

FAIR Data Science

Introduction

In this task you will become a (data) scientist. Your job requires integrating and analysing data with a use of software tools and visualisation frameworks. To get credit and progress quicker with your career you must not only publish world class reports, but also provide data and details on implementation of your analysis to ensure trust in your results. You have recently learned that the best way is to create a data management plan and to follow FAIR principles to make your research findable, accessible, interoperable, and reusable. Since you are also allergic to bureaucracy, you want to investigate ways on how to automate this process.

In this exercise you need to show off with your excellent data management skills and basic modelling skills. The exercise consists of three parts (not equally graded):

- use case - you pick an already existing experiment of yours or create one for this exercise;
- data management - you create a data management plan (DMP) and publish your experiment;
- machine-actionable DMP – you express your DMP using RDA DMP Common Standard.

You are supposed to work **individually**. For questions please use TUWEL forum. When there is no answer within few days (unlikely), please write to tmiksa@sba-research.org

Part 1 - use case

According to Jim Gray [1], computational research includes tasks ranging from "data capture and data curation to data analysis and data visualization". Hence, **your task is to choose an experiment** for which you will write a data management plan in the next part. The experiment can be any of the assignments you used in other courses. It can also be your bachelor or master thesis. For your own convenience do not come up with too complicated examples – building a large hadron collider at home may be too ambitious, but a python “hello world” script is not a scientific experiment. The experiment must fulfil following conditions:

- Data sourcing – experiment reuses data from external sources (not created by you), e.g. Kaggle data, open data, etc.
- Data transformation - data is filtered, processed, some decision is made, etc. Specific libraries or software is needed for computation.
- Data visualisation – the output of the experiment isn’t just a simple “it works”, but there are at least two outputs, e.g. raw data and visualisations (charts, histograms, etc.).

For the TUWEL submission (apart from the things requested in other parts), please provide a PDF with a summary of your experiment: 1-2 pages. Please include the following information:

- experiment overview (abstract),
 - what it is used for, what’s the input, what’s the output, etc.
- diagram explaining the experiment (e.g. data flow diagram),
- (screenshots/listings that may help in understanding it).

Part 2 – data management

The goal of this part is to organize your data and code in such a way that it is FAIR and reproducible. In this part you will: write a DMP, perform FAIR self-assessment, share your experiment in a FAIR compliant way.

Depending on the student number you must follow one of the options below.

Option 1-Odd student number

Go to <https://dmponline.dcc.ac.uk>, create an account, log in, and create a new DMP like depicted in Figure . Don't forget to tick the DCC guidance checkbox. Create the DMP for the EC Horizon 2020.

Create a new plan

Before you get started, we need some information about your research project to set you up with the best DMP template for your needs.

* What research project are you planning?

☒ mock project for testing, practice, or educational purposes

* Select the primary research organisation

- or - ☒ No research organisation associated with this plan or my research organisation is not listed

* Select the primary funding organisation

- or - ☐ No funder associated with this plan or my funder is not listed

Figure 1: DMP online – select Horizon 2020 template.

Follow the guidance and answer all questions. Your answers should be exhaustive but not too long. Be precise – answer with full sentences, write coherent text. It is important that you **provide evidence and/or explanation** for answering in a given way. You may need to go back to the lecture slides and also may need to look for additional information on metadata, repositories, etc.

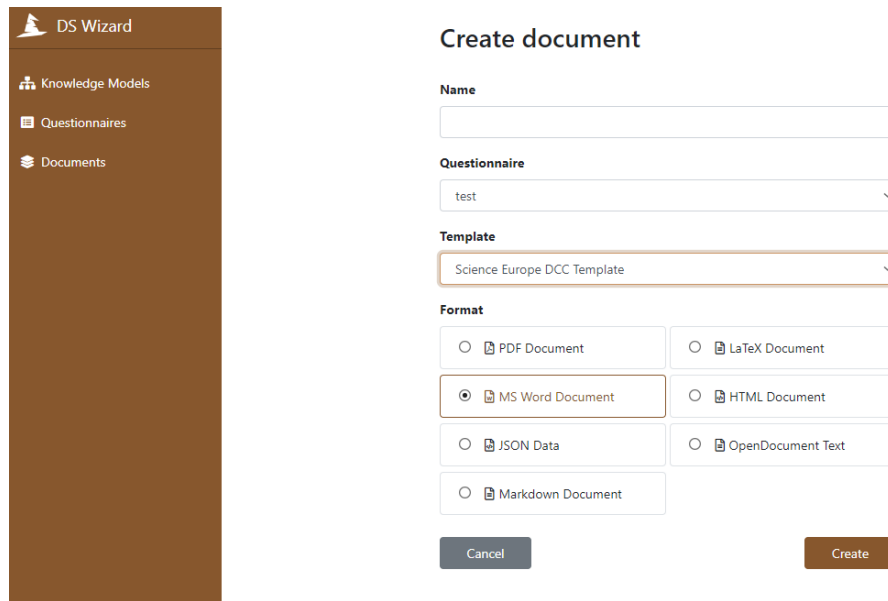
Use FAIR assessment tools to identify how to make your experiment as FAIR as possible. Follow their recommendations and update your DMP if needed. You can use any of the two tools:

- ARDC tool:
<https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/>
- DANS tool:
<https://docs.google.com/forms/d/e/1FAIpQLSf7t1Z9IOBoj5GgWqik8KnhtH3B819Ch6ID5KuAz7yn0I0Opw/viewform>

Filling out a DMP requires you to think and to check some external sources. Rushing quickly through questions is not a recommended way for solving this part.

Option 2 - Even student number

Go to <https://ds-wizard.org/get-started.html>. Create an account, log in. Create new Questionnaire. Select “Common DSW Knowledge Model”. Answer all questions. Check summary report to find out how FAIR you are. If it is needed, improve the way you manage your data to increase the score.



The image shows the 'DS Wizard' interface. On the left is a sidebar with a tree view containing 'Knowledge Models', 'Questionnaires', and 'Documents' (which is selected). The main area is titled 'Create document' and contains several form fields: a 'Name' text input, a 'Questionnaire' dropdown menu (currently showing 'test'), a 'Template' dropdown menu (currently showing 'Science Europe DCC Template'), and a 'Format' section with six radio button options: 'PDF Document', 'LaTeX Document', 'MS Word Document' (which is selected), 'HTML Document', 'JSON Data', and 'OpenDocument Text'. At the bottom of the form are two buttons: 'Cancel' and 'Create'.

Figure 1: DS Wizard – DMP export.

Based on the questions answered, the DS-Wizard can export a DMP. To do this, create a new Document using the questionnaire you have just filled out. Use “Science Europe DCC Template” and export your DMP to a format of your choice. Please make any necessary edits to the document if needed and save it as PDF. It is important that you provide evidence and/or explanation for answering in a given way. Filling out a DMP requires you to think and to check some external sources. Rushing quickly through questions is not a recommended way for solving this part.

Common for both options

When creating the DMP you were asked where the data produced by the experiment would be shared. In case you still haven’t made the experiment publically available you have to do it now. To enable reproducibility and to give others a chance to run the experiment you need to provide not only data produced by the experiment, but also source code, parameters, instructions how to run it, etc. In case your DMP needs changes then do them.

Good data management plans and FAIR experiments:

- have a good README and follow conventions! (check one of many tutorials)
- use ORCID to identify researchers (get yourself one)
- follow file naming convention and clear folder structure (have you mentioned that in the DMP? Refer to a specific one!)
- assign DOI for data produced in the experiment
- assign DOI for source code (check GitHub and Zenodo integration)
- use licenses for both code and data that allow reuse
- refer to and describe input data (did the license allow you to use it? What if there was no license?)
- enable verification of experiment (e.g. intermediate data used in processing)
- provide metadata (remember about Magpies in Australia?)

The most important – your results must be replicable. When grading, we will rely on information from the submitted DMPs. We will try replicating results of the experiments, because this is the best way to evaluate how good your DMPs are. You can get **bonus points** for providing Docker files (not images) together with your code.

Export your DMP into a PDF. Upload it to the *Data Stewardship 2020 – DMPs* community on Zenodo using this link: <https://zenodo.org/deposit/new?c=tuw-dmps-ds-2020> Please, make it open or embargoed (embargo end = deadline for this exercise). As a result, you and the other students will have a chance to compare and discuss your DMPs.

For the TUWEL submission (apart from the things requested in other parts), please provide:

- Text file with a DOI pointing to your DMP

Part 3 –maDMPs

In this part you will make your DMPs machine-actionable. The larger goal is to automate DMP creation and facilitate exchange of information between systems, thus reduce workload associated with creation of traditional DMPs. To make this happen, research funders (who enforce usage of DMPs) must accept machine-actionable DMPs. This can be achieved when they see that maDMPs contain the same information as traditional DMPs. In this part, you will convert your DMP into maDMP to show that this is possible.

Express your DMP using the *RDA DMP Common Standard for Machine-actionable Data Management Plans* [2]. The result of this ‘translation’ should be a JSON document that validates against this schema¹. Full documentation, simplified examples, FAQs, etc. can be found in the GitHub² repository of the standard.

Note that not all of the information in the DMP you created in Part 2 is machine-actionable. It is easy to map a license to a specific field in a standard. However, for open questions like “describe your quality assurance processes” there is often no specific field corresponding 1:1. This is because the standard is not bound to a specific questionnaire and there are many questionnaires around the world.... However, standard foresees ways to include such information – usually there is a field of String type can be used, or a more generic field “description”.

Upload your JSON document to the *Data Stewardship 2020 – DMPs* community on Zenodo using this link: <https://zenodo.org/deposit/new?c=tuw-dmps-ds-2020> Please, make it open or embargoed (embargo end = deadline for this exercise). As a result, you and the other students will have a chance to compare and discuss your maDMPs.

For the TUWEL submission (apart from the things requested in other parts), please provide:

- PDF report describing how you translated each part of the DMP into maDMP
 - Describe how you dealt with each question
 - Describe design decisions you had taken and why
 - Describe any particular challenges
- Text file with a DOI pointing to your maDMP

¹ <https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard/tree/master/examples/JSON/JSON-schema/1.0>

² <https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard>

Deadlines

The start for the exercise is 09.03.2020. The **deadline** is **20.04.2020 at 23:59** local Vienna time.

References

[1] Tony Hey, Stewart Tansley, and Kristin Tolle, editors. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, 2009.

[2] Miksa, T., Walk, P., & Neish, P. (2019). RDA DMP Common Standard for Machine-actionable Data Management Plans. <https://doi.org/10.15497/rda00039>