Facilitate Open Science Training for European Research

What to know about interdisciplinary research data management in order to assess DMPs

Martin Donnelly
Digital Curation Centre
University of Edinburgh

Research Executive Agency, FET-Open Unit
European Commission, Brussels
8 February 2018

Presentation outline

1. About the DCC
2. RDM and research
3. Good practice in RDM
4. Focus on Data Management Plans and Planning
5. The FAIR data principles
6. Reviewing DMPs
7. About FOSTER / questions

FOSTER

Presentation outline

1. **About the DCC**
2. RDM and research
3. Good practice in RDM
4. Focus on Data Management Plans and Planning
5. The FAIR data principles
6. Reviewing DMPs
7. About FOSTER / questions

FOSTER

The Digital Curation Centre (DCC)

- UK national centre of expertise in digital preservation and data management, est. 2004
- Principal audience is the UK higher education sector, but we increasingly work further afield (continental Europe, North America, South Africa, Asia…)
- Provide guidance, training, tools (e.g. DMPonline) and other services on all aspects of research data management and Open Science
- Tailored consultancy/training
- Organise national and international events and webinars (International Digital Curation Conference, Research Data Management Forum)

FOSTER

Presentation outline

1. About the DCC
2. **RDM and research**
3. Good practice in RDM
4. Focus on Data Management Plans and Planning
5. The FAIR data principles
6. Reviewing DMPs
7. About FOSTER / questions

FOSTER

RDM and research: the primary benefits

- RDM helps preserve, protect and proliferate the data behind scientific (research) discoveries and claims – first and foremost it is a **QUALITY** issue…
  - When research data is managed actively and responsibly, 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, and helps discourage and identify academic fraud
- A secondary benefit is **PROTECTION:** the rights and legitimate interests of data subjects and IP owners are mindfully protected
  - Active and responsible data management reduces the chances of inadvertent data leaks or loss

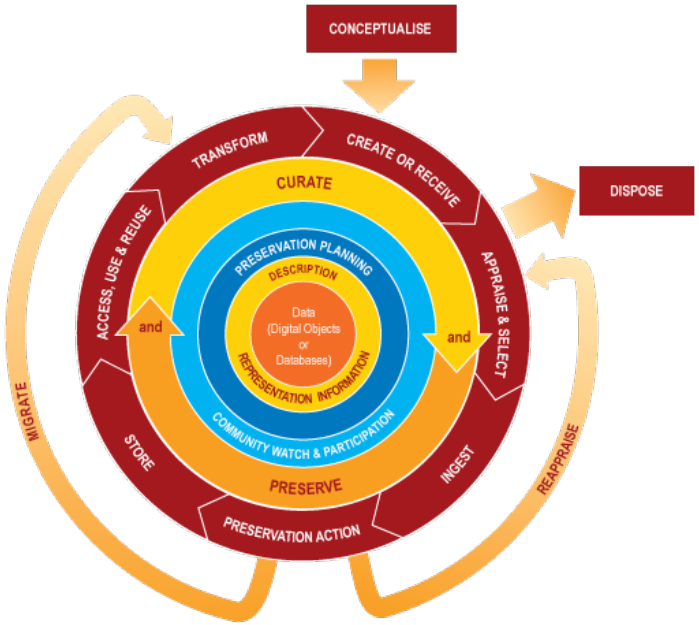
FOSTER

Other benefits of RDM

- It also has other benefits…
  - **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 outputs with minimal barriers to access
  - **SPEED:** The research process becomes faster
  - **IMPACT and LONGEVITY**: Data linked to publications receive more citations, over longer periods
  - **DURABILITY:** Simply put, fewer important datasets will be lost

FOSTER

RDM is an active process…



RDM is "the active management and appraisal of data over the lifecycle of scholarly and scientific interest"

Core activities include:

- **Planning** and **describing** data-related work before it takes place
- **Documenting** your data (and processing/workflows) so that others can find and understand it
- Choosing **open** (or at least standardised) **file formats** where possible
- **Storing** data safely during a project
- **Depositing** it in a trusted archive at the end of the research

FOSTER

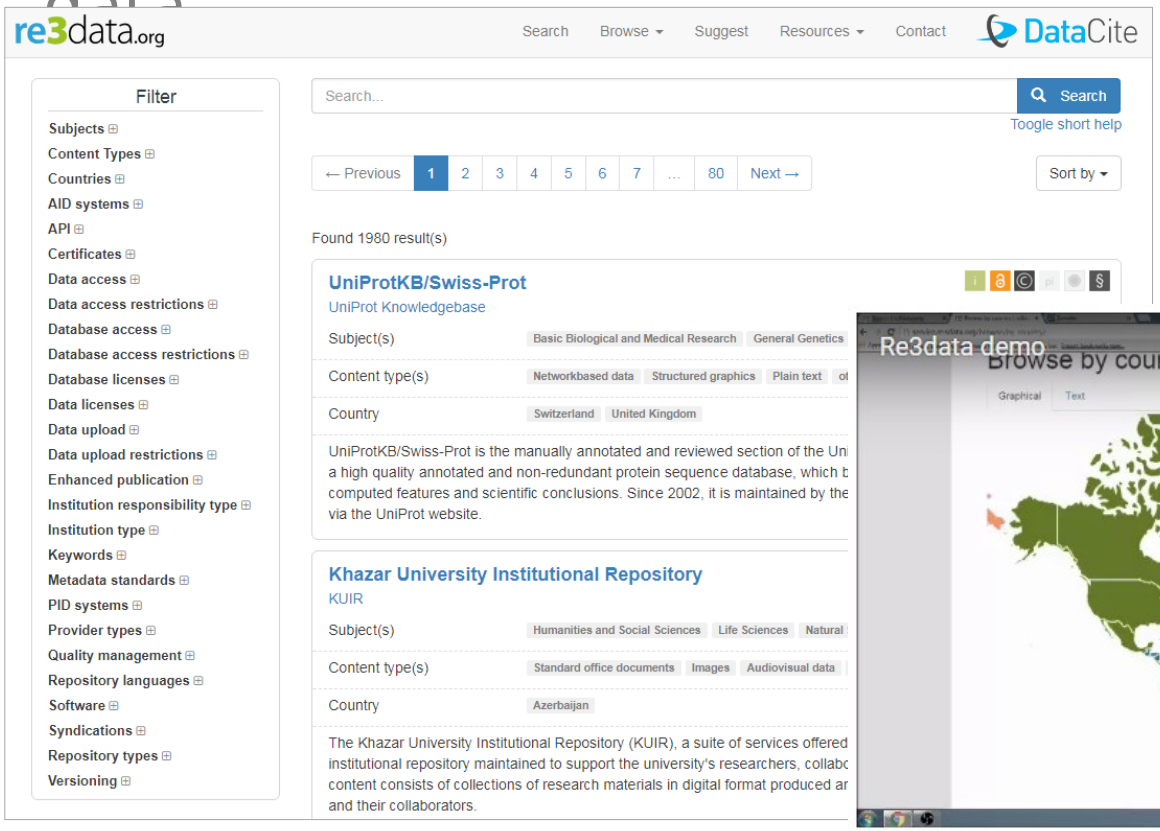
Presentation outline

1. About the DCC
2. RDM and research
3. **Good practice in RDM**
4. Focus on Data Management Plans and Planning
5. The FAIR data principles
6. Reviewing DMPs
7. About FOSTER / questions

FOSTER

Deposit in a data repository

The EC guidelines point to Re3data as one of the registries that can be searched to find a home for data

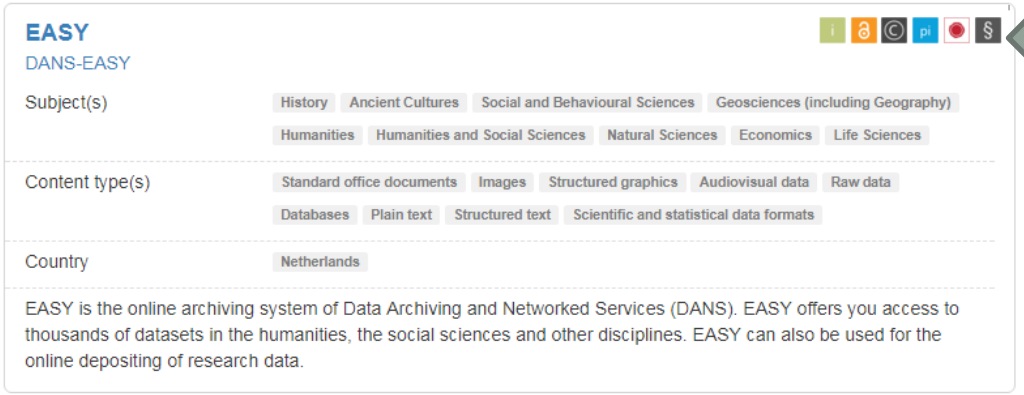


www.re3data.org

www.fosteropenscience.eu/content/re3data-demo

How to select a repository?

- Better to use a subject specific repository if available
- Check they match particular data needs e.g. formats accepted, mixture of Open and Restricted Access.
- Do they assign a persistent and globally unique identifier for sustainable citations and to links back to particular researchers and grants?
- Look for certification as a '*Trustworthy Digital Repository*' with an explicit ambition to keep the data available in long term.



Icons to note open access, licenses, PIDs, certificates…

FOSTER

Zenodo

Zenodo is a multi-disciplinary repository that can be used for the long-tail of research data

- An OpenAIRE-CERN joint effort
- Multidisciplinary repository accepting
  - Multiple data types
  - Publications
  - Software
- Assigns a Digital Object Identifier (DOI)
- Links funding, publications, data & software

www.zenodo.org

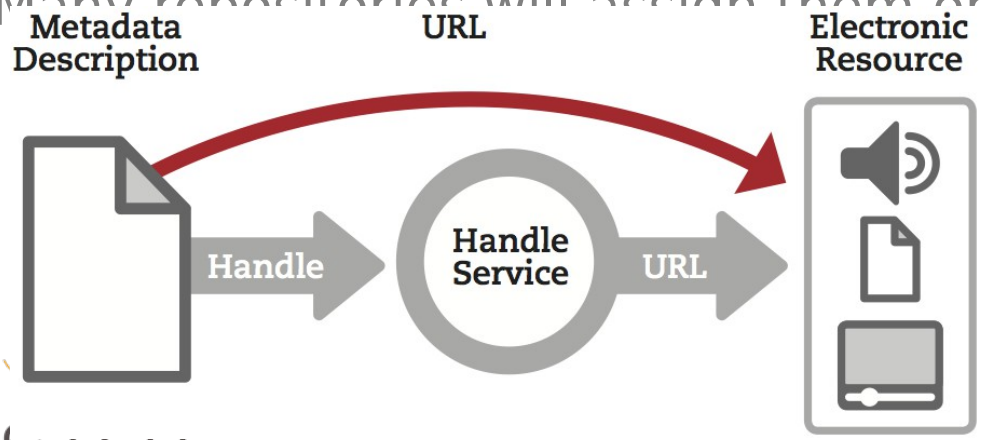
FOSTER

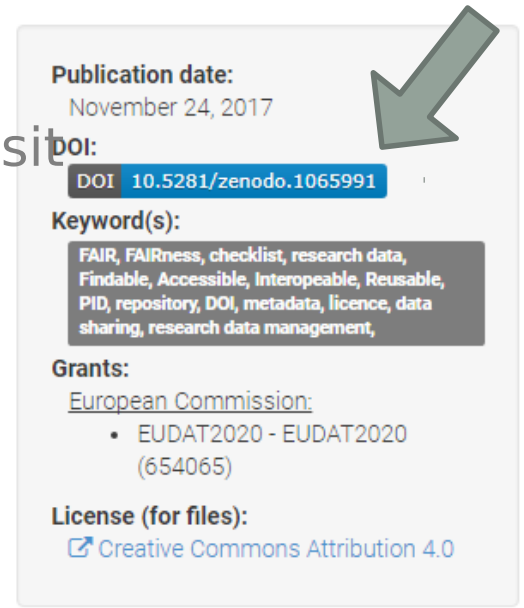
zenodo

Attach Persistent Identifiers (PIDs)

A PID is a long-lasting reference to a document, file or other object

- PIDs come in various forms e.g. ARK, DOI, URN, PURL, Handles...
- Typically they're actionable i.e. type it into web browser to access
- Many repositories will assign them on deposit





FOSTER

Create metadata

- At a basic level, metadata supports findability, disambiguation and citation
- Rich, specific metadata will support interoperability & reuse
- Standards should be used. These can be general – such as Dublin Core, or discipline specific
  - Data Documentation Initiative (DDI) – social science
  - Ecological Metadata Language (EML) - ecology
  - Flexible Image Transport System (FITS) – astronomy

FOSTER

Dublin Core metadata example

Creator: Donald Cooper
Role=Photographer

Subject: Shakespeare, William, 1564-1616, Antony and Cleopatra [LC]

Description: Vanessa Redgrave as Cleopatra

Date: 1973-08-09

Type: Image

Format: JPEG

Identifier:4150 [catalogue no]

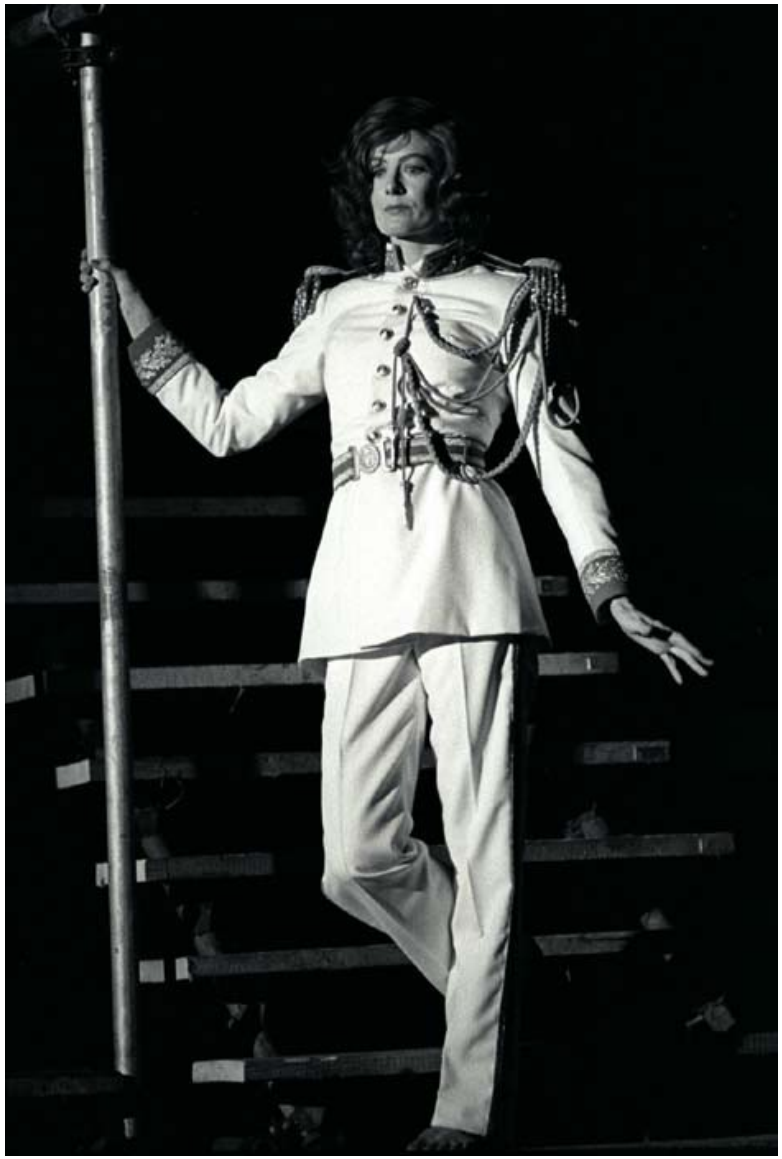
Source: negative no 235

Relation: Antony and Cleopatra: Thompson/73-8
IsPartOf

Coverage: Bankside Globe
Role=Spatial

Rights: Donald Cooper

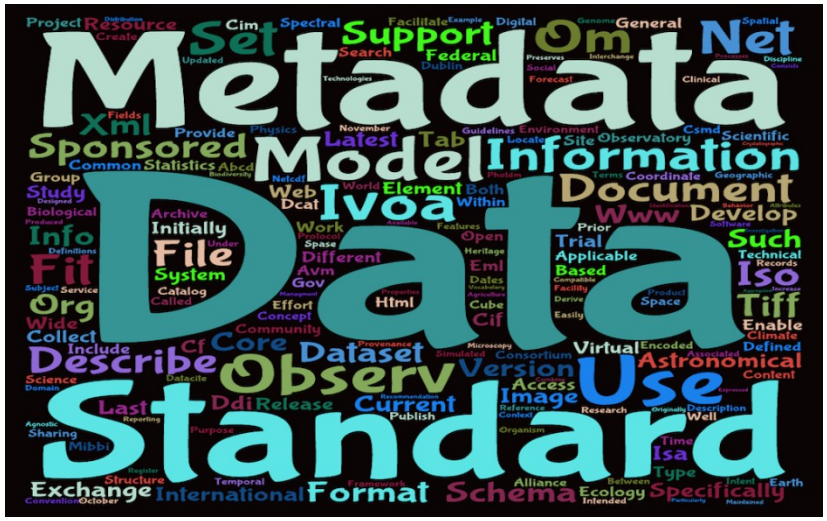
www.ahds.ac.uk/performingarts



Where to find relevant standards?

Metadata Standards Directory

Broad, disciplinary listing of standards and tools. Maintained by RDA group



http://rd-alliance.github.io/metadata-directory

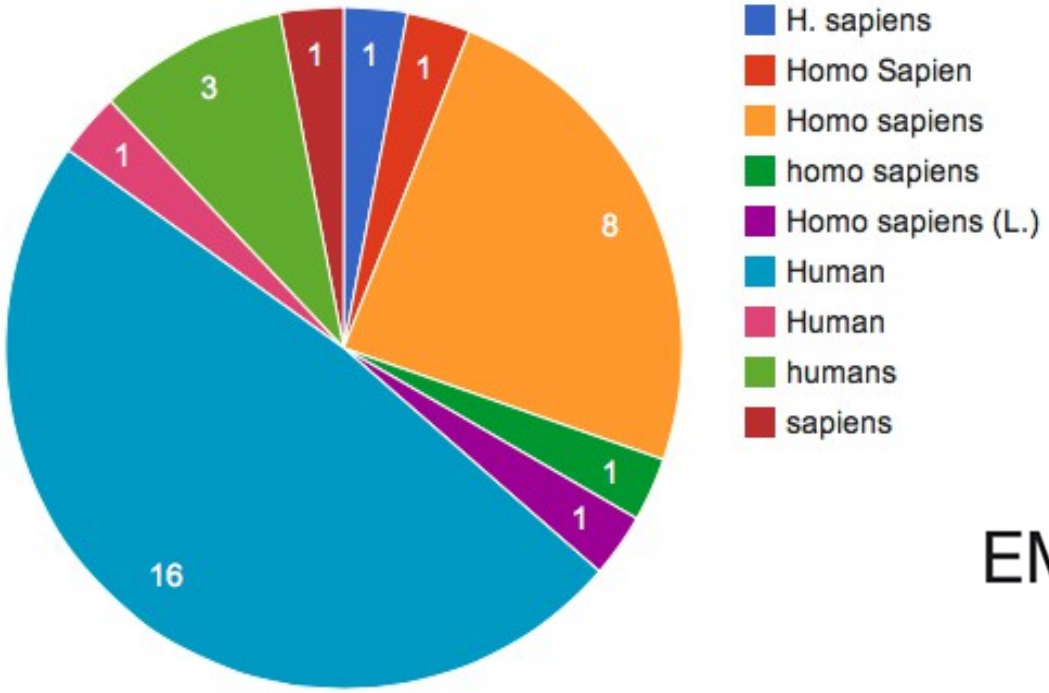
FAIRsharing

- A portal of data standards, databases, and policies
- Focused on life, environmental and biomedical sciences, but

https://fairsharing.org

Value of controlled vocabularies

"MTBLS1: A metabolomic study of urinary changes in type 2 diabetes in……"



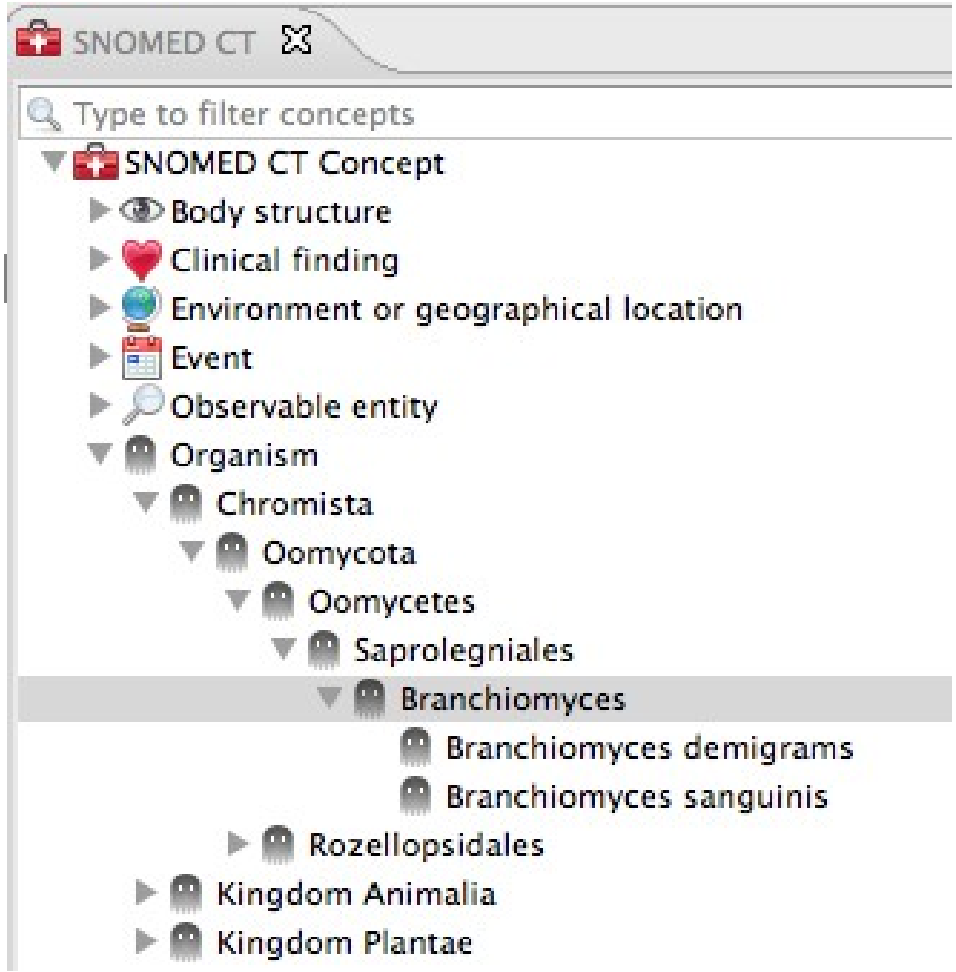
EMBL-EBI

FOSTER

Example courtesy of Ken Haug, European Bioinformatics Institute (EMBL-EBI)

Controlled vocabularies

- E.g. SNOMED CT (clinical terms) or MeSH
- Include ontologies as well
  - Defined terms + taxonomy
- Useful for selecting keywords to tag datasets



FOSTER

Create documentation

We recommend that a ReadMe be a plain text file containing:

- for each filename, a short description of what data it includes, optionally describing the relationship to the tables, figures, or sections within the accompanying publication
- for tabular data: definitions of column headings and row labels; data codes (including missing data); and measurement units
- any data processing steps, especially if not described in the publication, that may affect interpretation of results
- a description of what associated datasets are stored elsewhere, if applicable
- whom to contact with questions

FOSTER

DRYAD

http://datadryad.org/pages/readme

Choose appropriate file formats

Different formats are good for different things

- open, lossless formats are more sustainable e.g. rtf, xml, tif, wav
- proprietary and/or compressed formats are less preservable but are often in widespread use e.g. doc, jpg, mp3

One format for analysis then convert to a standard format

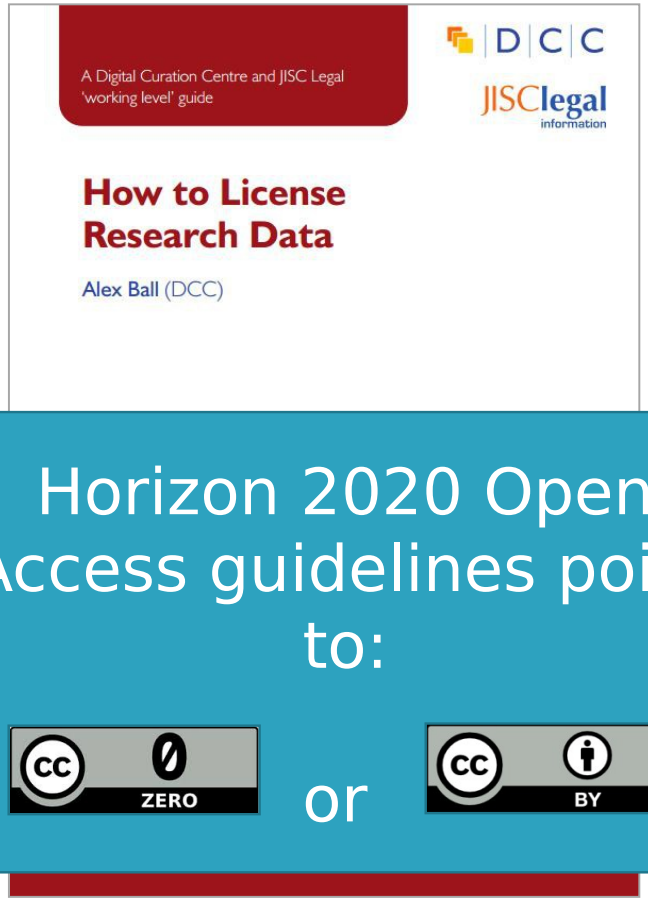
BioformatsConverter batch converts a variety of proprietary microscopy image formats to the Open Microscopy Environment format - OME-TIFF

Data centres may suggest preferred formats for deposit

www.data-archive.ac.uk/create-manage/format/formats-table

FOSTER

Attach appropriate licenses or waivers



Horizon 2020 Open Access guidelines point to:

CC 0 ZERO or CC BY

This DCC guide outlines the pros and cons of each approach and gives practical advice on how to implement your licence

CREATIVE COMMONS LIMITATIONS

NC Non-Commercial
What counts as commercial?

ND No Derivatives
Severely restricts use

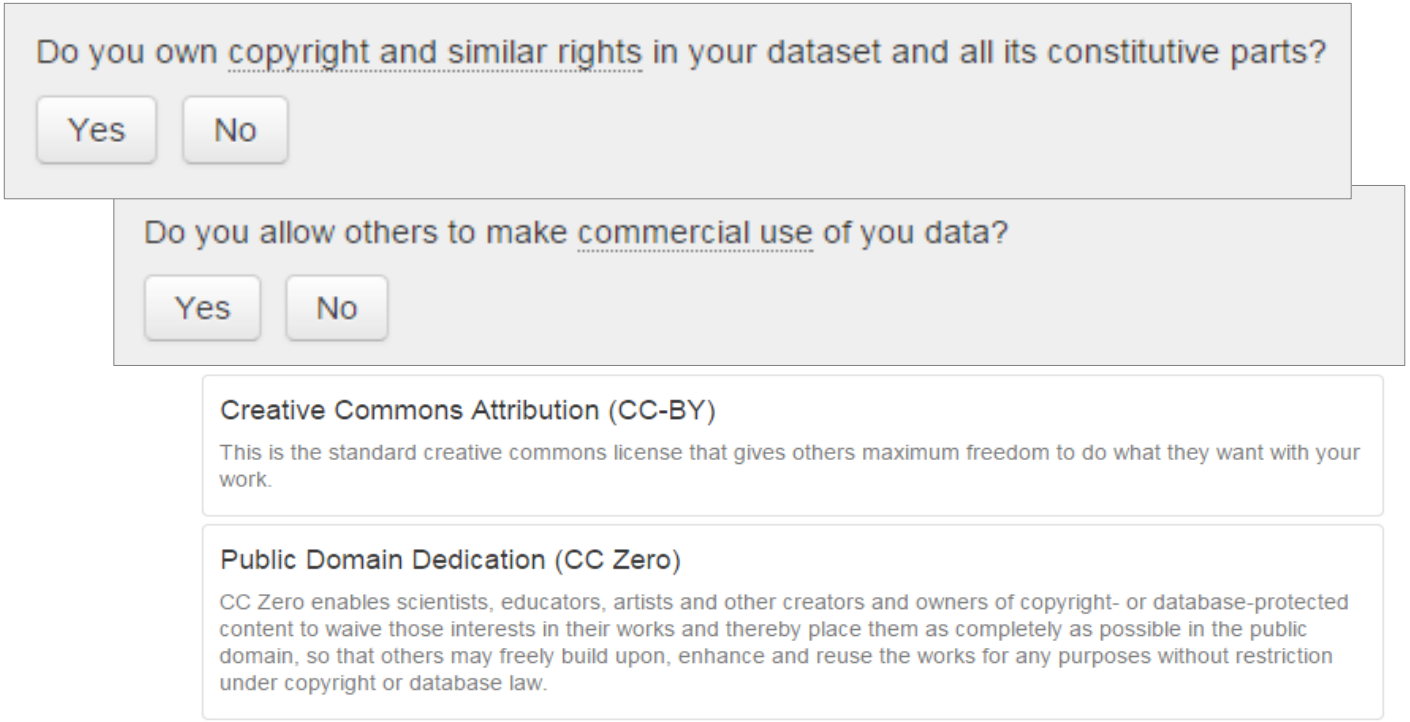
These clauses are not open licenses

FOSTER

www.dcc.ac.uk/resources/how-guides/license-research-data

EUDAT licensing tool

Answer questions to determine which licence(s) are appropriate to use



FOSTER

http://ufal.github.io/lindat-license-selector

Recap: a few do's and don'ts

DO	DON'T
Have a plan for your data	Make it up as you go along
Keep backups. Make this easy with automated syncing services like Dropbox, provided your data isn't too sensitive	Carry the only copy around on a memory card, your laptop, your phone, etc
Describe your data as you collect it. This makes it possible for others to interpret it, and for you to do the same a few years down the line	Leave this till the end. The quality of metadata decreases with time, and the best metadata is created at the moment of data capture
Save your work in open file formats, where possible, and use accepted metadata standards to enable like-with-like comparison	Invent new 'standards' where community norms already exist
Deposit your data in a data centre or repository, and link it to your publications	Be afraid to ask for help. This will exist both within your institution, and via national / European support organisations

FOSTER

Presentation outline

1. About the DCC
2. RDM and research
3. Good practice in RDM
4. **Focus on Data Management Plans and Planning**
5. The FAIR data principles
6. Reviewing DMPs
7. About FOSTER / questions

FOSTER

Focus on Data Management Planning

- Data Management Planning (DMP) **underpins and pulls together** the various strands of RDM 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
- Remember that Data Management Plans are **a means of communication**, with contemporaries and future re-users alike
  - This is especially beneficial in multi-partner, interdisciplinary research projects, where local and disciplinary norms may vary quite radically
  - The production of a DMP will likely involve input from many authors

FOSTER

DMPs are records of data-related activity

- A data management plan (or data management record) records how data was collected/created, processed, verified, described etc
  - In justifying decisions re. access, embargo, selection and appraisal… the list can be very long…
- It can be one of many useful documents
  - Projects may also attach more detailed metadata to describe the data, and help others to understand and re-use it
  - They may also attach workflows and README files
  - There may be multiple versions of datasets, software etc

Communication is crucial…

…and plans can and do change!

FOSTER

H2020 Data Management Plan

- The DMP should include information on:
  - the handling of research data during and after the end of the project
  - what data will be collected, processed and/or generated
  - which methodology and standards will be applied
  - whether data will be shared/made open access, and
  - how data will be curated and preserved (including after the end of the project)

- DMPs are submitted as deliverables – first version is due at the six-month stage
- Template and guidance is given in the Guidelines doc

FOSTER

Risks of not doing this, or getting this wrong

- **LEGAL** – sensitive data is protected by law (and contracts) and needs to be protected
- **FINANCIAL** – non-compliance with funder policies can lead to reduced access to income streams
- **SCIENTIFIC** – potential discoveries may be hidden away in drawers, on USB sticks or non-networked drives
- **OPPORTUNITY COST** – reduced visibility for research > lost opportunities for collaboration
- **QUALITY** – the scholarly record becomes less robust
- **REPUTATIONAL** – responsible data management is increasingly considered a core element of good scholarly practice in the 21st century

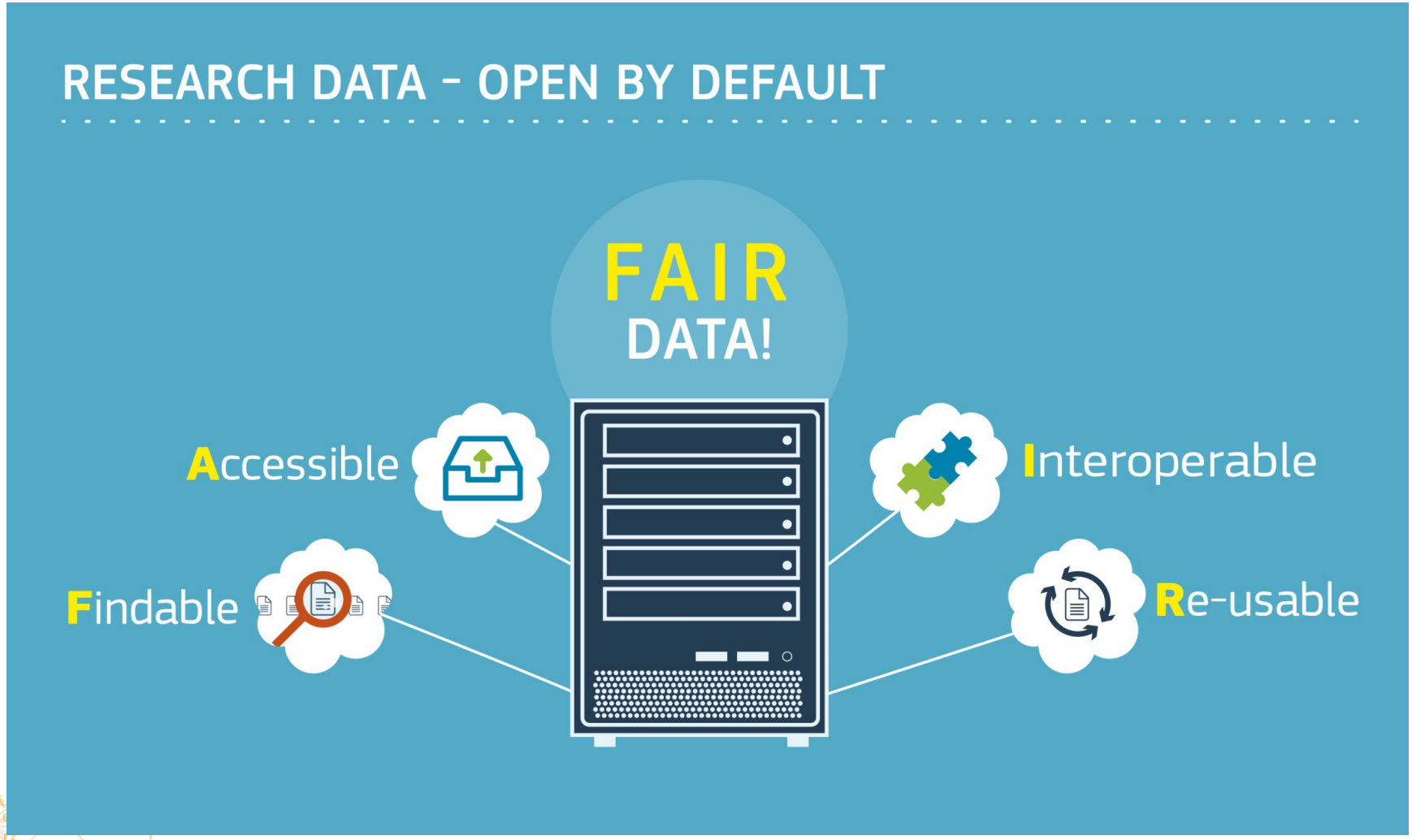
FOSTER

Presentation outline

1. About the DCC
2. RDM and research
3. Good practice in RDM
4. Focus on Data Management Plans and Planning
5. **The FAIR data principles**
6. Reviewing DMPs
7. Contacts and questions

FOSTER

FAIR data



FOSTER

The European Commission and FAIR data

- The EC has adopted FORCE11's 'FAIR' approach to research data management.
- These principles state that "One of the grand challenges of data-intensive science is to facilitate knowledge discovery by assisting humans and machines in their discovery of, access to, integration and analysis of, task-appropriate scientific data and their associated algorithms and workflows."
- To help achieve this, (meta)data should be…
  - **F**indable
  - **A**ccessible
  - **I**nteroperable
  - **R**eusable

FOSTER

The FAIR Data Principles (1/4)

FORCE11
The Future of Research Communications and e-Scholarship

To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

FOSTER

The FAIR Data Principles (2/4)

To be Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1. the protocol is open, free, and universally implementable.

A1.2. the protocol allows for an authentication and authorization procedure, where necessary.

A2. metadata are accessible, even when the data are no longer available.

The FAIR Data Principles (3/4)

FORCE11
The Future of Research Communications and e-Scholarship

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles.

I3. (meta)data include qualified references to other (meta)data.

FOSTER

The FAIR Data Principles (4/4)

FORCE11
The Future of Research Communications and e-Scholarship

To be Re-usable:

R1. meta(data) have a plurality of accurate and relevant attributes.

R1.1. (meta)data are released with a clear and accessible data usage license.

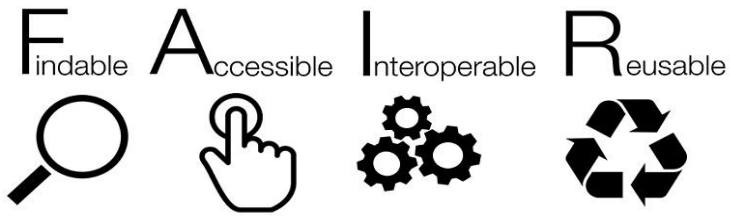
R1.2. (meta)data are associated with their provenance.

R1.3. (meta)data meet domain-relevant community standards.

FOSTER

FAIR and/or Open?

- Making data FAIR ensures it can be found, understood and reused – by the creator as well as others
- FAIR data does not have to be Open
  - Open data is a subset of all the data shared
  - Data can be shared under restrictions & still be FAIR
- *Approach: "As open as possible, as closed as necessary"*



FOSTER

FAIR data checklist

Findable

- Persistent ID
- Metadata online

Accessible

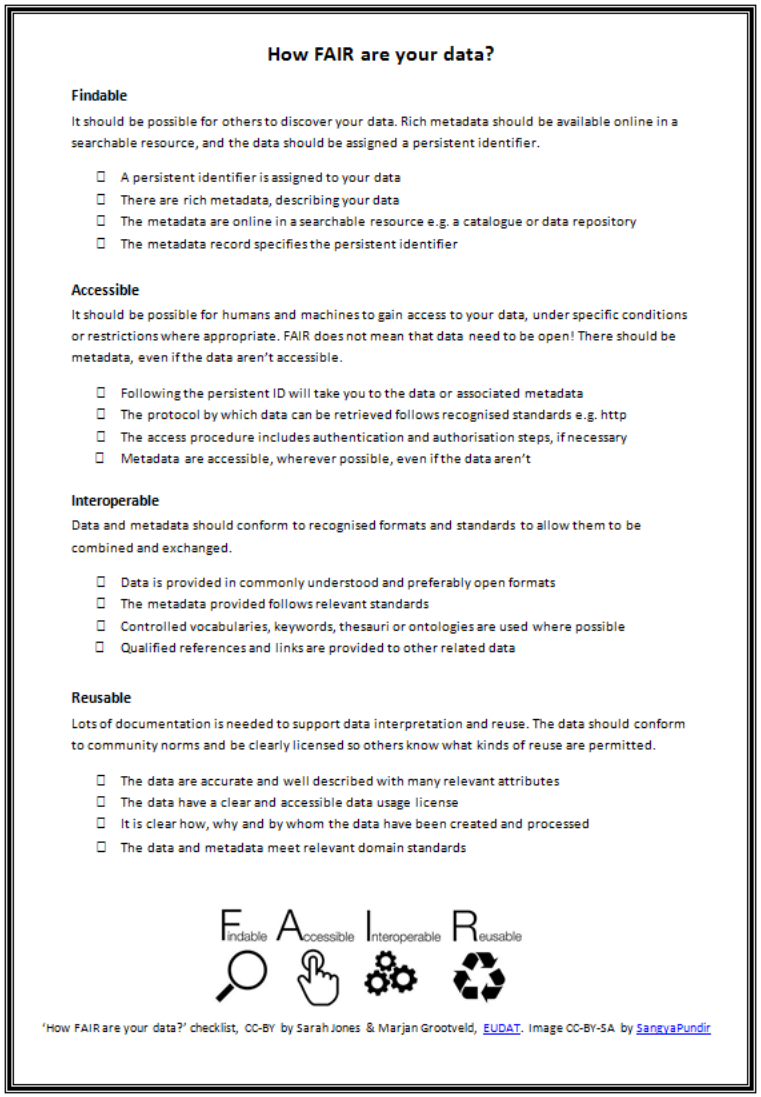
- Data online
- Restrictions where needed

Interoperable

- Use standards, controlled vocal
- Common (open) formats

Reusable

- Rich documentation
- Clear usage licence



How FAIR are your data?

Findable

It should be possible for others to discover your data. Rich metadata should be available online in a searchable resource, and the data should be assigned a persistent identifier.

- ☐ A persistent identifier is assigned to your data
- ☐ There are rich metadata, describing your data
- ☐ The metadata are online in a searchable resource e.g. a catalogue or data repository
- ☐ The metadata record specifies the persistent identifier

Accessible

It should be possible for humans and machines to gain access to your data, under specific conditions or restrictions where appropriate. FAIR does not mean that data need to be open! There should be metadata, even if the data aren't accessible.

- ☐ Following the persistent ID will take you to the data or associated metadata
- ☐ The protocol by which data can be retrieved follows recognised standards e.g. http
- ☐ The access procedure includes authentication and authorisation steps, if necessary
- ☐ Metadata are accessible, wherever possible, even if the data aren't

Interoperable

Data and metadata should conform to recognised formats and standards to allow them to be combined and exchanged.

- ☐ Data is provided in commonly understood and preferably open formats
- ☐ The metadata provided follows relevant standards
- ☐ Controlled vocabularies, keywords, thesauri or ontologies are used where possible
- ☐ Qualified references and links are provided to other related data

Reusable

Lots of documentation is needed to support data interpretation and reuse. The data should conform to community norms and be clearly licensed so others know what kinds of reuse are permitted.

- ☐ The data are accurate and well described with many relevant attributes
- ☐ The data have a clear and accessible data usage license
- ☐ It is clear how, why and by whom the data have been created and processed
- ☐ The data and metadata meet relevant domain standards

Findable Accessible Interoperable Reusable

'How FAIR are your data?' checklist, CC-BY by Sarah Jones & Marjan Grootveld, EUDAT. Image CC-BY-SA by SangyaPundir

https://zenodo.org/record/1065991

FOSTER

Presentation outline

1. About the DCC
2. RDM and research
3. Good practice in RDM
4. Focus on Data Management Plans and Planning
5. The FAIR data principles
6. **Reviewing DMPs**
7. About FOSTER / questions

FOSTER

Reflections on assessing H2020 DMPs

- It would be better if everyone followed the same template – the EC does provide one, but usage isn't (yet) mandatory
- A DMP doesn't need to tell everything there is to know about a project: brevity is a plus!
- Areas of frequent weakness: security (access and storage), ethical restrictions for data sharing, appraisal of long-term value/interest, quality assurance processes, costs
- Advice:
  - Be clear about the different between in-project and post-project data storage and archiving;
  - Don't just regurgitate the H2020 guidelines – reviewers pick up on that really quickly;
  - Try not to confuse publications and data (I have seen projects describe archived data as 'gold Open Access' which doesn't make much sense)

FOSTER

Things to look out for

1. Clarity
2. Regurgitation
3. Omissions

FOSTER

Point 1. Clarity

- **RDM is a hybrid activity**, involving multiple stakeholder groups…
  - The researchers themselves
  - Research support personnel
  - Partners based in other institutions, funders, data centres, commercial partners, etc
- Plans should be clear about who does what, when, how, etc
- No single person does everything, and it makes no sense to duplicate effort or reinvent wheels

FOSTER

Point 2. Regurgitation

- Look out for stock phrases which either directly quote or paraphrase the H2020 Guidelines.
  - This is often a sign that the project is only paying lip service to RDM, and has not thought things through for itself!
  - Remember: thinking things through is one of the main points of DMP
  - The planning process is every bit as important as the plans produced

FOSTER

Point 3. Omissions

- It’s very possible to read a DMP and get a feeling that they authors know what they are talking about – being lulled into a false sense of security.
  - This is particularly possible when a project utilises the services of professional scientific communicators
- Often it’s only when using the Assessment Grid that you realise they have failed to address some of the fundamental data-related issues
  - So while you may get a sense of confidence that data-related matters are in hand, it’s necessary to check off each issue against an assessment rubric – confidence AND confirmation

FOSTER

Recap: a few RDM rules of thumb

- Without intervention, data + time = no data
  - See 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
- Prioritise: could anyone die or go to jail?
  - Legal issues (e.g. protecting vulnerable subjects) are the most important
- Storage is not the same as management
  - Think of data as plants and the servers as a greenhouse
  - The plants still need to be fed, watered, pruned, etc… and sometimes disposed of
- Management is not the same as sharing
  - Not all data should be shared
  - Approach: “As open as possible, as closed as necessary”
- Remember that plans are just that – they are not contracts!

FOSTER

Presentation outline

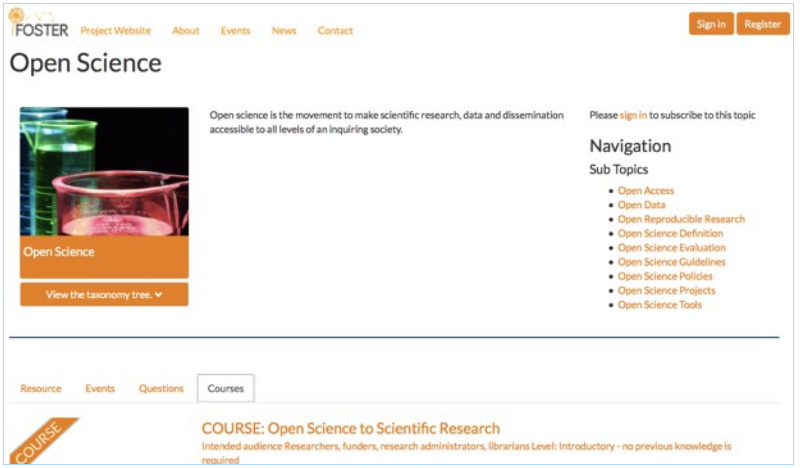
1. About the DCC
2. RDM and research
3. Good practice in RDM
4. Focus on Data Management Plans and Planning
5. The FAIR data principles
6. Reviewing DMPs
7. **About FOSTER / questions**

FOSTER

The FOSTER project

Facilitate Open Science Training for European Research

- Phase 1 (2014-2016): Spread the Seeds of Open Science and Open Access
- Creation of Open Science Taxonomy
- 2000+ training materials, categorized in the FOSTER Portal
- More than 100 f2f training events in 28 countries and 25 online courses, totalling more than 6300 participants



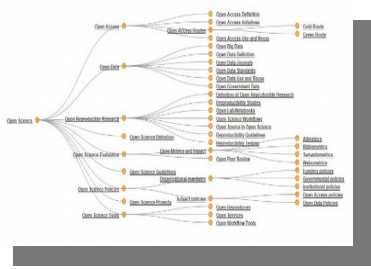
FOSTER

http://fosteropenscience.eu

The FOSTER project

Facilitate Open Science Training for European Research

- Phase 2 (2017-2019): Let the Flowers of Open Science Bloom
- Focus on:
  - Training for the practical implementation of Open Science (face to face and online) including RDM and Open Data
  - Developing intermediate/advanced level/discipline-specific training resources in collaboration with three disciplinary communities (ar related RIs): Life Sciences (ELIXIR), Social Sciences (CESSDA) and Humanities (DARIAH)
  - Update the FOSTER Portal to support moderated learning, badges and gamification
- In concrete terms:
  - 150 new training resources
  - Over 50 training events (outcome-oriented, providing participant with tangible skills) and 20 e-learning courses
  - Multi-module Open Science Toolkit
  - Trainers Network, Open Science Bootcamp, Open Science Training Handbook, and more...

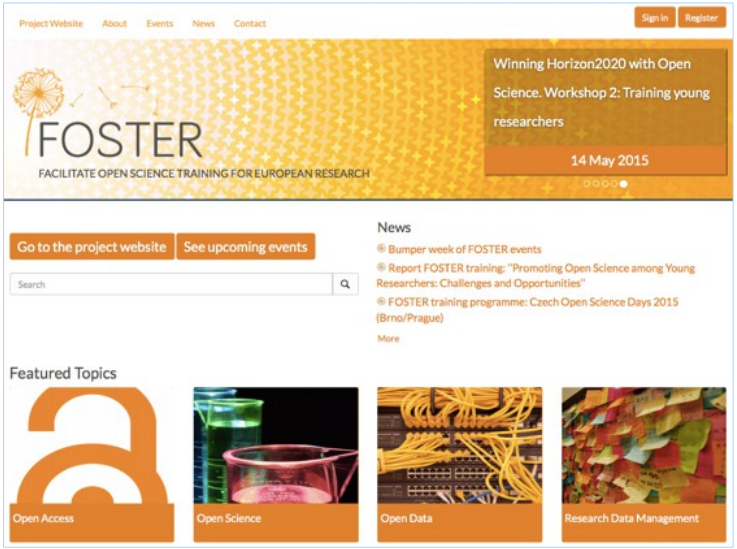


FOSTER

http://fosteropenscience.eu

Contact details

- 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)
  - Events: www.fosteropenscience.eu/events
  - Twitter: @fosterscience and #fosteropenscience



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

FOSTER

N.B. Some slides courtesy of Sarah Jones!