FOSTER

Facilitate Open Science Training for European Research

Research Data Management
& the H2020 Open Data Pilot

Martin Donnelly
Digital Curation Centre
University of Edinburgh

FOSTER event, University of Cyprus
Nicosia, 22 October 2015

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 tailored 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

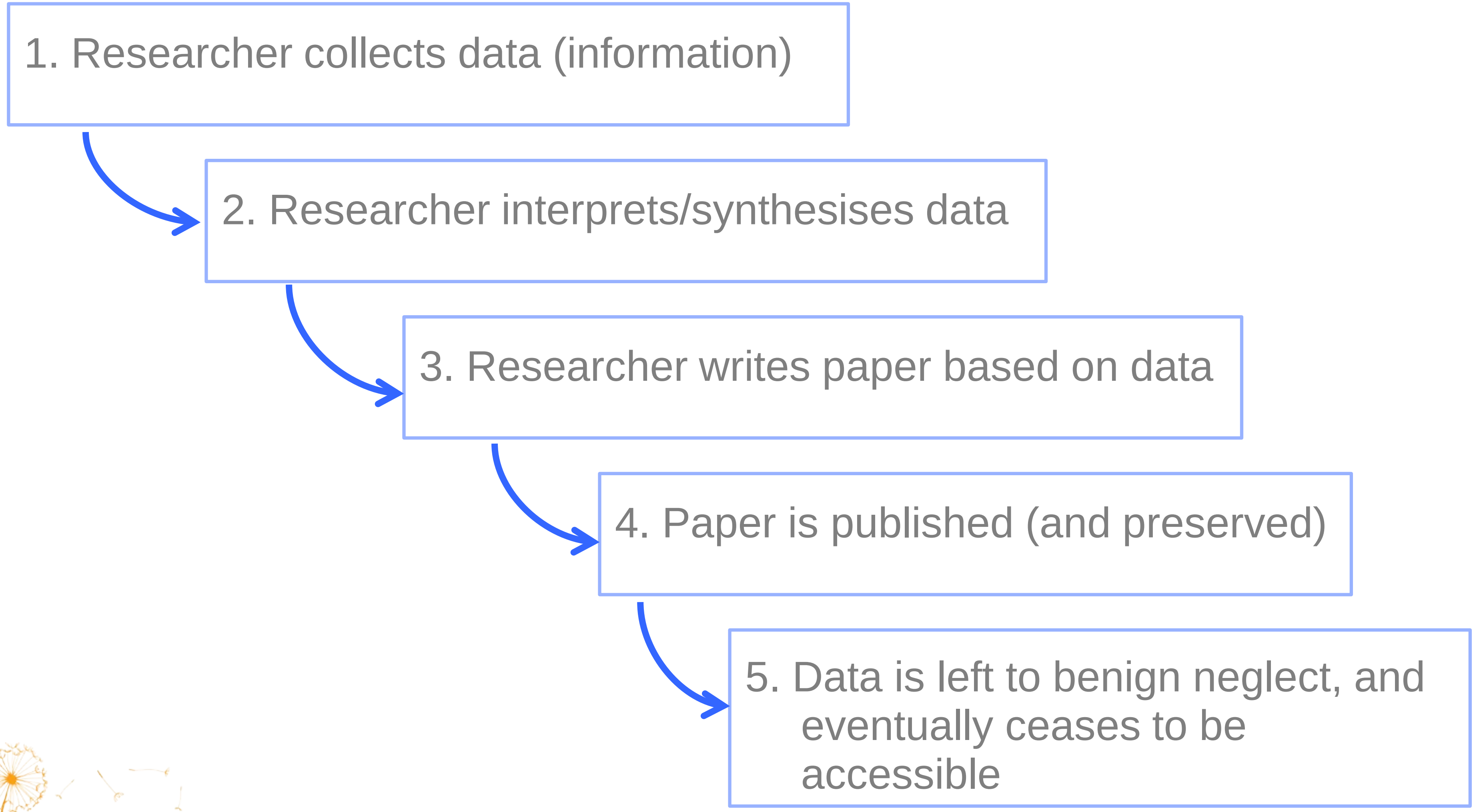
Overview

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

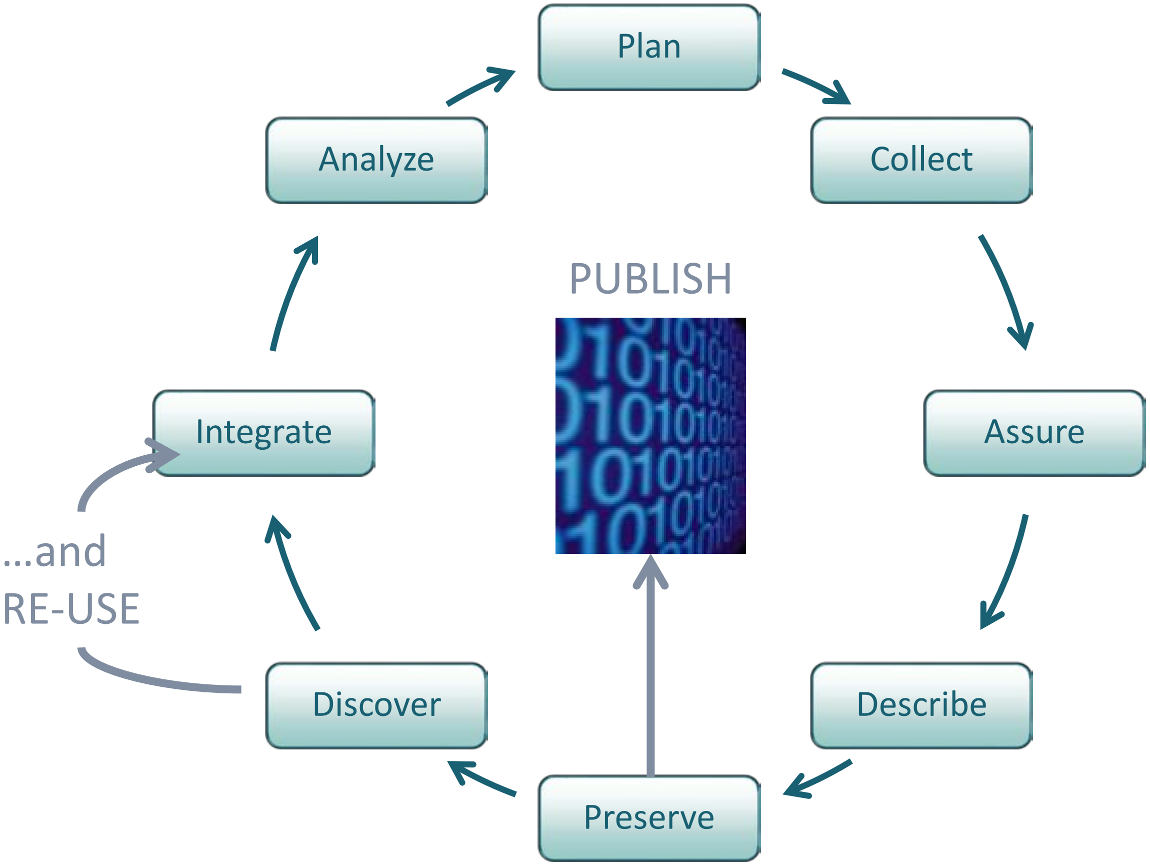
Vines et al. “examined the availability of data from 516 studies between 2 and 22 years old”

- **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



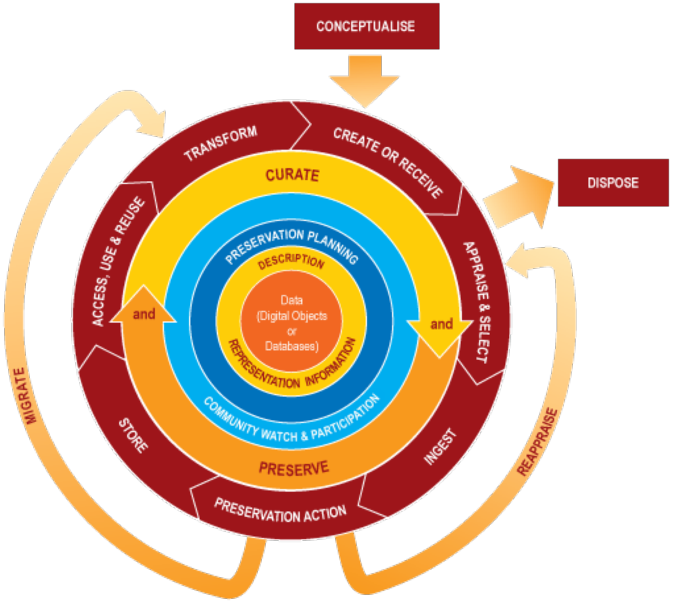
The DataONE lifecycle model

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

FOSTER

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?

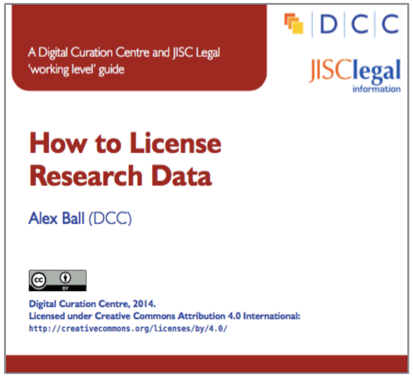
- **Planning** and **describing** data-related work before it takes place
- **Documenting** your data so that others can find and understand it
- **Storing** it safely during the project
- **Depositing** it in a trusted archive at the end of the project
- **Linking** publications to the datasets that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

FOSTER

(Aside: dealing with sensitive data)

MANAGEMENT
≠
SHARING

- Although the two terms are sometimes used interchangeably, **data management is not the same as data sharing**. In fact, when the data in question is sensitive, good management is the *opposite* of sharing
- Data may be sensitive for two main reasons: ethical and commercial…
  - Ethically sensitive data often corresponds to living human subjects, who have rights that are protected by law. Other areas where data should not be shared without careful thought include security, nuclear science, and infectious diseases (e.g. Ebola)
  - Commercially sensitive data can take many forms. At a basic level it may constitute a business's competitive advantage. Private sector data is not generally expected to be made open (except in (e.g.) clinical trials, where the public interest factor is strong.) But increasing amounts of research are carried out via partnerships between the public sector and private enterprise, so conflicts can arise here. The best policy will be to consider potential difficulties early in the process, via (e.g.) a data management plan
- There are various options for copyrighting or licensing data. The DCC guide on "How to License Research Data" can help you decide which approach is the best fit for your research



A Digital Curation Centre and JISC Legal 'working level' guide

DCC

JISClegal information

How to License Research Data

Alex Ball (DCC)

Digital Curation Centre, 2014.
Licensed under Creative Commons Attribution 4.0 International:
http://creativecommons.org/licenses/by/4.0/

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** the 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 also **a means of communication**, with contemporaries and potential future re-users alike...

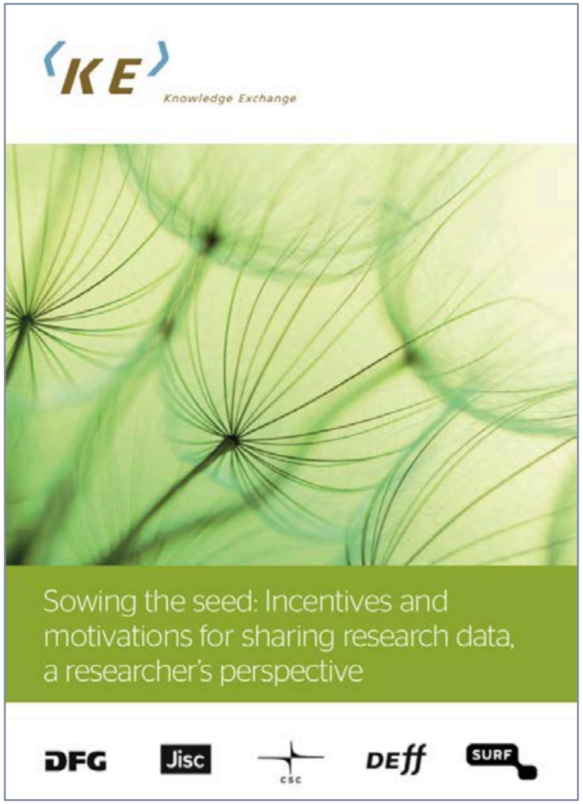
Benefits of RDM and data sharing

- **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
- **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
- **SPEED:** The overall research process becomes faster

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)



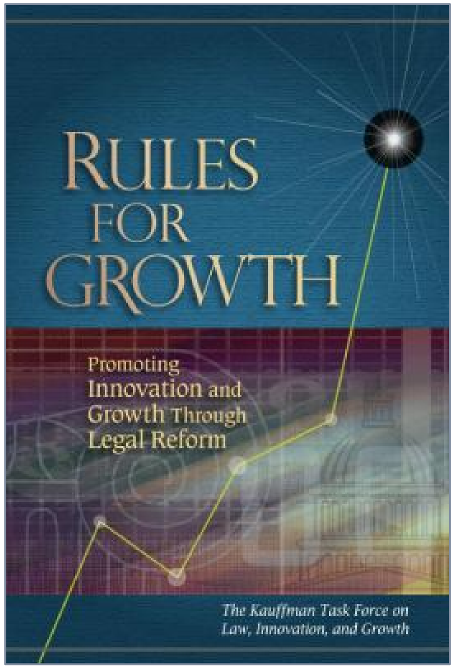
Sowing the seed: Incentives and motivations for sharing research data, a researcher's perspective

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

FOSTER

Benefits of RDM: Quality

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

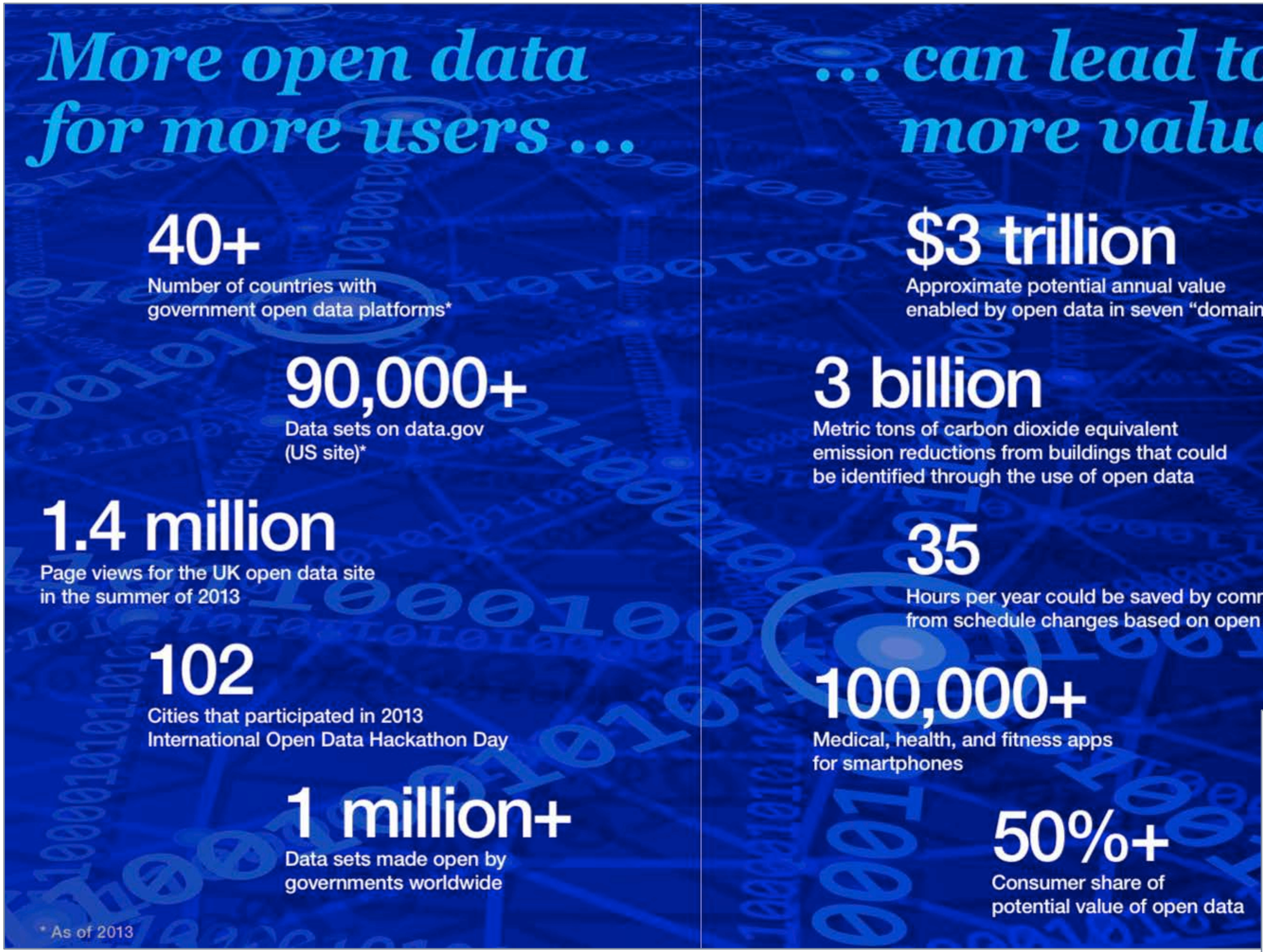

RULES FOR GROWTH
Promoting Innovation and Growth Through Legal Reform
The Kauffman Task Force on Law, Innovation, and Growth

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.](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&Company

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

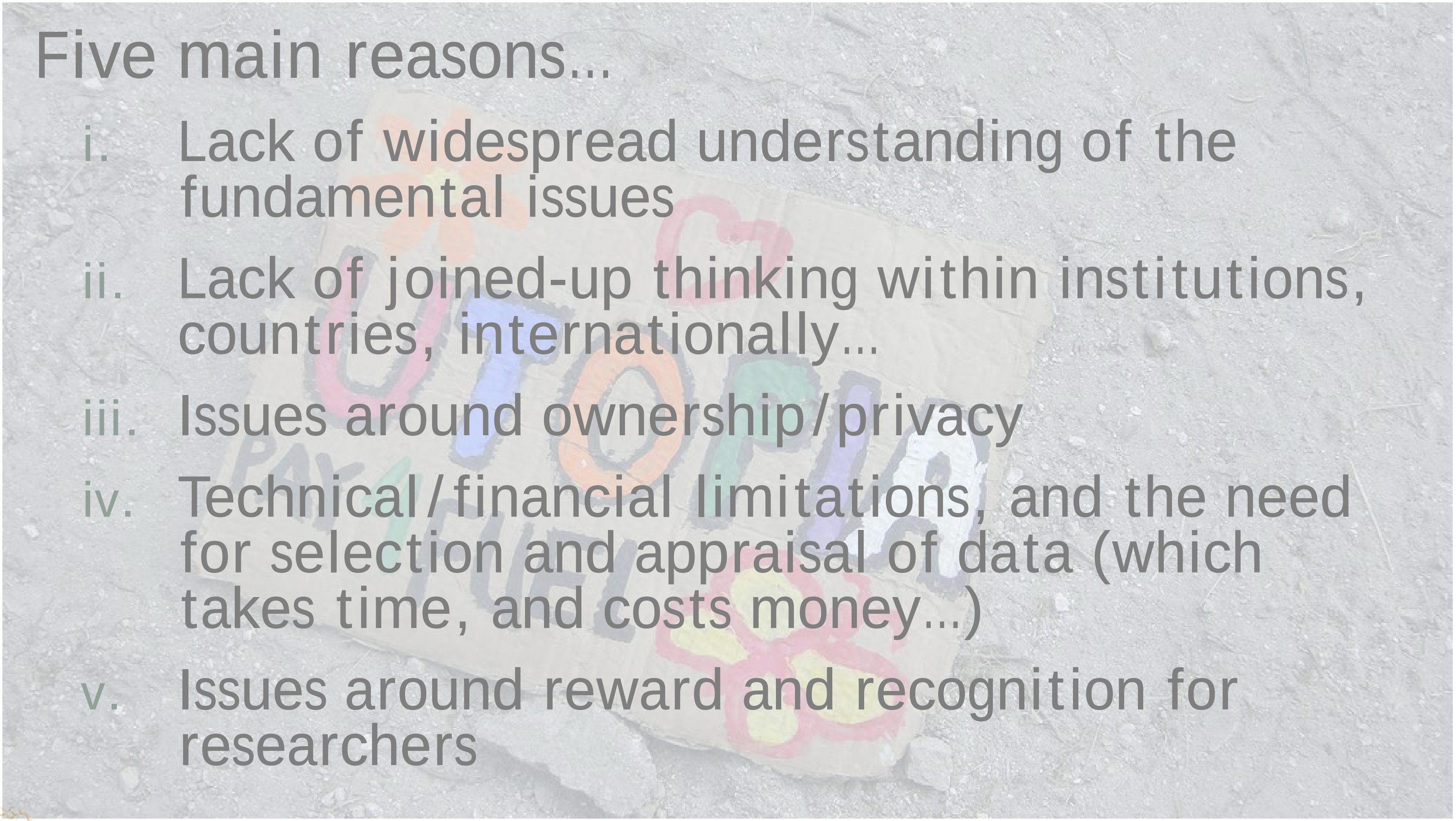
"If we are going to wait five years for data to be released, **the Arctic is going to be a very different place.**"

Bryn Nelson, Nature, 10 Sept 2009

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

https://www.flickr.com/photos/gsfc/7348953774/ - CC-BY

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



Five main reasons…

1. Lack of widespread understanding of the fundamental issues
2. Lack of joined-up thinking within institutions, countries, internationally…
3. Issues around ownership/privacy
4. Technical/financial limitations, and the need for selection and appraisal of data (which takes time, and costs money…)
5. Issues around reward and recognition for researchers

FOSTER

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

FOSTER

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 the funders' policies (e.g. EC H2020…)
2. Check 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 (e.g. 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 data repositories, such as Zenodo, will do this for you
7. Investigate EC infrastructure services and resources, e.g. EUDAT, OpenAIRE Plus, FOSTER, etc…

FOSTER

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

FOSTER

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.

FOSTER

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 (N.B. these should be declared at the proposal stage)

FOSTER

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

Data management planning resources (DCC)

DCC Checklist for a Data Management Plan

DCC

v3.0

A Digital Curation Centre 'working level' guide

DCC

JISC

How to Develop a Data Management and Sharing Plan

Sarah Jones (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

Other 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)

UK • DATA ARCHIVE

MANAGING AND SHARING DATA

UK • DATA ARCHIVE

BEST PRACTICE FOR RESEARCHERS

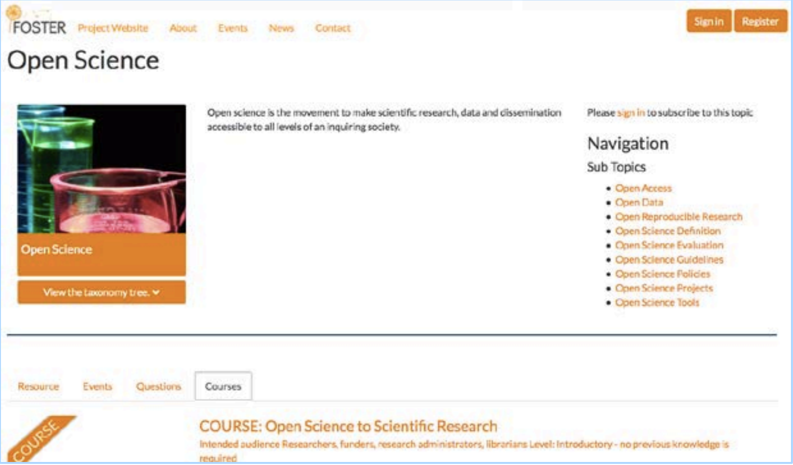
MAY 2011

The FOSTER project

Facilitate Open Science Training for European Research

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



Open Science

Open science is the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society.

Please sign in to subscribe to this topic

Navigation

Sub Topics

- Open Access
- Open Data
- Open Reproducible Research
- Open Science Definition
- Open Science Evaluation
- Open Science Guidelines
- Open Science Policies
- Open Science Projects
- Open Science Tools

COURSE: Open Science to Scientific Research

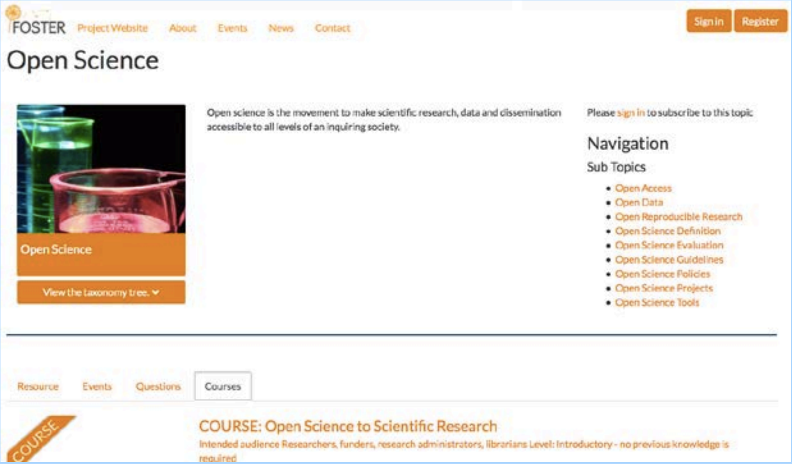
Intended audience Researchers, funders, research administrators, librarians Level: Introductory - no previous knowledge is required

The FOSTER project

Facilitate Open Science Training for European Research

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*



Open Science

Open science is the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society.

Please sign in to subscribe to this topic

Navigation

Sub Topics

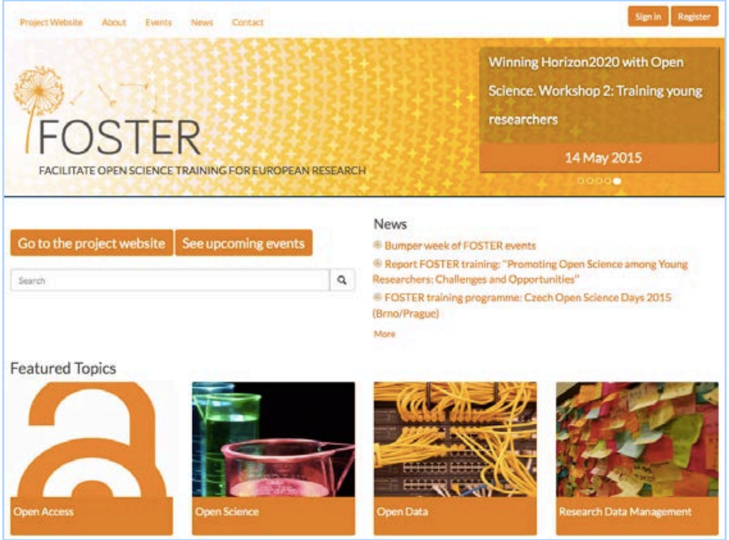
- Open Access
- Open Data
- Open Reproducible Research
- Open Science Definition
- Open Science Evaluation
- Open Science Guidelines
- Open Science Policies
- Open Science Projects
- Open Science Tools

COURSE: Open Science to Scientific Research

Intended audience Researchers, funders, research administrators, librarians Level: Introductory - no previous knowledge is required

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



Winning Horizon2020 with Open Science. Workshop 2: Training young researchers

14 May 2015

News

- Bumper week of FOSTER events
- Report FOSTER training: "Promoting Open Science among Young Researchers: Challenges and Opportunities"
- FOSTER training programme: Czech Open Science Days 2015 (Brno/Prague)

More

- My contact details:
  - Email: martin.donnelly@ed.ac.uk
  - Twitter: @mkdDCC
  - Slideshare: http://www.slideshare.net/martindonnelly