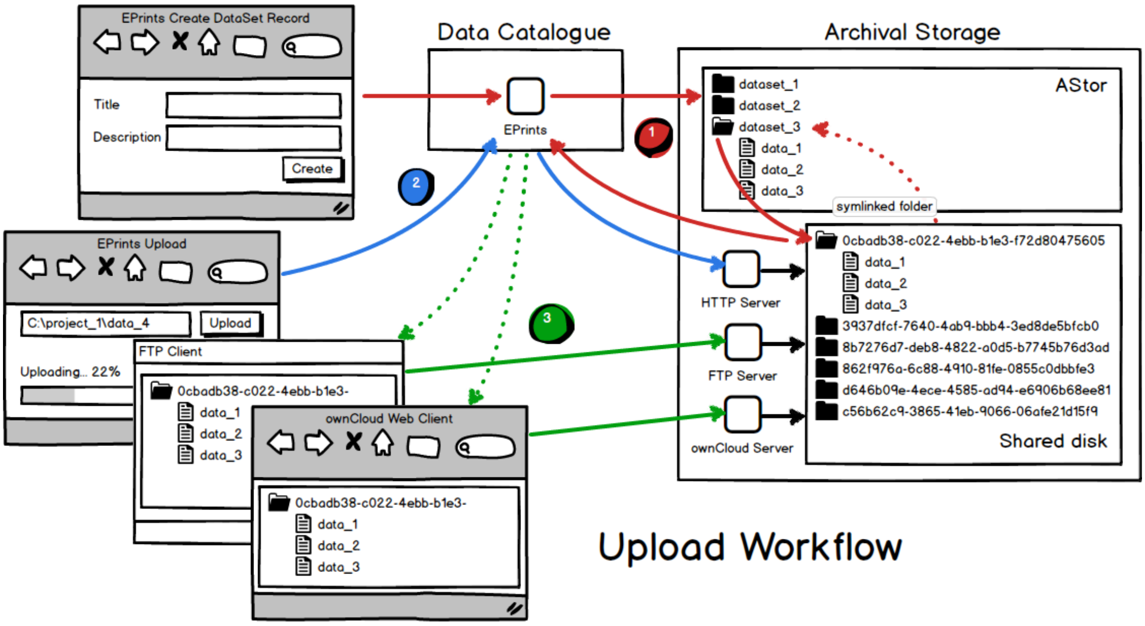


Figure 11. Researcher download user interface mock-up (Created by Tim Miles-Board at ULCC).
http://dx.doi.org/10.6084/m9.figshare.1477825

1. An external user requests a selection of files from a dataset - behind the scenes EPrints sends this information to the Arkivum appliance (the Arkivum Service) which creates a temporary download folder containing links to the actual files selected. The name of the temporary download folder is returned to EPrints.
2. Using the download folder name, EPrints can provide URLs to allow the external user to gain access to the folder over HTTP, FTP or via an ownCloud Web client.

## 5.5 Workflow for linking separate data and publication repositories (EPrints)

### Overview

Many institutions have an existing repository for their publications that has been in use for some time. Rather than modify the publication repository to accept research data, institutions sometimes choose to set up a separate data repository for research data and then link the publication and research data repositories together. The University of East London (UEL) and the London School of Hygiene and Tropical Medicine (LSHTM) are currently developing this model using EPrints for both the publication (ROAR at UEL [1], LSHTM Research Online [2]) and data repositories (data.uel at UEL [3], LSHTM Data Compass [4]). The benefits are that each repository can be implemented and optimized for the different characteristics of the content it holds and the community it serves. The challenges are linking the repositories together in a way that ensures consistency and has minimal burden on the researcher. This use case shows the workflows that are supported in the current implementation.

### Key points

- Adding a data repository to operate alongside an existing publication repository can be an attractive way to support research data management with minimal disruption to existing infrastructure and processes.
- For data and publications that are created and stored separately, it can be difficult to determine automatically what connections exist between the two. This knowledge is typically with Researchers and hence requires them to actively create the links. This is potentially a brittle process.
- Data may get deposited in multiple locations, e.g. in discipline specific or funder-mandated repositories as well as within an institution. For locations that automatically create DOIs [5] (Dyrad, Figshare, EPrints using DataCite plugin, discipline repositories etc.) there is the potential for DOI proliferation. In the absence of tools and databases for cross-referencing DOIs, manual checks need made by librarians or research support staff, or knowledge needs to be extracted from Researchers on where data has been deposited and which publications relate to it.

### Considerations for RDM as a service

- Data deposit and publication deposit may sometimes be done by different stakeholders, e.g. Researchers or Library staff. Workflows need to be supported to allow administrative or support staff to deposit on behalf of research staff.
- Synchronisation of metadata between repositories should be done in a way

---

1. http://roar.uel.ac.uk/
2. http://researchonline.lshtm.ac.uk/
3. http://data.uel.ac.uk/
4. Data Compass is currently under development. A test version is here: http://w01.lshtmdrtest.da.ulcc.ac.uk/
5. At the time of writing (11 June 2015) there are 1260 repositories registered with re3data.org. Of these, 388 (30%) are tagged as supporting some form of persistent identifier (ARK, DOI, HDL, PURL, URN, other). See DataCite schema for more information on identifier types (http://doi.org/10.2312/re3.007)

that is automatic, works in both directions, and is done transparently to the users of the system.

- The ability of Researchers to deposit data into multiple locations, e.g. an institutional repository and a discipline specific repository, can mean that multiple DOIs get generated. Automatic DOI checking and cross-referencing is needed to ensure copies of the data are linked and are recognised as manifestations of the same thing.

**Specific workflows**

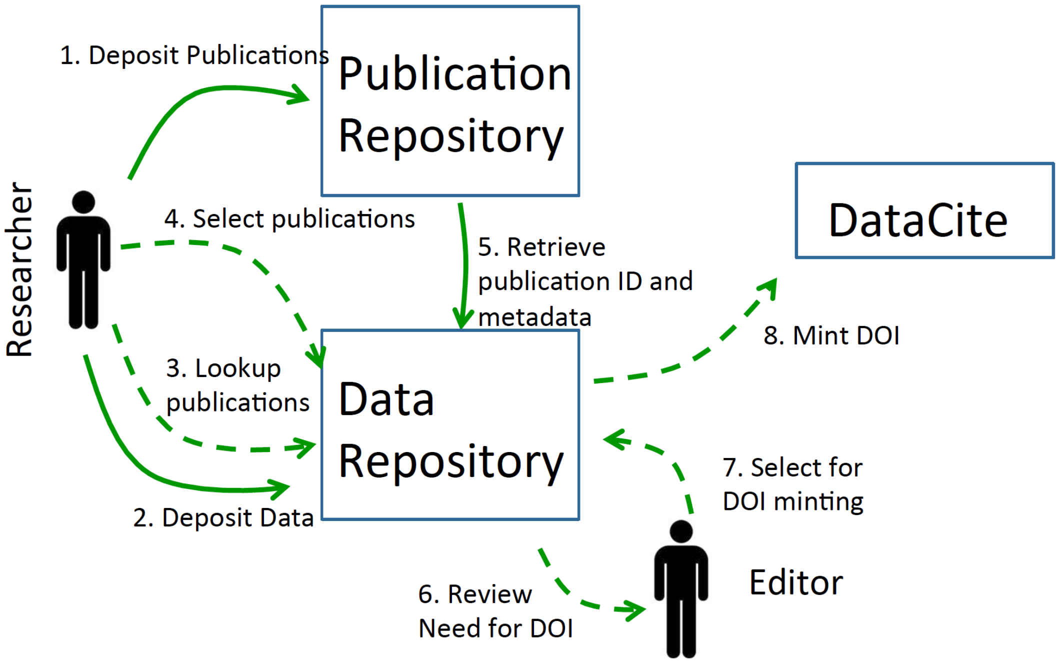


Figure 12. Deposit workflow. http://dx.doi.org/10.6084/m9.figshare.1476831

1. Researcher deposits one or more publications into the Publication Repository.
2. Researcher deposits data into the Data Repository
3. As part of the data deposit process, the Researcher can lookup relevant publications in the Publication Repository. This is done through the data repository. For example, the Researcher enters a publication title or keywords and the data repository will retrieve a list of matches from the Publication Repository.
4. The Researcher selects matching publications.
5. The Data Repository pulls metadata about the matching publications from the Publication Repository, e.g. Title and ID and stores these in the Data Repository as part of the metadata about associated publications.
6. The Editor of the Data Repository reviews the data and (potentially in collaboration with the Researcher) identifies if the data already has or needs a DOI.
7. If the data has no DOI then the data is selected to go through the DOI minting process. This selection is done on a per item basis.
8. DOIs are minted through the Data Cite plugin.

The steps above may not necessarily happen in this order. For example, data could be deposited into the Data Repository in advance of publications. In this case, the data record would be subsequently updated when publications have been made based on that data.

Some steps may be done by institution staff rather than the Researcher. For example, the Researcher may inform the institution about the existence of research data they have created and then administrative staff, e.g. in the Library, might enter that data into the data repository. This reduces the burden on the Researcher and helps kick-start population of the data repository.

Note that metadata about data sets relating to a publication is not added to the Publication Repository. The association between data and publication is held in the data repository. This means that the schema for the publication repository doesn't need to be modified, which is advantageous in a 'brown field' site where the publication repository may have been in operation for sometime.

If a researcher finds a publication in the publication repository and wants access to the associated research data then the workflow below is followed.

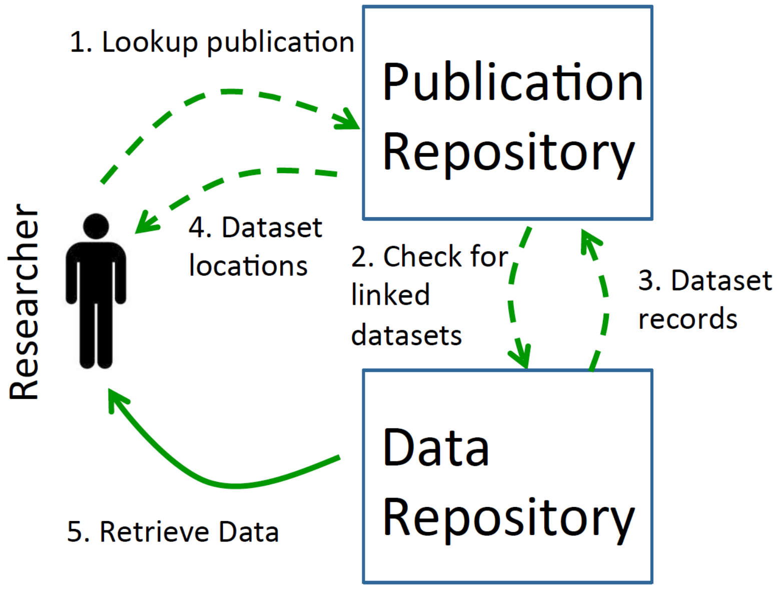


Figure 13. Data discovery and access workflow. http://dx.doi.org/10.6084/m9.figshare.1476831

1. Researcher finds publication(s) of interest in the Publication Repository, for example through a search or by using a DOI that resolves to the publication.
2. When the Researcher views the details of the publication, then the Publication Repository server contacts the Data Repository server and asks if there are any datasets associated with that particular publication.
3. The Data Repository server returns a list of matching datasets (it can do this because it holds a record of which publications are associated with which datasets).
4. The Publication Repository displays a list of matches in the web page presented to the Researcher. This includes links to the data in the Data Repository.
5. The Researcher accesses the data directly from the Data Repository.

The lookup of related datasets in the Data Repository is a dynamic process that happens in real-time when a Researcher is viewing the contents of the Publication Repository. This makes it seamless and transparent. It does however depend on the links between data and publications having first been made through the Data Repository.

## 5.6 Workflow for gathering and reporting usage metrics

### Overview

Institutions are interested in collecting and analysing metrics around the usage of their research data. Reasons can include: assessing and reporting the impact of the research, e.g. as may be required in future assessments of research quality following on from the RAE [1] and REF [2]; understanding if research outputs are compliant with funding body expectations and reporting to funders that they are demonstrably accessible and being accessed; and comparing usage statistics with other institutions to understand the effectiveness of an institutional repository or other access strategy. Statistics can be collected using a variety of tools and sources. These include Altmetrics [3], IRUS-UK [4], Google Analytics [5], repository plugins such as IRStats [6] for EPrints, publisher metric systems such as SciVal [7], and citation indexing such as Google Scholar [8], Scopus [9] from Elsevier, and Web of Science [10] from Thomson Reuters. The workflow below shows how some of these are supported and used in EPrints for the London School of Hygiene and Tropical Medicine.

### Key Points

- Metrics are available from multiple sources and more than one source is typically required to understand the real impact of providing access to research data as opposed simply knowing whether that research data has been accessed or not. For example, download statistics might be combined with citation counts.
- Metrics typically involves both harvesting information from external services, e.g. citation statistics, and pushing information to external services, e.g. IRUS.
- Metrics can take the form of detailed information on individual items, e.g. Altmetrics, or summary information on comparative performance, e.g. IRUS. Therefore, there needs to be support at all levels in a RDM solution, e.g. seeing the stats for a given item on the repository page for that item, or generating an overall usage report for internal monitoring of an RDM strategy within an institution.

### Considerations for RDM as a service

- Gathering and presenting metrics should be as automated and transparent as possible to avoid burden on Researchers or institution staff on collecting and collating metric information.
- Metric information needs to be made accessible to individual researchers and also institution staff, for example through customisable dashboards or

---

1. http://www.rae.ac.uk/
2. http://www.ref.ac.uk/
3. http://www.altmetric.com/
4. http://irus.mimas.ac.uk/
5. https://www.google.com/analytics
6. http://wiki.eprints.org/w/IRStats_2
7. http://www.elsevier.com/solutions/scival
8. https://scholar.google.co.uk/
9. http://www.scopus.com/
10. http://wok.mimas.ac.uk/

embedding statistics directly into repository web pages for research data items.

### Specific workflows

The diagram below shows how several metrics tools and services are incorporated into EPrints. Several of ULCC's EPrints customers are using will use all of these metrics.

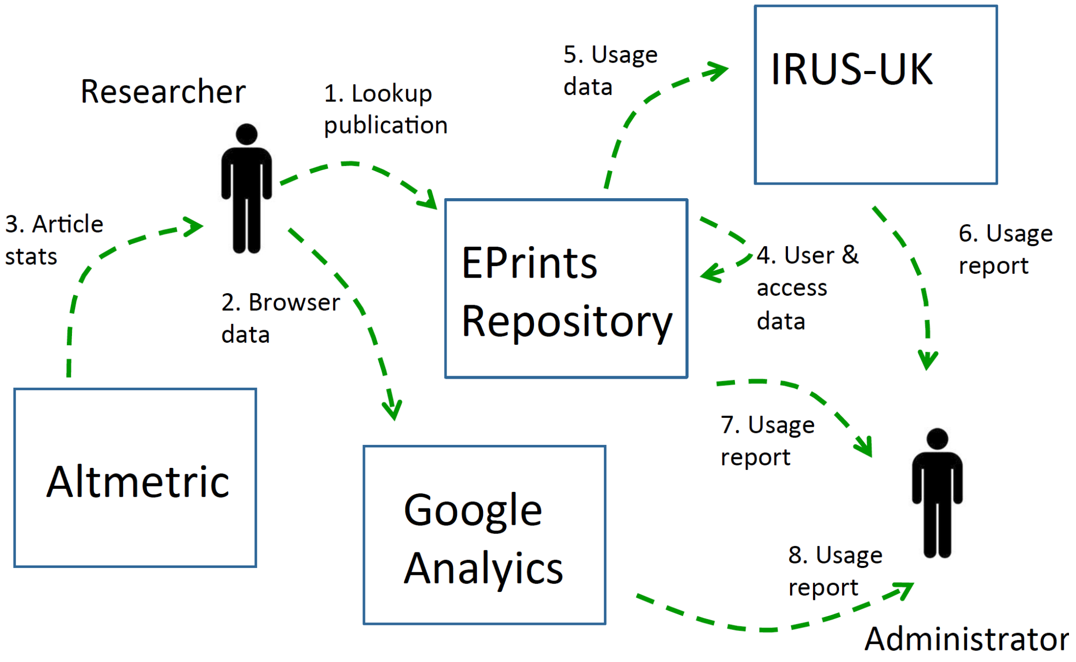


Figure 14. Metrics gathering and reporting. http://dx.doi.org/10.6084/m9.figshare.1476831

1. A Researcher looks up and retrieves a publication or dataset from the EPrints Repository
2. The Researcher's browser reports to Google Analytics that the publication or data has been accessed. This report includes details such as time of access, browser software being used, network address of the Researcher, and what page the Researcher was on before they accessed the data. The report doesn't include the Researcher's identity.
3. The Researcher retrieves statistics on the publication or data using Altmetrics. They do this by clicking on an Altmetrics 'badge' that is 'attached' to the publication/data. This triggers the Altmetrics service to provide a report on that specific item, e.g. in the form of pop-up window.
4. EPrints records that the publication or data has been accessed. This may include the actual identity of the Researcher if they have accessed the repository by logging in first.
5. EPrints reports summary usage information to the IRUS-UK service.
6. The Administrator, e.g. a member of the Library at an institution, accesses IRUS-UK to get a usage report for their repository and comparative statistics of usage compared to other repositories.
7. The Administrator uses the IRStats plugin to EPrints to get a report on repository downloads.

8. The Administrator uses Google Analytics to get a report on where downloads are being made from, where people are coming from before they get to the download (e.g. a DOI), and other client-side oriented statistics.

An extension to Steps 6-8 might be to combine them through the use of a ‘Dashboard’ (not currently available in EPrints) that puts all the information together on a single web page.

Step 3 might also be done by the Administrator and if they have an institutional account with Altmetrics then they can do this as a summary of all their research outputs rather than having to get the details of an item one at a time.

Altmetrics and Google Analytics work by embedding information about an object, e.g. its DOI or metadata, into a web page and then a script running in the user’s Browser that reports this information to the Altmetrics or Google server. In this way, reporting technically comes from the user and not from the EPrints server even though it might look like it is functionality being provided by EPrints.

## 5.7 Workflows for combining CRIS, IR, archive and data publication platforms.

### Overview

The University of Loughborough use Symplectic Elements for their CRIS system and have a publication repository based on DSpace. Researchers use Elements to deposit publications and this automatically creates a corresponding record in the DSpace repository. For data publication, Researchers find Figshare easier to use. In order to provide long-term protection and access to University research outputs, Loughborough use Arkivum for long-term research data storage. Rather than expecting Researchers to interact with all these systems separately, Loughborough have collaborated with Arkivum, Figshare and Symplectic (Figshare and Symplectic are both part of Digital Science) to create an integrated solution. From the Researcher perspective the objective is to simplify the workflow so that they can deposit research data and make it accessible using only one tool. Typically this is through Figshare as described below, but development is planned so that it can also be done through Symplectic Elements. Behind the scenes, the components of the system interact to ensure that metadata is exchanged, records are made in the institutional systems, and the University can track what is being done with the research data created by its Researchers.

### Key points

- Automated integration between systems allows Researchers to use the tools that they are familiar with (Figshare) and get immediate value from, whilst at the same time ensuring that the Institution can properly manage the research outputs of its Researchers.
- Incentives and benefits to Researchers, e.g. easy data sharing with others and fast DOI minting for publications, can provide a strong 'carrot' type of motivation. This operates alongside the need to comply with funder or institution policies, which can be perceived as more of a 'stick' by Researchers.

### Considerations for RDM as a service

- A 'one-stop-shop' for the Researcher with a simple user interface for data deposit makes it easier for the Researcher to simultaneously make their data accessible and conform to funder and institution policies on RDM.
- Integration between components needs to be 'bi directional' so that metadata synchronization happens both ways and more than one component can be used as a route for deposit, discovery or access to data. For example:
    a. Deposit data via Figshare or a CRIS.
    b. Discover data via an Institutional Repository or via Figshare.
    c. DOI linking from data in Figshare to an article in a journal plus link from the journal article back to the primary research data in Figshare.

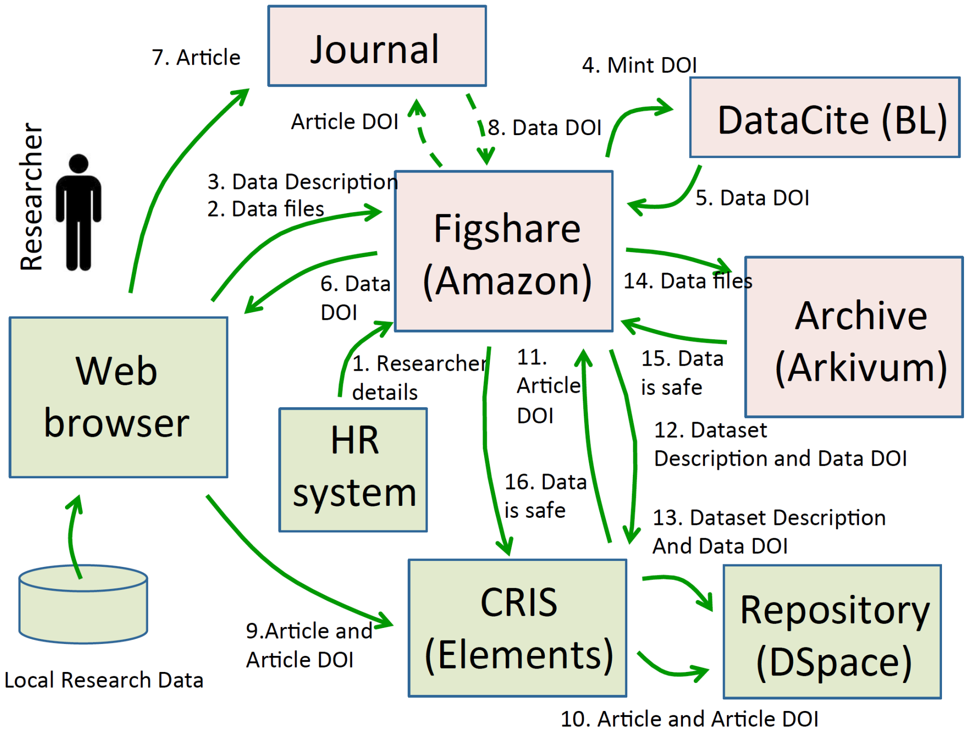


Figure 15. Figshare/Elements/DSpace/Arkivum deposit workflow. http://dx.doi.org/10.6084/m9.figshare.1476831

The green boxes show components that run at the University.

The pink components show externally hosted services or applications.

All interactions between the Researcher and the system is web-based and done through a web browser.

1. The University HR system exports user information to Figshare so that Figshare accounts can be set up automatically for each Researcher. This means the Researcher doesn't have to go through a registration/sign-up process.
2. The Researcher uploads data files to Figshare
3. The Researcher adds metadata to the Figshare dataset record.
4. Figshare requests a DOI for the dataset
5. A DOI is minted by DataCite through the British Library.
6. The DOI is retrieved by the Researcher so they can reference their data, e.g. in a journal article.
7. The Researcher publishes an article to a Journal. The Journal mints an Article DOI for the Article when published and this is made available to the Researcher.
8. The article includes the Data DOI to the underpinning dataset. The Data DOI resolves to Figshare. This allows readers of the article to get access to the data.
9. The Researcher adds an entry to the CRIS system for their publication. This includes a version of the article (e.g. accepted version) and the DOI to the article. The Article DOI resolves to the journal publisher.

10. The CRIS pushes the Article and the Article DOI to the Institutional Repository. This ensures that the article is accessible under the University's access policy. Users of the Repository can use the Article DOI to locate the Article in the Journal.
11. The Article DOI and Article metadata is harvested by Figshare. This allows the dataset record to be updated with a link (Article DOI) that points to the Journal. Therefore, people discovering the data in Figshare can use the Article to get extra context.
12. The metadata on the dataset and the Data DOI are sent to the CRIS. This allows the CRIS to be updated to include a record of the Researcher's published data.
13. The CRIS pushes metadata on the dataset including the Data DOI to the Institutional Repository. This ensures that the Repository has a publicly accessible record of the data for the Researcher so that University access policy is fulfilled.
14. Figshare makes a copy of the data files in the dataset to Arkivum so that they can be archived.
15. Arkivum replicates and safely stores the data. Arkivum reports back when the data archiving process is complete. This allows Figshare to update its metadata on the archive status of the data.
16. The CRIS system is updated with the archive status of the data so that the University has a record of archiving being completed and the location of the data within the Arkivum service. The Researcher can see this archive status information (or through Figshare).

There are many variations and extensions possible for the workflow above. For example

- Data might be initially uploaded to Figshare but access embargoed until the corresponding article is published. Figshare supports the minting of DOIs for data in advance of the data being made public.
- There could be a 'review/approve' type workflow to include quality control steps, for example when records are added to the CRIS system.
- Depending on the size of the dataset and frequency of access, the copy on Figshare might be removed after the data has been successfully archived. This still allows Figshare to analyse the data and create proxy/visualisation versions at the time of deposit, but data is then removed from Figshare to reduce ongoing storage costs. The data still remains accessible through Figshare at all times and is restored on demand from the archive.

Direct upload of data to Figshare is not necessarily the best workflow for large datasets or for datasets that do not need to be made immediately accessible. For example, a research project might create large volumes of primary data (e.g. output of scientific instruments, video recordings, multiple simulation runs) that isn't going to change and needs to be held securely for the duration of the project. The Researcher might want to archive this data as they go along, but only release it for public access towards the end of a project when summary research findings are available. In this case, the workflow might create a holding record in Figshare and upload the corresponding data files directly to the archive. An example is shown below.

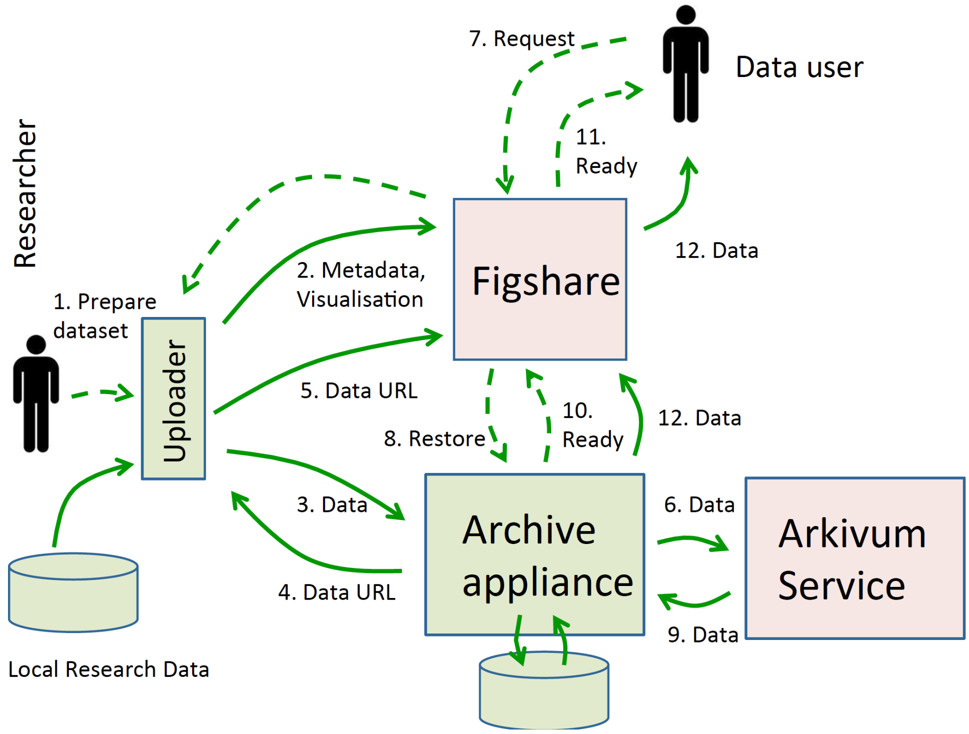


Figure 16. Deposit using an uploader followed by download through Figshare. http://dx.doi.org/10.6084/m9.figshare.1476831

The green boxes show components that run at the University.

The pink components show externally hosted services or applications.

1. The Researcher has a large dataset that they want to archive and make accessible.
    a. The Researcher creates a record in Figshare for the dataset. Instead of using their browser to send data directly to Figshare, they download a special Figshare 'desktop uploader' for handling large data files.
    b. The Researcher uses the desktop uploader to prepare a dataset. This includes selecting files on their local research data store and adding descriptive metadata.
2. The desktop uploader processes the files locally to create proxy versions/visualisations as required for presenting the dataset on Figshare. This is sent to Figshare along with the metadata to create a dataset record. The record might be embargoed or be made public at this point.
3. The desktop uploader sends the data files directly to the Archive appliance. The appliance is running locally at the institution. This allows the transfer to be done quickly over the local network. The appliance stores the files in its local cache.
4. The appliance returns a persistent URL to the files.
5. The desktop uploader sends the data URL to Figshare. This means that Figshare has a record of where it can access the data.
6. The appliance sends the data to the Arkivum service for long-term archiving. When the data is securely archived, the copy in the appliance local store is removed.

This completes the data deposit part of the workflow. Some of the details are omitted for simplicity, e.g. confirmation of archiving and updating the CRIS or IR.

The next part of the workflow (actually a separate workflow but shown on the same diagram) describes what happens when someone else wants to access and use the dataset.

7. A data user discovers the dataset through Figshare and wants to access the original data files from the archive. The files are not immediately accessible through Figshare at this stage and will take some time to retrieve from the archive. The data user makes a request to access the data set and is informed that they will need to wait until the data is ready.
8. Figshare makes a request to the Archive appliance for the data to be restored to the appliance cache.
9. The appliance retrieves the dataset from the Arkivum service and stores it in the local cache.
10. The appliance notifies Figshare that the data is ready for access (or Figshare periodically checks-in with the appliance to see if the data is ready).
11. Figshare notifies the user that the data is ready for access (or the user checks-in with Figshare periodically to check the availability of the data).
12. The user accesses the data. Access could be:
    a. The data is streamed directly from the appliance through Figshare and on to the user.
    b. The data is copied from the appliance into a Figshare cache and served from there
    c. The user accesses the appliance directly (e.g. if they are inside the same institution/group as the original data depositor).

If the data needs to be accessed repeatedly then it might be held in the Figshare cache rather than retrieved from the appliance each time.

## 5.8 Workflows for data presentation built upon data repositories (RDIVA)

### Overview

The use of institutional repositories, discipline specific repositories, or data sharing platforms such as Figshare don't always provide the necessary features to make research data as accessible or usable as the Researcher might like. One example is in the visual arts where a bespoke and often external website might be created around a specific research project and is used to present the results of that research in the best possible way online [1].

The 'presentation' website provides important context and in some cases may also be considered as part of the overall work. This brings with it several challenges. If the website is the primary means of access to the research outputs, then there is a risk is that it becomes the only home for that research data but the website may not have been designed with long-term sustainability in mind. If the website presents a 'dissemination' version of the content that is derived from the primary data rather than the primary data itself, e.g. web resolution images, transcoded video, compressed audio, then the researcher may 'forget' to archive the primary data. In addition comes the challenge of how to capture and preserve the website itself if the extra context adds long-term value to the data. Overall, unless carefully managed, the use of external websites as a primary means of access can put the long-term accessibility of data at risk if the data is 'outside of the control 'of the institution where the Researcher resides.

An alternative is to make an Institutional Repository the primary means of access to the data. This is also problematic because of the effort needed to customise the user interface to create all the functionality of a bespoke website. More simply, it may not be possible at all because the research output is part of a large collaboration, exhibition or other broader activity that is beyond the scope of the institutional repository.

One strategy used by institutions is to put data into an institutional repository and link-out to external websites. Examples can be seen at the University of the Arts London (UAL) where there are entries for research projects in the institutional repository [2] (UAL Research Online) which link to external websites that present the work in its full context and are the primary means of access [3]. Problems with this approach are: (a) ensuring all the data does get deposited into the repository by the Researcher, (b) consistency between the website presenting the data and the content of the institutional repository, which provides the long-term record, (c) there is typically no link from the website to the repository so the user can get to the definitive record of the primary data, and (d) how the institution can track and gather statistics on the impact of the research.

---

[1] For example, this scenario was discussed at the recent RDIVA workshop that looked at how RDM could be applied in the visual arts http://www.crest.ac.uk/gathering-rdiva-requirements/

[2] http://www.arts.ac.uk/research/current-research/

[3] Current Examples from the University of the Arts London (UAL) include the Textile Toolbox (http://www.textiletoolbox.com/) and Mark Making: The Arts in Dementia Care (http://markmaking.arts.ac.uk/). These are summarized in the UAL repository with outbound links to the full websites. The corresponding repository entries are http://www.arts.ac.uk/research/research-projects/current-projects/future-fashion/ and http://www.arts.ac.uk/research/research-projects/current-projects/arts-in-dementia-care/

An alternative approach is to build a website on top of the institutional repository so that the primary data is either accessed in a transparent way directly from the repository, or if a copy is held on the website then the website will report access to the repository, and in all cases there is a facility for the user to reference/access the repository version if they want to. A workflow to support this model is provided below.

### Key points

- In disciplines such as the visual arts where the presentational aspects of access to results outputs is very important, there is often the need to use bespoke external resources or services. The role of the institutional repository becomes one of holding the original research data so it can feed these external presentational vehicles rather than being the primary means of access.
- Making institutional repositories as easy to use and as attractive as possible for researchers in disciplines such as the visual arts is key to their adoption. In turn, this makes it more likely that they will be used to underpin other ways to present research results, e.g. external websites.
- Providing easy ways for Researchers to use/link content in institutional repositories when collaborating with external organisations, e.g. website developers, will help adoption of institutional repositories earlier in the research process.

### Considerations for RDM as a service

- The institutional repository needs to be easily extensible so it can be customised to support particular content types, e.g. image or video viewers, so it is easier to use in specific disciplines, e.g. visual arts.
- The institutional repository needs to facilitate reuse/export of content to external websites, e.g. by creating proxy resolution files, automatic link sharing, support for embedded players, and scripts for usage monitoring.
- The institutional repository needs to allow external sites to report back to the repository on the usage of external copies of research data, e.g. dissemination versions, especially if these are being accessed in lieu of the original.

**Specific Workflow** - In reference to Figure 17 below:

1. The Researcher is collaborating with a Website Developer on the presentation of their research work, for example on the design of a bespoke website.
2. The Researcher deposits the primary version of research data items (e.g. a video or an image) with their Institution’s Data Repository.
3. The Researcher provides the Website Developer with links (persistent) to the items in the Data Repository.
4. The Website Developer follows these links so they can access the data. The Website Developer might make dissemination copies for use in the website, e.g. lower resolution images or video, they might ‘embed’ a player or viewer that accesses the original direct from the repository, or they might include links directly to repository content.

5. In addition to extracting data files for use in the website, the Data Repository provides the Website Developer with ‘badges’, ‘metadata tags’, ‘javascript trackers’ and the like that the Website Developer can embed into their web pages. For example, this means that user access can be captured, or a user can be given a link to the original data, e.g. a DOI or URL.
6. A Consumer accesses the website, i.e. an end-user of the website.
7. Consumer access is tracked by the website and reported to the Data Repository, or, alternatively, the embedded scripts in the web pages cause the user’s browser to report access to the Data Repository (in a similar way to how Google Analytics works).
8. The Consumer decides they want to access the original data in the repository, e.g. to see a higher resolution version, or to see similar items that weren’t included in the website. For example, this information might be presented when they ‘hover’ over an image on the website. The Consumer follows the link to a landing page for the data or research project and makes a data access request In many cases, access requests may not be necessary, e.g. the data might already be open access because its been cleared for inclusion in the website.
9. An Administrator reviews the data access request (if appropriate).
10. The administrator grants access (again this might ne automatic).
11. The Consumer accesses the data.

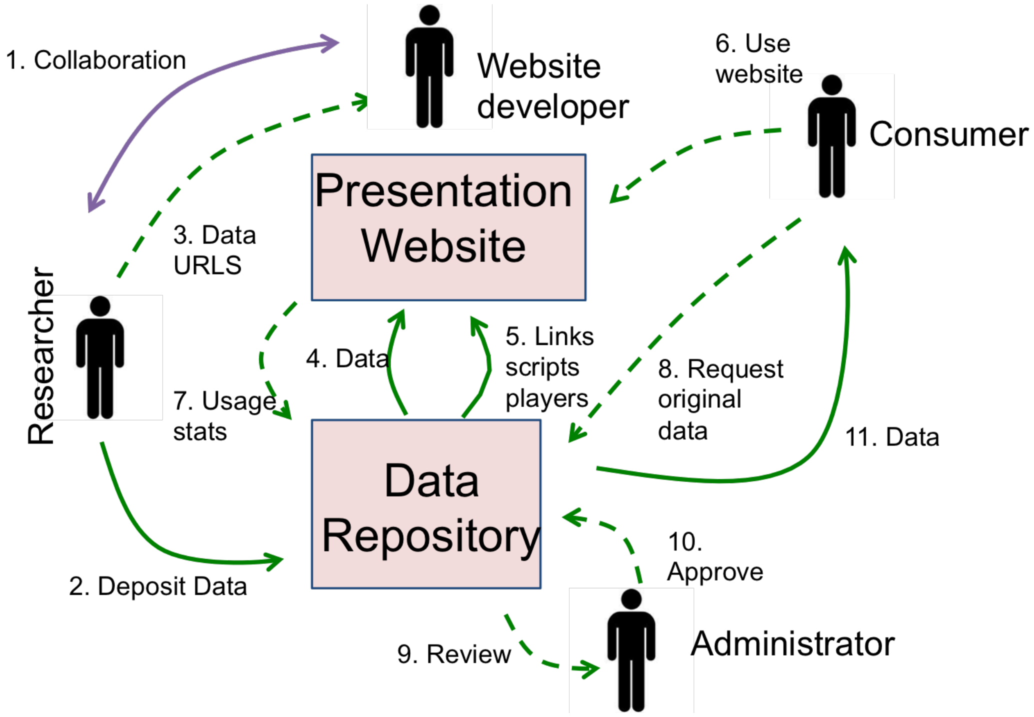


Figure 17. Possible workflow for building a presentation website for research data on top of an institutional data repository. http://dx.doi.org/10.6084/m9.figshare.1476831

The approach above ensures that the Researcher can still take advantage of using external websites to present their work whilst the institution still has some visibility of who is accessing the data. More importantly, the institution has a process that requires the Researcher to deposit data during the website creation process (or before) rather than at a later date when there is a higher risk of this being forgotten. This helps ensure that the repository has all the content being presented on the external website and that the repository and website versions are consistent.

## 5.9 Workflows for research data preservation

### Overview

Research data is potentially very long lived, especially where it is irreplaceable and supports long running research studies, for example climate data, astronomy observations, and population surveys. This data will only remain usable if it undergoes active digital preservation to ensure that the applications of tomorrow can successfully find, retrieve, and understand the research data of today. Long-term digital preservation can be a major challenge and there are dedicated organisations and resources available, including the Digital Preservation Coalition[1] and the Open Preservation Foundation[2], a wide range of tools for example as listed by COPTR[3], and frameworks for assessing maturing for example from the NDSA[4].

Digital preservation extends beyond data and includes the applications that create and consume that data. Digital Preservation[5] is "the series of managed activities necessary to ensure continued access to digital materials for as long as necessary" where access is "continued, ongoing usability of a digital resource, retaining all qualities of authenticity, accuracy and functionality deemed to be essential for the purposes the digital material was created and/or acquired for". In the context of RDM, research data is kept to ensure that any research outputs based upon it are repeatable and verifiable[6] and also because research data has value through sharing so it can be reused and re-purposed[7]. This in turn means preserving the software that creates and uses research data, with initiatives in this area including the Software Sustainability Institute[8] with its sustainability and preservation framework[9]. Institutional policies are starting to include software as well as data, for example the Imperial College policy states[10] "If software is developed as part of a research project, Principal Investigators must archive the particular version of the software used to generate or analyse the data in a repository and inform the Library of its location".

From a workflow perspective, there are increasingly well-defined workflows for doing preservation, especially for data. Examples include workflows based on the functional model of the Open Archive Information System (OAIS)[11], which can be manifested in the policies/procedures of an organisation, for example the Archive Training Manual from the UK Data Archive[12][13], or can embedded into software to automate these preservation processes, for example in Archivematica's[14] software workflow[15].

---

1. http://www.dpconline.org/
2. http://openpreservation.org/
3. http://coptr.digipres.org/Main_Page
4. http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_Levels_Archiving_2013.pdf
5. http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts
6. https://royalsociety.org/~/media/royal_society_content/policy/projects/sape/2012-06-20-saoe.pdf
7. http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report,_Jan14_v1-04.pdf
8. http://www.software.ac.uk/
9. http://www.software.ac.uk/resources/guides/sustainability-and-preservation-framework
10. https://workspace.imperial.ac.uk/researchservices/Public/Imperial%20College%20RDM%20Policy.pdf
11. http://public.ccsds.org/publications/archive/650x0m2.pdf
12. http://www.data-archive.ac.uk/curate/archive-training-manual
13. http://www.dcc.ac.uk/sites/default/files/documents/RDMF11/HERVE.pdf
14. https://www.archivematica.org/en/
15. https://wiki.archivematica.org/Workflows

However, there is a current challenge of how well digital preservation tools can support the specifics of research data and be embedded into an RDM infrastructure.

Workflows for using Archivematica as part of RDM provides a useful example with several UK Universities such as York and Hull actively testing out the tool in their institutions. The Jisc "Filling the Digital Preservation Gap" project[16] is considering the wider question of how Archivematica could be applied to research data[17] including some of the workflow aspects[18].

The benefits of using Archivematica stem from how it can be used to 'know what you have' and then being able to make informed decisions on what to do about different types of data.

In some senses, this is similar to Donald Rumsfeld's[19] description of 'known knowns' and most importantly 'known unknowns'. Rumsfeld is reported as saying[20]:

> *"Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tend to be the difficult ones."*

In this context, the uses of Archivematica are three-fold:

1. Unknown data formats (known unknowns). Archivematica has the ability to automatically identify a file format, e.g. by identifying it in the PRONOM registry[21]. However, for many research data formats, e.g. the outputs of laboratory instruments, Archivematica will know nothing about the data format because there is no entry in the registry. This is important information in its own right. This allows the institution to 'know what it doesn't know', i.e. know that there is an unknown data format being submitted to a repository. This might trigger the institution to ask a Researcher for more information, for example the name of a software that can read the data. Even if the format is unknown, Archivematica will also perform useful functions such as virus checking, generating checksums, and packaging files into 'bags' with manifests[22].
2. Known data formats (known knowns). If Archivematica does recognise the data format, then a record will be created of what the file is (ideally characterised and not just identified). This empowers the institution to assess (now or in the future) whether the format is in any way at risk. This is the bedrock of preservation. Even if an institution is unable to take any action or get further information from a researcher, then 'knowing what you know' is still a much better place to be.
3. Improved data formats (Doing better than known knowns). If Archivematica

---

[16] http://digital-archiving.blogspot.co.uk/2015/05/jisc-archivematica-project-update.html
[17] "Filling the Digital Preservation Gap, A Jisc Research Data Spring project. Phase One report - July 2015"
[18] http://digital-archiving.blogspot.co.uk/2015/06/the-second-meeting-of-uk-archivematica.html
[19] https://en.wikipedia.org/wiki/There_are_known_knowns
[20] http://www.defense.gov/transcripts/transcript.aspx?transcriptid=2636
[21] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
[22] https://en.wikipedia.org/wiki/BagIt

recognises the data format then it can also be configured to take action. This may require an institution to add support for specific formats to their Archivematica install. For example, Archivematica will characterise the data and then normalise it to a better format for preservation and/or access. The original is still kept. This is an additive way of protecting the value of the data whilst 'doing no harm' to the original. This is in effect a 'knowing what you know, and then knowing you are now in a better position'.

The benefits of all of these come at the cost of using and maintaining an Archivematica pipeline for research data. Whilst a lot of effort might be needed to achieve scenario 3, which may not be cost-effective, the use of a 'out the box' automated Archivematica system can drive the costs right down whilst still allowing benefits from 1 and 2.

Put simply, automated workflows that allow Archivematica to be incorporated into RDM can move an institution away from the dangerous 'unknown unknowns' where it doesn't know whether it has any understanding or not of the research data it holds. The aim is to move to a position of an institution knowing what it has and hence what it needs to do to ensure that the data is usable in the future.

### Key Points

- There is no 'one size fits all' for how preservation workflows should be incorporated into RDM infrastructures. The approach depends heavily on who is responsible for preservation, e.g. the Researcher or the Library, the type of data being preserved and whether it is well supported in existing tools, and generally the where the skills lie in an organisation on how handle and preserve data.
- Many data formats created in research are not well supported by current preservation tools. This is particularly true for scientific disciplines. This affects how and where preservation takes place within the overall RDM process, and, crucially, the extent to which Researchers need to be involved either at the start or on an on-going basis.

### Considerations for RDM as a service

- There needs to be support for preservation processes at different points in the data lifecycle, for example at point of ingest to a repository or as part of long-term curation and archiving.
- Preservation is an on-going activity and hence workflows should allow application of preservation actions over time, e.g. file format migrations, and not just 'one-shot' preservation at the time of deposit.
- Where researchers need to be involved, e.g. in data format conversion or software archiving, then the interfaces they use need to be simple, require minimal training, and ideally be embedded into the research tools and systems they use on daily basis.

### Specific Workflows

The Archivematica internal workflow involves one or more files being uploaded to an Archivematica server where they form a Submission Information Package (SIP). The SIP is then processed to create an Archive Information Package (AIP) and a

Dissemination Information Package (DIP) according to the OAIS model. Processing includes virus checking, identifying file formats, characterising the files, extracting metadata, conversion to normalised [23] formats for long-term preservation or access, generating checksums for fixity, adding user provided metadata, uploading the resulting DIP to a access repository such as AtoM [24] or an Institutional Repository, and transferring the AIPs to archival storage for long-term bit preservation.

The first example of how Archivematica could be used is shown below.

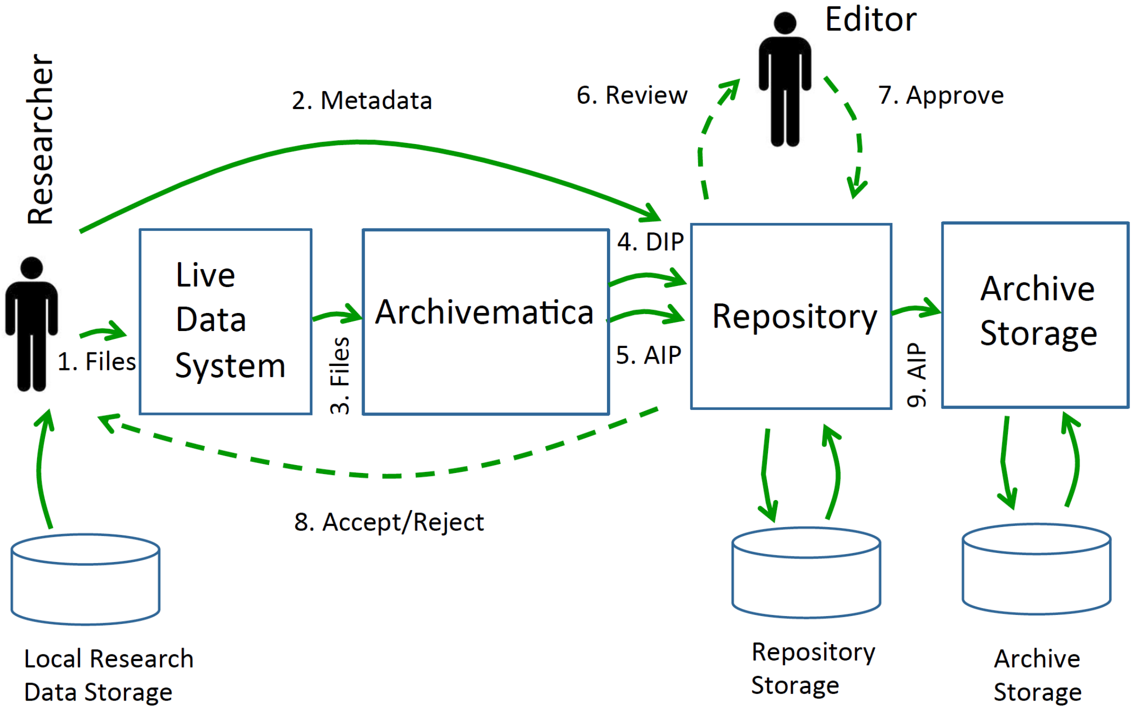


Figure 18. Archivematica used to prepare datasets for deposit into a Repository. http://dx.doi.org/10.6084/m9.figshare.1476856

1. Researcher data files are stored in whatever system (Live Data System) they use for storing that data during a project. This could be a document management system, a local storage server, institutional storage, HPC facilities, or something discipline specific such as a Media Asset Management (MAM) system or Laboratory Information Management System (LIMS). Files may be manually added to the Live Data System or they may be captured directly at source, for example from laboratory instruments or a Electronic Lab Notebook (ELN).
2. The Researcher creates a record in the Repository for a dataset. Alternatively, this might be the institution's CRIS system.
3. The files for the dataset are exported from the Live Data System and into Archivematica. This transfer might be automated through a Data Mover type tool, it might be done by support staff at an institution, or it might be done by copying data to a 'data holding area' within the institution and from there it into Archivematica. Due to the complexity of using Archivematica then it is unlikely that the Researcher will perform this activity themselves.
4. Archivematica creates a DIP and this is uploaded to the Institutional Repository

---

[23] https://wiki.archivematica.org/Format_policies
[24] https://www.accesstomemory.org/en/

and is stored in the Repository Storage. The DIP becomes part of the dataset record that the Researcher created.

5. Archivematica creates an AIP and this is uploaded to the Institutional Repository and is stored in the Repository Storage. The AIP becomes part of the dataset record that the Researcher created.
6. The Repository Editor reviews the contents of the AIP and DIP, e.g. to ensure they conform to the Repository minimum metadata requirements.
7. The Editor approves the submission. The DIP might then be used to create a record in the Repository and a publicly accessible version of the dataset. The AIP provides the definitive copy of the primary research data, including the original files as well as any normalised versions. This might also be made publicly accessible.
8. The Repository notifies the Researcher that the submission has been successful (or if not then it is rejected and the Editor and Researcher work to resolve any issues).
9. The AIP is copied to archive storage for long-term preservation. A copy might be retained in the Repository if the content is accessed frequently.

The workflow above gives the institution immediate visibility at time of deposit of whether the researcher's data is in a 'known' or 'unknown' format. For example, this allows them to work with the Researcher straight away on getting more information on the data. The workflow above might also be appropriate in cases where discipline specific input is needed into the preservation workflow and the tools used, for example normalisation of types of scientific data that are not in formats supported 'out of the box' by Archivematica's Format Policy Register (FPR)[25], i.e. not in FIDO[26] or PRONOM[27]. Archivematica in this case might be operated by a subject specific librarian or some other form of data expert, for example as part of technical support within a given research group or department.

The benefit of using Archivematica before content enters the Repository is that it pushes preservation actions closer to the source of the data where there is typically more expertise in the specific types of data involved and the data is 'fresh in the mind' of the researchers depositing it.

The downside is that this could create an extra burden on the researcher if they are asked to provide supplemental information, e.g. data format descriptions or software applications that can read the data. This creates extra barriers to deposit. These barriers can be partly lowered by integration with tools within the research environment, where this information may already be extant, for example in a scientific discipline this might be an Electronic Lab Notebook (ELN) or workflow engine (Taverna, Knime, Galaxy). In this way, data files and contextual metadata can at least be automatically gathered, albeit at the expense of implementing the integration.

An alternative workflow is to defer the use of Archivematica to later in the RDM process, for example before a dataset is permanently archived. This approach is shown below.

---

25. https://wiki.archivematica.org/Administrator_manual_1.0#Format_Policy_Registry_.28FPR.29
26. http://openpreservation.org/technology/products/fido/
27. http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

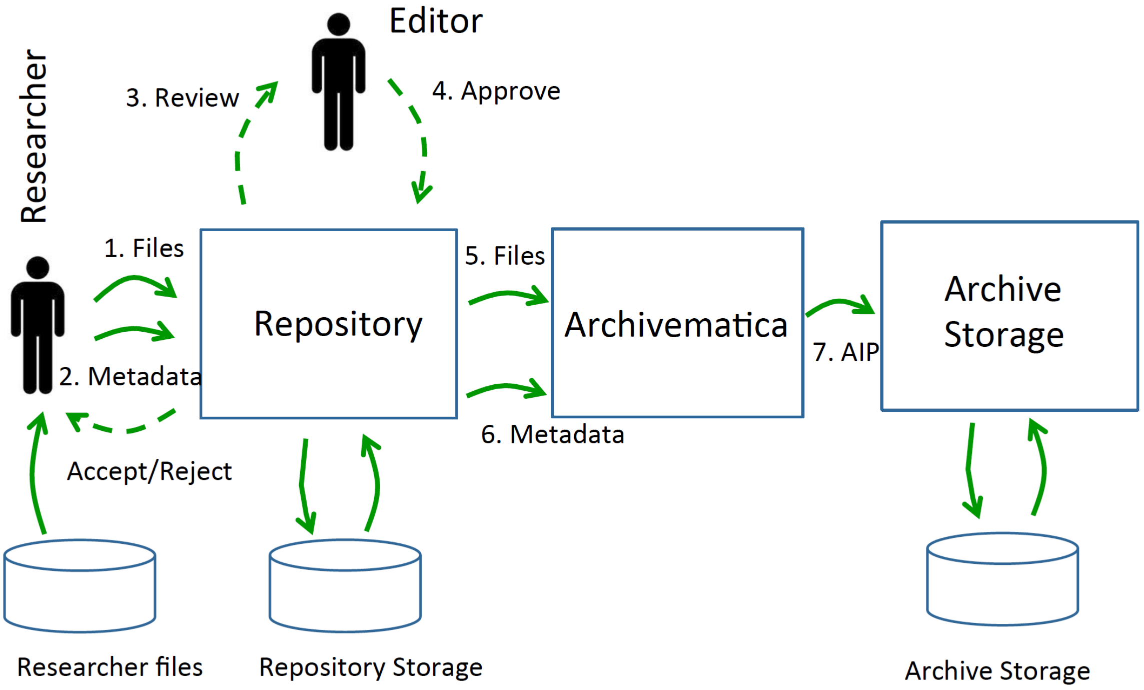


Figure 19. Use of Archivematica to create AIPs immediately before archiving. http://dx.doi.org/10.6084/m9.figshare.1476856

1. Researcher uploads data files to the Repository in the normal way. Alternatively, this might be the institution's CRIS system.
2. The Researcher adds descriptive metadata.
3. The Editor reviews the Researcher's dataset, e.g. against minimum repository requirements. This review process doesn't use Archivematica and is whatever standard procedure the institution might already operate.
4. The Editor approves the dataset and the Researcher is informed. At this point, the dataset might be made publicly accessible.
5. The Repository uploads the finalized version of the dataset to Archivematica
6. Metadata is added if necessary if not already in a Repository export.
7. Archivematica creates an AIP and this is set to the Archive for long-term storage.

In this approach, Archivematica is used to create the final preservation copy of the dataset which is then archived. File characterization, metadata extraction, and format normalization all help ensure that the dataset is usable in the future.

The benefit of this approach is that the normal Repository deposit workflow is unaltered and there is no additional burden on the Researcher. Archivematica is also only applied to approved datasets that have already undergone initial QC and have passed a decision to be archived.

The downside of this approach is that if there are problems with the dataset, especially if the point of archiving is considerably after the original deposit by the Researcher, then detailed knowledge of the dataset may no longer be available or repository staff may not have the expertise to solve the problems.

An intermediate approach in-between the preceding two workflows is shown below.

In this example, Archivematica is used within the Repository QC process and supports the Editor in checking that a deposited dataset is complete and correct.

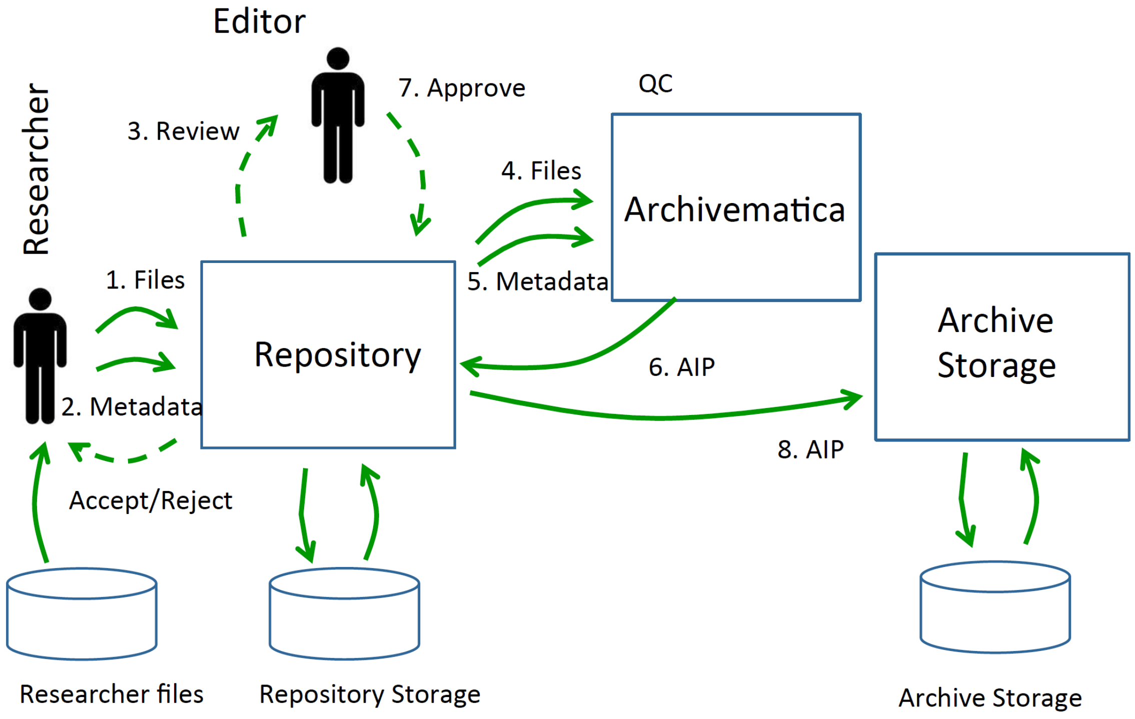


Figure 20 Archivematica used for research data quality control. http://dx.doi.org/10.6084/m9.figshare.1476856

1. Researcher uploads data files to the Repository in the normal way. Alternatively, this might be the institution's CRIS system.
2. The Researcher adds descriptive metadata.
3. The Editor reviews the Researcher's dataset, e.g. against minimum repository requirements.
4. As part of the review process, the data files are uploaded to archivematica
5. Metadata is added if necessary. Archivematica and the tools it applies is used to in effect perform quality control on the dataset, e.g. to flag any files that don't have identified file types or any files that don't conform to their file format specification.
6. Archivematica generates an AIP, which is returned to the repository and stored in Repository Storage.
7. The Editor reviews whether processing in Archivematica was successful and that the dataset is correctly represented by the AIP. The Editor then approves the Researcher's submission and the Researcher is notified.
8. The AIP is set to the Archive for long-term storage.

The benefit of using Archivematica within the QC process is that the Researcher isn't directly involved and hence their normal deposit workflow is not disrupted whilst at the same time ensuring that data passes through the Archivematica process very soon after deposit so that any issues can be resolved 'sooner rather than later' with the Researcher.

The downside of this approach is that the Editor using Archivematica will need experience of specific data formats and tools if there isn't support from Archivematica out of the box. A compromise is to use Archivematica as a way of detecting potential issues with a dataset, for example it contains files in an unknown format, but then to resolve these problems outside of Archivematica, for example by requiring the Researcher to add a description of the unknown file format to the metadata about the dataset that is then stored in the Repository.

Which of the three workflows above to use, or indeed a variation of them, depends on who is responsible in an institution for long-term preservation, e.g. the PI or the Research Office, and who understands the data and can perform preservation appropriately. For example, in STEM areas it will typically be scientists who understand the data and tools, including specific data formats, which tools to use for format validation and characterisation, and how to do normalization using open standards and community tools where possible. However, it is the Institution that may have long-term responsibility for the data, not least because it is an asset of the institution and not the 'property' of the researcher. This means the institution has a vested interested in data QA and QC and the can't rely on the scientist always being available or the scientists having the time and effort to support preservation activities (even if there is a strong argument that this is good research practice and should be part of the scientists' day-to-day work especially in data driven disciplines). The question becomes one of the cost of maintaining domain knowledge, or whether instead to 'do the best job at the time' and then put onus on the user to deal with any issues at some time in the future when they want to access and use the data. Different institutions will have different policies on this and hence will want to adopt different preservation strategies and workflows.

# 6 Acknowledgements

We would like to acknowledge support and input from the following people and organisations. This is not an exhaustive list – many of the workflows presented in this report have come from a wide range of ongoing discussions with institutions, user groups, and from Jisc/DCC/EPSRC events in the RDM area.

- Funding from the Jisc Research Data Spring.
- Inputs and discussions with project partners including CREST, Leeds Trinity University, University of the Creative Arts (UCA), and the University of London Computing Centre (ULCC).
- Feedback and suggestions from the members of CREST on earlier drafts and presentations of findings.
- Input, discussions, descriptions, clarifications and corrections from University of Southampton, Loughborough University, Imperial College London, University of York, University of Leeds, Figshare, Wiley, the UK Archivematica Users Group, and UK EPrints user group meetings.