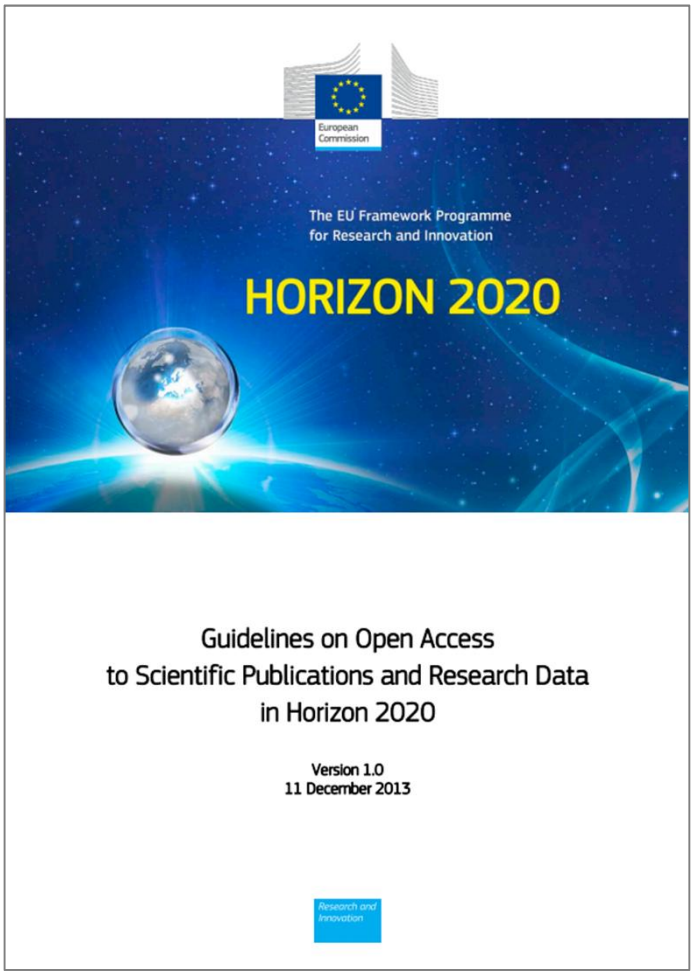
The Horizon 2020 Open Data Pilot

Sarah Jones
Digital Curation Centre, University of Glasgow
sarah.jones@glasgow.ac.uk
Twitter: sjDCC

Why open access and open data?



"The European Commission's vision is that information already paid for by the public purse should not be paid for again each time it is accessed or used, and that it should benefit European companies and citizens to the full."

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

What is research data?

‘Research data’ refers to information, in particular facts or numbers, collected to be examined and considered as a basis for reasoning, discussion or calculation.

In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. The focus is on research data that is available in digital form.

Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020
v.1.0, 11 December 2013, Footnote 5, p3

FOSTER

What is open data?

Openly accessible research data can typically be accessed, mined, exploited, reproduced and disseminated, free of charge for the user.

Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, p3

Tim Berners-Lee's proposal for five star open data - http://5stardata.info

- ★ make your stuff available on the Web (whatever format) under an open licence
- ★★ make it available as structured data (e.g. Excel instead of a scan of a table)
- ★★★ use non-proprietary formats (e.g. CSV instead of Excel)
- ★★★★ use URIs to denote things, so that people can point at your stuff
- ★★★★★ link your data to other data to provide context

FOSTER

H2020 OPEN DATA PILOT

Guidelines on Data Management in Horizon 2020

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

H2020 areas participating in the pilot

- Future and Emerging Technologies
- Research infrastructures – part e-Infrastructures
- Leadership in enabling and industrial technologies – Information and Communication Technologies
- Societal Challenge: 'Secure, Clean and Efficient Energy' – part Smart cities and communities
- Societal Challenge: 'Climate Action, Environment, Resource Efficiency and Raw materials' – except raw materials
- Societal Challenge: 'Europe in a changing world – inclusive, innovative and reflective Societies'
- Science with and for Society

Projects in other areas can participate on a voluntary basis

FOSTER

Why would researchers want to opt in?(1)

"It was a mistake in a spreadsheet that could have been easily overlooked: a few rows left out of an equation to average the values in a column.

The spreadsheet was used to draw the conclusion of an influential 2010 economics paper: that public debt of more than 90% of GDP slows down growth. This conclusion was later cited by the International Monetary Fund and the UK Treasury to justify programmes of austerity that have arguably led to riots, poverty and lost jobs."

... validation of results

The error that could subvert George Osborne's austerity programme

The theories on which the chancellor based his cuts policies have been shown to be based on an embarrassing mistake

Charles Arthur and **Phillip Inman**
The Guardian, Thursday 18 April 2013 21.10 BST

George Osborne says that Ken Rogoff, the man whose economic error has been uncovered, has strongly influenced his thinking. Photograph: Stefan Wermuth/PA

FOSTER

www.guardian.co.uk/politics/2013/apr/18/uncovered-error-george-osborne-austerity

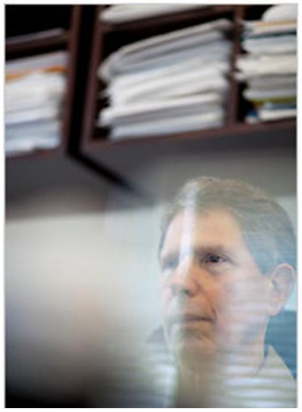
Why would researchers want to opt in?(2)

Sharing of Data Leads to Progress on Alzheimer's

By GINA KOLATA
Published: August 12, 2010

In 2003, a group of scientists and executives from the National Institutes of Health, the Food and Drug Administration, the drug and medical-imaging industries, universities and nonprofit groups joined in a project that experts say had no precedent: a collaborative effort to find the biological markers that show the progression of Alzheimer's disease in the human brain.

Enlarge This Image



Now, the effort is bearing fruit with a wealth of recent scientific papers on the early diagnosis of Alzheimer's using methods like PET scans and tests of spinal fluid. More than 100 studies are under way to test drugs that might slow or stop the disease.

And the collaboration is already serving as a model for similar efforts against Parkinson's disease. A $40 million project to look for biomarkers for Parkinson's, sponsored by the Michael J. Fox Foundation, plans to enroll 600 study subjects in the United States and Europe.

www.nytimes.com/2010/08/13/health/research/13alzheimer.html?pagewanted=all&_r=0

"It was unbelievable. Its not science the way most of us have practiced in our careers. But we all realised that we would never get biomarkers unless all of us parked our egos and intellectual property noses outside the door and agreed that all of our data would be public immediately."

Dr John Trojanowski, University of Pennsylvania

FOSTER

... scientific breakthroughs

Why would researchers want to opt in?(3)

“There is evidence that studies that make their data available do indeed receive more citations than similar studies that do not.”

Piwowar H. and Vision T.J 2013 "Data reuse and the open data citation advantage“ https://peerj.com/preprints/1.pdf

9% - 30% increase

... more citations

FOSTER

Exemptions – reasons for opting out

- If results are expected to be commercially or industrially exploited
- If participation is incompatible with the need for confidentiality in connection with security issues
- Incompatible with existing rules on the protection of personal 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 reason to not take part in the Pilot

Can opt out at proposal stage OR during lifetime of project.
Should describe issues in the project Data Management Plan.

FOSTER

Which data does the pilot apply to?

- Data, including associated metadata, needed to validate the results in scientific publications

- Other curated and/or raw data, including associated metadata, as specified in the DMP

Doesn't apply to all data (researchers to define as appropriate)

Don't have to share data if inappropriate – exemptions apply

FOSTER

Metadata and documentation

Metadata: basic info e.g. title, author, dates, access rights...
Documentation: methods, code, data dictionary, context...

Use **standards** wherever possible for interoperability

www.dcc.ac.uk/resources/metadata-standards

FOSTER

Requirements of the open data pilot

1. Develop (and update) a Data Management Plan
2. Deposit in a research data repository
3. Make it possible for third parties to access, mine, exploit, reproduce and disseminate data – free of charge for any user
4. Provide information on the tools and instruments needed to validate the results (or provide the tools)

FOSTER

1. Develop a Data Management Plan

Not a fixed document – should evolve and gain precision

- Deliver first version within initial 6 months of project
- More elaborate versions whenever important changes to the project occur. At least at the mid-term and final review.

Two templates provided (annex 1 & 2)

Note that the Commission does NOT require applicants to submit a DMP at the proposal stage. A DMP is therefore NOT part of the evaluation.

However, all project proposals submitted to "Research and Innovation actions", as well as "Innovation actions", include a section on research data management which is evaluated under the criterion 'Impact'.

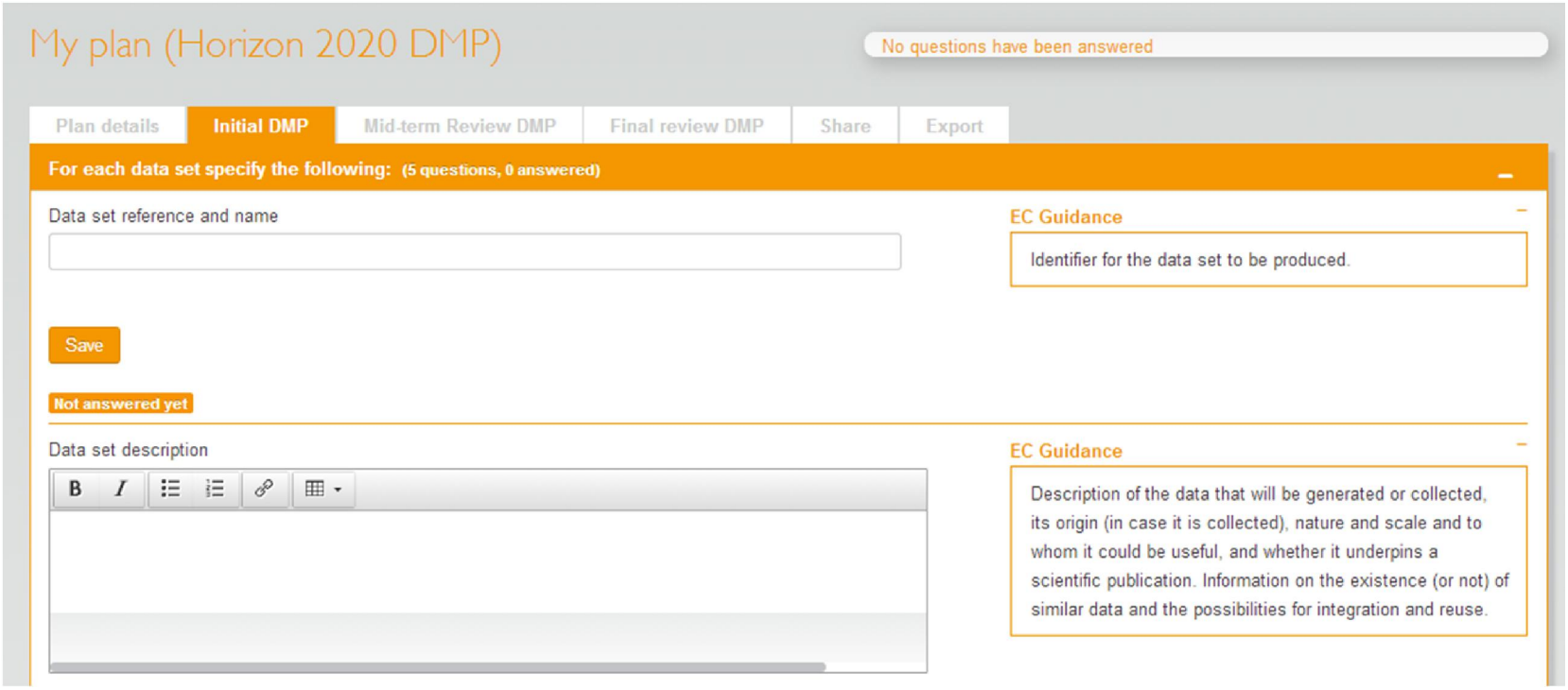
Guidelines on Data Management in Horizon 2020, v.1.0, 11 December 2013
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

FOSTER

DMPonline

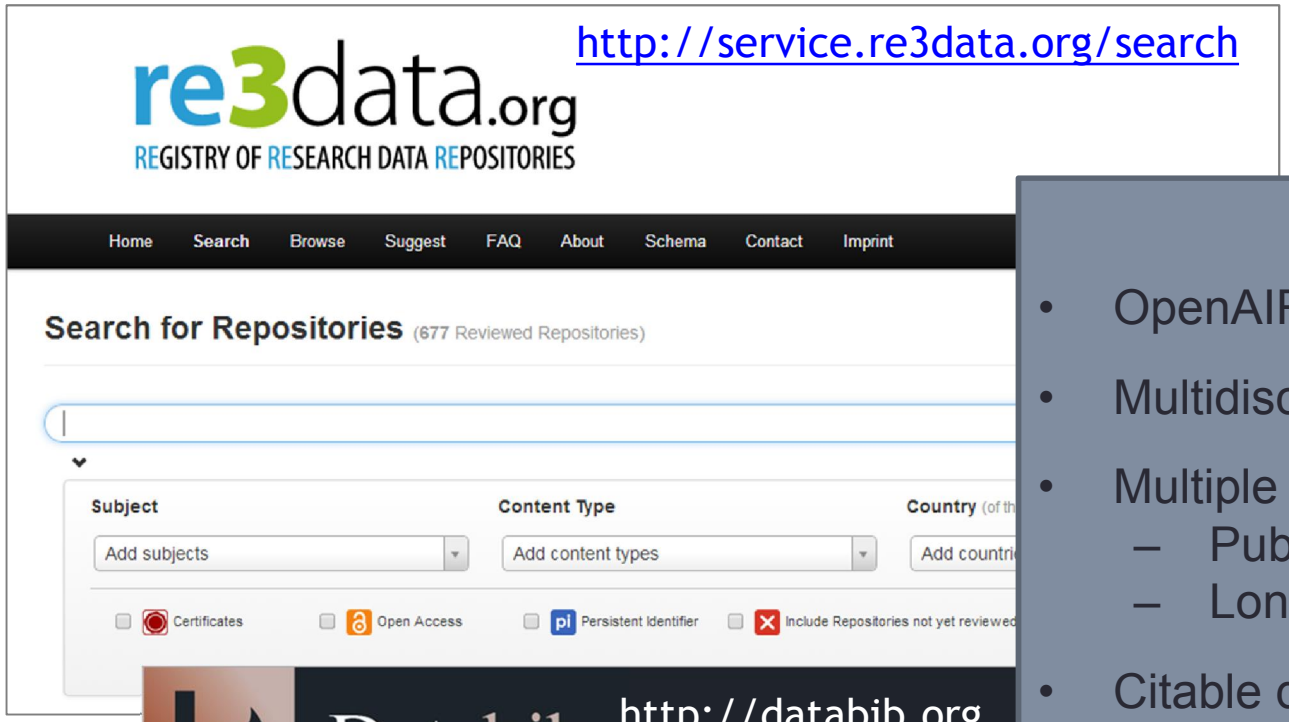
A web-based tool to help researchers write DMPs

Includes a template for Horizon 2020



https://dmponline.dcc.ac.uk

2. Deposit in a repository



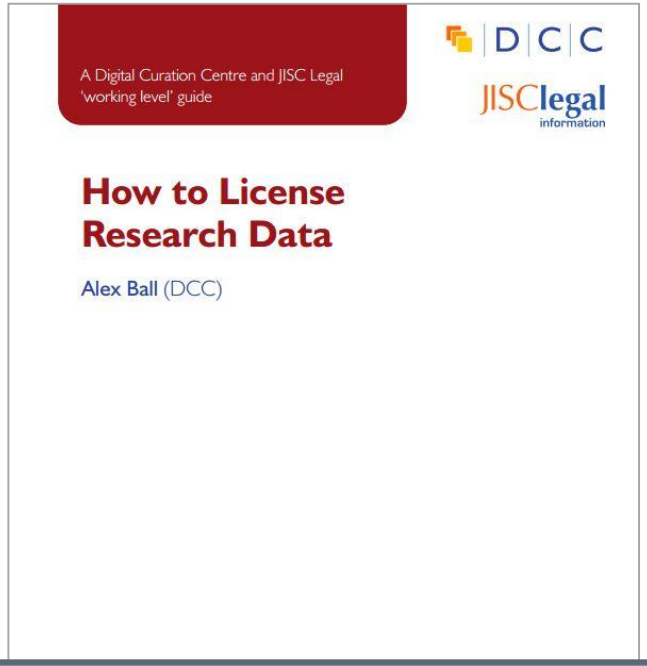
Zenodo

- OpenAIRE-CERN joint effort
- Multidisciplinary repository
- Multiple data types
  - Publications
  - Long tail of research data
- Citable data (DOI)
- Links to funding, publications, data & software

www.zenodo.org

FOSTER

3. License your data for reuse



Outlines 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?

SA Share Alike
Reduces interoperability

ND No Derivatives
Severely restricts use

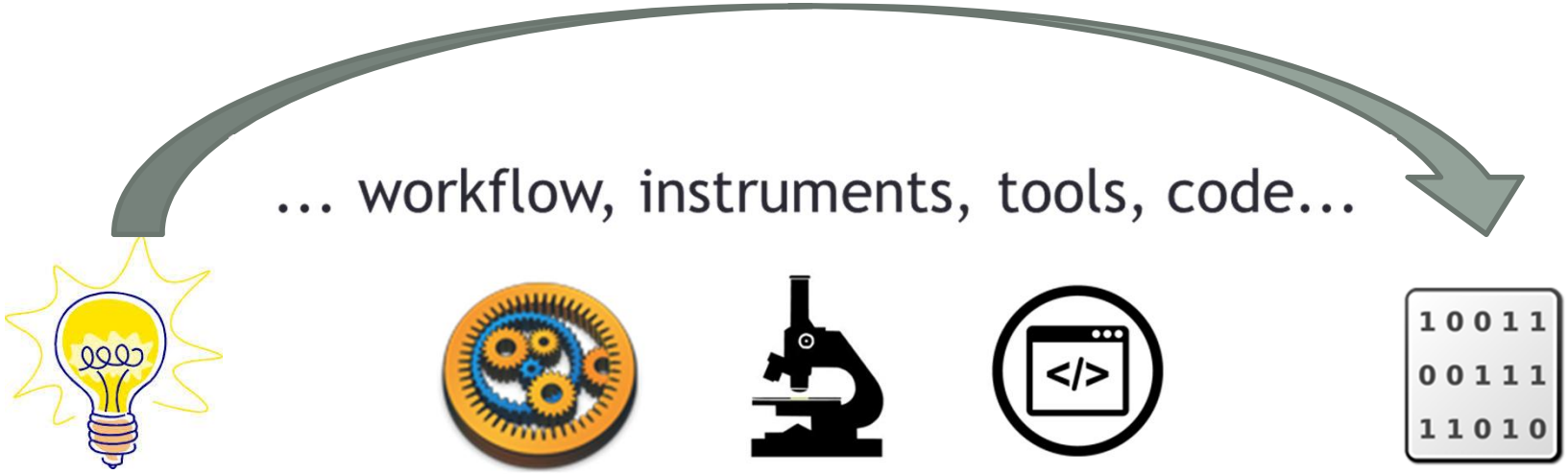
Horizon 2020 recommendation is to use

CC ZERO OR CC BY

p://www.dcc.ac.uk/resources/how-des/license-research-data

4. Provide info on tools needed for validation

Need to share much more than just the data
for research to be reproducible...



Difficult to validate data if you're missing info on the steps between the initial idea and end results

FOSTER

Useful links

- Open Knowledge Foundation (advocacy, training, services, handbook...) https://okfn.org
- MyExperiment and Taverna (sharing workflows) http://www.myexperiment.org and http://www.taverna.org.uk
- Software Sustainability Institute (UK-based) http://www.software.ac.uk
- School of Data (training to help people use open data) http://schoolofdata.org
- Digital Curation Centre (RDM guidance, tools and resources) http://www.dcc.ac.uk/resources

FOSTER

Discussion

- What concerns / misconceptions need to be overcome about open data?
- What guidance, tools and resources do you need to know about to support projects in the open data pilot?
- What assistance is needed to review DMPs and monitor the success of the pilot?
- What other issues or recommendations do you have?

FOSTER