



Big data: the potential role of research data management and research data registries

Joy Davidson, Sarah Jones and Laura Molloy
Digital Curation Centre (DCC)

www.dcc.ac.uk



Funded by:



The DCC Mission

“Helping to build capacity, capability and skills in data management and curation across the UK’s higher education research community”

Phase 3 Business Plan



RDM policy drivers for universities

ORGANISATION
FOR ECONOMIC
CO-OPERATION
AND DEVELOPMENT



declaration

data are a public good and should
be openly available



RESEARCH
COUNCILS UK

Common principles on data policy

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

EPSRC EXPECTATIONS

Engineering and Physical Sciences
Research Council

EPSRC has the following clear expectations of organisations in receipt of EPSRC research funding:

- i. Research organisations will promote internal awareness of these principles and expectations and ensure that their researchers and research students have a general awareness of the regulatory environment and of the available

The 'big data' challenge for universities



Often curated by large international projects, national facilities or disciplinary initiatives.



Some large data, but mostly a long-tail of small, distributed datasets with diverse needs.

Services to help unis



A web-based tool to help researchers write Data Management Plans

Jisc Research Data Registry and Discovery Service

A service to ensure that research data held within UK data centres and universities can be found, understood and reused



What is a data management plan?

A brief plan to define:

- how the data will be created?
- how it will be documented?
- who will access it?
- where it will be stored?
- who will back it up?
- whether (and how) it will be shared & preserved?

DMPs are often submitted as part of grant applications, but are useful whenever researchers are creating data.

Why create Data Management Plans?

- Plans help researchers to manage their data
 - Make informed decisions
 - Anticipate and avoid problems
- Many funders ask for DMPs in grant proposals
www.dcc.ac.uk/resources/data-management-plans/funders-requirements
- DMPs are often required by universities too
“All new research proposals must include research data management plans or protocols that explicitly address data capture, management, integrity, confidentiality, retention, sharing and publication.”



DMPonline

- Free to use
- Code is open source
- Helps researchers write DMPs
- Provides funder questions & guidance
- Provides help from universities
- Examples and suggested answers
- ...

<https://dmponline.dcc.ac.uk>

Registration

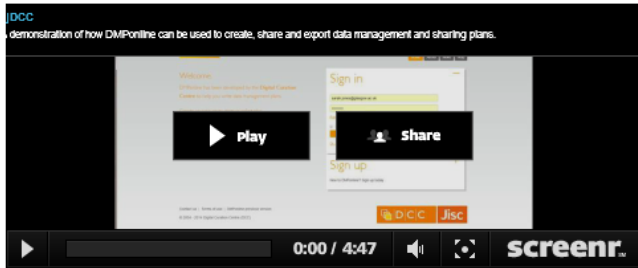


[Home](#) [About](#) [News](#) [Help](#)

Welcome.

DMPonline has been developed by the [Digital Curation Centre](#) to help you write data management plans.

Screencast on how to use DMPonline



Sign in



Sign up



New to DMPonline? Sign up today.

Email *

Organisation

Password *

Password confirmation *

I accept the terms and conditions *

Sign up

Sign up with your email address, organisation and password

Select 'other organisation' if yours is not listed

Creating a plan

DMP ONLINE

Signed in as Sarah Jones ▾

[View plans](#) [Create plan](#) [About](#) [News](#) [Help](#)

Create a new plan

Please select from the following drop-downs so we can determine what questions and guidance should be displayed in your plan.

If applying for funding, select your research funder.
Otherwise leave blank. [Not applicable/not listed.](#)

European Commission (Horizon 2020) ▾

To see institutional questions and/or guidance, select your organisation.
You may leave blank or select a different institution to your own. [Not applicable/not listed.](#)

University of Glasgow ▾

Tick to select any other sources of guidance you wish to see.

Generic guidance from the Digital Curation Centre

[Create plan](#)

Select funder (if any)

Select organisation for additional questions and guidance

Select other sources of guidance

Plan details: summary

This plan is based on:

Funder | Economic and Social Research Council

Answer questions

Export

The ESRC requires that all applicants seeking ESRC funding include a statement on data sharing in the relevant section of the Je-S application form. If data sharing is not possible, the applicant must present a strong argument to justify their case.

Sections	Questions
Existing data	<ul style="list-style-type: none">- An explanation of the existing data sources that will be used by the research project (with references)- An analysis of the gaps identified between the currently available and required data for the research
Information on the data that will be produced	<ul style="list-style-type: none">- Methodologies for data collection- Data volume and data type, e.g. qualitative or quantitative data- Data quality, formats, standards documentation and metadata
Planned quality assurance and back-up procedures (security/storage)	<ul style="list-style-type: none">- Quality Assurance- Back-Up
Management and archiving of collected data	<ul style="list-style-type: none">- Plans for management and archiving of collected data
Overcoming data sharing difficulties	<ul style="list-style-type: none">- Expected difficulties in data sharing, along with causes and possible measures to overcome these difficulties.
Consent, confidentiality, anonymisation and other ethical considerations	<ul style="list-style-type: none">- Explicit mention of consent, confidentiality, anonymisation and other ethical considerations
Copyright and intellectual property ownership of the data	<ul style="list-style-type: none">- Copyright and intellectual property ownership of the data
Responsibilities for data management and curation	<ul style="list-style-type: none">- Responsibilities for data management and curation within research teams at all participating institutions

Summary of the sections and questions in your DMP

Answering questions

My project (DCC Template) 1/13

Plan details **Generic DMP** Share Export

Data Collection (2 questions, 0 answered) +

Documentation and Metadata (1 question, 1 answered) -

What documentation and metadata will accompany the data?

B *I* Paragraph ☰ ☷ ☰☷

Metadata will be tagged in XML using the Data Documentation Initiative (DDI) format. The codebook will contain information on study design, sampling methodology, fieldwork, variable-level detail, and all information necessary for a secondary analyst to use the data accurately and effectively.

Save

Answered less than a minute ago by Sarah Jones

DCC Guidance -

Questions to consider:

- What information is needed for the data to be read and interpreted in the future?
- How will you capture / create this documentation and metadata?
- What metadata standards will you use and why?

Guidance:

Describe the types of documentation that will accompany the data to help secondary users to understand and reuse it. This should at least include basic details that will help people to find the data, including who created or contributed to the data, its title, date of creation and under what conditions it can be accessed.

Documentation may also include details on the methodology used, analytical and procedural information, definitions of variables, vocabularies, units of measurement, any assumptions made, and the format and file type of the data. Consider how you will capture this information and where it will be recorded. Wherever possible you should identify and use existing community standards.

Notes who has answered the question and when
Progress bar updates how many questions remain

Sharing plans

Withdrawal of services for young people

Plan details

ESRC Data Management Questions

Share

Export

You can share your plan to allow others to read or edit it. Please insert the email address of the person you wish to share it with. You can leave a note to explain why you are sharing the plan, or what you wish them to look at.

Collaborators

User name	Permissions	
Sarah Jones	Owner	
Laura Molloy	<input type="text" value="Read only"/>	Remove user access

Add collaborator

Permissions:

Add collaborator

Allow colleagues to read-only, read-write, or become co-owners

Co-writing DMPs

The screenshot shows the DMP ONLINE interface. At the top left is the logo for DMP ONLINE. At the top right, it says "Signed in as Laura Molloy". Below the logo are buttons for "View plans", "Create plan", "About", "News", and "Help". The main heading is "My project (DCC Template)" with a progress indicator showing "1/13". There are three tabs: "Plan details", "Generic DMP", and "Export". The "Generic DMP" tab is active. Below the tabs is a list of sections: "Data Collection (2 questions, 0 answered)", "Documentation and Metadata (1 question, 1 answered)", "Ethics and Legal Compliance (2 questions, 0 answered)", "Storage and Backup (2 questions, 0 answered)", "Selection and Preservation (2 questions, 0 answered)", "Data Sharing (2 questions, 0 answered)", and "Responsibilities and Resources (2 questions, 0 answered)". The "Documentation and Metadata" section is expanded and shows a message: "This section is locked for editing by Sarah Jones." Below this message is a question: "What documentation and metadata will accompany the data?" and an answer: "Metadata will be tagged in XML using the Data Documentation Initiative (DDI) format. The codebook will contain information on study design, sampling methodology, fieldwork, variable-level detail, and all information necessary for a secondary analyst to use the data accurately and effectively." The answer is attributed to "Sarah Jones" and is dated "7 minutes ago". At the bottom right, there is an "Export" button.

Signed in as Laura Molloy

View plans Create plan About News Help

My project (DCC Template) 1/13

Plan details Generic DMP Export

Data Collection (2 questions, 0 answered) +

Documentation and Metadata (1 question, 1 answered) -

This section is locked for editing by Sarah Jones.

What documentation and metadata will accompany the data?

Metadata will be tagged in XML using the Data Documentation Initiative (DDI) format. The codebook will contain information on study design, sampling methodology, fieldwork, variable-level detail, and all information necessary for a secondary analyst to use the data accurately and effectively.

Answered 7 minutes ago by Sarah Jones

Ethics and Legal Compliance (2 questions, 0 answered) +

Storage and Backup (2 questions, 0 answered) +

Selection and Preservation (2 questions, 0 answered) +

Data Sharing (2 questions, 0 answered) +

Responsibilities and Resources (2 questions, 0 answered) +

Export

Sections are locked for editing when they're being worked on by colleagues

Exporting DMPs

Can export as plain text, docx, PDF, html...

Withdrawal of services for young people

ESRC Data Management Questions

Existing data

Questions	Answers
- An explanation of the existing data sources that will be used by the research project (with references)	The ESDS archive has been systematically searched using a series of search terms related to and derivative of 'public service withdrawal', 'impacts', and/or 'children' and 'young people'. Our overall assessment is that there are no datasets that will adequately address the aims of this project. The following datasets are tangentially related to this project.
- An analysis of the gaps identified between the currently available and required data for the research	Given the contemporary nature of the proposed project, we know of no datasets that cover users' (especially young people's) views and experiences of austerity measures and service withdrawal. The proposed project will therefore capture new and unprecedented data, for which there is an evident demand among national and regional stakeholders, decisionmakers and service-providers (see 'Pathways to Impact' attachment). Moreover, whilst several datasets incorporate longitudinal data, none includes data gleaned from oral history and multigenerational family interview methods central to the proposed project. The proposed project therefore represents an extension to the methods and data quality of the tangentially-related projects listed in section 1. Finally, the novel use of 'impact' activities to generate data (e.g. workshops, mapping software) exceeds the scope of all extant and even tangentially-related datasets.

Information on the data that will be produced

Questions	Answers
- Data volume and data type, e.g. qualitative or quantitative data	The project shall generate new quantitative data (Bristol Online Survey outputs, SPSS data and outputs), qualitative data (digital audio files, audio transcripts, digital photographic and video data, workshop outputs, NVivo files), and mapping data (TIFF files). Metadata, in the form of pdfs and Excel spreadsheets, shall be used to facilitate the management and archiving of these data. Data shall be stored in password-protected folders on the host institution's secure servers. Data transfer between the PI and Co-I shall take place via face-to-face meetings.
	Quantitative data shall be generated from an anonymous online survey (target 10,000 responses). The survey will be administered via Bristol Online Surveys (BOS) software: a secure, quality-assured, widely-used online survey tool. Data will be exported to SPSS for analysis.

Unis can customise the tool....

- Add templates
- Add custom guidance
- Provide example or suggested answers
- Monitor usage in their organisation
- Offer foreign language versions
- ...

www.dcc.ac.uk/news/customising-dmponline-admin-interface-launches

Ways libraries could use DMPonline

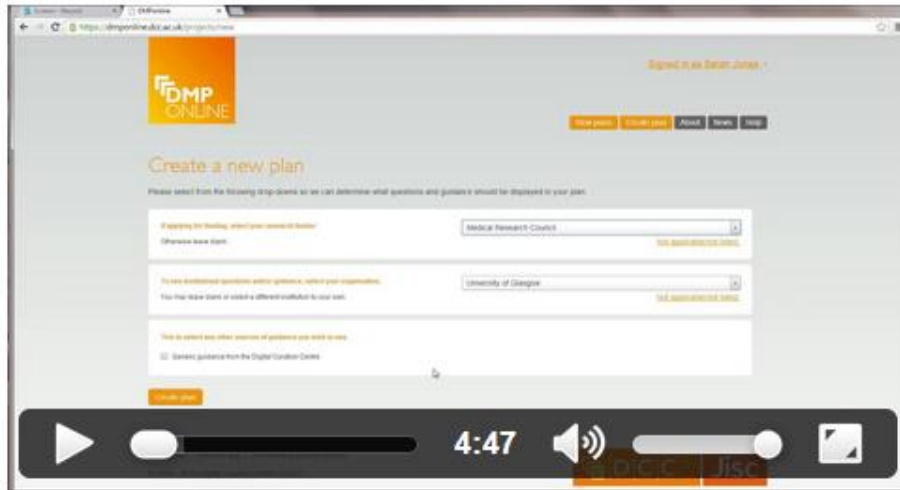
- To identify requirements for support
 - volume of data being created
 - what needs to be preserved and for how long
- Monitor awareness or uptake of services
 - how many DMPs say they will use repository?
- For outreach – advocate and improve practice



More information

Screencast on how to use DMPonline

<http://www.screenr.com/PJHN>



Customising DMPonline

www.dcc.ac.uk/news/customising-dmponline-admin-interface-launches



GitHub

Get the code, amend it, run a local instance, flag issues, request features...

<https://github.com/DigitalCurationCentre/DMPOne v4>

Research data repositories and registries



Image Credit: Hitachi Data Systems

Archiving – external data centres

Research funders' data centres...



British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL



Economic and Social Data Service

Structured databases



ChemSpider
The free chemical database



Registries of international data centres

Databib

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Disciplinary & community initiatives



GoGeo



Institutional data repositories



<http://datashare.is.ed.ac.uk>

Not intended to
replace national,
subject or other
established data
repositories



Research Data at Essex and
DataPool at Southampton



www.dspace.cam.ac.uk

Acknowledge hybrid
environment



<https://databank.ora.ox.ac.uk>

Data catalogues

- Oxford is developing its DataFinder tool <http://blogs.it.ox.ac.uk/damaro>
- Research Data @ Essex has developed a profile based on DataCite, Inspire and DDI standards www.data-archive.ac.uk/media/395364/rde_march2013_repositoryoutputs.pdf
- C4D is developing a research data extension to the CERIF standard - <http://cerif4datasets.wordpress.com>
- CKAN is being explored by various projects <http://ckan.org>

Research Data Registry and Discovery (RDRDS) pilot

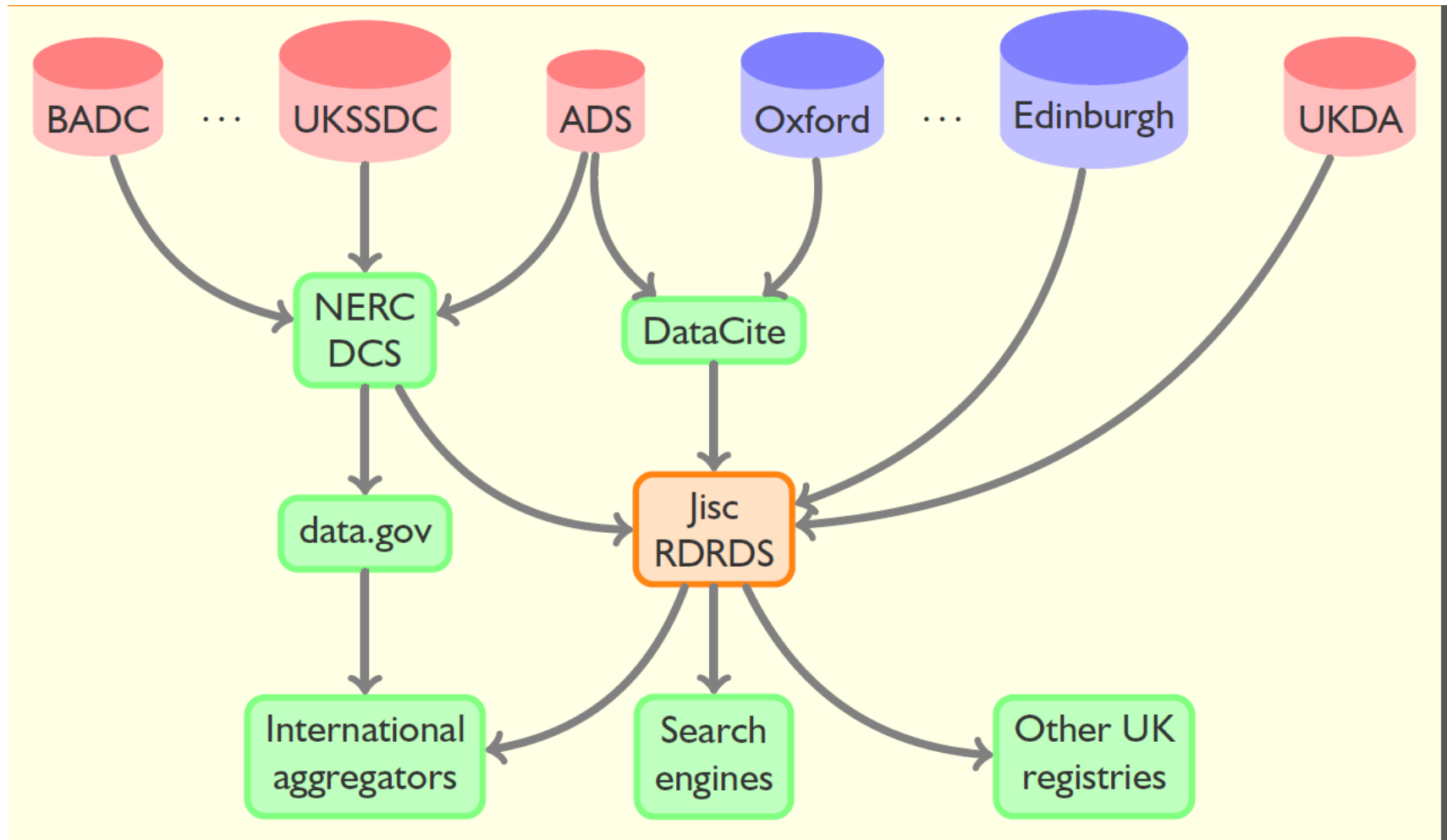
The screenshot shows the RDRDS pilot website interface. At the top, there is a navigation bar with the title "Research Data Registry & Discovery" and links for "About", "Collections", "Parties", "Activities", "Services", and "Themes". A "D|C|C" logo is also present. Below the navigation bar is a search bar with the placeholder text "Search for Research Data" and a magnifying glass icon. To the right of the search bar are two buttons: "Browse by Subject Area" with a tree icon and "Browse by Map Coverage" with a map icon. Below the search bar is a link for "Advanced Search".

The main content area is divided into two columns. The left column is titled "What's in the Research Data Registry and Discovery Service" and contains four categories:

- Collections (49)**: Research datasets or collections of research materials. (Icon: folder)
- Parties (36)**: Researchers or research organisations that create or maintain research datasets or collections. (Icon: two people)
- Activities (0)**: Projects or programs that create research datasets or collections. (Icon: flask)
- Services (0)**: Services that support the creation or use of research datasets or collections. (Icon: laptop with gears)

The right column is titled "Spotlight on research data" and features a large orange-bordered box containing the URL <http://rdrds.cloudapp.net/>. Below this is a section titled "Who contributes to the Research Data Registry and Discovery Service?" which states that 5 research organisations from around the UK contribute information to Research Data Australia. A "See All" link is provided. A "Share" button is located at the bottom right of the page.

Where does RDRDS fit in?



RDRDS metadata

- Description/Abstract
- Dataset identifier
- Subject
- URL of landing page
- Creator (+ID)
- Release date
- Rights information
- Spatial coverage
- Temporal coverage
- Publisher

Jisc RDRDS Pilot Partners

Data centres:

- ▶ UK Data Archive
- ▶ NERC Data Catalogue Service
 - ▶ BADC
 - ▶ BODC
 - ▶ EIDC
 - ▶ NEODC
 - ▶ NGDC
 - ▶ PDC
 - ▶ UKSSDC
 - ▶ ADS

Universities:

- ▶ Edinburgh
- ▶ Glasgow
- ▶ Hull
- ▶ Lincoln
- ▶ Leeds
- ▶ Oxford
- ▶ Oxford Brookes
- ▶ St Andrews
- ▶ Southampton


RDRDS Phase 2

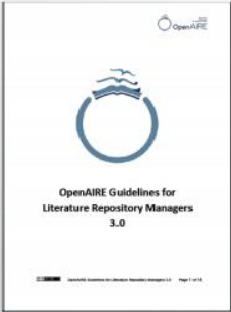
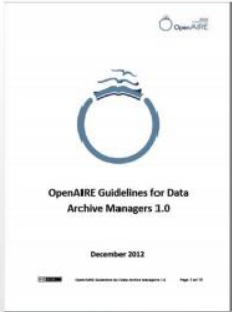
- ▶ Define a set of clear use cases and workflows.
- ▶ Compare different possible platforms for the service and assess their suitability.
- ▶ Establish a working instance of the system, involving all UK data centres and university data repositories.
- ▶ Establish a simple workflow for adding more data sources to the service, adapting to changes in existing data sources, and avoiding duplication.
- ▶ Test the system for usability.
- ▶ Produce recommendations for quality and standardisation of metadata records.
- ▶ Evaluate the costs and benefits of the system.


Short-term opportunities for ongoing research and development

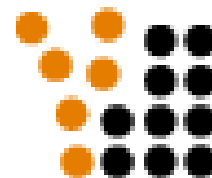


Harvesting metadata from CRIS to institutional data catalogues

 **OpenAIRE Guidelines for**

- 1 Literature Repository managers (version 3.0)**

- 2 Data Archive Managers**

- 3 CRIS Managers**


 <http://guidelines.openaire.eu>



Include equipment data into data management plans and institutional catalogue records

equipment.data

[Compliance](#) [Status](#) [FAQ](#) [Uniquip](#)

Enabling access to UK HE research equipment

 [Follow us](#)

Search for equipment:



It's good to share!

Welcome to equipment.data, your one stop shop for accessing UK wide HE research equipment. Use this searchable database to locate a piece of equipment or facility within your own or at another institution. equipment.data could be your starting point for some valuable collaborations.



Icon: CC2.5 Oily Holovchenko, 2013 <http://handdrawngoods.com>

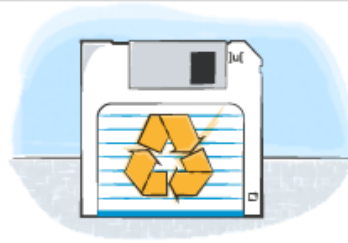
<http://equipment.data.ac.uk/>

Publish DMPs



DATASETS

Data papers highlight openly archived data with high reuse potential, and provide recognition for the producers of the data.



SOFTWARE

Software papers help you to locate openly archived, reusable code relevant to your research, and provide a mechanism for citing its use.



RESEARCH

Research reports provide concise summaries of key developments in a field that the community do not otherwise have access to.

[Journal of Open Archaeology Data](#)

[Journal of Open Psychology Data](#)

[Open Health Data](#)

[Journal of Open Research Software](#)

Use of data registries to gather usage statistics: institutions

REF2014
Research Excellence Framework

[Publications](#) | [Submissions](#) | [Expert panels](#) | [Research users](#) | [Equality & diversity](#) | [Background](#) | [Timetable](#) | [FAQs](#) | [Contact](#)

Research Excellence Framework

The Research Excellence Framework (REF) is the new system for assessing the quality of research in UK higher education institutions (HEIs). It will replace the [Research Assessment Exercise](#) (RAE) and will be completed in 2014.

The REF will be undertaken by the four UK higher education funding bodies. The exercise will be managed by the REF team based at HEFCE and overseen by the REF Steering Group, consisting of representatives of the four funding bodies.

The primary purpose of the REF is to produce assessment outcomes for each submission made by institutions:

- The funding bodies intend to use the assessment outcomes to inform the selective allocation of their research funding to HEIs, with effect from 2015-16.
- The assessment provides accountability for public investment in research and produces evidence of the benefits of this investment.
- The assessment outcomes provide benchmarking information and establish reputational yardsticks.

The REF is a process of expert review. HEIs will be invited to make submissions in [36 units of assessment](#). Submissions will be assessed by an expert sub-panel for each unit of assessment, working under the guidance of four main panels. Sub-panels will apply a set of [generic assessment criteria and level definitions](#), to produce an overall quality profile for each submission.

REF 2014 Latest

HEFCE provide [information](#) on the use of REF impact case studies.

The REF outcomes will be published on **18 December 2014**. [Further information](#) is available.

The REF submission deadline passed on 29 Nov 2013. A summary of [submissions data](#) is available.

See the [panel membership](#) lists for details of new appointments.

<http://www.ref.ac.uk/>

Use of data registries to gather usage statistics: funders

Home Digital Curation Centre | because good research needs good data
www.dcc.ac.uk

 **RESEARCH COUNCILS UK** Gateway to Research

 Technology Strategy Board
Driving Innovation

Welcome to the RCUK gateway to publicly funded research
Search for and analyse information on the latest innovative research in the UK

Please enter a search term.

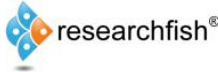
Data

The data on this website provides information about publications, people, organisations and outcomes relating to research projects

APIs

The data is accessible through three application programmes and CERIF

<http://gtr.rcuk.ac.uk/>

 **researchfish®** Request new password

Username Password

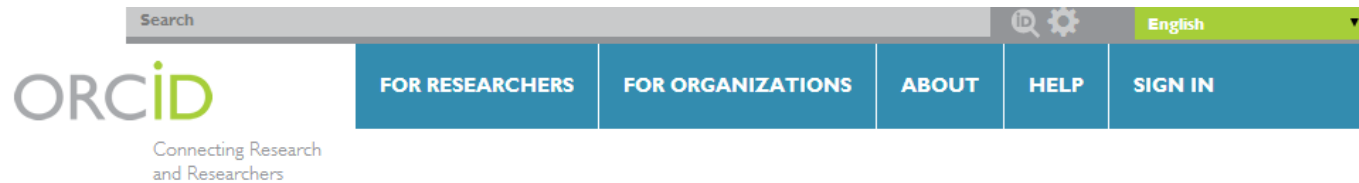
[Home](#) [About Us](#) [Our Members](#) [Help](#)

Researchfish is an easy-to-use research outcomes system for researchers and funding organisations.

Researchers and Delegates <input type="button" value="Find Out More"/>	Funding Organisations <input type="button" value="Find Out More"/>	Research Organisations/HEIs <input type="button" value="Find Out More"/>
--	--	--

<https://www.researchfish.com/>

Use of data registries to gather usage statistics: researchers



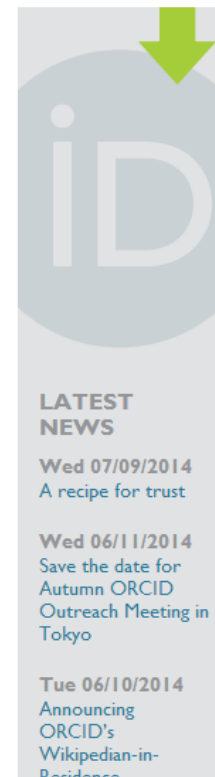
DISTINGUISH YOURSELF IN THREE EASY STEPS

ORCID provides a persistent digital identifier that distinguishes you from every other researcher and, through integration in key research workflows such as manuscript and grant submission, supports automated linkages between you and your professional activities ensuring that your work is recognized. [Find out more.](#)

1 REGISTER Get your unique ORCID identifier [Register now!](#)
Registration takes 30 seconds.

2 ADD YOUR INFO Enhance your ORCID record with your professional information and link to your other identifiers (such as Scopus or ResearcherID or LinkedIn).

3 USE YOUR ORCID ID Include your ORCID identifier on your Webpage, when you submit publications, apply for grants, and in any research workflow to ensure you get credit for your work.



Longer-term challenges



Providing tools to analyse and manipulate the data

The screenshot shows the website for the Centre for Environmental Data Archival. The header includes the organization's logo, name, and affiliations (Science and Technology Facilities Council, Natural Environment Research Council). It also features social media icons for RSS, Facebook, and Twitter, along with links for 'Contact Us' and 'CEDA News'. A search bar is present with a 'Go' button. A navigation menu lists 'About CEDA', 'Data Centres', 'Services', 'Projects', 'For Academics', 'For Business', 'Help', and 'Contact Us'. The main content area is titled 'Services' and contains a paragraph explaining that CEDA provides various services for finding, accessing, visualising, and processing data. Below this is a table with two columns: 'Service' and 'Brief Description'. The table lists several services, including myCEDA, Data Catalogue Search, Archive access services, Data preparation services, Data Analysis Environments, CEDA Document Repository, CEDA Document Repository Service, and Visualisation Services.

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Contact Us CEDA News

Home Services Search Go

About CEDA Data Centres Services Projects For Academics For Business Help Contact Us

Services

CEDA provides a range of services to assist users in finding, accessing, visualising and processing. Details of each of these services can be found following the links below.

Service	Brief Description
myCEDA	CEDA user account service
Data Catalogue Search	Services to find data held in CEDA archives
Archive access services	Details on accessing CEDA data via FTP, HTTP, PyDAP etc.
Data preparation services	Details of data preparation tools available through CEDA's online Web Processing Service
Data Analysis Environments	CEDA provides the JASMIN and CEMS environments for large scale data analysis, plus details about the LOTUS service.
CEDA Document Repository	Online, persistent repository for useful presentaitons, reports, flight logs etc. relating to CEDA archive data, data curation and environmental science. Content from both CEDA staff and the wider community welcome.
CEDA Document Repository Service	
Visualisation Services	Details of the data visualisation service supported by CEDA

<http://www.ceda.ac.uk/services/>

Common agreement on attribution stacking

DC¹

Data Citation Principles

<https://www.force11.org/datacitation>



Thanks – any questions?

DCC guidance, tools and case studies:

www.dcc.ac.uk/resources

Follow us on twitter:

@digitalcuration and #ukdcc



D|C|C

because good research needs good data