



Facilitate Open Science Training for European Research

Open Access to Research Data: Challenges and Solutions

Martin Donnelly
Digital Curation Centre
University of Edinburgh

FOSTER event, National Library of Latvia
Riga, 20 October 2015

Decorative dandelion seed heads are located in the bottom left and bottom right corners of the slide, rendered in a light orange color.

The Digital Curation Centre



- The UK's centre of expertise in digital preservation and data management, established 2004
- Provide guidance, training, tools and other services on all aspects of research data management
- Organise national and international events and webinars (International Digital Curation Conference, Research Data Management Forum)
- Principal audience is the UK higher education sector, but we increasingly work further afield (Europe, North America, South Africa...)
- Now offering paid consultancy/training services

Overview

1. Background and context
2. An introduction to Research Data Management (RDM)
 - a. What is RDM?
 - b. What are the main benefits?
 - c. What are the main problems?
3. RDM in practice
 - a. What does it mean for researchers?
 - b. Research data policies
4. Some useful resources



Overview

1. **Background and context**
2. An introduction to Research Data Management (RDM)
 - a. What is RDM?
 - b. What are the main benefits?
 - c. What are the main problems?
3. RDM in practice
 - a. What does it mean for researchers?
 - b. Research data policies
4. Some useful resources



Background and context

- Research data management exists within a context of **ever greater transparency, accessibility and accountability**
- The impetus for openness in research comes from two directions:
 - **Ground-up** - Open Access began in the High Energy Physics research community, which saw benefit in not waiting for publication before sharing research findings (and data / code)
 - **Top-down** - Government/funder support, increasing public and commercial engagement with research
- The main goals of these developments are to **lower barriers to accessing** the outputs of publicly funded research (often called ‘science’ for short), to **speed up** the research process, and to strengthen the **quality, integrity and longevity** of the scholarly record...

The old way of doing research

1. Researcher collects data (information)

2. Researcher interprets/synthesises data

3. Researcher writes paper based on data

4. Paper is published (and preserved)

5. Data is left to benign neglect, and eventually ceases to be accessible

Without intervention, data + time = no data

Vines et al. “examined the availability of data from 516 studies between 2 and 22 years old”

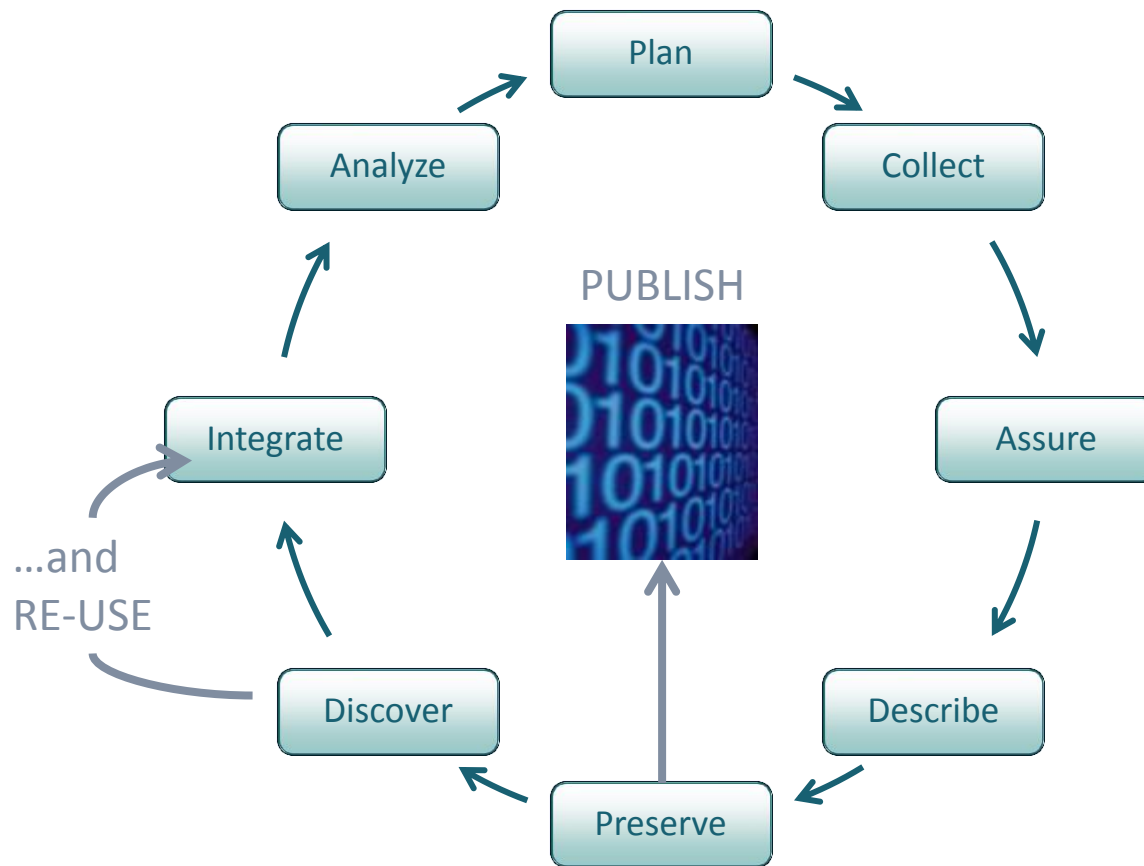
- The odds of a data set being reported as extant fell by 17% per year
- Broken e-mails and obsolete storage devices were the main obstacles to data sharing
- Policies mandating data archiving at publication are clearly needed

“The current system of leaving data with authors means that almost all of it is lost over time, unavailable for validation of the original results or to use for entirely new purposes” according to Timothy Vines, one of the researchers. This underscores the need for intentional management of data from all disciplines and opened our conversation on potential roles for librarians in this arena. (“80 Percent of Scientific Data Gone in 20 Years” *HNGN*, Dec. 20, 2013, <http://www.hngn.com/articles/20083/20131220/80-percent-of-scientific-data-gone-in-20-years.htm>.)

Vines et al., The Availability of Research Data Declines Rapidly with Article Age, *Current Biology* (2014), <http://dx.doi.org/10.1016/j.cub.2013.11.014>



The new way of doing research



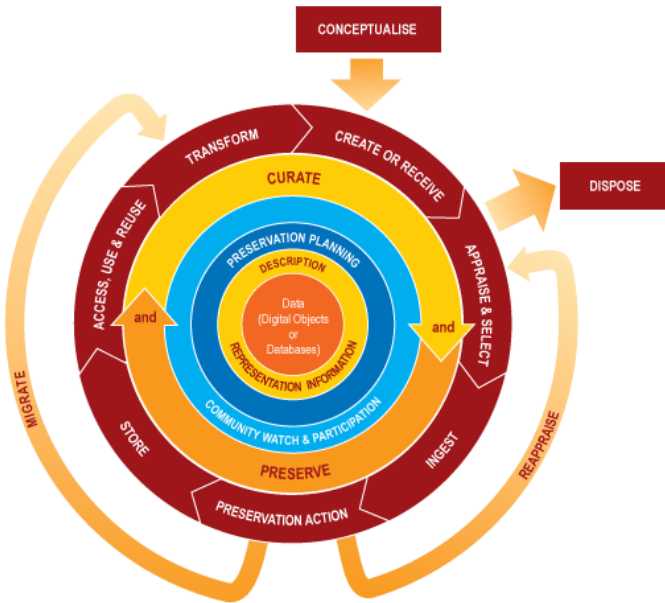
The DataONE
lifecycle model

Overview

1. Background and context
2. **An introduction to Research Data Management (RDM)**
 - a. What is RDM?
 - b. What are the main benefits?
 - c. What are the main problems?
3. RDM in practice
 - a. What does it mean for researchers?
 - b. Research data policies
4. Some useful resources



What is RDM?



“the **active** management and appraisal of data over the lifecycle of scholarly and scientific interest”

Data management is a part of good research practice.

- RCUK Policy and Code of Conduct on the Governance of Good Research Conduct

What sorts of activities?

- **Planning** and **describing** data-related work before it takes place
- **Documenting** your data so that others can find and understand it
- **Storing** it safely during the project
- **Depositing** it in a trusted archive at the end of the project
- **Linking** publications to the datasets that underpin them

RDM: who and how?

- RDM is a hybrid activity, involving multiple stakeholder groups...
 - The researchers themselves
 - Research support personnel
 - Partners based in other institutions, commercial partners, etc
- Data Management Planning (DMP) underpins and pulls **together** different strands of data management activities. DMP is the process of **planning, describing and communicating** the activities carried out during the research lifecycle in order to...
 - Keep sensitive data safe
 - Maximise data's re-use potential
 - Support longer-term preservation
- Data Management Plans are a **means of communication**, with contemporaries and potential future re-users alike...

Benefits of RDM and data sharing

- **SPEED:** The research process becomes faster
- **EFFICIENCY:** Data collection can be funded once, and used many times for a variety of purposes
- **ACCESSIBILITY:** Interested third parties can (where appropriate) access and build upon publicly-funded research resources with minimal barriers to access
- **IMPACT and LONGEVITY:** Open publications and data receive more citations, over a longer period
- **TRANSPARENCY and QUALITY:** The evidence that underpins research can be made open for anyone to scrutinise, and attempt to replicate findings. This leads to a more robust scholarly record

Benefits of RDM: Impact and Longevity

“In genomics research, a large-scale analysis of data sharing shows that studies that made data available in repositories received **9% more citations**, when controlling for other variables; and that whilst self-reuse citation declines steeply after two years, reuse by third parties **increases even after six years.**”

(Piwowar and Vision, 2013)

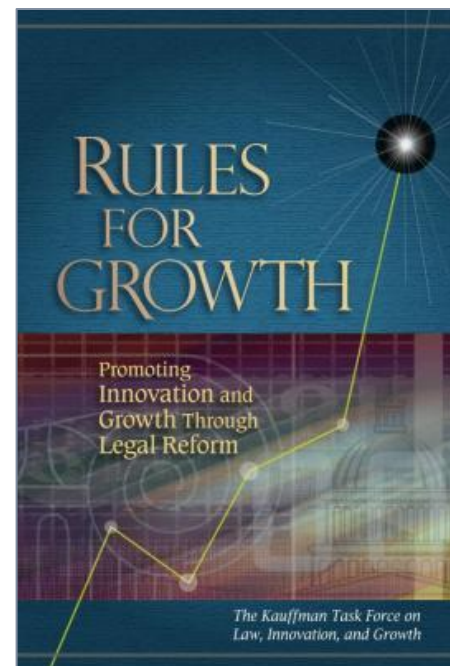


Van den Eynden, V. and Bishop, L. (2014). Incentives and motivations for sharing research data, a researcher's perspective. A Knowledge Exchange Report,

http://repository.jisc.ac.uk/5662/1/KE_report-incentives-for-sharing-researchdata.pdf

Benefits of RDM: Quality

“Data is necessary for reproducibility of computational research, but an equal amount of concern should be directed at code sharing.”



Victoria Stodden, “Innovation and Growth through Open Access to Scientific Research: Three Ideas for High-Impact Rule Changes” in Litan, Robert E. et al. Rules for Growth: Promoting Innovation and Growth Through Legal Reform. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, February 8, 2011. <http://papers.ssrn.com/abstract=1757982>.

Benefits of RDM: Financial



“Conservatively, we estimate that the value of data in Australia’s public research to be at least \$1.9 billion and possibly up to \$6 billion a year at current levels of expenditure and activity. Research data curation and sharing might be worth at least \$1.8 billion and possibly up to \$5.5 billion a year, of which perhaps \$1.4 billion to \$4.9 billion annually is yet to be realized.”

- “Open Research Data”, Report to the Australian National Data Service (ANDS), November 2014 - John Houghton, Victoria Institute of Strategic Economic Studies & Nicholas Gruen, Lateral Economics

More open data for more users ...

40+

Number of countries with
government open data platforms*

90,000+

Data sets on data.gov
(US site)*

1.4 million

Page views for the UK open data site
in the summer of 2013

102

Cities that participated in 2013
International Open Data Hackathon Day

1 million+

Data sets made open by
governments worldwide

* As of 2013

... can lead to more value

\$3 trillion

Approximate potential annual value
enabled by open data in seven "domains"

3 billion

Metric tons of carbon dioxide equivalent
emission reductions from buildings that could
be identified through the use of open data

35

Hours per year could be saved by commuters
from schedule changes based on open data

100,000+

Medical, health, and fitness apps
for smartphones

50%+

Consumer share of
potential value of open data

McKinsey Global Institute
McKinsey Center for Government
McKinsey Business Technology Office



October 2013

Open data: Unlocking
innovation and performance
with liquid information



J. Manyika et al. "Open data: Unlocking innovation
and performance with liquid information" McKinsey
Global Institute, October 2013

Benefits of RDM: Speed

“If we are going to wait five years for data to be released, the Arctic is going to be a very different place.”

Bryn Nelson, Nature, 10 Sept 2009

<http://www.nature.com/nature/journal/v461/n7261/index.html>



<https://www.flickr.com/photos/gsfcr/7348953774/>

- CC-BY

Why don't we live in a data sharing utopia?

Five main reasons...

- i. Lack of widespread understanding of the fundamental issues
- ii. Lack of joined-up thinking within institutions, countries, internationally...
- iii. Issues around ownership/privacy
- iv. Technical/financial limitations, and the need for selection and appraisal of data (which takes time, and costs money...)
- v. Issues around reward and recognition for researchers

Overview

1. Background and context
2. An introduction to Research Data Management (RDM)
 - a. What is RDM?
 - b. What are the main benefits?
 - c. What are the main problems?
3. **RDM in practice**
 - a. **What does it mean for researchers?**
 - b. **Research data policies**
4. Some useful resources



What does it mean for researchers?

- A disruption to previous working processes
- Additional expectations / requirements from the funders (and sometimes their home institutions and publishers too)
- But! It provides opportunities for new types of investigation
- And leads to a more robust scholarly record



What do researchers need to do?

1. Understand funders' policies (e.g. EC H2020...)
2. Check your intended publisher's OA policy (e.g. via Sherpa Romeo)
3. Create a data management plan (e.g. with DMPonline)
4. Decide which data to preserve using the DCC's How-To guide and checklist, "Five Steps to Decide what Data to Keep"
5. Identify a long-term home for your data (e.g. via re3data.org)
6. Link your data to your publications with a persistent identifier (e.g. via DataCite)
 - N.B. Many repositories, including Zenodo, will do this for you
7. Investigate EC infrastructure services and resources, e.g. EUDAT, OpenAIRE Plus, FOSTER, etc...

UK funders

The screenshot shows the Research Councils UK website. The header includes the logo and the tagline 'Excellence with Impact'. The main content area is titled 'RCUK Common Principles on Data Policy' and contains a paragraph about making research data available, a section for 'Principles' with a bulleted list, and a search bar on the right. The left sidebar contains a navigation menu with various categories like 'Research and Funding', 'Research Careers', 'Public Engagement with Research', etc.

RESEARCH COUNCILS UK Excellence with Impact

Home > Research and Funding > RCUK Common Principles on Data Policy

RCUK Common Principles on Data Policy

Making research data available to users is a core part of the Research Councils' remit and is undertaken in a variety of ways. We are committed to transparency and to a coherent approach across the research base. These RCUK common principles on data policy provide an overarching framework for individual Research Council policies on data policy.

Principles

- Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property.
- Institutional and project specific data management policies and plans should be in accordance with relevant standards and community best practice. Data with acknowledged long-term value should be preserved and remain accessible and usable for future research.
- To enable research data to be discoverable and effectively re-used by others, sufficient metadata should be recorded and made openly available to enable other researchers to understand the research and re-use potential of the data. Published results should always include information on how to access the supporting data.
- RCUK recognises that there are legal, ethical and commercial constraints on release of research data. To ensure that the research process is not damaged by inappropriate release of data, research organisation policies and practices should ensure that these are considered at all stages in the research process.
- To ensure that research teams get appropriate recognition for the effort involved in collecting and analysing data, those who undertake Research Council funded work may be entitled to a limited period of privileged use of the data they have collected to enable them to publish the results of their research. The length of this period varies by research discipline and, where appropriate, is discussed further in the published policies of individual Research Councils.
- In order to recognise the intellectual contributions of researchers who generate, preserve and share key research datasets, all users of research data should acknowledge the sources of their data and abide by the terms and conditions under which they are accessed.
- It is appropriate to use public funds to support the management and sharing of publicly-funded research data. To maximise the research benefit which can be gained from limited budgets, the mechanisms for these activities should be both efficient and cost-effective in the use of public funds.

Search:

This website
 All Research Councils

Home
Research and Funding
Research Funding
Areas of Research
Cross-Council Research Themes
Research Infrastructure
Research Priorities
Peer review
Eligibility for Research Council funding
How to apply for research funding
Applications which may cross Research Council remits
International
Terms and Conditions of Research Council fEC Grants
Press and Media
Terms and Conditions of Research Council Training Grants
Publications
Open Access
RCUK Common Principles on Data Policy
About
Efficiency

1. Public good
2. Preservation
3. Discovery
4. Confidentiality
5. First use
6. Recognition
7. Public funding

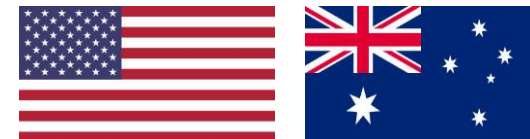
Six of the seven RCUK councils require data management plans (or equivalent), as do Wellcome Trust, Cancer Research UK, and more...

EXPECTATIONS

1. Research organisations will promote internal awareness of these principles and general awareness of the regulatory environment and of the available exemptions for research data;
2. Published research papers should include a short statement describing how research data is handled;
3. Each research organisation will have specific policies and associated procedures for data holdings and of requests by third parties to access such data; all of these should be consistent with research organisation policies in this area or, in exceptional circumstances, with EPSRC expectations;
4. Publicly-funded research data that is not generated in digital format will be made available in digital format or access to the data being received (this expectation could be satisfied by information in the metadata);
5. Research organisations will ensure that appropriately structured metadata (describing the data being generated) and made freely accessible on the internet; in the metadata, research data exists, why, when and how it was generated, and how to access it; it is expected that the metadata will include use of a robust digital object identifier (<http://datacite.org>).
6. Where access to the data is restricted the published metadata should also be restricted. For example 'commercially confidential' data, in which a business is subject to a suitable legally enforceable non-disclosure agreement.
7. Research organisations will ensure that EPSRC-funded research data is secure and that 'privileged access' period expires or, if others have accessed the data, that reasonable steps will be taken to ensure that publicly-funded data is not subject to more stringent protection than are available in the UK
8. Research organisations will ensure that effective data curation is provided for research data as defined by the Digital Curation Centre. The full range of responsibilities for data curation within the research organisation, and where research data is subject to more stringent security controls; research organisations will particularly ensure that the responsibilities for data curation are clearly defined and assigned to individuals;
9. Research organisations will ensure adequate resources are provided to support data curation allocated from within their existing public funding streams, whether received from higher education Funding Councils as block grants.

1. **INTERNAL AWARENESS** - of principles, expectations, regulatory environment, possible exemptions
2. **ACCESS STATEMENT** - included within research papers
3. **POLICIES AND PROCESSES** - covering maintenance and access requests
4. **NON-DIGITAL DATA** - strategy for access / digitisation
5. **METADATA PUBLICATION** - within 12 months of data generation
6. **RESTRICTIONS** - list these within metadata
7. **PRESERVATION** - 10 years from date of last access
8. **CURATION** - maintenance and security
9. **RESOURCING** - from existing funding streams

RDM in other countries (i)



USA

- The National Science Foundation (NSF) announced a DMP requirement in 2010, effective 2011
- White House Office of Science and Technology Policy requirement for DMPs announced March 2013 (programmes awarding >\$100m annually). White House requirements include mechanisms covering compliance with plans and policies, and also cover costs of implementing plans

AUSTRALIA

- In 2014 The Australian Research Council (ARC) released new instructions for applications for Laureate Fellowships and Discovery Grants. Both include the following requirements when describing a proposal...
 - **COMMUNICATION OF RESULTS:** Outline plans for communicating the research results to other researchers and the broader community, including scholarly and public communication and dissemination
 - **MANAGEMENT OF DATA:** Outline plans for the management of data produced as a result of the proposed research, including but not limited to storage, access and re-use arrangements

RDM in other countries (ii)



SOUTH AFRICA

- Announced in January 2015 that (from March 2015) “authors of research papers generated from research either fully or partially funded by NRF, when submitting and publishing in academic journals, should deposit their final peer-reviewed manuscripts that have been accepted by the journals, to the administering Institution Repository with an embargo period of no more than 12 months.”
- In addition, the data supporting the publication should be deposited in an accredited Open Access repository, with the provision of a Digital Object Identifier for future citation and referencing.
- The NRF encourages its stakeholder community, including NRF’s Business Units and National Research Facilities, to:
 - Formulate detailed policies on Open Access of publications and data from its funded research;
 - Establish Open Access repositories; and
 - Support public access to the repositories through web search and retrieval according to international standards and best practice.

RDM in Europe



- Horizon 2020 (FP8) features an Open Research Data pilot, and it seems likely that it will become an across-the-board requirement in FP9...
- It applies to data (and metadata) needed to validate scientific results, which should be deposited in a dedicated data repository
- The Horizon 2020 Open Research Data pilot covers “Innovation actions” and “Research and Innovation actions”, and involves three iterations of Data Management Plan (DMP)
 - 6 months after start of project, mid-project review, end-of-project (final review)
- DMP contents
 - Data types; Standards used; Sharing/making available; Curation and preservation
- There are certain opt-out conditions

H2020 Open Data Pilot: specifics (ii)



STEP 1

- The data should be deposited, preferably in a dedicated research data repository. These may be subject-based/thematic, institutional or centralised.
- EC suggests the Registry of Research Data Repositories (www.re3data.org) and Databib (<http://databib.org>) for researchers looking to identify an appropriate repository
- Open Access Infrastructure for Research in Europe (OpenAIRE) will also become an entry point for linking publications to data.

STEP 2

- So far as possible, projects must then take measures to enable for third parties to access, mine, exploit, reproduce and disseminate (free of charge for any user) this research data.
- EC suggests attaching Creative Commons Licence (CC-BY or CC0) to the data deposited (<http://creativecommons.org/licenses/>, <http://creativecommons.org/about/cc0>).
- At the same time, projects should provide information via the chosen repository about tools and instruments at the disposal of the beneficiaries and necessary for validating the results, for instance specialised software or software code, algorithms, analysis protocols, etc. Where possible, they should provide the tools and instruments themselves.

H2020 Open Data Pilot: specifics (iii)



COSTS

Costs relating to the implementation of the pilot will be eligible. Specific technical and professional support services will also be provided (e-Infrastructures WP), e.g. EUDAT and OpenAIRE, alongside support measures such as FOSTER.

OPT-OUTS

Opt outs are possible, either totally or partially. Projects may opt out of the Pilot at any stage, for a variety of reasons, e.g.

- if participation in the Pilot on Open Research Data is incompatible with the Horizon 2020 obligation to protect results if they can reasonably be expected to be commercially or industrially exploited;
- confidentiality (e.g. security issues, protection of personal data);
- if participation in the Pilot on Open Research Data would jeopardise the achievement of the main aim of the action;
- if the project will not generate / collect any research data;
- if there are other legitimate reasons to not take part in the Pilot (to be declared at proposal stage)

Overview

1. Background and context
2. An introduction to Research Data Management (RDM)
 - a. What is RDM?
 - b. What are the main benefits?
 - c. What are the main problems?
3. RDM in practice
 - a. What does it mean for researchers?
 - b. Research data policies
4. **Some useful resources**



DMPonline



- Web-based tool to help researchers write and maintain DMPs
- Provides funder questions and guidance
 - Includes templates for all RCUK funders, and Horizon 2020
- Provides tailored help from universities
- Can include examples and suggest responses
- Free to use
- Mature (v1 launched April 2010)
- Code is Open Source (on GitHub)

• <https://dmponline.dcc.ac.uk>

EUDAT



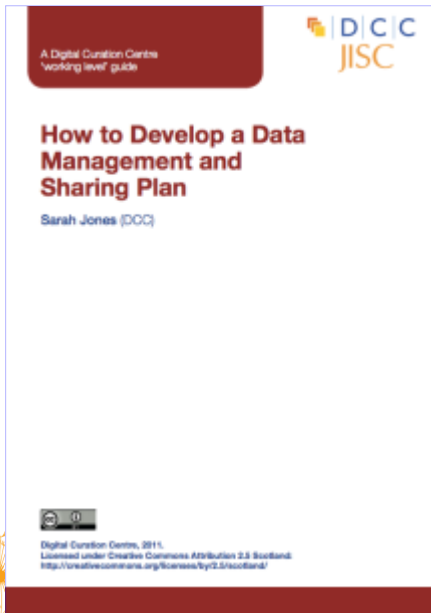
- EUDAT offers **common data services** through a geographically distributed, resilient network of 35 European organisations. These **shared services and storage resources** are distributed across 15 European nations and data is stored alongside some of Europe's most powerful supercomputers.
- The EUDAT services address the full lifecycle of research data, covering both access and deposit, from informal data sharing to long-term archiving, and addressing identification, discoverability and computability of both long-tail and big data
- The vision is to enable European researchers and practitioners from any academic discipline to preserve, find, access, and process data in a trusted environment, as part of a Collaborative Data Infrastructure (CDI) conceived as a network of collaborating, cooperating centres, combining the richness of numerous community-specific data repositories with the permanence and persistence of some of Europe's largest scientific data centres
- Seeks to bridge the gap between research infrastructures and e-Infrastructures through an active engagement strategy, using the communities in the consortium as EUDAT beacons, and integrating others through innovative partnership approaches
- Jisc and DCC are partners, and we're working to embed DCC's DMPonline tool within the EUDAT suite of services / infrastructure

Zenodo



- Zenodo is a **free-to-use data archive**, run by the people at CERN
- It accepts **any kind of data**, from any academic discipline
- It is generally preferable to store data in a disciplinary data centre, but **not all scholarly subjects are equally well served** with data centres, so this may make for a useful fallback option
- See <http://zenodo.org/> for more details

Other data management resources (DCC)



- Book chapter
 - Donnelly, M. (2012) “Data Management Plans and Planning”, in Pryor (ed.) *Managing Research Data*, London: Facet
- Guidance, e.g. “How-To Develop a Data Management and Sharing Plan”
- DCC Checklist for a Data Management Plan:
<http://www.dcc.ac.uk/resources/data-management-plans/checklist>
- Links to all DCC resources via
<http://www.dcc.ac.uk/resources/data-management-plans>

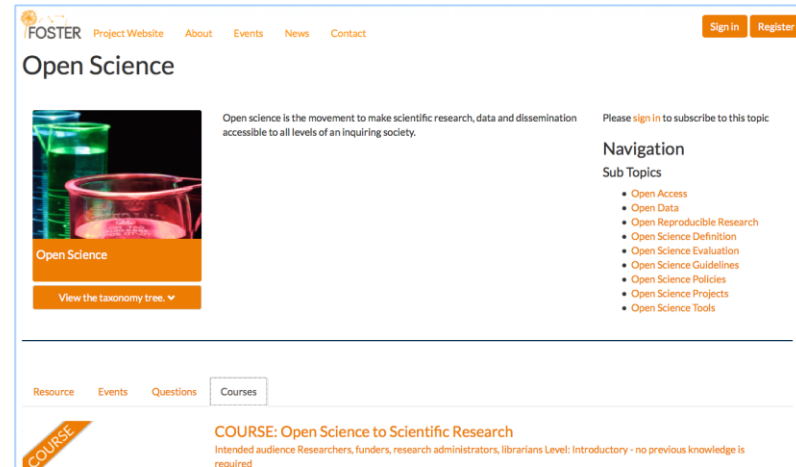
Data management resources (non-DCC)

- UKDA guidance and book (<http://data-archive.ac.uk/media/2894/managingsharing.pdf>)
- Guidance from funders (ESRC and NERC are particularly strong)
- Resources from other universities, e.g. Bath, Bristol, Cambridge, Edinburgh, Glasgow, Oxford (to name but a few)



OBJECTIVES

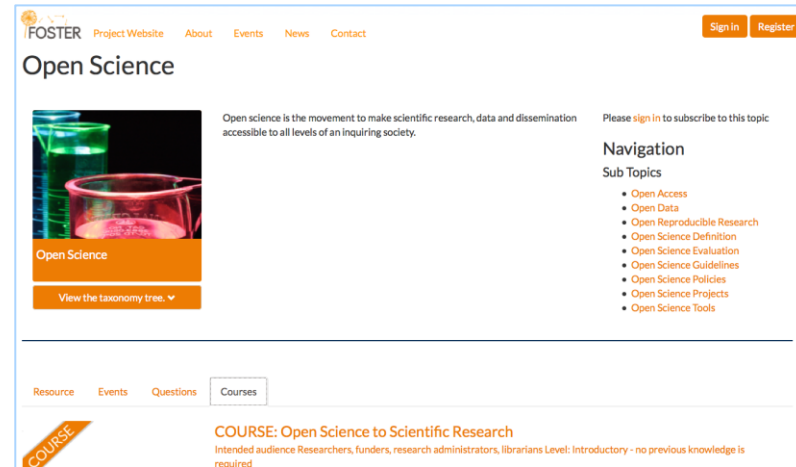
- To support different stakeholders, especially younger researchers, in adopting open access in the context of the European Research Area (ERA) and in complying with the open access policies and rules of participation set out for Horizon 2020
- To integrate open access principles and practice in the current research workflow by targeting the young researcher training environment
- To strengthen institutional training capacity to foster compliance with the open access policies of the ERA and Horizon 2020 (beyond the FOSTER project)
- To facilitate the adoption, reinforcement and implementation of open access policies from other European funders, in line with the EC's recommendation, in partnership with PASTEUR4OA project



The screenshot shows the FOSTER Project Website interface. At the top, there is a navigation menu with links for "Project Website", "About", "Events", "News", and "Contact". On the right side of the header, there are "Sign In" and "Register" buttons. The main content area is titled "Open Science" and features a large image of laboratory glassware (a beaker and a flask) with a blue glow. Below the image is a button labeled "View the taxonomy tree." To the right of the image, there is a text block defining open science: "Open science is the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society." Further right, there is a "Please sign in to subscribe to this topic" prompt and a "Navigation" section with "Sub Topics" listed: Open Access, Open Data, Open Reproducible Research, Open Science Definition, Open Science Evaluation, Open Science Guidelines, Open Science Policies, Open Science Projects, and Open Science Tools. At the bottom of the page, there is a "COURSE" banner for "COURSE: Open Science to Scientific Research" with a description: "Intended audience: Researchers, funders, research administrators, librarians Level: Introductory - no previous knowledge is required".

METHODS

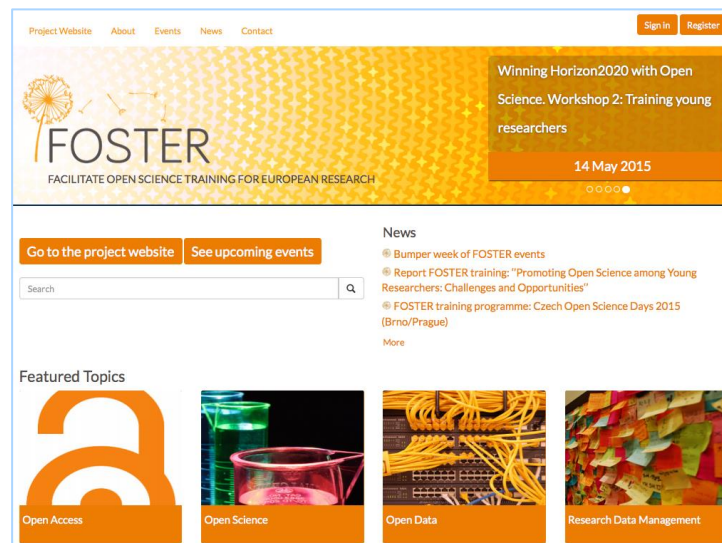
- Identifying already **existing content** that can be reused in the context of the training activities and repackaging, reformatting them to be used within FOSTER, and developing/creating/enhancing contents as required
- Developing the **FOSTER Portal** to support e-learning, blended learning, self-learning, dissemination of training materials/contents and a Helpdesk
- Delivery of **face-to-face training**, especially **training trainers/multipliers** who can deliver further training and dissemination activities, within institutions, nations or disciplinary communities
 - *The EC is also funding other specific technical and professional support services via the e-Infrastructures WP, e.g. EUDAT and OpenAIRE*



The screenshot shows the FOSTER Project Website interface. At the top, there is a navigation bar with links for 'Project Website', 'About', 'Events', 'News', and 'Contact', along with 'Sign In' and 'Register' buttons. The main heading is 'Open Science'. Below this, there is a featured image of laboratory glassware (a beaker with green liquid and a flask with red liquid) with the text 'Open Science' and a button 'View the taxonomy tree, v'. To the right of the image, there is a definition: 'Open science is the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society.' Further right, there is a 'Please sign in to subscribe to this topic' prompt and a 'Navigation' section with 'Sub Topics' including: Open Access, Open Data, Open Reproducible Research, Open Science Definition, Open Science Evaluation, Open Science Guidelines, Open Science Policies, Open Science Projects, and Open Science Tools. At the bottom, there is a 'COURSE' banner for 'COURSE: Open Science to Scientific Research' with the text 'Intended audience Researchers, funders, research administrators, librarians Level: Introductory - no previous knowledge is required'. A secondary navigation bar at the bottom includes 'Resource', 'Events', 'Questions', and 'Courses'.

Thank you

- For more information about the FOSTER project:
 - Website: www.fosteropenscience.eu
 - Principal investigator: Eloy Rodrigues (eloy@sdum.uminho.pt)
 - General enquiries: Gwen Franck (gwen.franck@eifl.net)
 - Twitter: @fosterscience
- My contact details:
 - Email: martin.donnelly@ed.ac.uk
 - Twitter: @mkdDCC
 - Slideshare: <http://www.slideshare.net/martindonnelly>



This work is licensed under the Creative Commons Attribution 2.5 UK: Scotland License.

