DCC Spring Training:
The role of good RDM in accelerating scientific progress

Presenters: Joanne Yeomans & Kristina Hettne, Leiden University Libraries
Facilitator: Margriet Miedema, LCRDM
Guest: Barbara Magagna, independent semantic expert, knowledge facilitator and expert on FAIR data management

THIS GOOGLE DOCUMENT WILL BE CLOSED FOR COMMENTS AT THE END OF JUNE AND THE PADLET WILL BE SAVED AND CLOSED.

9th June 2022, 9:30 - 12:30

Contents of this document:

* (when you have time)

Agenda:

9:30	👋 **Welcome and introductions (Margriet)**
	Part 1: Reproducibility
9:35	Introduction to the session (Joanne) / Explanation breakout room exercises (Joanne) Set up breakout rooms (10 rooms of 3-4 people) (Margriet)
9:45	Activity 1 in breakout room (All)

	• Breakout rooms will contain 3 or 4 people. • Everyone opens the padlet and everyone can type: https://edu.nl/838xr • One person can share their screen if it helps. 💬 Round of introductions within breakout room: (1 minute each). 📝 Brainstorm "How to make research results reusable?" (15 minutes) Answer the following question in the same padlet: 1. How can researchers make their results, data and code reusable?
10:00	💬 Summary (5 min) (Joanne)
10:05	Introduction to Reprohack and how it works (Kristina). • Presentation
10:35	☕ **Break 15 minutes**
10:50	**Part 2: FAIR (meta)data in practice**
10:50	FAIR (meta)data basics (Kristina)
11:20	Developing the I-ADOPT standard (Barbara Magagna) • Presentation • Questions
11:55	☕ **Break 5 minutes**
12:00	Activity 2 in breakout room (All) • set up breakout rooms (10 rooms of 3-4 people) (Margriet) • Find your group in the GoogleDoc • One person can share their screen if you want. 📝 Brainstorm: Improving the metadata of a dataset (All)

12:10	💬 Summary (10 min) (Kristina)
12:20	💬 **Wrap up:** Feedback (one up, one down) & Additional Resources (Kristina & Joanne) ● What's next? Look out for training opportunities (DCC, GO FAIR, funders). In Leiden we are developing training as well. ● As a basic: help researchers with their metadata
12:30	🙏The end

Code of Conduct

- Be inclusive and constructive
- Be patient with the technology/each other/yourself
- Let everyone have a chance to speak
- Be respectful of different viewpoints and experiences
- Gracefully accept constructive criticism
- Show courtesy and respect towards other community members
- Give, as well as take

What do you wish to learn from this session?

-

Improving the metadata of a dataset for reusability

These links relate to a research project about the chemical analysis of the ultramarine pigment in historical paints.

Step 1: Go to the publication:

https://www.science.org/doi/10.1126/sciadv.aay8782

Read the abstract of the paper to get an idea about the topic.

Step 2: Take a look at the README file for the code to get an idea of the contents and to read a summary of the datasets:

https://github.com/alessaan/rhapsody-in-blue

Step 3: Go to the data: https://doi.esrf.fr/10.15151/ESRF-DC-186933507 You will need to 'log in as Anonymous' in order to reach the dataset. You should see the following page:

Step 5: Click on the big blue "Access data" button. You should see the dataset list as below.

Explore the data by clicking on the magnifying glass on the left hand side for the first dataset (dataset_43).

There are three tabs: Summary, Files and Metadata List. See screenshot below.

Write down your ideas for improving the metadata about dataset 43. What do you miss in order to understand what the data is about?

For example:
1. Is it clear what you are looking at and how to navigate the files?
2. Do you recognize these file types and know how to open them?
3. Can you say something about the use of standard vocabularies?
4. Does the ESRF archive system enable you to give clearer metadata? Do you have thoughts about how the archive itself could be better designed?

Step 4: Find a csv file and try downloading it.

1. Are the column headers clear?
2. Does the filename give you any information that could be useful?
3. Does the code documentation in github help you to understand what you're looking at? For example, dataset 43 is used to create figures in the paper, see this codebook documentation: https://github.com/alessaan/rhapsody-in-blue/blob/master/17_FigurePaint_Treatment_2018.ipynb. Is the information in the codebook enough for you to understand the csv files?

Please make some notes of your thoughts below:

Group 1
   1.

Group 2
   1.

Group 3
   1.

Group 4
   1.

Group 5
   1.

Group 6
   1.

Group 7
   1.

Group 8
   1.

Group 9
   1.

Group 10
   1.

One up, one down feedback

Please tell us what you liked and what can be improved with this session:

One up

- Good presentations, interesting topics
- Reprohack was new to me
- Interesting questions from the audience
- Great to share experiences with other people with different backgrounds

- Nice to have a quite practical perspective of research supporters, this is very valuable and applicable to my own work.
- Great topics and presentations. Also enough room for discussion
- Very good and clear presentations (I'm definitely going to use the Reproducible/Robust/ etc. diagram)
- All of the items mentioned above……………. → I second this.
- Like the initiative

One down

- I really enjoyed the small group discussions and the different views/outcomes, maybe induce a couple more during the training.
- The practical implication on how to make data FAIR (especially the I) is still difficult to grasp, maybe for future use examples of variables that a lot of researchers collect (like age, gender, bmi) → agreed (I second this)

- Give some more info about what a datamanager can do with the reprohack or I-ADOPT, in practice. Both were new for me. How can I use it to help the researchers
- My breakout group only had two members so our discussion was a bit limited (but very "gezellig"!)
- Giving only one up is impossible!

- Would like to have more hands-on tips

Please feel free to share with everyone additional tips or resources that you are aware of, or comments.

Pre-registration information (shared in 2021)

- VU Data Conversations: "Personal Journeys into Preregistration" https://zenodo.org/record/4455208#.YMs0_GgzaUk
- Pre Registration https://osf.io/
- jupyter notebooks are heavily used in astronomy
- The Turing Way: A Handbook for Reproducible Data Science https://zenodo.org/record/3233986#.YMs3Z2gzaUk

A selection of data archiving solutions (shared in the 2021 training):
- YODA: https://www.uu.nl/en/research/yoda
- CLARIN: https://www.clarin.eu/faq/what-clarin
- Genbank: https://www.ncbi.nlm.nih.gov/genbank/

Resources useful for training (shared in the 2021 training):
- Data sharing and management snafu https://www.youtube.com/watch?v=N2zK3sAtr-4
- Slides and a guide for running a reproducibility workshop: https://www.repro4everyone.org/resources
- Reproducibility self-assessment tool: https://www.protocols.io/reproducibility

Please add your own suggestions above or below..

Other interesting resources:

-

NOTE: