tmiksa@sba-research.org

FAIR Data Science – Part 1

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. You have recently learned that the best way is to create a data management plan and follow FAIR principles to make your research findable, accessible, interoperable, and reusable.

In this exercise you need to show off with your excellent data management skills and basic software engineering skills. The exercise consists of two consecutive parts:

- part 1 (described in this document)
    - data science use case - you design your own experiment that uses open data;
    - data management - you create a data management plan and publish your experiment;
- part 2 – to be announced.

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

Part 1.1 – data science 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 implement an experiment** that must consist of the following parts:

- Data sourcing – data is downloaded from **two different open data repositories**. You can find plenty of data about Austria on data.gv.at. You can also use any other repository that provides open data (earth observation[1], climate change[2], bioinformatics, astronomy, financial, weather, etc.).
- Data transformation - data is filtered, processed, some decision is made, etc. Specific libraries or software is needed for computation. This can be some simple statistical computation, but can also include more advanced transformations or machine learning. It's up to you. You can use R, Python, Java, or any other scripting or programming language. You can call some external services for processing. You can also use Jupyter notebooks and any other cool stuff you want to play with.
- Data visualisation – the output of the experiment isn't just a simple "it works", but provides:
  - raw data,
  - visualisations (charts, histograms, etc.). Take a look at plot.ly or Highcharts, but feel free to use also other libraries.

Figure 1: Part 1 – overview of the experiment.

Don't be afraid to be creative here, but for you 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 main aim of this part is to provide you with a good use case for the second part. Feel free to customise any existing projects or real experiments. Check this link for some inspiration: http://www.tylervigen.com/spurious-correlations. Be creative, the "experiments" do not have to make scientific sense – you can make us laugh!

[1] https://apps.sentinel-hub.com/sentinel-playground
[2] https://www.ccca.ac.at/datenzentrum/

tmiksa@sba-research.org

For the TUWEL submission (apart from the things requested in other parts), please provide summary of the experiment – one, max two, A4 pages. Please include the following information:

- experiment overview
    - what it is used for, what's the input, what's the output, etc.
- diagram explaining the experiment (e.g. UML diagrams)

For this subpart (1.1) you will receive 30% of points for Part 1 of the first exercise.

Part 1.2 – data management

You are now ready to create a DMP. Go to https://dmponline.dcc.ac.uk, create an account, log in, and create a new DMP like depicted in Figure 2. Don't forget to tick the DCC guidance checkbox.

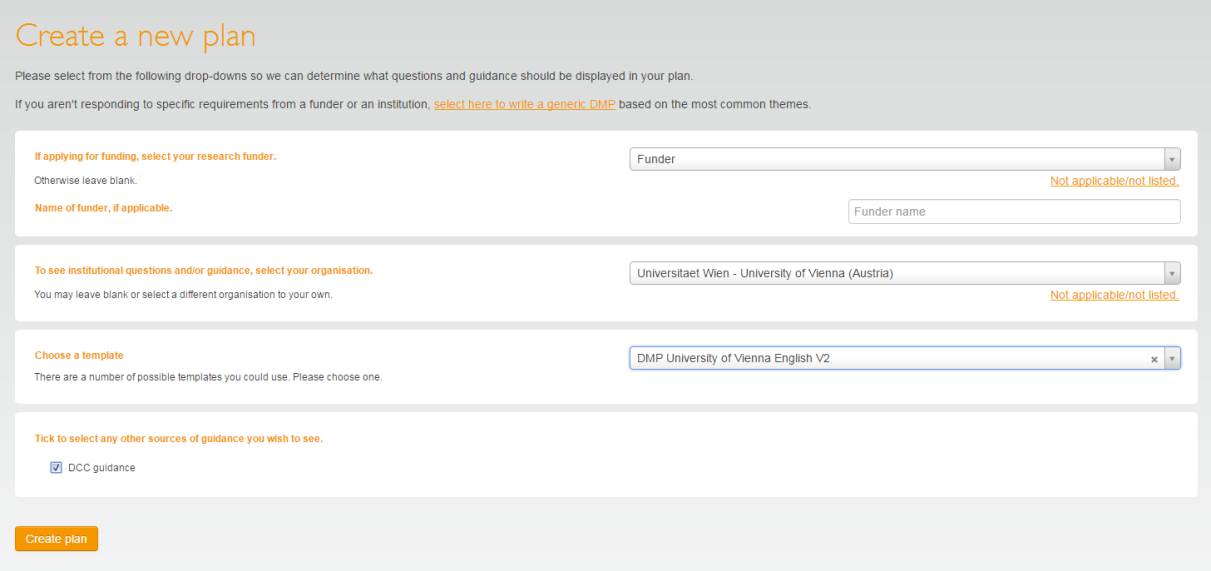

Figure 2: DMP online – select University of Vienna template with DCC guidance.

Following the guidance 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 can provide evidence for answering in a following way and provide sufficient explanation. You may need to go back to the lecture slides and also may need to look for additional information on metadata, repositories, etc. Once you have created the DMP, please **export it into PDF and include in your submission**. (If you don't like the DMP Online tool, you can use any text editor and create your own document that covers the same information as requested by the tool)

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 and instructions how to run it. Using a correct folder structure and a proper naming convention is obvious to an expert like you. In case your DMP needs changes then do them.

Good data management plans and FAIR experiments:

- use ORCID to identify researchers (get yourself one)
- follow file naming convention and clear folder structure
- deposit experiment results in a data repository
- assign DOI for data produced in the experiment
- assign DOI for source code (check GitHub and Zenodo integration)
- use licenses which allow reuse
- refer to input data
- enable verification of experiment (intermediate data used in processing must be available)
- 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.

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

- PDF with a Data Management Plan

Part 2 (not covered in this document, will be announced later)

Deadlines and Teams

The start for the exercise is 08.03.2017. The **deadline** for Part 1 (described in this document) is **29.03.2018 at 23:59** local Vienna time. You are supposed to work in groups of two students.

References

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