IJDC | *Conference Paper*

Reproducible and Attributable Materials Science Curation Practices: A Case Study

Ye Li
Massachusetts Institute of Technology

Sara Laura Wilson
Massachusetts Institute of Technology

Micah Altman
Massachusetts Institute of Technology

Abstract

While small labs produce much of the fundamental experimental research in Material Science and Engineering (MSE), little is known about their data management and sharing practices and the extent to which they promote trust in and transparency of the published research. In this research, a case study is conducted on a leading MSE research lab to characterize the limits of current data management and sharing practices concerning reproducibility and attribution. The workflows are systematically reconstructed, underpinning four research projects by combining interviews, document review, and digital forensics. Then, information graph analysis and computer-assisted retrospective auditing are applied to identify where critical research information is unavailable or at risk.

Data management and sharing practices in this leading lab protect against computer and disk failure; however, they are insufficient to ensure reproducibility or correct attribution of work, especially when a group member withdraws before the project completion. Therefore, recommendations for adjustments in MSE data management and sharing practices are proposed to promote trustworthiness and transparency by adding lightweight automated file-level auditing and automated data transfer processes.

Submitted 9 February 2024 ~ *Accepted* 22 February 2024

Correspondence should be addressed to Micah Altman and Ye Li, MIT Libraries, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: escience@mit.edu and yelibrarian@gmail.com

This paper was presented at the International Digital Curation Conference IDCC24, 19-21 February 2024

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Background

Increasing Attention to Reproducibility, Openness, and Attribution in Science

Reproducibility is a foundation of science. However, over the last two and half decades, mounting evidence has called into question the reproducibility of findings in a continually expanding set of fields, leading to regular calls to assess reproducibility and improve scientific practice systematically (National Academies of Science Engineering and Medicine, 2019).

More recently, there have been high-profile calls and initiatives by research societies, funders, and publishers to make scientific practice and data more open and transparent (National Academies of Sciences, Engineering, and Medicine, 2018) and to develop systematic attribution standards (McNutt et al., 2018) and practices for contributors to scientific publications and outputs.Science stakeholders increasingly realize that a scientific discipline's reproducibility needs to be empirically evaluated, not simply assumed. A hallmark study by the National Academies of Science Engineering and Medicine (NASEM) (2019)reviewing the state of knowledge on scientific transparency finds that the evidence base of non-replicability across all science and engineering research is incomplete.

Scientific replication practices are neither sufficiently consistent nor sufficient to make confident statements about the replicability rate in most fields. However, the major empirical studies on replication failure conducted in the natural, clinical, and social sciences have yielded failure rates from somewhat lower than 20% to higher than 80%. Further, this study found an uneven awareness of issues related to replicability practices and awareness across fields and within science and engineering.

Similarly, although many fields have widespread norms or even stated policies on research transparency (e.g., making data available after publication) and appropriate attribution of contributors, these policies are unreliable predictors of practice (for example, Savage & Vickers, 2009). Empirical evaluation is needed to understand how and where these practices are followed and what effects they yield.

Studies of Practices in Experimental Materials Science and Engineering

Schechtman's Nobel-winning discovery of quasi-crystals stands as a particular occurrence (and eventual resolution) of the classic "file drawer" problem (Timmer, 2011) that is highlighted by open science advocates; however, this is one illustration with a happy ending and cannot establish a pattern. Few published studies describe or evaluate practices related to replication, transparency, and attribution in Materials Science and Engineering (MSE).

A more recent study suggests a rosier picture; an analysis of retractions in MSE publications finds a relatively low rate (0.03%) (Coudert, 2019). However, while a high retraction rate signals problems, most non-replicable research is generally not retracted; therefore, a low retraction rate does not strongly suggest replicability. Another recent study examining data-sharing practices in small MSE labs (Wilson et al., 2019) revealed that while many researchers in materials science embrace the idea of open science, reproducible research, and data-sharing, they are frustrated with the inadequate infrastructure, tools, and practice guidelines. This finding suggests the potential for gaps between aspiration (for reproducibility and openness) and practice. Perhaps most concerning; however, is a recent set of case studies (Han et al., 2019) published in the Annual Review of Chemical and Molecular Engineering that found a high (20%) rate of reproducibility failure in the two research areas, the properties of metal–organic frameworks (MOFs) and the synthesis of crystalline nanoporous materials, were targeted for study. A 2017 study on isotherm measurements in MOFs also revealed a similar level of irreproducible rate (Park et al., 2017).

Experimental materials science typically does not generate large quantities of data in coordinated or collective studies compared with geology, genomics, and some disciplines within economics. In

MSE, experimentalists generate materials property data in their "small labs" individually and have not developed a shared practice of data-sharing as in many other "big data" disciplines. Moreover, gaps in experimental data availability have been identified as a barrier to computational materials science since the early 1980s (Westbrook & Rumble, 1983) and remain a significant obstacle to progress.

Rapid progress in data science and the ever-increasing number of demonstrated applications of data science approaches in data-rich fields produce optimism that data science could be productively applied to materials science (Tinkle, et al., 2013). Significant progress in this direction requires significant data resources. Pioneering studies highlight the difficulty in assembling large quantities of experimental materials science data that can be the basis for valuable and insightful inferences (Raccuglia et al., 2016).

The renewed promise of machine learning and its applications in materials science has made the need for Findable Accessible Interoperable & Reusable (FAIR) experimental data more urgent (Blaiszik et al., 2016) -- especially with the acceleration of machine learning methods in the second decade of the new millenium. Further, applying machine learning and artificial intelligence (AI) to materials science at scale has been identified as a significant challenge for the discipline. It depends on robust tools and practices for data-sharing and replicable workflows (Stein & Gregoire, 2019). In 2022, the National Science Foundation (NSF) Division of Materials Research underscored the importance of transparent access to data by issuing specific policy guidance for the field (National Science Foundation, 2022).

Data resources can grow through open science practices, such as sharing data generated across the research life cycle. However, experimental materials science lacks the norms, standards, and tools to make this widespread, especially in academic labs. There have been notable efforts to develop infrastructure, standards, and tools to enable experimental reproducible workflow management and data-sharing in materials science (Hill et al., 2018; Himanen et al., 2019). For example, the 4Ceed Project (Nguyen et al., 2017) developed a cloud framework and associated curation services for the real-time capture of materials data from instruments based on a survey they carried out among experimentalists (4Ceed Design Team, 2016). The Materials Data Facility (MDF) service launched in 2016 (Blaiszik et al., 2016) was designed to provide an interconnection point for data-sharing, discovery, access, and analysis. The MDF, sponsored by the National Institute of Standards (NIST) and the Center for Hierarchical Materials Design (CHIMaD), now hosts about 578 data sets (116 experimental data sets) and indexes over 970,000 records on materials data from other repositories as of December 2021. Other recent efforts include infrastructure for a federated registry of information resources for materials science (Plante et al., 2021), a proposed controlled vocabulary and metadata schema for materials discovery (Medina-Smith et al., 2021), and a new experimental infrastructure under development for the integration of Electronic Lab Notebooks (ELNs) and data archiving systems with materials science workflows (Brandt et al., 2021). In industry, software platforms (e.g., the Citrine Platform (Informatics, 2022)) that combine the data management infrastructure and AI-based tools facilitating materials design provide customizable solutions for corporate labs, which have more consistent pipeline workflows and can afford the resource-intensive infrastructure. The FAIR Data Infrastructure for Physics, Chemistry, Materials Science, and Astronomy e.V. (FAIR-DI), a European-originated effort, aims to build a reliable infrastructure for data from materials science, engineering, and astronomy that follows FAIR principles (FAIR-DI, 2022). The FAIR-DI launched the NOMAD repository1 in 2014 and has been developing data management and sharing support. Their recent FAIRmat hands-on tutorial series (FAIR-DI, 2022) is designed to provide connections between the existing infrastructure and researchers' daily practices.

More recently, as a paradigm shift rooted in the exponential growth of computing power, integrated systems of AI-based predictions and experimental automation via robotics are explored and examined to accelerate materials discovery with the promise of replacing the manual and human-intensive material discovery process (Pyzer-Knapp et al., 2022). For example, a technology roadmap was outlined to articulate the hardware and software infrastructure requirements and demonstrate a re-imagined role of humans when ensuring data is appropriately managed, aggregated, standardized, and shared (Delgado-Licona & Abolhasani, 2023). The analysis of the potential to apply accelerated

[1] NOMAD repository https://nomad-lab.eu/

materials discovery in clean energy highlighted insufficient experimental data sets for AI model training as a limitation for clean energy as a relatively new technology (Maleki et al., 2022).

Despite these particular efforts and the overall progress in developing tools, standards, and practices, the adoption of these infrastructure and tools by individual "small" labs remains limited. No direct solutions have been provided for individual labs to streamline their workflows and efficiently prepare their data for sharing throughout the research life cycle. Instead, these labs use informal sociotechnical workflows that combine documented procedures, undocumented conventions, semi-automated tools, and manual processes. This research determines the informal workflows operating within a top material science lab. They are documented and described using a formal workflow graph notation and analyzed using qualitative and mathematical graph analysis.

Research Questions

This research aims to identify potential gaps and challenges for small lab MSE research replicability (trustworthiness), data availability (transparency), and attribution with an in-depth analysis of the practices supporting workflow and data management at a leading lab.

To identify the gaps and opportunities in the current research practices for these improvements, this study is designed to answer the following research questions probing the trustworthiness and transparency of MSE data curation:

1. To what extent does research depend on manual processes for information management?
2. Explicit processes:

   a) what processes concerning data and research workflow management are documented?

   b) to what extent are documented processes consistent with practice?
3. To what extent are documentation processes complete enough to support another person's replication of a result within the lab (without further communication with the original researcher)?
4. To what extent are data management processes robust enough to survive the departure of a project member or the loss of an individual's personal computer or storage?
5. To what extent are workflow data, outputs, and documentation sufficient to describe responsibility (or support attribution) for published results?

Data and Methods

Overview

This research focuses on practices within the research group for several reasons. First, internal data management is a prerequisite for external data-sharing and transparency. If research information created by one researcher becomes unavailable, uninterpretable, or irreproducible for a close team member, there is little hope it can be made meaningfully available for external reuse and review. Second, MSE relies largely on internal processes to guarantee replicability; there are no formal processes for external validation, systematic studies of replicability conducted across the field, or systematic reporting guidelines for reporting failures. Further, null results and those deemed uninteresting may end up in the file drawer and, therefore, unavailable for any external examination. Moreover, even published results of sufficient commercial value for an enterprise to attempt them in

production may fail and be discarded without any subsequent reporting. Third, MSE relies almost entirely on internal processes to ensure the appropriate attribution of work.

This approach employs a purposive case study design. A leading MSE research lab is selected, and selected lab members are interviewed in detail about their most important research project. As illustrated in Figure 1, the collected interview data were coded to create a standardized description of each task conducted for the project step, which was used to construct a formal workflow process graph. By tracing across and within these graphs, the overall collaboration and information flow patterns are characterized, and the extent to which information shared with the group is sufficient for replication and attribution is evaluated.

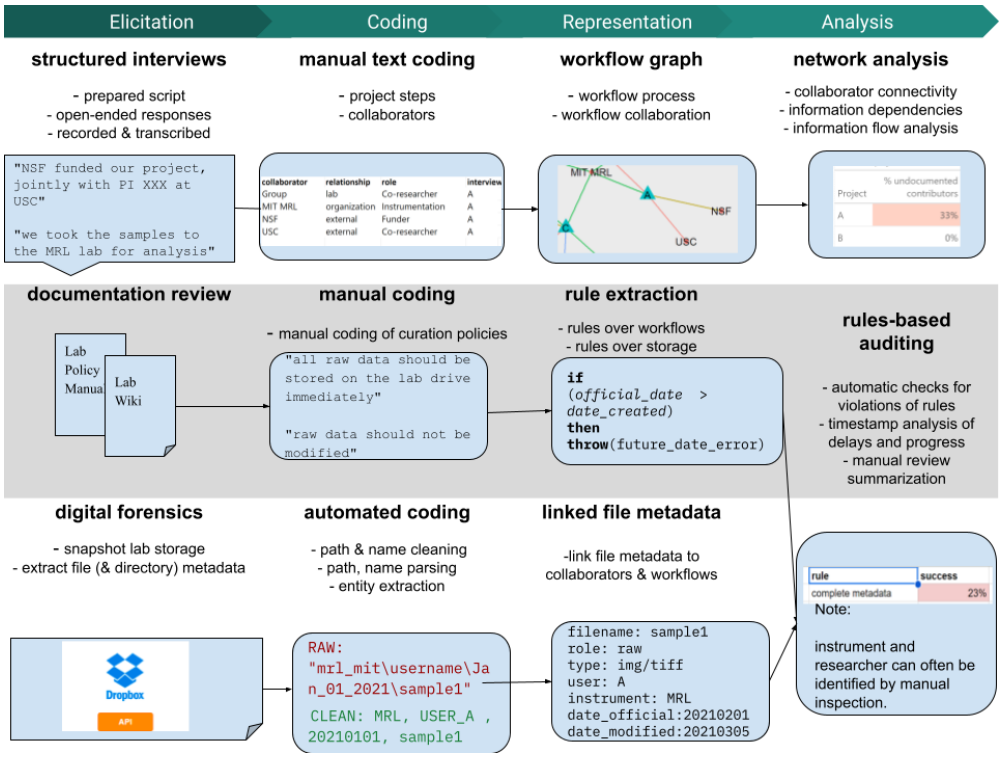

Figure 1. Overview of data sources and methods.

These workflow graphs are supplemented with information gained by a review of lab documentation and an audit of the lab's digital repository. With manual documentation coding, rules are extracted, which could be tested using extracted file-level metadata (digital forensics) collected by the audit.

Case and Interviewee Selection

The use of "small lab" is common in the literature; however, it is often unaccompanied by a precise definition. Within this paper, the term "small lab" refers to a set of researchers that: 1) self-identified as a research collective, 2) aims to conduct research and produce scholarly communications, 3) is substantially responsible for identifying its research agenda, design, and methods, 4) contains under 20 people, and 5) conducts experiments.

Although it is impossible to precisely determine the number of "small labs" in science, in general, because no comprehensive survey of research groups exists, however, past research into research group size in selected disciplines and countries (e.g., Brandt et al., 2021; Cook et al., 2015; Qurashi, 1984; Seglen & Aksnes, 2000) suggest that "small" research groups are a common or the predominant form of organization within the natural and applied sciences.

The total number of research groups in MSE is unknown. However, public university rankings establish that at least 750 academic materials science programs exist worldwide, and a substantial proportion of these probably include small MSE labs.

Professor Rafael Jaramillo's group conducts experimental materials science within the Department of MSE at the Massachusetts Institute of Technology (MIT) --iin Cambridge, Massachusetts, USA. Their research focuses on the synthesis, properties, and application of electronic materials. Each research project in the group generates many experimental data sets and can be supplemented by computational studies to reveal mechanisms or analyze structures. This typical type of workflow bonds the four elements of MSE research: 1) structure and composition, 2) synthesis and processing, 3) properties, and 4) performance (Flemings, 1999).

The Jaramillo lab meets all the criteria of a small MSE lab, it has a scientific aim, collects its data from local experiments, is composed of less than 20 Full Time Equivalent staff (FTEs), provides its scientific direction, and oversees its methods and infrastructure. However, the lab is unlikely to be statistically representative of small materials science labs for several reasons.

The external rankings of MIT's materials science department place it in the top five schools worldwide. The MIT is a well-resourced institution, and the MIT faculty are typically well-supported. The MIT faculty and Professor Jaramillo specifically successfully obtain external research support. Professor Jaramillo is interested in reproducible research and open science: he has published in this area, his group has developed related software prototypes and grant proposals, and has advocated for reproducible and open science practice within his institution and discipline. Therefore, this lab should be considered a near-best case for FAIR data workflows in small materials science labs. It is implausible that many other small experimental materials science groups have the resources, experience, or interest to perform substantially better in this area.

Synergistic collaboration between group members, including graduate students and postdoctoral fellows, allows for the continuous monitoring of lab equipment. This collaboration is facilitated by shared information repositories, including a group ELN in LabArchives, a group Dropbox account, and Google Drive. Access to all the cloud-based storage and services is provided to the group via MIT campus-wide site licenses. The protocol for saving and sharing information is specified in a group manual, which all researchers in the group are encouraged to follow for the group repositories and personal data storage systems. Therefore, this lab is representative of good practices for data-sharing because individual data from one researcher is, ideally, stored in a format that is comprehensible and a location that is accessible by all members of the lab. Consequently, reproducibility of research is possible in the absence of the originator of the research.

Investigating the workflow of four researchers within the Jaramillo group highlights which practices are most essential to open and reproducible research; these practices appear to be standardized across the researchers in the lab despite idiosyncrasies due to personal preference. Identifying these practices allows other "small academic labs" to formulate and adopt the most effective structure for their data storage framework.

Data Collection Methods

Structured interviews were conducted with four graduate students in Jaramillo's group to obtain the specifications of their workflow, data profile, and challenges in daily practices. This study (Exempt ID: E-2317) was exempt from further review by the Committee on the Use of Humans as Experimental Subjects (COUHES) at MIT on June 2, 2020.

Two researchers interviewed each graduate student: one served as the interviewer and the other as the transcriptionist. The interview audio was recorded and reviewed, and the transcribed notes were compared post-interview for completeness and accuracy.

The interview protocol (e.g., Appendix I) consisted of three sections: 1) interviewee background, 2) top priority project background, and 3) top priority project workflow. The protocol was a guideline for the interviewer to construct the most complete narrative of each student's workflow. Each question was explicitly asked or indirectly answered with the student's response to a different question.

For the last section (i.e., top priority project workflow), it was evident that the most natural interview process was one in which the student first described the overall workflow for the selected project and then was prompted to recall each operational step. The interviewer then asked follow-up questions to fill in gaps and probe for additional detail.

Interview Coding

The interview coding process aimed to describe each workflow step in a systematic structure database. A wide range of existing formal models for provenance and workflow (e.g., Jandre et al., 2020). However, most of these are designed for automated execution and contain much more detail than is feasible to elicit during a standard interview. Therefore, a simplified coding approach was used in which the actions that each described an action were labelled with their objective, task, and sub-task. The sequence, actor, input, output, data source, data target, level of automation, type of action, equipment, and methods used were recorded using standardized codes (e.g., *Appendix II* for the coding method and dictionary.) Then, this tabular data was used to impute collaboration and data-dependency graphs (e.g., the following *Results* section).

Workflow Representation

Workflows were represented as formal graphs and social network analysis methods were applied. This follows a common approach to interpreting workflows, first documented by Tan et al. (2010). The graph systematically describes all process, informational, and collaboration dependencies elicited by the interview process. By analysing these graphs, workflow gaps were visually and analytically identified, processes concerning the stated policy were evaluated, and potential interventions were probed.

This core workflow graph was augmented in several ways. First, a collaborator network graph was created by coding the interviews directly for any mentions of collaborations. Second, separate dependency, collaboration, and information flow graphs were derived directly from the workflow process graph. Finally, graphical methods from network and social network analysis (Carrington et al., 2005; Horwitz & Reps, 1992; Sharir, 1981) were applied to probe questions related to collaboration (e.g., analysis of connectivity and centrality), attribution (e.g., comparing the explicitly elicited collaboration graph with its workflow-induced counterpart), and replicability (e.g., dependency and subcomponent analysis of the information flow graph) (Appendix III-A).

Digital Forensics

During the interview process, the lab used a shared folder in Dropbox as its official repository for collected data and documentation. With permission from the Principal Investigator (PI), a snapshot of the repository contents was cloned, and all of the file system metadata for use in a digital forensic analysis were collected. For consistency with lab policies and the timeline of the projects evaluated, the analysis was restricted to files deposited between January 2019 and January 2022.

Deposit patterns demonstrated that the repository was actively used, with deposit rates varying seasonally (Appendix V). All information was collected for 31,929 files, including file names, paths, content hashes, client-side modification, and deposit times. Additional internal timestamps and content metadata were obtained for image files, which are a common set of raw data formats used in this lab. Finally, many files used a naming convention to embed additional information, such as creation date, creating user, and creating instrument; therefore, regular expression-based cleaning and parsing to extract this information was used where possible. This information set was then used to check for inconsistencies with documented information organization practices, as described in the following results section.

Post-Analysis Validation

Appendix III-B gives details on post-analysis validation with follow-up interviews and documentation review.

Results

Workflow Overviews

For context, the workflows for each of the four projects are described as follows.

Interviewee A was involved in generating each output with the workflow sequence: sample preparation, synthesis, characterization, and analysis. The workflow sequence was iterative, so the analysis phase results informed how the next iteration's synthesis process was tuned. They use a personal LabArchives notebook to record observations. Metadata from equipment is recorded in the group Dropbox, and pre- and post-processed data are saved to the group Dropbox.

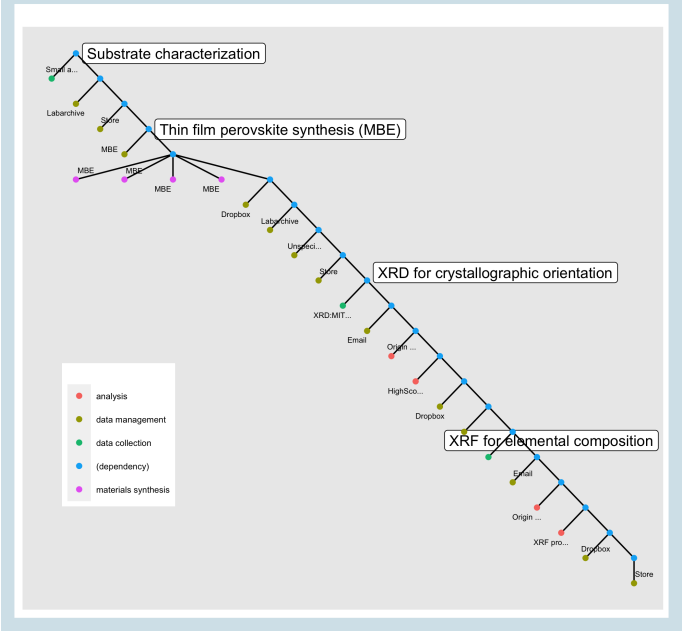

Figure 2A. Workflow overview: Project A (workflow steps by phases, description, and type). MBE: Molecular Beam Epitaxy; XRD: X-ray Diffraction; XRF: X-ray Fluorescence.

Interviewee B received the synthesized sample from a collaborator. They were responsible for preparing the sample for analysis, characterizing it, and analyzing the data. They used a personal LabArchives notebook as an ELN; therefore, any conditions needed to interpret and replicate a process were recorded. The group LabArchives notebook records measurements on lab tools that are shared to maintain a consistent tool log (required by the professor). The group Dropbox is used to save raw data directly from instruments. Post-processed data is saved to a personal Dropbox.

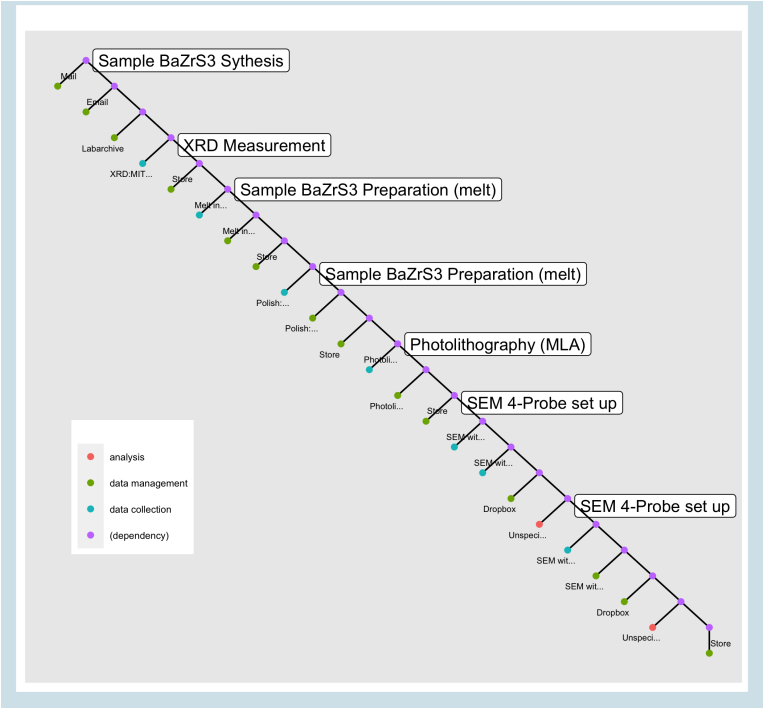

Figure 2B. Workflow overview: Project B (workflow steps by phases, description, and type). XRD: X-ray Diffraction; MLA: Multilayer; SEM: Scan Electron Microscopy.

Interviewee C received the synthesized materials from a collaborator. They prepare the acquired sample, characterize it, and analyse the results. Finally, they transform the sample via a laser setup; this process is iterative as the transformed sample is characterized. A personal OneNote notebook is used for experimental notes. OneNote is manually synchronized to the group LabArchives notebook. The group Dropbox records all sample notes, raw data, and analysed data.

Interviewee D is directly involved in each sequence step, which includes sample preparation, characterization, analysis, and simulation. A personal LabArchives notebook is used to write details of each experiment and record measurements from the equipment used during the synthesis process. The group Dropbox saves equipment metadata and raw or lightly processed data from characterization. A personal Dropbox is used for processed data.

Of note, each workflow is hierarchical; each project does not interact (there are no connecting branches), and the work can be represented as a set of independent, self-contained tasks (summary graph statistics are shown in Appendix V, Table A1.) Most of the tasks contain one atomic action. In addition, there is a rhythm across each workflow in which the type of task at each step alternates.

All four interviewees used instruments or equipment outside their lab, either at a shared facility or a collaborator's lab. Each interviewee saved a copy of the raw data from those instruments in the group Dropbox; however, they had different practices when transferring data. Each in-house instrument in the lab is overseen by an unofficially designated group member for maintenance. Regular maintenance notes for each in-house instrument are recorded in the shared LabArchives notebook folder.

Group members regularly use equipment outside the lab and MIT, which interviewees indicated creates additional data transfer and documentation challenges. Interviewees noted that equipment within the MSE department is locatable through an internal wiki; however, no other central documentation or standardization around equipment configuration, data transfer, network access, or acknowledgment of equipment use exists.

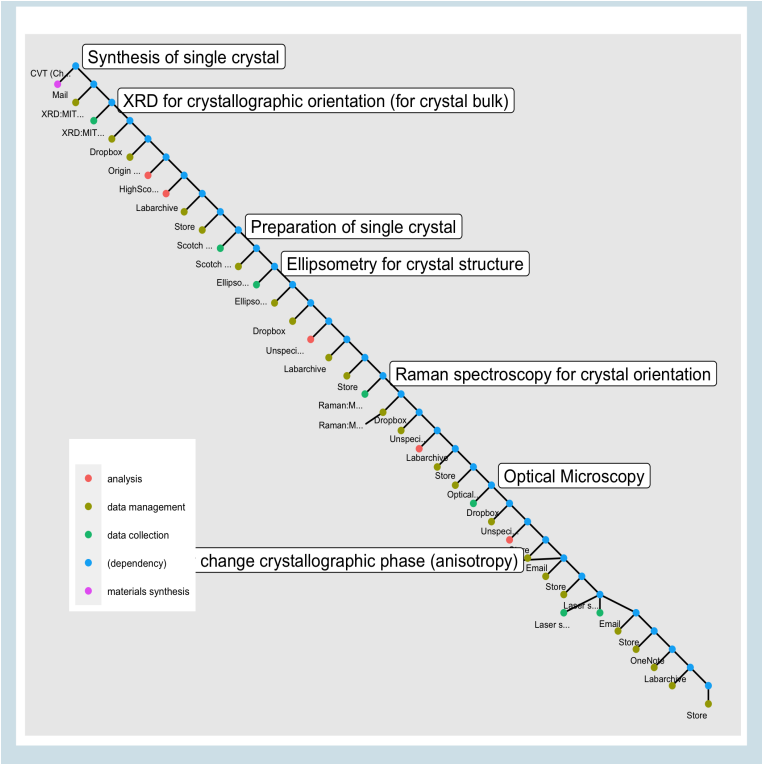

Figure 2C. Workflow overview: Project C
(workflow steps by phases, description, and type). XRD: X-ray Diffraction.

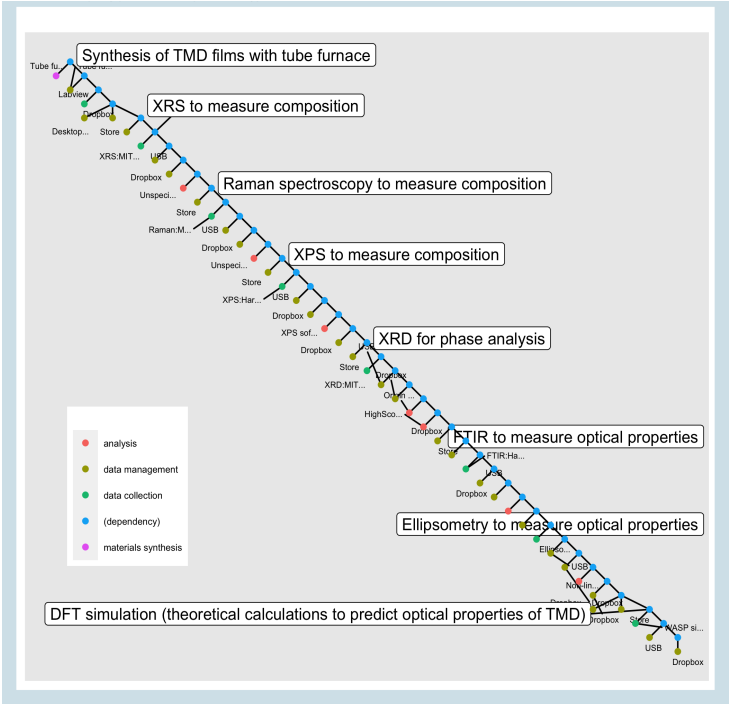

Figure 2D. Workflow overview: Project D (workflow steps by phases, description, and type). TMD: Transition-metal dichalcogenide; XRS: X-ray Spectroscopy; XPS: X-ray Photoelectron Spectroscopy; XRD: X-ray Diffraction; FTIR: Fourier Transform Infrared Spectroscopy; DFT: Density Function Theory.

Workflow Automation

Limited systematic research on the rate and frequency of human errors in scientific research generally (or MSE specifically) exists; however, a long history of research in human performance and reliability engineering suggests that human error rates are substantial in the absence of well-engineered monitoring and error-mitigation regimes (Reason 1995; Jacobs 1995). For example, over the last 15 years, human error in medicine has been a focus of study. Systematic reviews demonstrate the high level of harmful and avoidable human error and the efficacy of error, reduction processes, such as the adoption of automated recording systems and the use of explicit checklists and logs for manual procedures (Rodziewicz et al., 2020; Institute of Medicine, 2000).

During the interviews, information was collected about the automation associated with each instrument, storage facility and analysis method. As described in the methodology section, each workflow step was coded for the type of action performed and level of automation used: 1) an "automated" if the step is initiated automatically following the prior step, 2) "partially" automated if the operation was launched manually but was entirely described by digital metadata (e.g., configuration files), or 3) as "manual" if the step depended on manual initiation and manually configuration for correct operation.

Figure 3 summarizes the selected characteristics of the workflow. Overall, this figure reveals that workflow is dominated by manual activities. Figure 3 provides an answer to the first research question concerning automation, i.e.

1. To what extent does research depend on manual processes for information management?

Automation is not a panacea and can increase system complexity or decrease local transparency in ways that increase errors across a broader system. However, automation is often recommended for tasks not involving complex judgment (e.g., file transfers) and are not otherwise associated with a specific performance, audit, or quality assurance procedure. Further, targeted automation enables people to shift their efforts to tasks where judgment is required, reducing the cost and effort of logging and auditing. Therefore, where errors occur, they are more readily detected.

Further, explicit communication and documentation are relatively infrequent. There is a high level of reliance on manual transmission of information (e.g., for instrument setup or contextualization of the analysis) and a substantial incidence of email and portable media for information storage. Together, this suggests a significant opportunity for human error in data management and organization.

Collaboration and Information Flow

In the previous section, the workflows were used to show the dependencies between steps in the research process. The same workflow graphs were used to derive the dependency graph for each analysis and, in conjunction with interview data, to derive the collaboration networks.

The information connections within and across projects (Figure 4) are densely interconnected within projects, contrasting with the linearity of the process used to produce the information. Further, most information flow is implicit with a shared context. Information rarely flows through direct input–output. There is no information flow between projects.

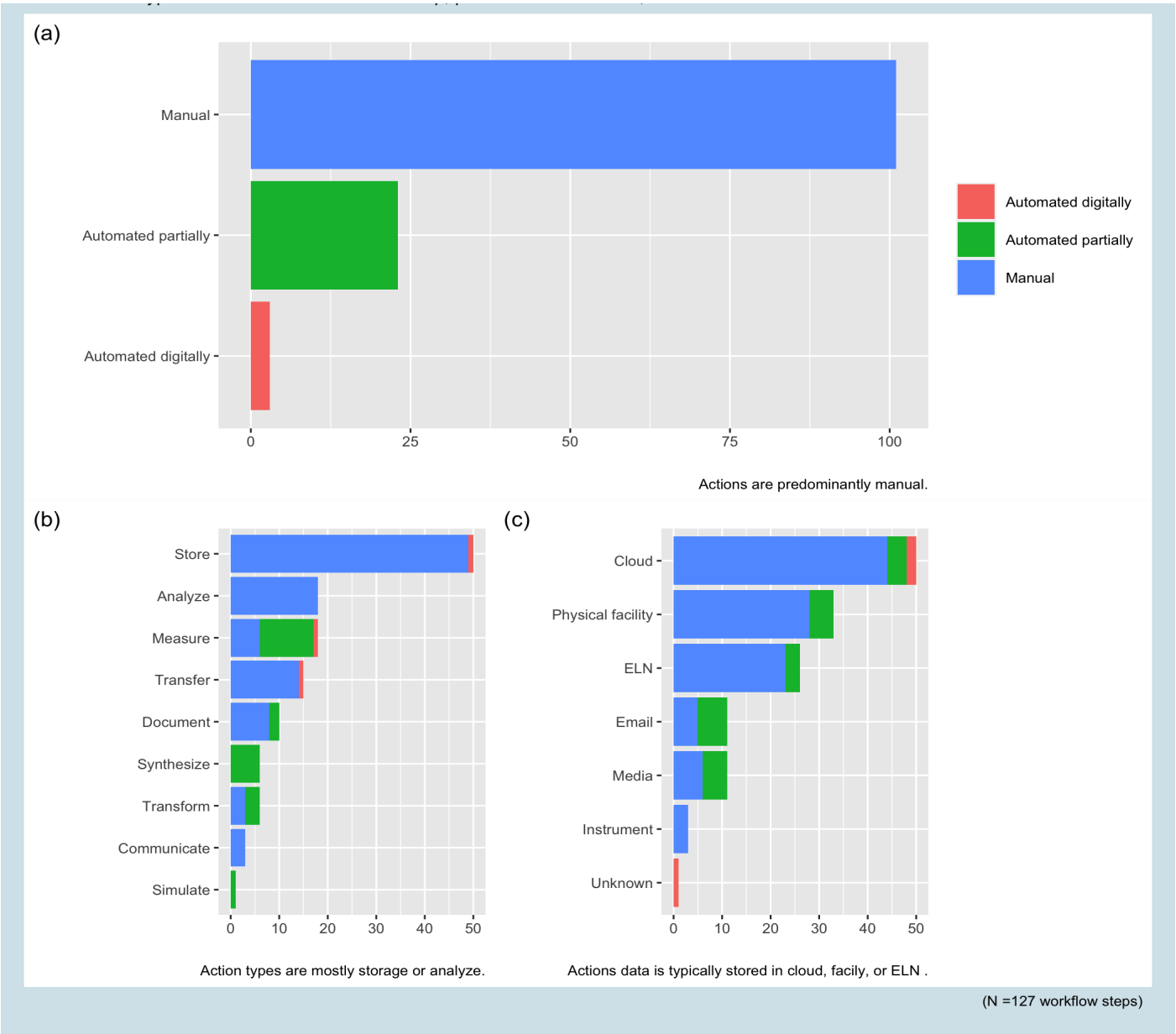

Figure 3. Selected characteristics of the workflow steps (summarizes the type of action described in each step, proximate data source, and level of automation).

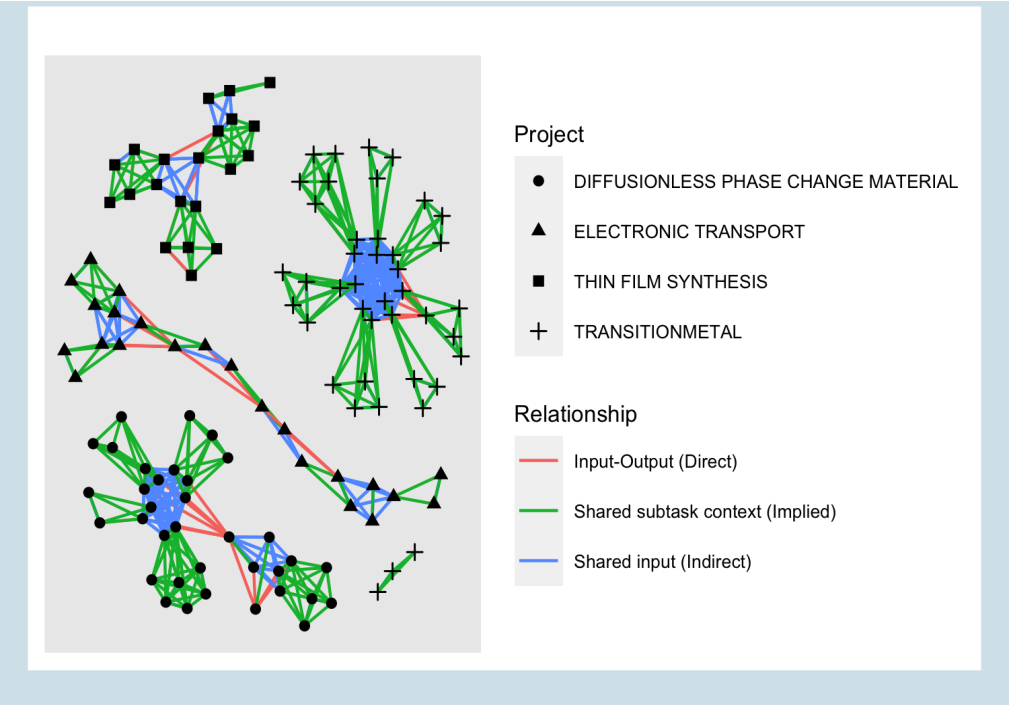

Figure 4. Project information exchange (implicit and indirect information exchange occurs frequently within projects but does not connect projects).

In addition, collaboration networks (Figure 5) are partitioned by project and workflow. The size of the grid varies substantially across projects.

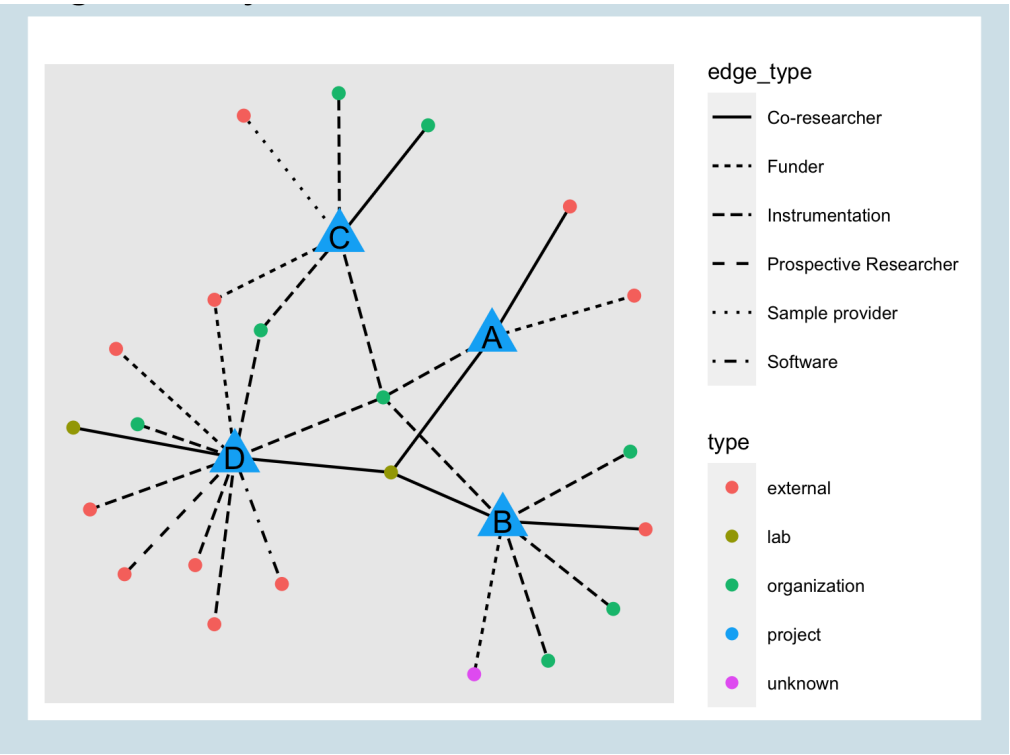

Figure 5. Project collaboration.

Lab Practices

Characterizing documented data practices

The second research question concerns documented practices:

2(a). What data and research workflow management processes are documented?
2(b). To what extent are documented processes consistent with practice?

The practices were identified with direct interview questions administered during face-to-face interviews of the PI and group members. Then, copies of the documented processes were obtained from the subjects to characterize these.

During the interviews, one interviewee, who self-identified as a founding member of the group, mentioned a document "Jaramillo group new member checklist," which described shared computing resources, lab safety, and access, and data management practices. The document and group wiki site were reviewed to summarize the documented practices. In addition, the documented practice was compared with practices elicited by interviews for relevance to the research questions, and to inform the interpretation of the workflow networks as follows.

The practices that were most relevant to data management and scientific workflow management were identified and grouped into three categories: 1) information sharing, 2) security, and 3) organization (Appendix IV.)

Consistency of documented versus recalled and observed practices

Four strategies were employed to evaluate the consistency of the recalled practices with documented practices:

1. In general, all instances were identified where subjects explicitly referred to documented or established practices during the interview, either during the description of their project or separately.

2. For information sharing practices, network analysis of the workflows was used to identify where each information object was stored and compared this against documented policy.

3. For information security processes, whenever the network analysis identified information as stored in a non-group location, the subject verified whether the location was backed up using CrashPlan or an equivalent MIT service.

4. For information organization, the group Dropbox file listing was reviewed to confirm practices.

The results are summarized in Table 1.

The most significant deviation with formal documented practice is in the area of information organization. Of the 31,929 deposited over 2 years, less than a quarter (23%) provided could be readily assigned a collection date, researcher, and instrument. Furthermore, where collection dates were assigned, they often (53.5 %) pre-dated and, in rare cases, post-dated the modification time stamps provided by the researcher's computers when they were delivered to Dropbox (or by the image creation software, where applicable). In the absence of more systematic processes when maintaining the provenance and authenticity of digital records, this discrepancy raises the possibility that data files could have been modified after collection.

Table 1. Comparison of documented group practice with recalled individual practices.

Documented procedures	Inconsistencies with practice
Information sharing	Practices are predominantly consistent with documentation, although occasional lapses occur
Information security	Practices are consistent with documentation
Information organization	Practices are frequently inconsistent with documentation; however, the instrument, username, data, and sample can often be identified by human inspection of the file and directory name

The following assessment section provides more details on information sharing.

Process Robustness Assessment

To address the remaining research questions, the mathematical graphs describing the workflow process, information, and collaboration were measured and compared.

Internal replicability

The next research question concerns internal replicability:

3. To what extent are documentation processes complete enough to support another person's replication of a result within the lab (without further communication with the original researcher)?

In general, a documentation process might be implicit or explicit, and the documentation might be integrated with analytic outputs or stored separately. As noted in the previous sub-section, the documented practice in this lab does not include active replication of results before publication, nor does it require that materials and instructions sufficient to replicate published articles are available. Follow-up interviews (discussed at the end of this section) revealed that some projects have since adopted an informal local practice of depositing replication materials to the group drive after publication.

The group exhibits documentation practices during the data collection and analysis process to aid future replication. The interviews and workflow analysis demonstrate the use of multiple documentation strategies. For example, some data (and analysis) formats and systems provide the capability to store information on how the data (or analysis) was produced and how it is to be interpreted. When this capability is used, the documentation is described as integrated into the data (e.g., the data could be termed "self-documenting.")

However, much of the time documentation is stored separately from the outputs produced by measurement, experiment, and analysis. The researcher can manually add this separate documentation (e.g., a lab notebook entry or notes file). Alternatively, documentation may be implied by a previous step (e.g., when a measurement process is controlled by a configuration file already recorded).

The workflow information graph was used to identify when data or analysis was produced. Then, the graph was analysed to match each output to potential documentation based on the following:

- Outputs were coded as having "manual" documentation based on an analysis of the workflow graph to determine that data and documentation objects were produced during the same substage, or supplementary statements in the interviews that a specific output was manually documented.
- Outputs were coded as having "integrated" documentation when the output format matched a specific format confirmed through the interviews to be part of a general self-documentation process.

- Outputs were coded as having "implicit" documentation when they were derived from processes that were (semi-)automated and where either log files or generating scripts were also stored. Table 2 summarizes these categories of documentation.

Table 2. Documentation of outputs (missing documentation obstructs reproducibility).

	Integrated	Manual	Implicit	Missing
Processed data	0 (0%)	8 (88.89%)	1 (11.11%)	0 (0%)
Analysis	0 (0%)	8 (44.44%)	0 (0%)	10 (55.56%)
Raw data	7 (50.00%)	5 (35.71%)	2 (14.29%)	0 (0%)

Of note, the existence of documentation necessary for unassisted replication is not sufficient. The completeness of the documentation, if it existed was not evaluated, only its presence. However, analysis documentation was missing in over half of the cases examined. This obstructs future replication of results and publications, which must rely on communication with the researcher who conducted this analysis (and on their memory) and trial-and-error.

Robustness of storage practices

The following research question concerns the robustness of storage practices:

4. To what extent are data management processes robust enough to survive the departure of a project member or the loss of an individual's personal computer or storage?

The workflow information graph was used to probe this question to identify all collected data (digital and physical samples created as part of each scientific workflow), metadata, and analysis results. Then, the process in the graph was used to trace the flow of these objects across tasks and into storage locations. From this set of traces, the content of the designated group storage location post-analysis was inferred. The results are summarized in Table 3.

Table 3. Proportion of output in managed storage, by type (a substantial portion of highlighted outputs are at risk).

Type of research outputs	Percentage
Metadata	44
Analysis	44
Raw data	79
Processed data	100

Note: processed data includes derived, linked, and cleaned data; metadata includes configuration files, output logs, and manual documentation

On the positive side, almost all data objects (with exceptions) are deposited into institutionally managed shared group storage by the process end. This is consistent with the documented lab policy and is necessary for the work to support future data-sharing and for the workflow to be robust to the loss of an individual computer.

However, over half of the metadata and documentation and half of the analysis produced is never copied or transmitted to a group location but remains accessible solely from individually owned media, computers, or accounts. This will decrease the utility of data-sharing, because most data is not self-

documenting, and threatens the replicability of analysis. If a group member were to depart, there is insufficient information available to ensure that the work can be replicated or re-validated, even internally. Further, in a small number of cases (2), raw data were stored outside the group storage, contrary to documented policy.

The same approach described previously was used to identify when analyses depend on manual information transfer rather than being automated. Given the high frequency of manual operations documented in the previous section, it is unsurprising that 100% of the analyses relied on manual information management at an earlier step in the experiment and measurement process.

Serendipitously, file forensics data collected from the lab-shared storage system provides a glimpse of the reliability of manual transfer processes. The delay between data creation and deposit by comparing the manually recorded date in the path with the automatically recorded date in the shared file system can be measured. For half of these files, the delay is relatively small (40% of these files were deposited within 1 day). However, a substantial percentage were considerably delayed (25% of files were deposited after a delay exceeding 95 days). Of note, two mechanisms could produce significant delays. First, where raw data is collected and transferred by hand, errors, interruptions, or forgetfulness can contribute to the delayed deposit. Second, during the validation interviews, it was identified that some projects adopted an informal process of adding files associated with a publication, after that publication had been accepted. These added files can include processed data files and descriptions of data collection and analysis processes. It is impossible to determine the proportion of lag attributable to each mechanism because of the inconsistent use of documented naming practices and the variation in undocumented practices.

In addition, the 2 years of files examined included a substantial number (863) of image files that contained internal creation-time metadata produced by the original software. By comparing this time-stamp with the shared file system time-stamp, the elapsed time between creation and deposit could be computed. The delay is quite small for most of these files; less than a workday (75% of these files were deposited within 5 hours). However, the distribution of deposit latencies has a long tail, with some files not deposited until months (3,008 hours) after creation.

Attribution robustness

The final research question concerns attribution:

5. To what extent are workflow data, outputs, and documentation sufficient to describe responsibility (or support attribution) for published results?

To examine the final question, respondents were interviewed to elicit lists of all the collaborators on the project and their general collaborative relationships. This list includes active collaborators (e.g., actors who supply material, perform an analysis, or contribute to writing for publication) and passive collaborators (actors who provide access to equipment or software). From the interviews, it was confirmed that there were no written or standard processes or policies concerning recording or acknowledging collaborators. In assigning attribution, interviewees reported relying primarily on memory rather than written documentation and outputs.

A partial exception to the reliance on memory is an informal practice discovered during the file forensics analysis. A common practice was to structure the directory trees so that data produced by a specific instrument was contained under a folder named for the principal investigator. Where this practice was followed with a particular instrument this was coded as documentation of the collaboration.

Workflows may document collaborations explicitly (e.g., with entries in a lab notebook or an author line in an analysis document) or indirectly (with an email correspondence history). To quantify the degree to which attribution relies on memory, the list of collaborators stated by interviewers was compared with a list of collaborators that could be detected from the workflow outputs and documentation. To achieve this, direct and implied collaborators were extracted from each workflow step (e.g., when another person was recorded as doing the analysis when the interviewee sent someone an analysis by email, or an analysis when an external instrument was used).

As expected from the interviews, many collaborators are omitted from workflow documentation or action altogether. Table 4 summarizes these omissions.

Table 4. Undocumented collaborators (types of collaborations that were recalled but not documented in project work).

Project	Undocumented types	Percentage of undocumented contributors
A	Co-researcher	33
B	None	0
C	Co-researcher, sample provider	40
D	Instrumentation, prospective researcher, co-researcher, software	40

As shown in Table 4, a significant proportion of the collaborations could not be associated with the work process, information used in it, or the analysis produced by it.

Discussion: Toward More Reproducible and Attributed Practices

In summary, the workflow, documentation, and digital forensics analyses revealed the strengths and limitations of current practice. Practices in the lab mitigate the risk of data loss resulting from the failure of an individual's computer and ensure access to the raw data collected for the lab research.

The preservation of the data is necessary but insufficient for trustworthiness and transparency. Lab data curation practices often deviate from the stated policy and vary across projects, especially for the metadata and documentation needed to contextualize and analyse the collected data. Moreover, the policy and practice are insufficient for the attribution, replication, or verification of the labs' published results.

Therefore, the integrity and continuity of lab research are threatened if an individual fails to maintain private records of attribution and data provenance or withdraws from the research group. The following improvements are required.

In general, several general strategies could be employed to address the workflow gaps and should be considered as an approach to the gaps discussed previously:

- The addition of processes to regularly audit and validate ongoing projects for reproducibility and attribution.
- Changes in the research infrastructure (defined broadly) to automate the capture, transfer, or storage of critical information, preferably in standardized formats with necessary metadata.
- Changes to the lab policies regarding requirements for those activities are performed manually.

Recommendations for Auditing

It is a truism that manual processes and policies must be regularly audited and verified to be effective. Auditing and verification should evaluate the use of documented practice and the achievement of desired outcomes.

Recommendation 1

For the documented practices, minimal automated audits, in support of sanity checks, could verify that documented naming conventions are being followed and that systems are running backup software. For outcomes, less frequent (e.g., semi-annual) manual audits could be used to validate that

the current analytic results from each project could be reproduced from (or at least traced back to) data and metadata curated in the group storage.

Recommendation 2

Automated processes sometimes fail or are misconfigured. Automated validation could be used to detect system failures and flag unusual activity patterns for further investigations. For example, automated analysis of group storage could be used to flag the absence of data collection and processing for purportedly active projects. Automated analysis of deposits could provide evidence of the "liveness" of projects and individuals. In addition, automated analysis could correlate the timing of lab notebook updates with the timing of data deposits into the group storage system, substantial data changes/updates without corresponding lab notebook signal a potential threat to reproducibility.

Recommendations for Upgrading Infrastructure

Where feasible, automated infrastructure is attractive because they do not require people to change behavior–which is often costly, difficult to assess, error-prone, and requires consistent focus to maintain. While a fully automated infrastructure for materials science remains currently too expensive and immature for many labs, more minor changes in infrastructure and tooling have the potential to mitigate a number of the gaps identified by the workflow analysis.

Recommendation 3

All of the reported workflows involved the extensive use of personal portable storage to transfer data from experimental instruments manually. In addition, the file forensic analysis showed that the delays between file creation and deposit could be quite significant. In addition, no systems or processes are in place that would detect common categories of human errors that occur at this stage, such as erasing or overwriting local files, loss or replacement of the storage device, failure to delete files after the transfer is complete, or transfer of the files to an incorrect destination (e.g., user's personal computer or cloud) should these occur. This suggests that manual data transfer and operations will increase errors.

The portable storage is typically a simple off-line USB "flash drive." Alternative USB-compatible mobile storage devices, including built-in wireless networking and data synchronization capabilities, are readily available. Although researchers would need to transport these storage devices with a network connection to the instruments and plug them in, the manual data transfer to cloud storage could be automated, reducing the risk of reproducibility. This portable storage device would not introduce more security risks for instruments in shared facilities than an off-line USB "flash drive" would. During the analysis validation interviews, participants noted that enacting this recommendation would require agreement and action from the equipment or facility owners to align information security policies.

Recommendation 4

Similarly, most workflows involved a significant amount of regular transfer from personal cloud (e.g., Dropbox) to a group cloud storage. When multiple independently managed locations are unnecessary for data processing, analysis, and backup, eliminating the use of multiple storage locations would lower the risk of introducing inconsistency. When multiple independently managed services are necessary, services are readily available that could monitor target folders in one storage system and replicate or synchronize them with another. Using these tools and a more systematized practice of folder organization for work products maintained in personal storage would enable more reliable and robust data lab practices without sacrificing the convenience of personal cloud storage accounts.

Recommendations for Refining Practices

Although infrastructure and audition of current practices could be expected to facilitate the workflow gaps identified in this research, additional refinements to lab practices could be required in two areas.

Recommendation 5

Develop explicit practices around collaborator attribution. Practices are needed to identify the contributions of collaborators systematically. This might include: 1) enhancing existing workflow project documentation (e.g., in the lab notebook) to identify when the researcher uses externally contributed resources, borrowed equipment, or information received from a collaborator, 2) explicitly saving contributed data, analyses, and comments from collaborators in the group storage, rather than in personal emails, and 3) defining contributor roles according to taxonomy, such as the Contributor Roles Taxonomy (CRediT),2 in group documentations.

Recommendation 6

Develop explicit practices around reproducibility beyond the stage of raw data:

- Make documentation for standard practices at commonly used equipment in external locations (i.e., other MIT facilities, such as MRL and DMSE). Consistent practices at these facilities would allow for the comprehensible transfer of data between researchers within the lab.
- Establish a group-shared location for metadata (especially equipment parameters), which is essential to reproducibility. Monitor the progress in open data standards in the field and start to adopt them.
- Encourage analyses to be conducted in a framework that builds reproducibility, for example, using executable scripts or notebooks stored in cloud storage rather than spreadsheets transmitted by email.

Future Research

In this research, gaps were identified in an exemplar set of materials science workflow processes and characterized approaches to address those gaps. However, the effectiveness of specific practices and approaches is an open question: Empirical evidence, preferably from designed interventions, is needed to reliably measure how better practices can improve reproducibility and research attribution. (Altman & Cohen, 2021; NASEM, 2018)

Moreover, these practices are embedded in and responsive to a much broader system of scientific incentives, institutional and organizational collaboration, and professional training (Altman & Bourg, 2018) – research is needed on how effective practices can be aligned with incentives, training, institutional coordination, and infrastructure improvement.

Intrinsically, recognizing the value of FAIR data-sharing and computational use of experimental data for the research community in general and their study could further motivate individual researchers and their teams.

Hiring data curators or research workflow facilitators to provide discipline-specific support for particular groups and departments could further enable researchers to overcome the barriers to starting new practices.

The improvement in interfaces for human–computer interaction, accessibility, and security of cloud-based systems could be the key to lowering individual groups' barriers to fully adapting the digital workflows recommended, especially when shared instrument facilities are often inseparable components of the infrastructure.

With the improvement in research infrastructure for MSE that could integrate experimental data management and sharing and AI-based materials design, it will become critical to study how "small academic labs" could adapt to this infrastructure cost-effectively for open and reproducible research when maximizing creativity.

[2]CRediT https://credit.niso.org/

Data Availability

The deidentified data sets generated during or analysed during the current study, as well as the R scripts used for analysis and generating the research report, are available in the Zenodo repository under CC BY 4.0 license.[3]

Acknowledgments

The authors describe contributions to this article using a standard taxonomy. (Allen et al., 2014) All authors equally shared the core formulation of the research goals and aims. All authors co-developed the research design and the interview instrument and plan. M.A. led the formal analysis and visualization. Y.L. led administration, funding acquisition, and supervision. S.W. and Y.L. led the data curation (including collecting) and investigation. All authors shared writing the original manuscript and further refinement with commentary, review, editing, and revision.

The authors thank Professor Rafael Jaramillo at MIT for his commentary and for enabling access to lab records, as well as members of the Jaramillo research group for participating in interviews.

The authors thank MIT Libraries for the special fund and support for the project.

References

4CeeD Design Team, User Study and Survey on Material-related Experiments. (2016). Retrieved from https://www.ideals.illinois.edu/handle/2142/94738

Allen, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014). Publishing: Credit where credit is due. *Nature*, *508*(7496), 312–313. doi.10.1038/508312a

Altman, M., & Bourg, C. (2018). A Grand Challenges-Based Research Agenda for Scholarly Communication and Information Science. *MIT Grand Challenge Participation Platform*. doi.10.21428/62b3421f

Altman, M., & Cohen, P. N. (2021). *The scholarly knowledge ecosystem: Challenges and opportunities for the field of information*. doi./10.31235/osf.io/ctdb9

Blaiszik, B., Chard, K., Pruyne, J., Ananthakrishnan, R., Tuecke, S., & Foster, I. (2016). The Materials Data Facility: Data Services to Advance Materials Science Research. *JOM*, *68*(8), 2045–2052. doi.10.1007/s11837-016-2001-3

Brandt, N., Griem, L., Herrmann, C., Schoof, E., Tosato, G., Zhao, Y., Zschumme, P., & Selzer, M. (2021). Kadi4Mat: A Research Data Infrastructure for Materials Science. *Data Science Journal*, *20*. doi.org/10.5334/dsj-2021-008

Carrington, P. J., Scott, J., & Wasserman, S. (2005). *Models and methods in social network analysis* (Vol. 28). Cambridge university press.

[3] CC BY 4.0 https://doi.org/10.5281/zenodo.7158715

Cook, I., Grange, S., & Eyre-Walker, A. (2015). Research groups: How big should they be? *PeerJ*, *3*, e989. doi./10.7717/peerj.989

Coudert, F.-X. (2019). Correcting the Scientific Record: Retraction Practices in Chemistry and Materials Science. *Chemistry of Materials*, *31*(10), 3593–3598. doi.10.1021/acs.chemmater.9b00897

Delgado-Licona, F., & Abolhasani, M. (2023). Research Acceleration in Self-Driving Labs: Technological Roadmap toward Accelerated Materials and Molecular Discovery. *Advanced Intelligent Systems*, *5*(4), 2200331. doi.10.1002/aisy.202200331

FAIR-DI. (2022). *FAIR-DI e.V. - FAIR Data Infrastructure for Physics, Chemistry, Materials Science, and Astronomy*. Retrieved from https://www.fair-di.eu/about/info.

Flemings, M. C. (1999). What next for departments of materials science and engineering? *Annual Review of Materials Science*, *29*(1), 1–23. doi.10.1146/annurev.matsci.29.1.1

Han, R., Walton, K. S., & Sholl, D. S. (2019). Does chemical engineering research have a reproducibility problem? *Annual Review of Chemical and Biomolecular Engineering*, *10*(1), 43–57. doi.10.1146/annurev-chembioeng-060718-030323

Hill, J., Mannodi-Kanakkithodi, A., Ramprasad, R., & Meredig, B. (2018). *Materials Data Infrastructure and Materials Informatics* (D. Shin & J. Saal, Eds.; pp. 193–225). Springer International Publishing. doi.10.1007/978-3-319-68280-8_9

Himanen, L., Geurts, A., Foster, A. S., & Rinke, P. (2019). Data-driven materials science: Status, challenges, and perspectives. *Advanced Science*, *6*(21), 1900808. doi.10.1002/advs.201900808

Horwitz, S., & Reps, T. (1992). The use of program dependence graphs in software engineering. *Proceedings of the 14th International Conference on Software Engineering - ICSE '92*. doi.10.1145/143062.143156

Informatics (2022). What is the Citrine Platform? *Citrine Informatics*. Retrieved from https://citrine.io/product/what-is-the-citrine-platform/.

Institute of Medicine. *2000. To Err Is Human: Building a Safer Health System.* . (2000). National Academies Press. doi.10.17226/9728

Jacobs, P. (1995). *Human Reliability and Safety Analysis Data Handbook* by David I. Gertman & Harold S. Blackman 1994, 448 pages, $69.95 New York: John Wiley & Sons ISBN 0-471-59110-6. *Ergonomics in Design: The Quarterly of Human Factors Applications*, *3*(2), 33–34. doi.10.1177/106480469500300209

Jandre, E., Diirr, B., & Braganholo, V. (2020). Provenance in collaborative in silico scientific research. *ACM SIGMOD Record*, *49*(2), 36–51. doi.10.1145/3442322.3442329

Maleki, R., Asadnia, M., & Razmjou, A. (2022). Artificial intelligence-based material discovery for clean energy future. *Advanced Intelligent Systems*, *4*(10), 2200073. doi.org/10.1002/aisy.202200073

McNutt, M. K., Bradford, M., Drazen, J. M., Hanson, B., Howard, B., Jamieson, K. H., … & Verma, I. M. (2018). Transparency in authors' contributions and responsibilities to promote integrity in scientific publication. *Proceedings of the National Academy of Sciences*, *115*(11), 2557–2560. doi.10.1073/pnas.1715374115

Medina-Smith, A., Becker, C. A., Plante, R. L., Bartolo, L. M., Dima, A., Warren, J. A., & Hanisch, R. J. (2021). A controlled vocabulary and metadata schema for materials science data discovery. *Data Science Journal*, *20*. doi.10.5334/dsj-2021-018

National Academies of Science Engineering and Medicine (NASEM). (2018). *Open science by design*. Washington, DC.: National Academies Press. https://doi.org/10.17226/25116

National Academies of Science Engineering and Medicine (NASEM) (2019). *Policy, Global Affairs, Board on Research Data, Information, Division on Engineering, Physical Sciences, Committee on Applied, Theoretical Statistics, Board on Mathematical Sciences and Division on Earth,*. (A Consensus Study Report). National Academies Press. doi.10.17226/25303

National Science Foundation (2022). *Dear colleague letter: Effective practices for making research data discoverable and citable (data sharing)*. Retrieved from https://www.nsf.gov/pubs/2022/nsf22055/nsf22055.jsp

Nguyen, P., Konstanty, S., Nicholson, T., O'Brien, T., Schwartz-Duval, A., Spila, T., Nahrstedt, K., … & Paquin, N. (2017). *2017 17th IEEE/ACM international symposium on cluster, cloud and grid computing (CCGRID)*. 11–20. doi.10.1109/CCGRID.2017.51

Park, J., Howe, J. D., & Sholl, D. S. (2017). How reproducible are isotherm measurements in metal? *Chemistry of Materials*, *29*(24), 10487–10495. doi. 10.1021/acs.chemmater.7b04287

Plante, R. L., Becker, C. A., Medina-Smith, A., Brady, K., Dima, A., Long, … & Hanisch, R. J. (2021). Implementing a Registry Federation for Materials Science Data Discovery. *Data Science Journal*, *20*. doi.10.5334/dsj-2021-015

Pyzer-Knapp, E. O., Pitera, J. W., Staar, P. W. J., Takeda, S., Laino, T., Sanders, D. P., … & Curioni, A. (2022). Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials*, *8*(1), 1–9. doi.10.1038/s41524-022-00765-z

Qurashi, M. M. (1984). Publication rate as a function of the laboratory/group size. *Scientometrics*, *6*(1), 19–26. doi.10.1007/bf02020110

Raccuglia, P., Elbert, K. C., Adler, P. D. F., Falk, C., Wenny, M. B., Mollo, A., … & Norquist, A. J. (2016). Machine-learning-assisted materials discovery using failed experiments. *Nature*, *533*(7601), 73–76. doi.10.1038/nature17439

Reason, J. (1995). Understanding adverse events: human factors. *Quality and Safety in Health Care*, *4*(2), 80–89. doi.10.1136/qshc.4.2.80

Rodziewicz, T. L., & Houseman, J. E., Benjamin Hipskind. (2020). *Medical error reduction and prevention*. StatPearls. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK499956/

Savage, C. J., & Vickers, A. J. (2009). Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. *PLoS ONE*, *4*(9), e7078. https://doi.org/10.1371/journal.pone.0007078

Seglen, P. O., & Aksnes, D. W. (2000). *Scientometrics*, *49*(1), 125–143. doi.10.1023/a:1005665309719

Sharir, M. (1981). A strong-connectivity algorithm and its applications in data flow analysis. *Computers & Mathematics with Applications*, *7*(1), 67–72. doi.10.1016/0898-1221(81)90008-0

Stein, H. S., & Gregoire, J. M. (2019). Progress and prospects for accelerating materials science with automated and autonomous workflows. *Chemical Science*, *10*(42), 9640–9649. doi.10.1039/c9sc03766g

Tan, W., Zhang, J., & Foster, I. (2010). Network analysis of scientific workflows: A gateway to reuse. *Computer*, *43*(9), 5461.

Tinkle,S. McDowell, D.L., Barnard, A, Gygi, F. , Littlewood, P.B., Technology: Sharing data in materials science. (2013). *Nature*, *503*(7477), 463–464. doi.10.1038/503463a

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., …& Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, *6*(6), e21101. doi.10.1371/journal.pone.0021101

Timmer, J. (2011). *Symmetry free quasicrystals given the Nobel Prize in chemistry*. Retrieved from https://arstechnica.com/science/2011/10/symmetry-free-quasicrystals-given-the-nobel-prize-in-chemistry/

Westbrook, J. H., & Rumble, J. (1983, January 1). *Computerized materials data systems*. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1160497

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1). doi.10.1038/sdata.2016.18

Wilson, S. L., Altman, M., & Jaramillo, R. (2019). *Methods for open and reproducible materials science*. doi.10.31235/osf.io/ag8zu

Appendix

Appendix I: Interview Protocol

The interview protocol used in this research is included as file in the replication package: *Appendix_I_InterviewProtocol.pdf*

Appendix II: Coding Method and Coding Dictionary

The coding of this research workflow was conducted in four phases.

1. The first phase involved the direct translation of each interviewee's narration. During this phase, the steps in the sequence that were explicitly stated were recorded.
2. The second phase was an interpretation: the intended meaning of each statement was derived by assessing what the researcher implied but did not explicitly state. Each step of the workflow sequence had a series of sub-sequences that occurred before and after the main objective. For

example, when a physical material is placed in storage, it is implied that the next step involving it requires its removal from storage. The first and second phases were completed for all interviewees before progressing.

3. The third phase was inference. The same synthesis, characterization, and analysis techniques were often used across interviewees, and each lab member was subject to the same regulations to achieve each objective. Therefore, knowledge of one interviewee's workflow could be derived from what is known from another's workflow. This was used mainly for details, such as the names of analysis software and data output formats.
4. The fourth phase was extrapolation. The primary coder of this data was a materials scientist who conducted research in the same facilities used by the interviewees. This familiarity allowed for inferring implied steps from the workflow narrative that might not have been uncovered during the interview.

No additional assumptions were made during the coding. Any gaps in information that could not be acquired from these four steps were left blank.

The coding dictionary used in the study is included as a file: *Appendix_II_CodeDictionary_deidentified.ods*

Appendix III: A process of creating workflow graphs

The process of creating the workflow graph is summarized as follows. For replication purposes, all de-identified and coded interview data, the software code necessary to construct the graph in detail, and all the code needed to reproduce all figures and tables have been placed in a public archive:

- A node on the graph represents each atomic action (step) in the workflow. The node documents all characteristics of that single action;
    - Process dependencies are represented through sequences and sub-sequences linked by "process" edges.
    - Actions performed by the same person, in a required sequence, for a single goal, and over a continuous period are represented by "sequence" nodes. Edges link each sequence to one or more child sub-sequences.
    - Actions performed within a sequence (e.g., by the same person) and practically simultaneous (they have no natural order and occur during a brief period) are represented by sub-sequences. Edges link sub-sequences to one or more child steps.
- Informational dependencies are represented by augmenting the graph with "informational" edges. An edge is created when one of the following conditions holds:
    - When nodes share common data inputs; this represents passive information sharing.
    - When the output of one node is the input of another, this represents active information sharing.
    - When a single person conducts nodes during a continuous time (i.e., they are part of the same sequence), this represents implicit information sharing.
- Collaboration (attribution) dependencies are represented by augmenting the graph with typed nodes and edges:
    - Collaborator nodes represent individual or organizational collaborators.
    - Edges are created from workflows to collaborators when either the collaborator is explicitly referenced in the action (e.g., sending results to a collaborator, receiving samples from a collaborator) or, by implication, when the action involves some instrument (or other tool) provided by a collaborator.

Appendix III-B: Post-analysis validation

After the interviews were completed and their data were coded and analysed, the results were validated with follow-up interviews and a documentation review. In the follow-up interviews, for each subject, the gaps presented by the preliminary analysis were reviewed, confirmed whether the subject believed the gap to exist, and addressed the gap in the workflow. When an action was not noted in the original interview, the gap was addressed differently. Where these discussions pointed to workflow steps omitted during the initial interview, the workflow graphs were updated to include these additions.

In addition, the content of the existing group storage systems was reviewed (specifically, names, directories, and file types) to characterize data storage patterns per project and compare these to the patterns implied by the workflow analysis. In addition, content analysis was used to compare information organization naming practices with the documented lab policies.

Semi-structured follow-up interviews were conducted with all participants to assess the strength of (dis)agreement with the analysis described previously, its main conclusions, and the recommendations. In addition, it probed for additional comments, reflections, and recommendations. Participants consistently agreed with the analysis and confirmed the existence of the gaps that were noted.

Further, a number of participants reflected that since the initial interviews, they had noted some of these gaps and adopted informal practices within their project to address them. For example, one project had a local, undocumented, but intentional practice of, on the occasion of formally publishing an article, depositing into lab storage all analysis scripts necessary to reproduce the analysis in the article.

In addition, participants agreed with all areas of recommendations. One caveat: most participants noted that they faced institutional challenges when automating data collection from instruments outside the lab.

Appendix IV. Summary of Documented Data Practices

Information sharing

- P1. Shared data storage and management resources include a shared group account in Dropbox, a Group wiki, a shared group lab notebook in LabArchives, and a group Zotero account for sharing literature references.
- P2. All raw data (defined as "data as-recorded by the measurement instruments") must be stored in the group Dropbox folder and should never be modified. All internal lab computers are configured to save data to the group Dropbox folder automatically. Data collected outside the group lab must be manually transferred to the group Dropbox folder. Examples of raw data are JPG from a microscope, TXT from a probe station, or files in a proprietary format such as RAW from XRD.
- P3. Group members can store their analysis results wherever is most convenient.

Information security

- P4. Group members must use the MIT Enterprise version of CrashPlan to keep group-owned and individual computers backed up, especially the directories containing data or codes.
- P5. Group members are requested not to store raw data outside group-managed storage.

Information organization

- P5. The group Dropbox folder should be organized using the folder structure: instruments\username\YYMMDD\samplename.
- P6. Samples must be named consistently with a given scheme, including YYMMDD and a serial number.

Appendix V: AdditionalTables and Figures

Table A1. Information exchange graph statistics.

Direct connections	Graph diameter	Mean distance between steps
1,660	5	0.2172682

Table A2. Collaboration network

Direct connections	Graph diameter	Mean distance between steps
29	1	1

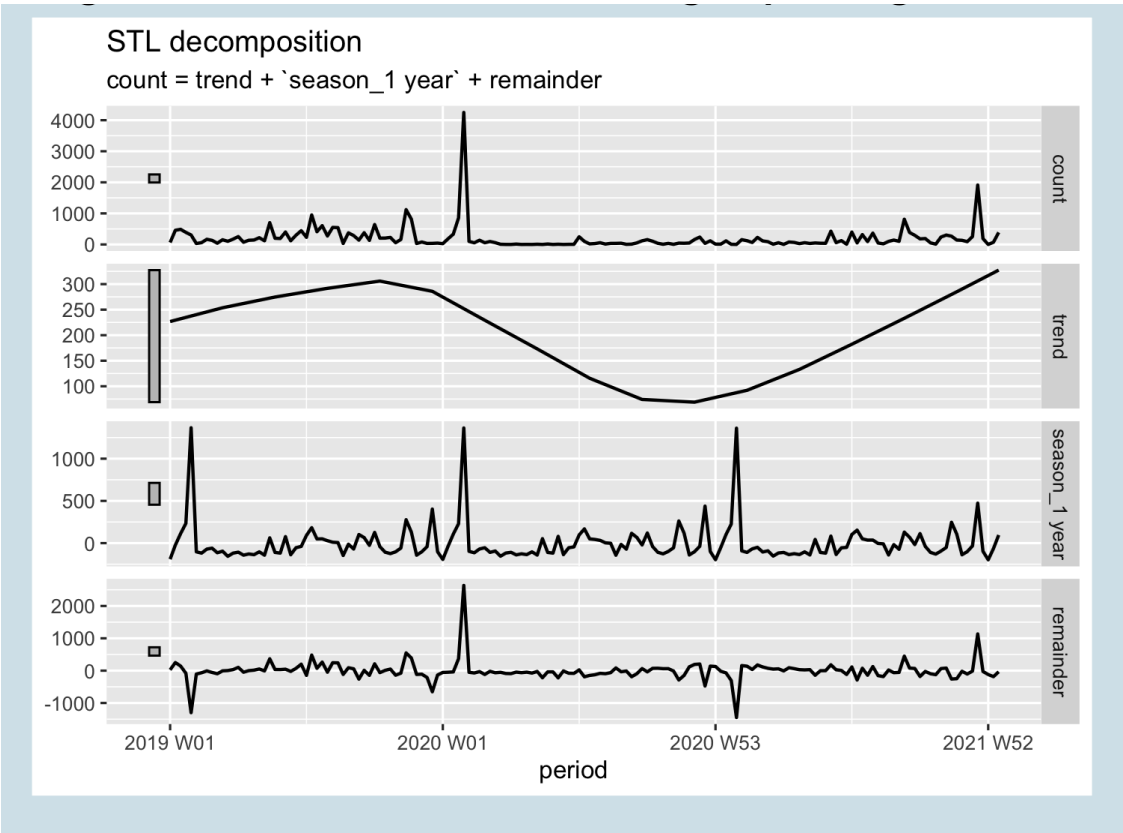

Figure A1. Trends in file creation in group storage.