Appendix I: Interview Protocol

Deep Dive into the Research Workflow and Data Profile in Material Science and Engineering

Version 0.1.1 (6/4/2020)

Interviewee:

Interviewer:

Date of interview:

Survey sections used:

- ❏ Interviewee Background
- ❏ Top Priority Project Background
- ❏ Top Priority Project Workflow
- ❏ Reflection

Survey version used:

Other Topics Discussed:

Documents Obtained:

Post Interview Comments or Leads:

Introductory Protocol

Thank you for participating in our project on the research workflows and data profiles in Professor Rafael Jaramillo's group. Our goal is to identify specific gaps and challenges in the current data practices as well as recommend guidelines and strategies to improve the practices from data collection to dissemination. The resulting guidelines can be adopted by "small labs" in Material Science and Engineering for cultivating FAIR (Findable, Accessible, Interoperable, Reusable) data practices as well as open and reproducible research.

At this time, we are trying to learn more about your daily work in the lab for a specific research project. We have planned this interview to last no longer than one hour. During this time, we have several questions that we would like to cover.

To facilitate our note-taking, we would like to record our conversations on Zoom today. For your information, only researchers on the project will be privy to the recordings. In addition, you must sign a form devised to meet our human subject requirements. The link to the form is at https://mit.co1.qualtrics.com/jfe/form/SV_3UUojXHIeJOg7IP . Essentially, this document states that: (1) all information will be held confidential, (2) your participation is voluntary and you may stop at any time if you feel uncomfortable, and (3) we do not intend to inflict any harm. Thank you for agreeing to participate.

1. Interviewee Background

a. How long have you been working in the group?

b. What is your role in the group?

2. Top Priority Project Background

a. Which research project(s) are you currently working on?

b. Which one is your top priority (if you have multiple projects)?

For your top priority project,

c. What is the source of research funding for this project?

d. Briefly describe the goals, objectives, and scope of this project.

e. Who, in and outside the research group, are collaborating with you on the project? What is each person's role in the project?

f. How long have you worked on the project and when do you expect the project to conclude?

g. What are the most important recent or forthcoming research publications from this project?

3. Top Priority Project Workflow

a. In material science research, a project often involves sample preparation, measurement of material composition and properties, and also simulations or other computational studies. For your top priority project, which of the above elements are involved? Does it involve other elements not mentioned here? Please specify.

b. Please help us walk through the details of your research process step by step. Please distinguish between what's been done, what is current, and what is in planning.

> [*For each step, please specify the instruments, what and how the data were measured and where they were stored, how and where the process were recorded, how and where the data were processed and analyzed, and what data will be published and when.*]

Checklist for details:
Make sure the following details are included in the description.

- ❏ Material sample preparation:
  - ❏ What instruments have been used and where?
  - ❏ How does the process (such as research plan, instrument calibration, instrument operation) get recorded? How is it this information made available within your group?
  - ❏ (If any), where and how does the sample preparation data get stored and backed up? How is this information made available within your group? How is it identified and versioned?
- ❏ Composition and property measurement:
  - ❏ What measurements are needed for this sample?
  - ❏ Why was this measurement type chosen?
    For each measurement:
  - ❏ What instruments have been used and where?
  - ❏ How did the measuring process get recorded?
  - ❏ How and where the data generated from the measurement are stored and backed up?
  - ❏ How are these measurements made available within your group?
  - ❏ What is the raw format of the data measured?
  - ❏ How and where the measured data get processed and analyzed? What software and tools are being used?
  - ❏ What are the intermediate data types and formats?

        - ❏ What are the final forms of data to present in publications based on this measurement? What post-processing or analysis do you have to apply?
        - ❏ When and how would you decide if this dataset will be presented in a publication?
        - ❏ When this study gets published, will the datasets be shared? In what format and where?
    - ❏ Simulations or computational studies: (if any)
        - ❏ What is the goal of the computational study?
        - ❏ Is the computational study a theory-based simulation or a machine learning study? Or both?
        - ❏ Do you need an initial dataset to start the computational study? If so, how and where do you get the initial datasets?
        - ❏ What tools and software do you use to perform the computational study?
        - ❏ How are these tools made available within your group?
        - ❏ How and where do you perform the computational study?
        - ❏ How did the computational process get recorded?
        - ❏ How is the process and its outputs made available within your group? How are the outputs identified and versioned?
        - ❏ What are the intermediate data types and formats?
        - ❏ What are the final forms of data to present in publications based on this computational study? What post-processing or analysis do you have to apply?
        - ❏ When and how would you decide if this dataset will be presented in a publication?
        - ❏ When this study gets published, will the datasets and the computation codes be shared? In what format and where?

4. Reflection

Considering your research workflows for your top priority project,

Work efficiency

a. **Which phases or activities are most time-consuming or effortful? ***

b. Which activities would you consider as busy work that could be automated or more streamline?

Gaps and disconnections

a. **Do you see any disconnections of your workflow digitally?** For example, do you often need to manually synchronize the same set of data at multiple locations? *

b. Are there any tools, platforms, or infrastructure particularly helpful or difficult to use for your digital workflow of this project? For example, do you find LabArchives, your data storages and backups helpful?

c. Do you find any practices that often lead to confusions? For example, do you have mixed naming conventions, file structures or documentations that's hard for yourself or others to decipher?

Collaboration and sharing

a. Among collaborators for this project, how do you pass along or share data and experiment records to each other?

b. If you were to leave the group tomorrow, who would pick up this work? What information would they need? How would they find it? *

c. If you are aware of, what are the requirements, obligations, or standards for sharing data and codes from the institution, publishers, or funders of this project? *

d. For the final publications from this project, how and when would you determine the authorship and credits among the collaborators?