Data Management Plans

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

Data Management Plan (DMP) workshop, e-infrastructures Austria, Vienna, 17 November 2016

What is a DMP?

A DMP is a brief plan to define:

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

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

Why manage data?

NON PECUNIAE INVESTIGATIONIS CURATORE

SED VITAE FACIMUS PROGRAMMAS DATORUM PROCURATIONIS

(Not for the research funder, but for life we make data management plans)

- Make your research easier
- Stop yourself drowning in irrelevant stuff
- Save data for later
- Avoid accusations of fraud or bad science
- Write a data paper
- Share your data for re-use
- Get credit for it

Don’t undervalue research data

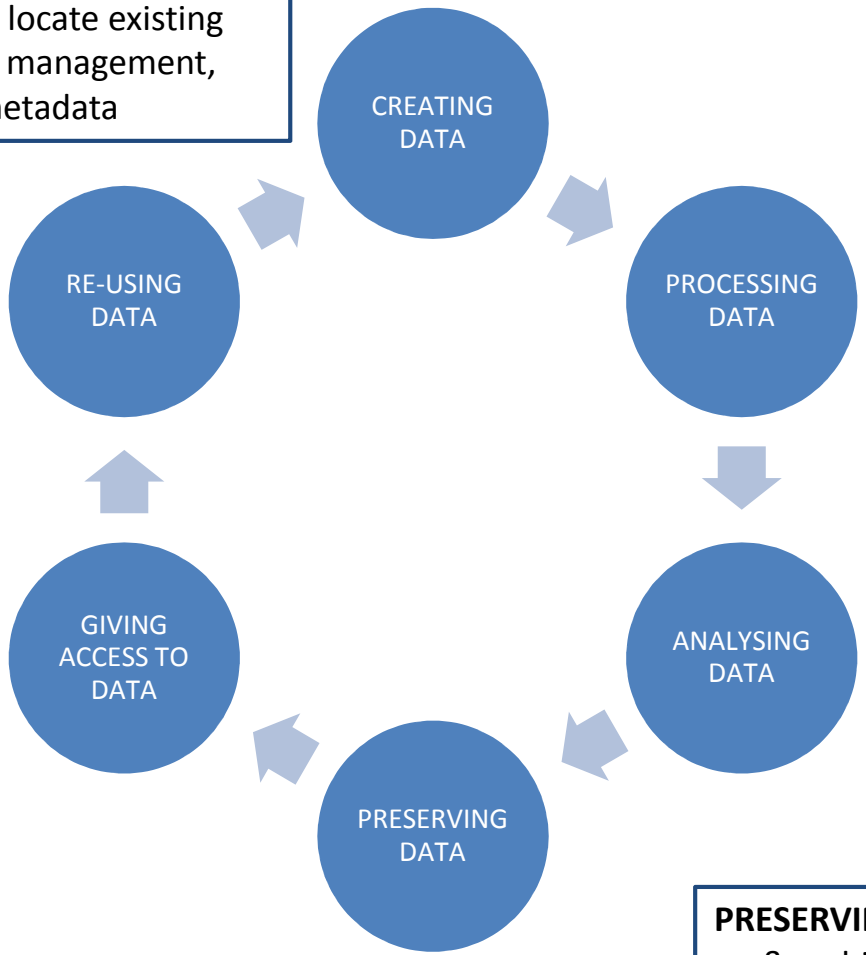
PUBLICATIONS AND DATA

Benefits of DMPs for institutions

- Opportunity to engage with researchers and improve RDM practice
- Raise awareness of support available
- Collate information to inform service delivery
- Ensure the University is not exposed to risk
- Ability to recover costs via grants

Research data lifecycle

CREATING DATA: designing research, DMPs, planning consent, locate existing data, data collection and management, capturing and creating metadata



PROCESSING DATA: entering, transcribing, checking, validating and cleaning data, anonymising data, describing data, manage and store data

RE-USING DATA: follow-up research, new research, undertake research reviews, scrutinising findings, teaching & learning

ANALYSING DATA: interpreting, & deriving data, producing outputs, authoring publications, preparing for sharing

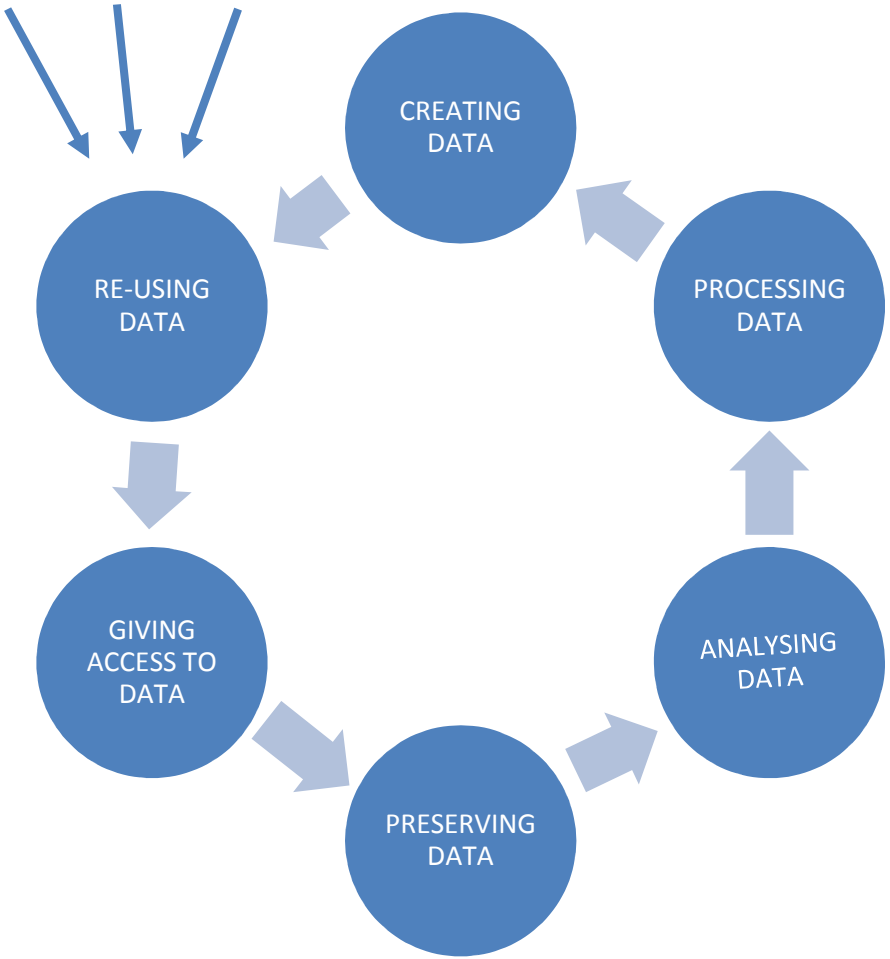
ACCESS TO DATA: distributing data, sharing data, controlling access, establishing copyright, promoting data

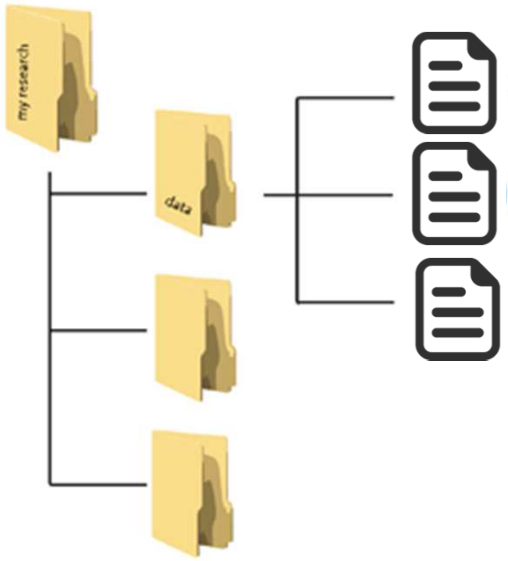
PRESERVING DATA: data storage, back-up & archiving, migrating to best format & medium, creating metadata and documentation

Ref: UK Data Archive: http://www.data-archive.ac.uk/create-manage/life-cycle

Planning trick 1: think backwards

What data organisation would a re-user like?





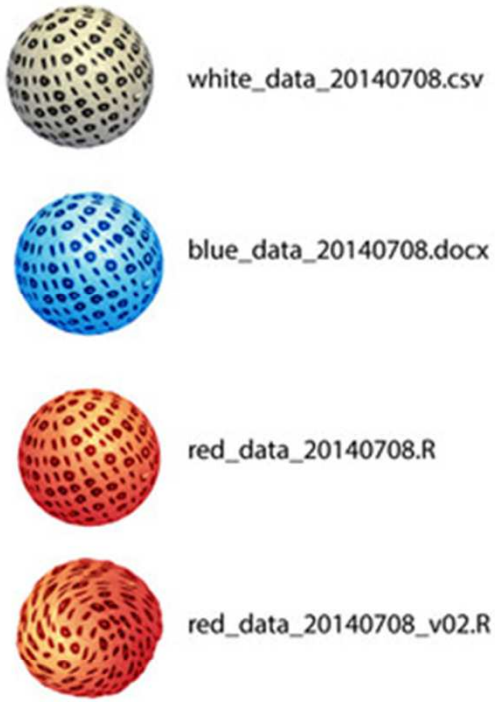
Data organisation

Meaningful file names

Below are tips on meaningful and consistent file names. Read more in *'Choosing a file name'*. (2)

- Make sure to use consistent file names. When you use a date in the file name, choose a notation (for instance, YYYYMMDD of yymmdd).
- Do not use strange characters like ?\!@*%{[<> in the file name.
- Use traceable file names, such as Project_Instrument_locatie_YYYYMMDD.ext.
- Make sure to only use each file once in the folder structure. If you store a file in more than one place, several versions of the same file can unwillingly be created.
- See also version management.

It is good practice to note the file naming and its meaning in a readme.txt.



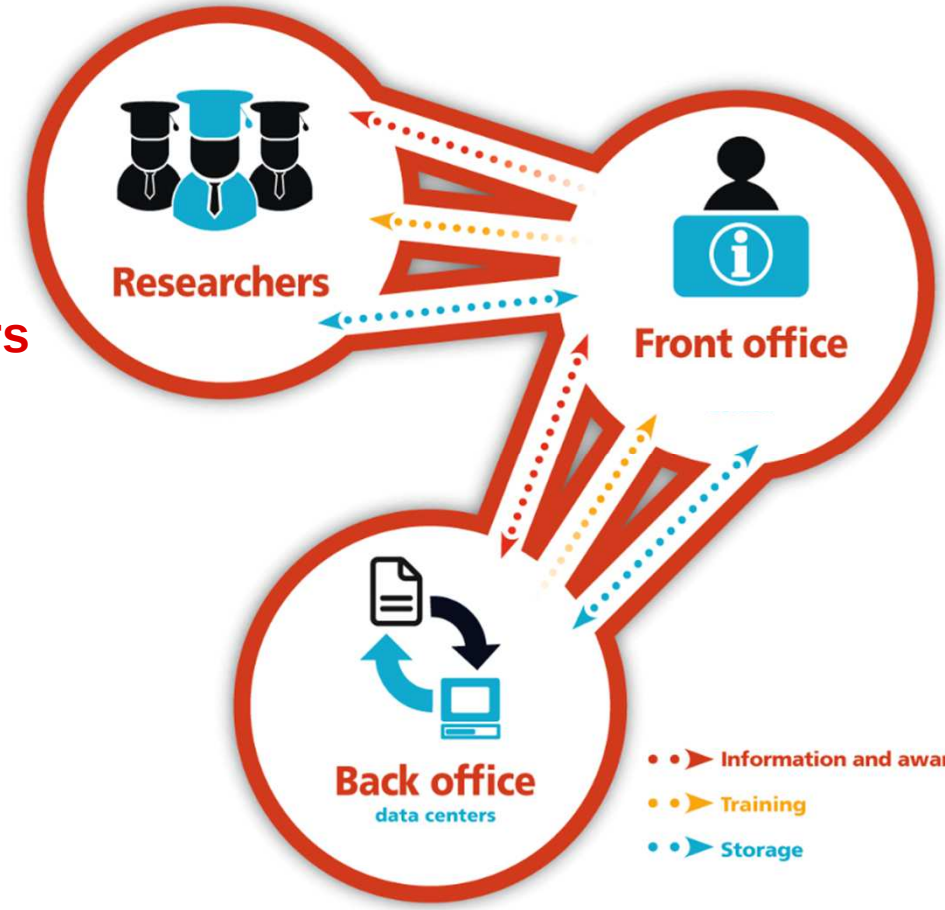
File naming and version management

Even if a researcher is well underway with his project consistent file naming is still an option by using a *bulk file rename utility*. (3) It is important, however, to check if this bulk renamer delivers on its promises.

http://datasupport.researchdata.nl/en/start-de-cursus/iii-onderzoeksfase/organising-data

Planning trick 2: include RDM stakeholders

Commercial partners



Institution

RDM policy

Facilities

Publishers

Data Availability policy

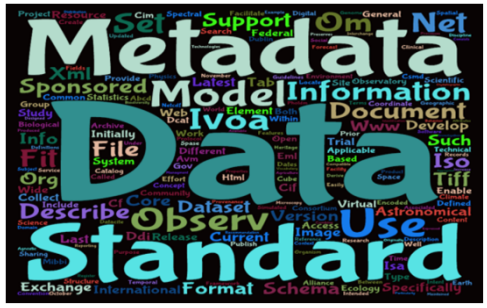
Research funders

www.openaire.eu/briefpaper-rdm-infonoads

Planning trick 3: ground your plan in reality

Base plans on available skills, support and good practice for the field – show it's feasible to implement

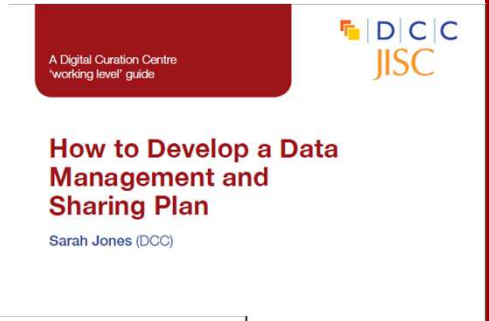
re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES



DCC support on DMPs

- Webinars and training materials
- How-to guides and other advisory documents
- Checklist on what to cover in DMPs
- Example DMPs
- DMPonline

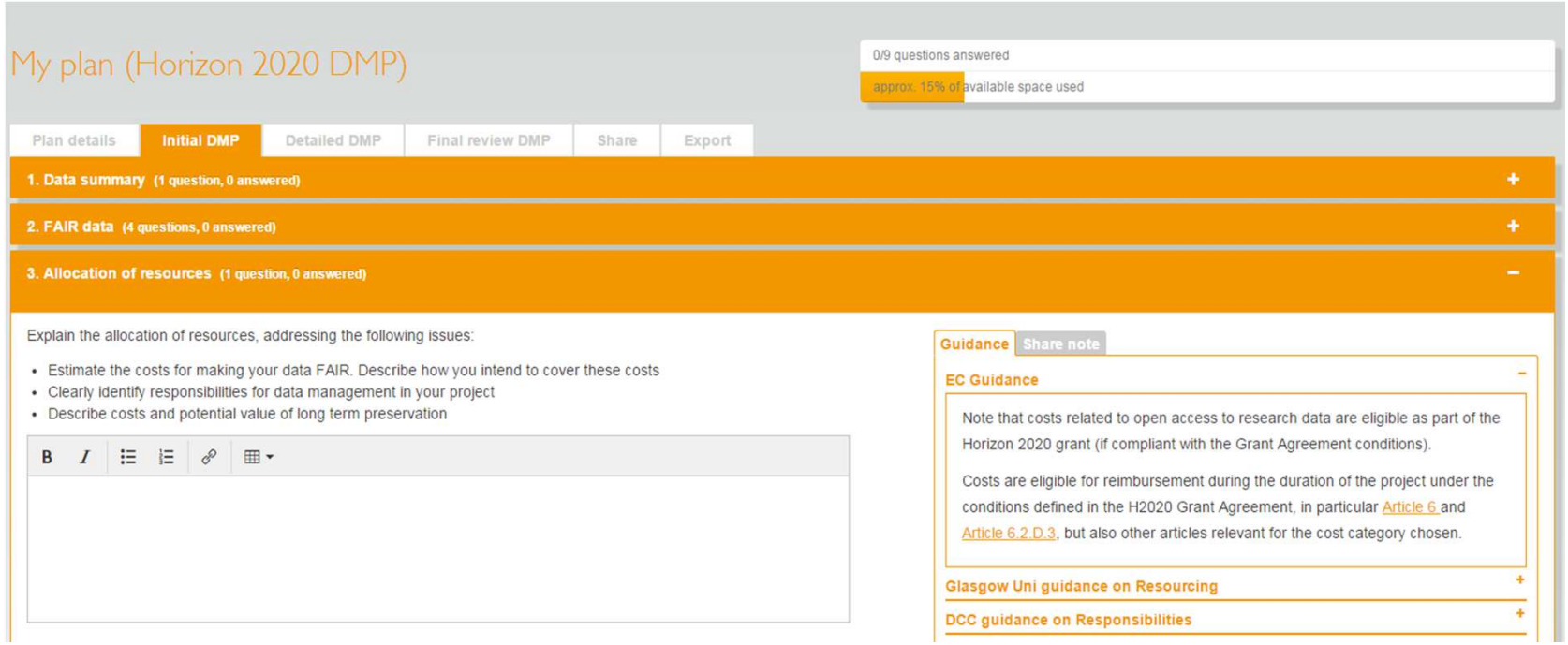
www.dcc.ac.uk/resources/data-management-plans



What is DMPonline?

A web-based tool to help researchers write DMPs

Includes a template for Horizon 2020

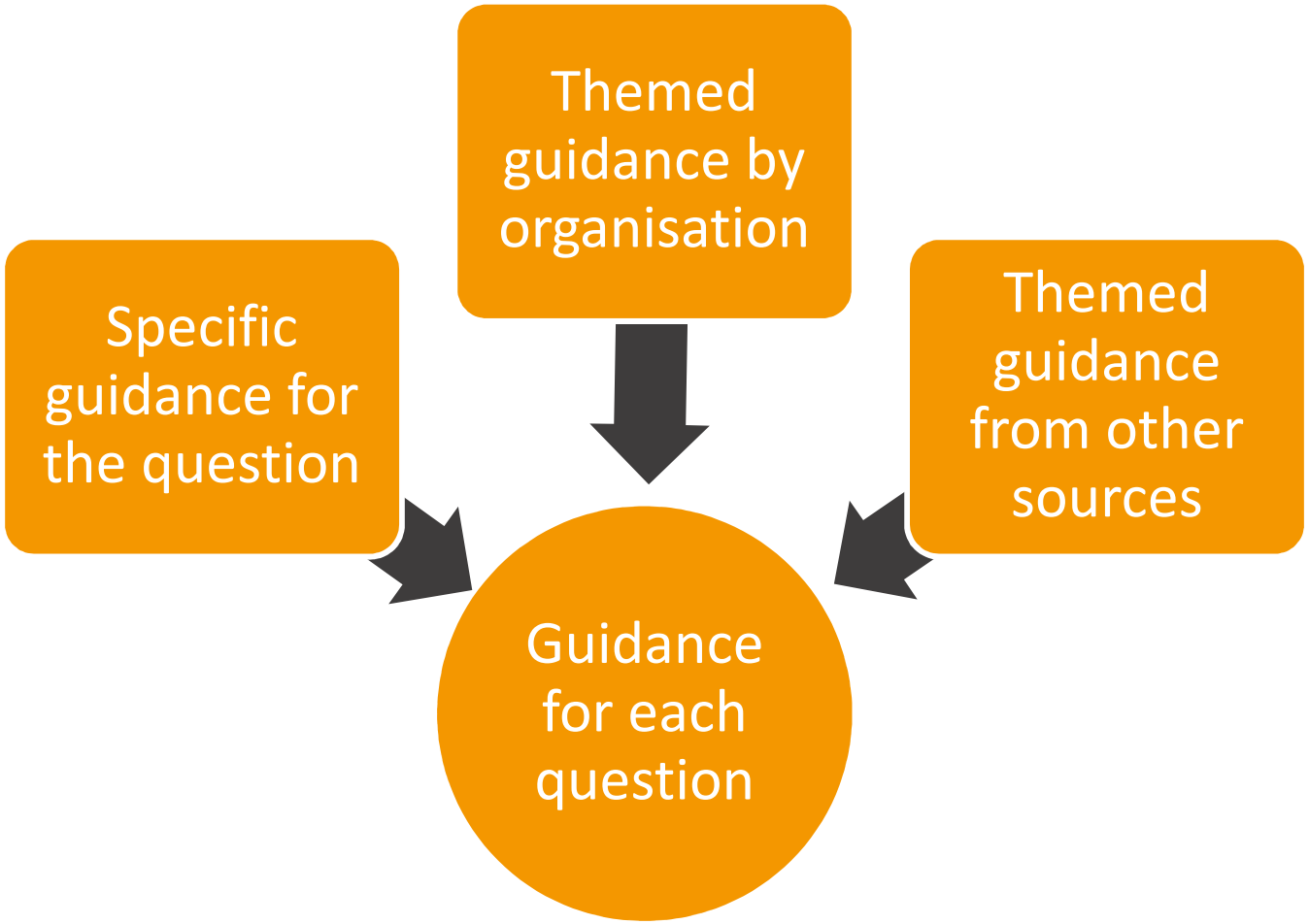


https://dmponline.dcc.ac.uk

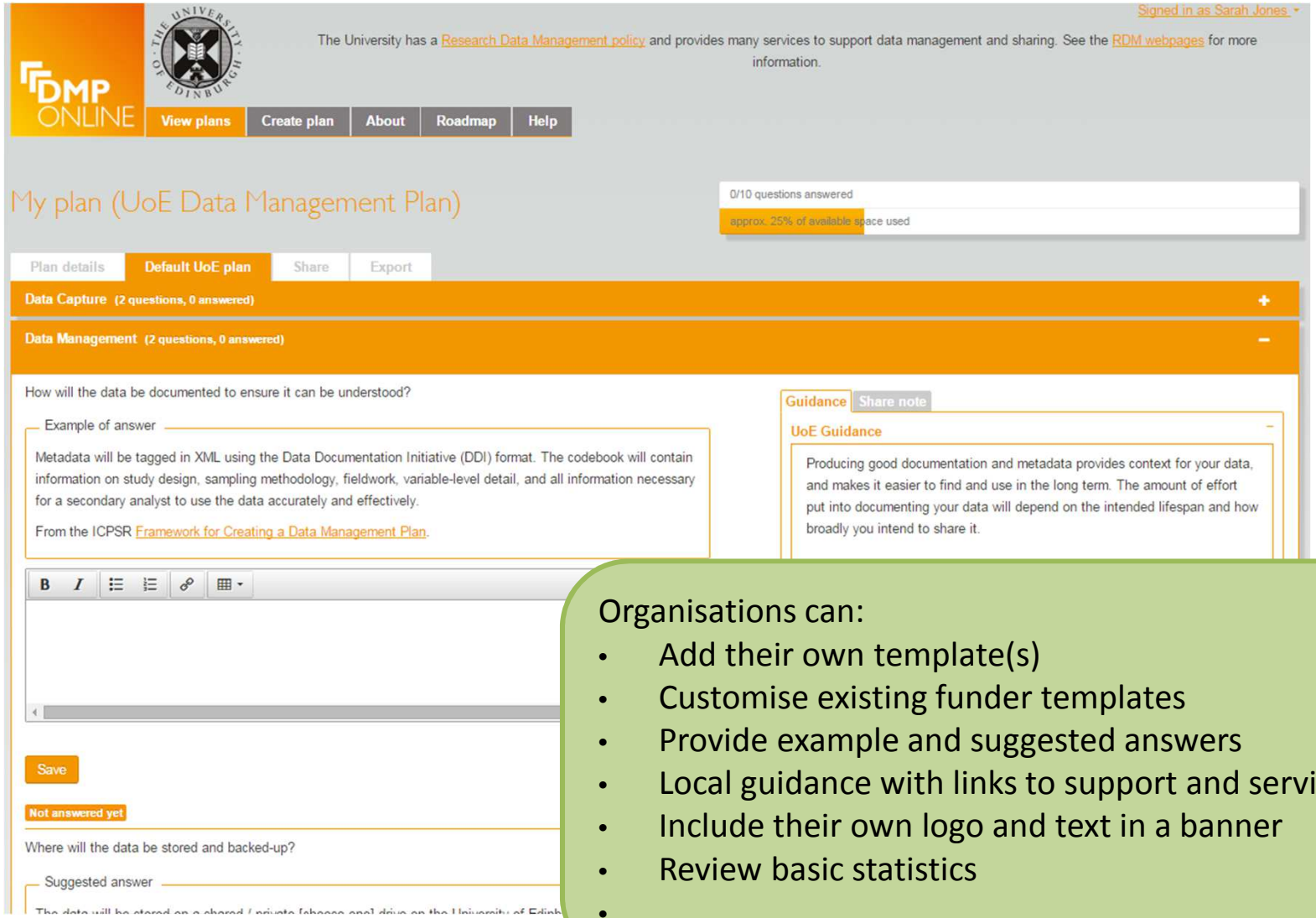
Main features in DMPonline

- Templates for different requirements (funder or institution)
- Tailored guidance (funder, institutional, discipline-specific etc)
- Ability to provide examples and suggested answers
- Supports multiple phases (e.g. pre- / during / post-project)
- Granular read / write / share permissions
- Customised exports to a variety of formats
- Shibboleth authentication → eduGAIN

Guidance in DMPonline



Options for unis to customise DMPonline



Organisations can:

- Add their own template(s)
- Customise existing funder templates
- Provide example and suggested answers
- Local guidance with links to support and services
- Include their own logo and text in a banner
- Review basic statistics
- …

A single platform for all things DMP

Agreed to converge on a single codebase, based on DMPonline with additional features from DMPTool

Bring together features and strengths of each tool

Co-manage, co-develop and issue joint roadmap

DMPRoadmap: https://github.com/DMPRoadmap

A FAIR approach to DMPs

Findable

– *Assign persistent IDs, provide metadata, register in a searchable resource...*

Accessible

– *Retrievable by their ID using a standard protocol, metadata remain accessible even if data aren't...*

Interoperable

– *Use formal, broadly applicable languages, use standard vocabularies, qualified references...*

Reusable

– *Rich metadata, clear licences, provenance, use of community standards...*

www.force11.org/group/fairgroup/fairprinciples

H2020 template

1. Data summary
2. FAIR data
   - 2.1 Making data findable, including provisions for metadata
   - 2.2 Making data openly accessible
   - 2.3 Making data interoperable
   - 2.4 Increase data re-use (through clarifying licences)
3. Allocation of resources
4. Data security
5. Ethical aspects
6. Other issues

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

Key differences in H2020

- The Commission does NOT require applicants to submit a DMP at the proposal stage. It's a deliverable (due by month 6).
- A DMP is therefore NOT part of the evaluation
- Optional section on data management in proposal is worth doing, especially to help justify costs
- A DMP is a living or "active" document that should be updated

Example H2020 DMPs in Zenodo

Helix Nebula – High Energy Physics example

https://zenodo.org/record/48171#.WATexnriF40

Tweether – engineering (micro-electronics) example

https://zenodo.org/record/55791#.WATei3riF40

AutoPost – ICT example

https://zenodo.org/record/56107#.WATefXriF40

Example: OpenMinTed

OpenMinTed aims to create an infrastructure for Text and Data Mining (TDM) of scientific and scholarly content

Have adopted their own structure to create a ‘Data and Software Management Plan’

1. Table of Contents

openMINTED
Open Mining Infrastructure for Text & Data

http://openminted.eu

Example: OpenMinTed – Data chapter

Six high-level datasets identified:

1. Scholarly publications
2. Language and knowledge resources
3. Services and workflows
4. Automatically and manually generated annotations
5. Consortium publications
6. Metadata

Described in a table per dataset (see illustration)

Table 1 – Scholarly Publications

Information Needed	Answer
Data set reference and name	ID or PID (Persistent Identifier) or DOI as a long-lasting reference to a publication or set of publications, depending on usage (eg. temporary data used for processing should be assigned a simple OpenMinTeD ID)
Data set description	Collections of publications stored for searching and sharing as well as input to be processed by workflows and services
Standards and metadata	To be decided, in accordance with our Interoperability guidelines (See also Table 6.)
Data sharing	Available under permissive licenses, (CC-BY 4.0, CC-0 or comparable) but certain conditions (e.g. Noncommercial use=NC) and/or exceptions may also apply
Archiving and preservation	To be decided. Depending on usage, data used for processing will be stored in PITHOS, persistent storage provided by GRNET and follow its respective archiving and preservation policy

OpenMinTed – Software examples

4.3 Intellectual Property

Information Needed	Answer
What type of software license have you chosen?	We plan to release our code under the Apache Software License 2.0 (http://www.apache.org/licenses/LICENSE-2.0) if possible. Alternative OSI-approved open source licenses if necessary due to external constraints, e.g. due to copyleft conditions on used third-party libraries. In some cases, external constraints may also mandate the release under proprietary licenses.
Have you chosen an OSI-approved open source license	yes
Is the license valid under your national laws	yes
Is your license acceptable to all partners	yes
Where will you publish your license	A copy of the license must be included with every released artifact, e.g. in e Java JAR file, in a ZIP archive containing the software. Furthermore, each source file shall bear a suitable license header if this is technically feasible

How you will track who does and has done what	GitHub's issue tracker also provides full tracking and history of changes is also available at any time. Due to the distributed nature of Git, copies of the full history are also typically maintained in clones of the canonical repository on Github, e.g. in the copies that individual developers keep on their machines or on the build server. Strong cryptographic measures ensure that changes to history (if made at all) do not go unnoticed.

Helix Nebula: access policy

"The 4 LHC experiments have policies for making data available, including reasonable embargo periods, together with the provision of the necessary software, documentation and other tools for re-use."

"The meta-data catalogues are typically experiment-specific although globally similar. The "open data release" policies foresee the available of the necessary metadata and other "knowledge" to make the data usable"

"Re-use of the data is made by theorists, by the collaborations themselves, by scientists in the wider context as well as for education and outreach."

Helix Nebula: open data

"Data releases through the CERN Open Data Portal (http://opendata.cern.ch) are published with accompanying software and documentation. A dedicated education section provides access to tailored datasets for self-supported study or use in classrooms."

"All materials are shared with Open Science licenses (e.g. CC0 or CC-BY) to enable others to build on the results of these experiments. All materials are also assigned a persistent identifier and come with citation recommendations."

"The data behind plots in publications has been made available since many decades via an online database: http://hepdata.cedar.ac.uk"

Plan to share data from the outset

Decisions made early on affect what you can do later

- Negotiation on licenses and consent agreement may preclude later sharing if not careful
- Costings can’t be included retrospectively
- Useful to consider data issues at the consortium negotiation stage to make sure potential issues are identified and sorted asap

Key messages

- Data management is part of good practice whether you plan to make the data open or not
  - **it benefits you!**
- The process of planning is the most important aspect. Think about the desired end result and plan for this.
- Approach DMPs in whatever way best fits your project. Don’t just let funder requirements drive things.

Thanks for listening

DCC resources on DMPs

www.dcc.ac.uk/resources/data-management-plans

Follow us on twitter:

@DMPonline and #DMPonline