Volume 124, Article No. 124034 (2019) https://doi.org/10.6028/jres.124.034
Journal of Research of the National Institute of Standards and Technology

Laboratory Information Management Systems for Electron Microscopy: Evaluation of the 4CeeD Data Curation Platform

June W. Lau, Rachel F. Devers, Marcus Newrock, and Gretchen Greene

National Institute of Standards and Technology,
Gaithersburg, MD 20899, USA

june.lau@nist.gov
rach.devers@gmail.com
marcus.newrock@nist.gov
gretchen.greene@nist.gov

An evaluation of the feasibility and the requirements associated with a facility-wide deployment of a laboratory information management system (LIMS) at an electron microscopy facility was conducted. 4CeeD, an open-source LIMS, was selected for the focus study. This report summarizes data infrastructure prerequisites, critical and desirable features, and lessons learned from using and interacting with 4CeeD, and broader LIMS adoption recommendations for this facility.

Key words: data management; laboratory information management system; microscopy data.

Accepted: November 15, 2019

Published: December 17, 2019

https://doi.org/10.6028/jres.124.034

1. Facility and Project Background

Electron microscopes (EMs) are cross-disciplinary tools useful for the examination of a myriad of specimen types, such as metals, polymers, and biological materials. In addition to images, EM can also produce spectroscopy data, hyperspectral images (where each pixel consists of a unique spectrum), multimodal images (images derived from unique signal sources from the same analysis area), and hyperdimensional images (where each image pixel embeds higher-dimensional information). EM data are also diverse in size and acquisition rate; they range typically from megabytes to terabytes and can be acquired at rates up to about 0.5 TB/s. Such a class of instrumentation provides a rich backdrop of sample and experimental conditions, user behaviors, and data practices, and consequently, a good proving ground for research data management. Currently, each step of the data path from our EM instruments, to our researchers, to publication, and to the archive, results in increased dispersion of data, metadata, details about the sample, and experimental intent along the way, and this seems to be a common issue in the broader EM community. This work on a laboratory information management system (LIMS) was motivated by the recognition that much value can be gained in the EM data life cycle with the proper data infrastructure in place, and that there are emergent technologies and philosophies that can support this paradigm change.

How to cite this article:
Lau JW, Devers RF, Newrock M, Greene G (2019) Laboratory Information Management Systems for Electron Microscopy: Evaluation of the 4CeeD Data Curation Platform. *J Res Natl Inst Stan* 124:124034. https://doi.org/10.6028/jres.124.034

In the National Institute of Standards and Technology (NIST) Material Measurement Laboratory's Materials Science and Engineering Division, the electron microscopy facility (EM Nexus) consists of three scanning electron microscopes (SEMs) and four transmission electron microscopes (TEMs), as well as two light optical microscopes. Figure 1 is a schematic representation of data flow within the Nexus today. Of the seven electron microscopes, two instruments are connected to a NIST research equipment network (REN), four are connected to an older local area network (LAN) network, and one relies on "sneakernet," *i.e.*, making physical copies with a removable drive. For the networked instruments, post-instrument acquired data are stored on a central file server.

Regardless of the data path, the terminal point of most EM data is usually on the researchers' office personal computers (PCs). Several emergent trends are providing momentum for a comprehensive data reform. The first issue is the immediate urgency for finding and organizing data. When postdoctoral and guest researchers leave, their data are often left behind in hard drives. These hard drives have been found in drawers and shelves, sometimes unlabeled, sometimes forgotten, but almost never reused. The second factor is the "reproducibility crisis" in scientific research, where substantial portions of research across various disciplines have not been successfully replicated [1–3]. Finally, there is the opportunity to capitalize on the burgeoning field of machine learning (ML) and artificial intelligence (AI). These techniques show tremendous promise to speed discovery by eliminating rote procedural work. However, the full potential of ML has yet to be realized in EM, because many of the requisite elements in the data ecosystem, such as collocation of data sets, digital records of sample information, and experiment intent, do not exist today. The FAIR data principles (findable, accessible, interoperable, reusable) [4–5], which are a conceptual framework designed to alleviate the collective data crises in retrieval, preservation, reproducibility, and accessibility, are a key motivator for this activity. This work set out to examine the characteristics of a LIMS that can successfully encompass and integrate the typical workflow in the NIST EM Nexus.

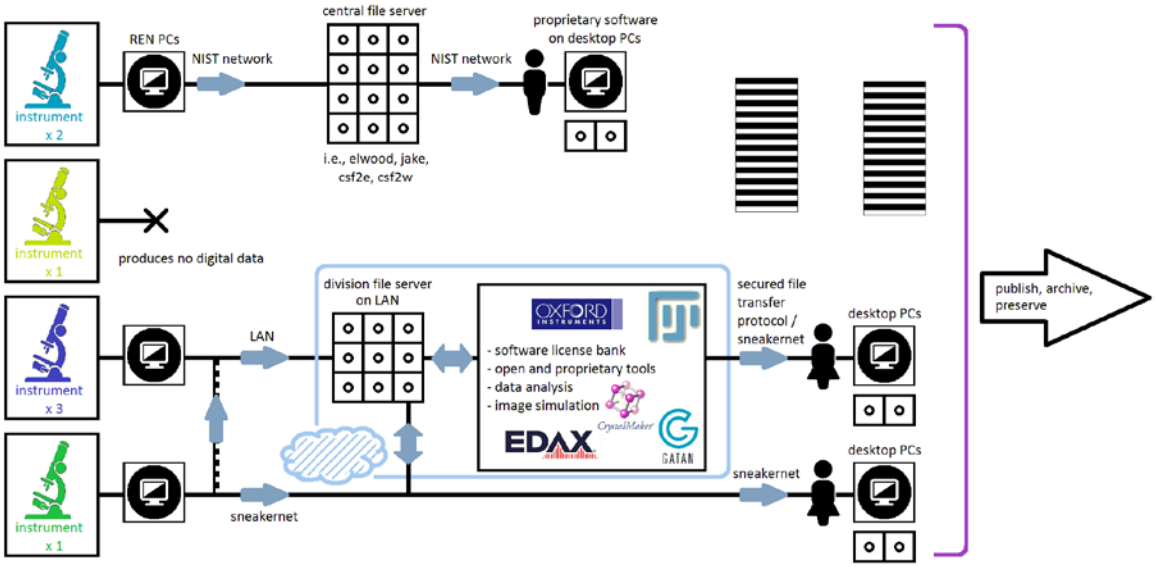

Fig. 1. Data flow of the EM Nexus today.

2. Specifying the Requirements

We began this project with the goal of evaluating existing LIMSs that might work well with EM data, and implementing internal infrastructural changes required for the adoption of an acceptable system. Since a key requirement is for the LIMS to be network-deployable, it is helpful to simultaneously evaluate both the network needs and the LIMS requirements. Additionally, because different types of users would likely interact with the LIMS differently due to differences in purpose and tenure durations, the categories of users considered included facility users, facility managers, project managers, postdoctoral researchers, and guest researchers. Further along through the deployment and testing cycle, we realized that it would be helpful to distinguish between user requirements and data management requirements. This means being deliberate about the intended purpose of each LIMS subcomponent. For example, can the LIMS simultaneously display a limited number of metadata fields of interest to the researcher and keep the full, Application programming interface (API)-searchable metadata record in the back-end? Table 1 is an accumulated list of the EM Nexus LIMS requirements, compiled through the life of this project. It attempts to group the requirements into three broad requirement categories: LIMS and network infrastructure, user capabilities, and data management. Many of the requirements, primarily those marked by (–) symbols, were added after testing with 4CeeD, one of the open-source LIMS products. It is worth pointing out that 4CeeD has several important security features that are likely to be critical under different use cases. For example, encrypted data transmission from instrument to user-access end point (PC, mobile) would be critical for data ecosystems that are not entirely contained within a firewall, as it was in our case. Another is the restriction on the directionality of data flow to arrest malware transmission to instrument controller PCs; at NIST, the REN serves this function.

3. Updating the Instrument Network

During our LIMS evaluation, we learned that certain network characteristics are highly desirable (see Table 1, subsection on network backbone for LIMS). At present, our instrument data acquisition (DAQ) control PCs are located on two different, networks (LAN and REN). DAQ PCs tend to be older, and they tend to be running legacy operating systems (OSs) that do not support newer security features. Furthermore, older OS versions have unpatched security vulnerabilities, many of which can be exploited. These protected instrument networks are meant to protect both the greater information technology (IT) network and the vulnerable DAQ PCs by isolating the DAQ PCs Additionally, these PCs are configured to guarantee proper instrument functionality by the vendor, and these configuration states are not necessarily consistent or compatible with both OS and network security requirements. To harness the full usefulness of an LIMS, our microscopy data infrastructure required several crucial updates. Historically, when we only had a few microscopes, we had used a LAN as our instrument network. It would be very difficult to scale up a LAN for storage capacity and backup for raw data, derived data, processed data, and potentially terabyte-scale data sets from high data-production-rate experiments. Instead, we can take advantage of the existing NIST-level central file server by migrating DAQ PCs to the REN. The REN–central file server combination satisfies the list of requirements in Table 1 under “network backbone.” As a result, we have recently completed the migration of the MMF–“sneakernet” hybrid instrument (Fig. 1, green icon) to the REN.

Table 1. List of the LIMS requirements for EM Nexus in three broad requirements categories: LIMS and network infrastructure, user capabilities, and data management. (+) symbols signify items satisfied by 4CeeD, (o) symbols signify items partially satisfied by 4CeeD, (–) symbols signify items not yet satisfied by 4CeeD at the time of product evaluation.

1. LIMS and Network Infrastructure Requirements		
LIMS General	+	File-server mountable
	+	Deployable as a single instance as a shared resource (so that multiple instances do not need to be updated and managed)
	+	Able to reconcile data from different instruments
	+	Able to support metadata extraction from open and proprietary data formats
	+	Customizable access control for individuals and user groups
	o	Able to support import and export of open data formats
	o	Able to offer feasible path for long-term extraction service and ability to manage extraction tools for new detector/upgrades/new capabilities
	–	Able to support automated data and metadata harvesting (from file servers for data, and from electronic lab notebooks (ELN), tool reservation/management systems, user calendar for metadata)
	–	Able to offer feasible path for archival capability
	–	Able to incorporate file metadata extractors that can be broadly reused or shared with other LIMSs
	–	Provide meaningful bundling of data and metadata from multiple sources
	–	Desirable to authenticate through NIST's active directory
Network Backbone for LIMS	+	Allow collocation of data from different instruments
	+	Support connection to multiple instruments in multiple buildings, and connection from different desktops from multiple locations
	+	Scalable for addition of instruments and servers to the LIMS area network
	+	Support an on-network computation environment so that raw data, derived data, processed data, and potentially terabyte-scale data sets from high data-production-rate experiments do not have to be downloaded to local machines
	+	Provide flexibility in allowing updates to operating system (OS) for machines in network
	+	Support legacy OS crucial for instrument operations
	+	Support flexible user policy
2. User Capabilities		
Data Files	+	Provide ease of access and management
	o	Facilitate search and discovery
	–	Enable uploading data sets from older or completed projects
Project Organization	+	Enable hierarchical views of data collection
	+	Enable collaborative access (model for internal and external collaborators)
	+	Have Metadata views
User Experience	+	Require minimal user intervention
	+	Support different user types (facility user, facility manager, project manager, postdocs, guest researchers)
	+	Support local file export
	+	Support customizable metadata extraction
	–	Workflow and interface customization
Data Analysis	+	Provide access of data from local work station
	+	Compatible with cloud-based data analysis
	o	Promote remote analysis/hot-linking to analysis environment
	–	Retain analysis codes used
3. Data Management Requirements		
	+	Manage raw data files
	+	Manage hierarchical data structures
	+	Compatible with large (TB) data sets
	–	Support automated data harvesting from instrument computer or network drive
	–	Track provenance for derived/processed data sets
	–	Have well-defined archival strategy (LIMS shall be compatible with preservation and archival for a predefined time period)

4. 4CeeD as a LIMS Candidate

At the onset of this project, three open-source LIMS options were presented: Hyperthought [6] (formerly, ICE) from the Air Force Research Laboratory, 4CeeD [7] (and its underlying infrastructure, Clowder [8]) from University of Illinois–Urbana Champaign, and an LIMS developed specifically by and for the Materials and Chemical Science and Technology directorate at the National Renewable Energy Laboratory [9]. After a rapid trial period of several months, 4CeeD was down-selected as the pilot system because its functionalities were most closely matched with the data needs of electron microscopy for materials science, and because a clone deployment is straightforward. Our instance of 4CeeD was deployed on a networked virtual machine. A user would open 4CeeD and perform authentication through a web browser, and then, ideally (more on this in Sec. 4.5), the transfer of microscope data can be a simple click and drop operation from the microscope PC to the web browser. 4CeeD then interacts with a data set by automatically generating image thumbnails and extracting its metadata. During the trial period of approximately one year, the EM Nexus coordinated data-upload sessions, partially reorganized the flow of data from selected instruments, and focused on two proprietary file formats for metadata extraction. These tasks demonstrated the feasibility of deploying a LIMS at a small facility.

The 4CeeD project is funded through the National Science Foundation Data Infrastructure Building Blocks program [10], and it is conducted through the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign [11]. Detailed information about the software, and a link to the code itself can be found at the 4CeeD GitHub page [12]. Features we found most useful for the EM Nexus were: customizable open-source metadata extraction, image thumbnails of proprietary format files, hierarchical project views, data sharing, and access control with collaborators. Additional requirements not within the scope of 4CeeD include automated harvesting (ingesting), and provenance tracking of derived data. We discuss these components and other considerations in depth below.

4.1 Display of Proprietary Files, in Thumbnail Previews and Page Views

From a user-experience perspective, the standard file browser in the OS is often the first point of entry for the screening and viewing of acquired data. After the file name, the next clue to the content of a file is the file extension. Many SEM images are stored in the TIFF format, which many file browsers support, rendering the content as a thumbnail image. Many TEM images and spectroscopy data, however, are held in formats, sometimes proprietary, that the file browsers do not render. Therefore, an important task for an LIMS is to generate thumbnails and/or full displays to facilitate data browsing for the researcher. 4CeeD produces image previews for many commonly used microscopy file formats using a default image preview extractor. This extractor parses encoded image information contained within proprietary image files and renders the preview and thumbnail in a common image file format. This allows the image thumbnail and preview to be displayed in a web browser, without the use of dedicated software. Figure 2 shows a side-by-side comparison of file browsing with Windows 7[1] and with 4CeeD.

[1] Certain commercial equipment, instruments, or materials are identified in this paper to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

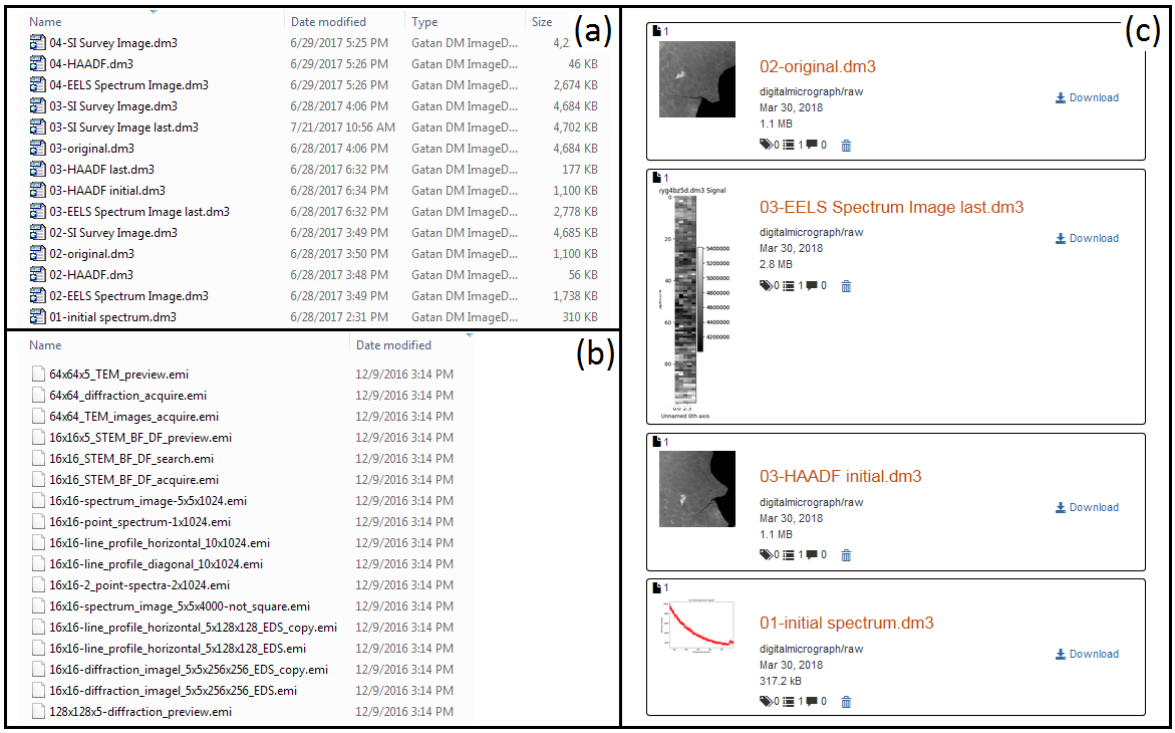

Fig. 2. Data browsing experience with (a, b) Windows 7 and (c) 4CeeD. A copy of Digital Micrograph was installed on this researcher's desktop computer, and, therefore, (a) the OS associated .dm3 files as Gatan DM Image Documents, displayed the Digital Micrograph icons, and retrieved information on the file size. However, the Tecnai Imaging & Analysis software was not installed on this computer, (b) so the only clues about these files are the modification dates, the file names, and the .emi file extensions. In contrast, with 4CeeD, (c) a researcher has immediate access to file names, dates, sizes, types, and image thumbnails through a web browser. If this was a shared data set, an in-browser download of the shared data (individually or by batch) could be initiated.

4.2 Customizable Open-Source Metadata Extraction

Metadata are embedded in each data file, and they contain pertinent information about the experimental conditions. These metadata can include, but are not limited to, specimen information, instrument operator, optical and mechanical configuration of the instrument, the calibration file(s) used, and the detectors in use and their configurations. However, these metadata are often not searchable, and they are viewable only if the proprietary software is installed. Because of these barriers, metadata remain an afterthought for the typical researcher. Here, we have an opportunity to promote metadata to the forefront, as an enabling agent for FAIR data. We developed several open-source Python 3–based [13] custom metadata extraction scripts using the HyperSpy [14] library (and the native loading and export functionalities within) for the Gatan[1] image formats (.dm3 and .dm4) and the Tecnai Imaging Analysis/Emispec[1] data formats (.ser and .emi). The extraction outputs are .txt files, which may be indexed and searchable by the OS. Much of the metadata output, however, is designed to be machine-readable rather than for human inspection. These outputs can be quite extensive (*e.g.*, one file's metadata occupied 44 printed pages with an 8 point font) and offer little satisfaction in terms of the human experience. To solve this problem, our metadata extractors can be integrated to work in tandem with the extensible 4CeeD framework. 4CeeD is deployed as a group of services, and each service runs as a separate Docker[1] [15] container. The flexible nature of the Docker platform enables the deployment of customized instances of 4CeeD; *i.e.*, we built a separate Docker image for our custom extractor and ran it in parallel with the native 4CeeD containers, which consist of the default extractors and framework services. Such a process allows

for the seamless integration of our custom extractor with the native 4CeeD platform. This is the key enabling feature for customized metadata views with proprietary image previews in a single web-browser window, which provides meaningful collocation of data with human-readable metadata. Figure 3 shows the 4CeeD rendering of a custom metadata structure tree; the tree form was borrowed directly from the Gatan hierarchy (*e.g.*, microscope info, image data, image tags, annotation group list, and image display info). In principle, metadata can be sorted into different tree structures, and displays can be customized for individual researchers.

In addition to the metadata extractors that we developed, the 4CeeD platform natively supports the automatic extraction of metadata from a wide range of detectors. However, metadata pulled from detectors generally do not describe broader experimental context, such as information about the sample (*e.g.*, origin and history, preparation steps), experimental intent, analysis rationale, codes used for analysis, and the researcher's interpretation. Meeting the reusability criterion within FAIR without inclusion of contextual metadata would be difficult for any LIMS.

Metadata extractors can be costly for one person or even one organization to develop. Electron microscopes from different vendors are supported by dissimilar software environments, which are often proprietary. Each microscope can have several cameras and spectrometers made by different manufacturers, with vastly dissimilar data and metadata outputs. Factors compounding the problem include the periodic software updates and the lack of agreement in metadata vocabulary and header standards. Therefore, we recommend adopting an extensible crowd-sourced approach to metadata extraction that facilitates researchers writing their own extractor services as needed for their EM equipment. Furthermore, a community extractor repository should be an integral part of a healthy LIMS ecosystem.

4.3 Hierarchical Project Viewing, Sharing, and Distribution

In 4CeeD, the highest-level data container is a *space*. Appropriate headings for a *space* might be "all of my work," "facility manager overview," or "additive manufacturing collaboration." Below a *space*, there is a *collection*. *Collections* serve the same function as folders in our desktop PCs, and *collections* may be nested indefinitely. The lowest-level container is the *dataset*. All files go into *datasets*; *datasets* are terminal and thus cannot be nested. 4CeeD organizes hierarchical data with *spaces*, *collections*, and *datasets*. For example, in a typical *space*, there might be a *collection* called "additive manufacturing (AM)," which might have two *subcollections*: "metal AM" and "polymer AM." Within "metal AM," there might additional *subcollections* called "17-4 Fe," "18-8 Fe," and so on. Finally, let us suppose a researcher has a collection with three *datasets*: "17-4 untreated," "17-4 500 C," and "17-4 1000 C." A benefit of hierarchical organization is that it is easy to associate related data, in this case, indicating different temperature treatments.

Another benefit of hierarchical organization is that it also makes data sharing straightforward. Continuing from the above example, suppose Coworker X is looking at Fe alloys, and Coworker Y is looking at Co alloys produced by the same process, and the two coworkers wish to compare their *datasets*. The two coworkers can create a new shared space called "additive manufacturing collaboration" in which Coworker X shares everything under the *collection* for "metal AM," but not "polymer AM," and Coworker Y shares everything under the Co alloy *collection*. In this configuration, not only can the two coworkers freely browse each other's contribution to the shared space, but they can also freely download anything within the shared space with the right permission settings. This eliminates the need to email or make physical copies of data sets between two collaborators on the same instance of 4CeeD.

One feature of automation not yet supported by 4CeeD that would go a long way towards fulfilling the interoperable requirement of FAIR is the ability to download stored data sets in an open format such as the hierarchical data format version 5 (HDF5). In response, we developed a Python shell script outside of the 4CeeD framework that can convert .dm3, .dm4, .emi, .ser, and .mcr format files to HyperSpy's HDF5 format.

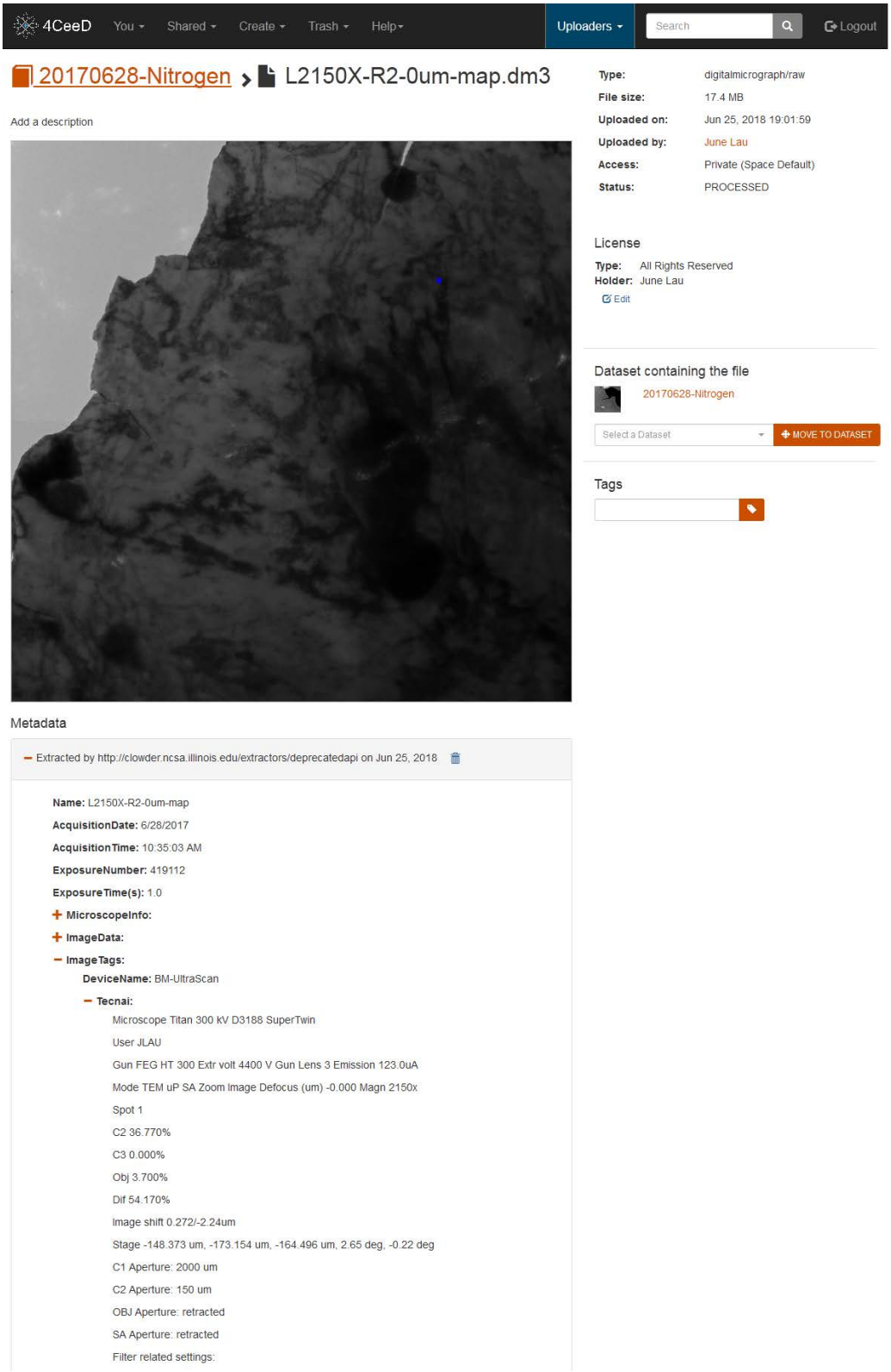

Fig. 3. Image saved in proprietary format accompanied by customized metadata view.

4.4 Searching and Tagging

4CeeD has a *search* box as a part of the user interface. While this search box performs as expected when searching through names of *files*, *datasets*, and *collections*, it does not search through metadata extracted natively or through custom means. The interface has a *tag* feature that allows researchers to tag whole images (rendered by 4CeeD) or specific features within an image. The *tag* feature has great potential because these features are postacquisition, human-annotated metadata fields. Researchers can label common features within an image data set or collection and make those features human and machine searchable, but this feature is also not yet supported by *search*. In addition to searchable metadata and tags, a faceted (*i.e.*, multiple filter selections) search capability would also be a highly desirable feature.

4.5 Harvesting

The EM Nexus is atypical in that it is operated as a cooperative facility, as opposed to the more familiar cost-recovery model typical for facilities. This mode of operation brings unique challenges because not all researchers are equally likely to comply with top-down data management policies. Therefore, we intended that an LIMS for the Nexus shall require minimal researchers' input and time commitment yet deliver tangible benefits and improve productivity. To achieve this goal, we embraced automated processes such as metadata extraction and harvesting (or ingestion) into a searchable database. While 4CeeD excels at handling the automatic extraction of images and spectroscopic metadata, human intervention (as of version 17.11) is required for the initial data ingest into 4CeeD. For 4CeeD to pull metadata from a file, the researcher must first manually upload the data set to 4CeeD as a drag-and-drop operation through a web browser. This is particularly problematic because many DAQ PCs run legacy OSs and cannot support newer web browsers. For example, such a drag-and-drop operation failed on an instrument running Windows XP[1]. We deployed Microsoft SyncToy[1] [16] as a work-around for this instrument.

The *Zip Uploader* is a utility within 4CeeD that enables manual batch file upload. This is a promising feature, and it becomes important should a researcher wish to upload data sets from previously completed projects. The *Zip Uploader* accepts a .zip file generated by the researcher. When uploading the .zip file, all file folder structure and substructures are replicated, and so the data hierarchy from the source is preserved. However, as of this software version, 4CeeD does not render thumbnails or page view images, nor does it extract metadata for files uploaded through the *Zip Uploader*.

4.6 Provenance

Microscopy data can sometimes generate multigenerational daughter (derived) data, and therefore provenance tracking is a crucial LIMS feature. Consider, for example, an original image called *Eve*. Coworkers X and Y are on a project together, and they wish to evaluate *Eve* separately. Coworker X creates a new daughter image (*Eve_FFT*) by applying a fast Fourier transform (FFT) to Eve. Next, Coworker X picks three different vector decompositions of *Eve_FFT* and creates three daughters for *Eve_FFT* (or granddaughters for *Eve*): *Eve_iFFT_A*, *Eve_iFFT_B*, and *Eve_iFFT_C*, through an inverse FFT operation. Suppose Coworker Y repeats essentially everything Coworker X did, except that the initial FFT was performed on a subregion of *Eve*. In Coworker Y's workflow, she names *Eve's* daughter *Eve_FFTpartial*, and the granddaughters are *Eve_iFFTpartial_A*, *Eve_iFFTpartial_B*, and *Eve_iFFTpartial_C*.

In this example, while 4CeeD accepts daughter *datasets* in a hierarchical structure under individual *spaces*, or a *shared space* between coworkers X and Y, there is no formal mechanism to link branched lineages (*i.e.*, *Eve_iFFT_C* and *Eve_iFFTpartial_C* share the same grandparent), or even a mechanism to document intrabranch (*Eve_iFFT_C* is the inverse FFT of the [220] vector from *Eve_FFT*) and interbranch (*Eve_FFTpartial* comes from a 512 × 512 pixel section of *Eve*, whereas *Eve_FFT* originated from the full-size, 2048 × 2048 *Eve*) relationships. Furthermore, under the 4CeeD framework, manual entries to the *comments* field are the only way we found to track the codes or operations used to produce the daughter

data. Tracking becomes even more challenging if the operations were performed with different software on different data storage locations (*e.g.*, Coworker X used Digital Micrograph on a desktop, and Coworker Y used HyperSpy through a Jupyter notebook interfacing with a network drive).

4.7 4CeeD Architecture Considerations

The 4CeeD application relies on Docker containers for system configuration. Docker is a lightweight software virtualization system used for modularizing applications and services on multiple platforms through *containers*. In addition to the image and metadata extractors mentioned previously, Docker containers also support uploads, databases, web interfaces, and task queuing. While it was recommended that Kubernetes [17], a container orchestration system for scaling and redundancy, be used with deployment, we deployed a stand-alone 4CeeD with container orchestration performed by Docker Compose instead. Docker Compose is listed as an acceptable alternative to Kubernetes by the developers, and our evaluation was shaped in part by this experience. The Docker Compose deployment revealed that some services, such as task queuing for uploads (freezes), and the HyperSpy container for generating thumbnails (consistent rendering required frequent Docker container redeployment), were vulnerable to instability.

Our experience with the 4CeeD deployment also provided several lessons on data integrity when the data and the application were collocated on the same server. 4CeeD uses MongoDB [18], a NoSQL database that is installed as a service in Docker, to store data and metadata, and the proper configuration of the MongoDB container is critically important. Orchestration tools such as Docker Compose are necessary to designate local volumes for 4CeeD. Without assigning volumes for MongoDB contents, application restarts have led to the loss of all uploaded data. Fortunately, this problem can be mitigated by storing data on a central file server rather than collocating data on a MongoDB instance in a virtual machine. 4CeeD provides the flexibility to link entries to a central file server location, which is our preferred mode of data storage. This allows researchers access to their data both within 4CeeD and from their own workstations.

5. Unsolved Issues

Microscopy data rates have ballooned with the availability of hyperdimensional and direct electron detectors. It is now possible to buy a detector that generates data at a rate of 0.5 TB/s. How do we make unprocessed raw data useful when they are being generated at these rates? New data management paradigms to treat raw data at the point of generation will likely be required.

Data acquired at terabyte-scale rates lead, of course, to terabyte-scale data sets. How might we effectively couple near-term activity with long-term data curation, considering the typical research life cycle? Specifically, how might an LIMS manage TB data sets after capture, and what happens to those data sets as they age out of the research life cycle?

At the onset of this LIMS pilot, we did not have an effective method of handling the large volume of postprocessed/analysis data that can sometimes accompany EM data, both in terms of deciding what to store, and assigning provenance. Towards the end of this study, some concepts in this area are beginning to emerge. For example, some researchers can now fully document all analysis steps taken to produce the results in the form of a Jupyter notebook (source code) that would only require the original raw data (downloadable) to reconstruct. This is a concept that the 4CeeD developers fully appreciate and are planning to support in a later version.

The EM data stream is also becoming more heterogeneous. In addition to traditional detectors, newer specimen holders, and various environmental sensors may also be outputting data related to an experimental session. The future LIMS must be able to intelligibly reconcile and seamlessly merge these disparate data streams. Finally, ML and AI are increasingly interwoven into the fabric of scientific research. The future LIMS will need to integrate well with new ML/AI libraries and tools that are dynamic and scalable in order to enable deeper data exploration.

6. Conclusions

The ideal LIMS should be a software solution that can view, store, and/or dynamically link data and metadata associated with all steps of data generation with contextual information in ways that satisfy the FAIR data principles. We conducted a needs assessment, an analysis of tools and infrastructure in place, and the data system gaps for an electron microscopy facility (the EM Nexus) within NIST. Following this assessment, we attempted to implement 4CeeD in our facility. We conclude that 4CeeD is an excellent first attempt for a microscopy LIMS tool. There are limitations, but there are also features of great value to microscopists and collaborators.

In addition to our specific experiences with 4CeeD, we have general observations about the LIMS and its place in the broader microscopy data life cycle. Considering currently available open-source LIMS options: Each performs somewhat different functions and excels at different tasks, but none provides an end-to-end solution. Therefore, we advocate that an LIMS should be modular and designed to integrate well with other data management components within the EM data ecosystem. An EM-compatible LIMS must also take into consideration the available data transfer infrastructure, and the size and storage duration of data and derived data products. For example, if the end strategy is to store, analyze, and archive on the cloud, the bandwidth and/or latency of the network connections may not be compatible with high data-production-rate (TB/s) data generation. Finally, we advocate that the LIMS architecture should exhibit full duality in supporting human and machine-readable contents because these two manifestations generally serve different end goals.

4CeeD is presently deployed internally at NIST. NIST staff and associates may contact the corresponding author for additional details about the test deployment.

Acknowledgments

We are grateful to Zach Trautt for assistance in the early phase of this project, Ray Plante for help with piecing together the microscopy data ecosystem, Will Osborn for helping us think through our data network, Andy Herzing and Mike Katz for uploading data and debugging the 4CeeD workflow, John Bonevich for building, maintaining, and providing key insights on the MMF network, and Klara Nahrstedt , Steve Konstanty, Todd Nicholson, Phuong Nguyen, and Tarek Elgamal for the valuable assistance with and support for 4CeeD. Additionally, we are grateful for the continuing support of NIST's Office of Data and Informatics, student funding support from the Montgomery County Community College internship program, and the NIST Summer Undergraduate Research Fellowship (SURF) program.

7. References

[1] Begley CG (2015) Improving the standard for basic and preclinical research. *Circulation Research* 116(1):116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819
[2] Baker M (2016) 1,500 scientists lift the lid on reproducibility. *Nature* 533(7604):452–454. https://doi.org/10.1038/533452a
[3] National Academies of Sciences, Engineering, and Medicine (2016) *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop* (The National Academies Press, Washington, DC). https://doi.org/10.17226/21915
[4] *FORCE 11*. Available at: https://web.archive.org/web/20180919134552/https://www.force11.org/fairprinciples, date accessed: September 19, 2019.
[5] Wilkinson MD (2016) The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3:160018. https://doi.org/10.1038/sdata.2016.18
[6] *HYPERthought.* Available at: https://web.archive.org/web/20180919135439/https://www.icemaker.afrlmakerhub.com/login/, date accessed: September 19, 2019.
[7] 4CeeD stands for Capture, Curate, Coordinate, Correlate, and Distribute Your Data and is a product of the Timely and Trusted Curation and Coordination (T2C2) project at the University of Illinois–Urbana Champaign, funded by the National Science Foundation's Data Intensive Building Blocks (DIBBs) program, K. Nahrstedt, principal investigator.
[8] Marini L, Gutierrez-Polo I, Kooper R, Satheesan SP, Burnette M, Lee J, Nicholson T, Zhao Y, McHenry K (2018) Clowder: Open source data management for long tail data. *Proceedings of the Practice and Experience on Advanced Research Computing (PEARC '18)* (ACM, New York), Article 40, 8 p. https://doi.org/10.1145/3219104.3219159

[9] *Handling Large and Complex Data in a Photovoltaic Research Institution Using a Custom Laboratory Information Management System*. Available at: https://web.archive.org/web/20180919140040/https://arxiv.org/ftp/arxiv/papers/1403/1403.2656.pdf, Date accessed: September 19, 2019.
[10] National Science Foundation Division of Advanced Cyberinfrastructure (ACI) 1443013, project title "CIF21 DIBBs: T2-C2: Timely and Trusted Curator and Coordinator Data Building Blocks."
[11] *T2C2: Timely and Trusted Curation and Coordination Primary Navigation*. Available at: https://web.archive.org/web/20180919134926/http://t2c2.csl.illinois.edu/, date accessed: September 19, 2019.
[12] *4CeeD: Capture, Curate, Coordinate, Correlate, and Distribute Your Data*. Available at: https://web.archive.org/web/20180919141159/https://4ceed.github.io/, date accessed: September 19, 2019.
[13] *Python*. Available at: https://web.archive.org/web/20180919141358/https://www.python.org/, date accessed: September 19, 2019.
[14] *HyperSpy Multi-Dimensional Data Analysis*. Available at: https://web.archive.org/web/20180919141529/http://hyperspy.org/#, date accessed: September 19, 2019.
[15] *Docker*. Available at: https://web.archive.org/web/20180919141703/https://www.docker.com/, date accessed: September 19, 2019.
[16] *SyncToy 2.1*. Available at: https://web.archive.org/web/20180919141800/https://www.microsoft.com/en-us/download/details.aspx?id=15155%20, date accessed: September 19, 2019.
[17] *Kubernetes: Production-Grade Container Orchestration*. Available at: https://web.archive.org/web/20180919142009/https://kubernetes.io/, date accessed: September 19, 2019.
[18] *MongoDB Atlas Database as a Service*. Available at: https://web.archive.org/web/20180919134023/https://www.mongodb.com/, date accessed: September 19, 2019.

About the authors: *June W. Lau is a physicist in the Materials Science and Engineering Division at NIST, and she manages the EM Nexus, an electron microscopy facility in that division. Rachel F. Devers was a full-time undergraduate at the University of Maryland, and she was a student intern with the Materials Science and Engineering Division at NIST. Marcus Newrock is a computer scientist in the Office of Data and Informatics. Marcus performs operations support for informatics infrastructure and scientific resources. Gretchen Greene is a group lead for data science in the Office of Data and Informatics at NIST. Gretchen leads scientific engagement projects for MML research data management and NIST open access to research data infrastructure with focus on data dissemination and public access. The National Institute of Standards and Technology is an agency of the U.S. Department of Commerce.*