Title: Laboratory Forensics: An Assessment Approach to Data Management in Research Labs

Instructor: Armel Lefebvre, PhD candidate in Research Data Management, Utrecht University, Netherlands; https://www.linkedin.com/in/armell/

Description:

Many research data management initiatives from funders, publishers, and universities focus on the open dissemination of research data to advance reproducibility and reusability of scientific results. At the same time, researchers find data dissemination unrewarding, time-consuming, and unnecessary if it is not mandatory to get articles published. What makes the dissemination of research data a challenge for researchers may also, partially, be explained by the basic storage technology frequently used in research laboratories: hierarchical file systems with digital files and folders.

This course will explore how locally preserved experimental evidence can be reconstructed from the corresponding publication(s). We will also explore how incomplete or fuzzy evidence left by experimentation processes in laboratories can potentially be corrected by engaging with researchers about the issues identified during the investigation of their storage systems. Therefore, research data management and data availability can benefit from “laboratory forensics” to improve critical aspects of preserved experimental evidence.

At the end of the course, participants will have basic knowledge of how to use forensics techniques to explore digital file systems.

Level: All levels

Intended audience: This course is suitable for beginners, and everyone is welcome. For the first part, participants will receive material such as lists of file paths and Python scripts (in Jupyter Notebooks). The course requires virtually no knowledge of Python, although students with advanced Python skills will have the opportunity to explore extra material and assignments to keep them interested and busy. The second part focuses more on organizational issues of research data management, which everyone can discuss based on their own experience in data management, stewardship, curation, and publication.

Requirements: A preselected set of publications will be used, so no preparatory work is required. However, participants who have access to real data from their research groups will be invited to use that data during the course if possible. Students need to bring a laptop with Python installed. WinPython (for windows users) and Anaconda both have a Jupyter Notebook server installed.

Course Learning Objectives

By the end of this course, participants will be able to:

- Collect and analyze meta-data from storage systems with Python.

- Match digital evidence to publications.
- Understand how to assess research data management in laboratories.
- Get experience with designing metrics for data management assessment.

Course Topics

This course will be presented over two days for 3 hours each afternoon and will cover these topics:

- Digital Forensics: The core of the techniques introduced on Day 1 are (a lightweight) version of digital forensic technique. The context differs from criminal investigations, though, as the main objective here is to reconstruct experimental events based on experimental reports, i.e., publications
- Metrics for research data management (RDM): how to measure and improve storage practices in laboratories

Course Schedule

Day 1

Fundamentals of digital forensics, scientific experimentation and, meta-data retrieval to extract knowledge from files and their corresponding scientific articles:

- Mini lecture: Digital forensics methods (30 minutes)
- Practice: Set-up and basic digital forensic techniques (30 minutes)
- Break: 10 minutes
- Mini lecture: The ingredients of scientific experiments (10 minutes)
- Practice: Extracting material from scientific articles (60 minutes)
- Break: 10 minutes
- Focus group: Applicability of digital forensics (30 minutes)

Day 2

Development of a core set of indicators to assess the state of research data preservation in laboratories:

- Mini lecture: General introduction to data management assessment (RDM) in research and industry (50 minutes)
- Break: 10 minutes
- Practice: Application of digital forensics and RDM assessment (60 min)
- Break 10 min
- Focus group: Reporting digital forensics findings with RDM indicators (50 min)

Course Materials and Supplies Required

This part will be updated on July 1. We will make sure to keep the required material to a minimum.

Students are expected to have downloaded this material before the course begins:

- The path2insight Python package available with Python pip. See here for more information: https://path2insight.readthedocs.io/en/latest/
- Practice notebook available in the course folder
- Good news! We have set up an online environment where the necessary material can be accessed from the browser. Nevertheless, we recommend you have python installed on your laptop as a backup option. The notebooks will be made available at the beginning of the course.

Students are expected to have read this material before the course begins:

Reproducibility metrics

- Recommendations from the ACM to evaluate software artifacts underlying scientific articles : https://www.acm.org/publications/policies/artifact-review-badging

Reproducibility concepts

- Chen, X., Dallmeier-Tiessen, S., Dasler, R., Feger, S., Fokianos, P., Gonzalez, J. B., … Neubert, S. (2019). Open is not enough. *Nature Physics*. https://doi.org/10.1038/s41567-018-0342-2: https://rdcu.be/bHz5q

Other Resources

- Defining reproducibility: Schloss, P. D. (2018). "Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research." MBio, 9(3), e00525-18.
- https://altmetrics.org/altmetrics12/iorns/
- https://metrics.stanford.edu/research/reproducibility
- State of the art of open science metrics. Methodology report for evaluating Open Science (April 2019): https://ec.europa.eu/info/files/open-science-monitor-methodological-note_en
- Defining reproducibility: Goodman, S. N., D. Fanelli and J. P. A. Ioannidis. (2016). "What does research reproducibility mean?" *Science Translational Medicine*, *8*(341), 341ps12-341ps12.
- Mabey, M., A. Doupé, Z. Zhao and G. J. Ahn. (2018). "Challenges, opportunities and a framework for web environment forensics." In: *IFIP Advances in Information and Communication Technology* (Vol. 532, pp. 11–33). Springer, Cham.

Other Helpful Information

A part of the course will be dedicated to the evaluation of lab forensic methods by practitioners (= the participants). During the course, you will have the opportunity to discuss the forensic

method and metrics for reproducible data management with the instructor. The format of the evaluation sessions will be a focus group. The focus-group format keeps the feedback sessions lively and interesting for all participants. The course is part of a broader design science project where we aim at evaluating lab forensics artifacts with (potential) practitioners and stakeholders in academia.