


Original Article

NexusLIMS: A Laboratory Information Management System for Shared-Use Electron Microscopy Facilities

Joshua A. Taillon^{1*} , Thomas F. Bina^{2,3,†}, Raymond L. Plante³, Marcus W. Newrock³, Gretchen R. Greene³ and June W. Lau²

¹Office of Data and Informatics, Material Measurement Laboratory, National Institute of Standards and Technology, Boulder, CO 80305, USA; ²Materials Science and Engineering Division, Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA and ³Office of Data and Informatics, Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

Abstract

This work introduces NexusLIMS, an electron microscopy laboratory information management system designed and implemented by the Office of Data and Informatics and the Materials Science and Engineering Division at NIST for a multi-user electron microscopy co-op facility. NexusLIMS comprises network infrastructure, shared information technology resources, a custom software package to harvest and extract experimental information and construct experimental metadata records, and an intuitive web-based user-facing interface for searching, browsing, and examining research data. These metadata records conform to the *Nexus Experiment* schema, which is introduced in this work. The NexusLIMS suite of tools requires minimal input and adjustments to user behavior, instead relying on existing organizational procedures and the collection of information from a multitude of sources to construct a complete picture and record of a research experiment. The underlying infrastructure and design considerations for a multi-user data management system are discussed. The core codebase for NexusLIMS is made publicly available alongside this work, and its modular design encourages the adaptation of the presented methods at other research organizations.

Key words: data management, laboratory information management, metadata extraction, open source software, user facility, electron microscopy

(Received 28 October 2020; revised 31 December 2020; accepted 25 February 2021)

Introduction

Many challenges currently face scientists and research facility managers. Perhaps most prominent among these is the management and processing of laboratory information and the vast amounts of research data produced on modern scientific instrumentation. Electron microscopy (EM) researchers produce data using a myriad of instruments, and this data are commonly analyzed using expensive and proprietary software (typically with restrictive licensing terms). Critically, even if users have licensed access to these software packages on their personal workstations, the data produced (and associated instrumental/experimental metadata) can frequently be viewed *only* using these commercial software packages [although modern open source projects such as HyperSpy are alleviating some of this burden (de la Pena et al., 2017)]. Users are often forced to implement their own strategies to curate their personal research data, relying on basic file metadata, naming conventions, and notes/memory to identify the

significance of each dataset. Individual strategies become incompatible when collaborating with other researchers or over long timespans, leading to “abandoned” datasets that are forgotten and lost after a manuscript has been published or a researcher has continued on to another position. Like many research institutions, certain research areas at the National Institute of Standards and Technology (NIST) suffer from a lack of centralized and automated data management, leading to lost productivity, unnecessary replication of experiments, and limited experimental reproducibility.

To address these challenges, the NIST EM Nexus and the NIST Office of Data and Informatics have recently co-developed NexusLIMS, a laboratory data management system (LIMS; Gibbon, 1996) for EM data in the discipline of materials science. The EM Nexus is a shared-use instrument co-op within NIST; each staff member responsible for a particular instrument can agree to share their instrument time with the EM Nexus user group in exchange for centralized data management, access to the other instruments within the facility, and a centralized scheduling platform. Such a model facilitates collaboration between the EM researchers at NIST, as well as sharing of certain costs and funding opportunities. The methods discussed in this article are expected to be broadly applicable at other institutions, but due to the specifics of the NIST network configuration, it will be

*Author for correspondence: Joshua A. Taillon, E-mail: joshua.taillon@nist.gov

†Current address: Department of Biomedical Engineering, Columbia University, New York, NY, USA.

Cite this article: Taillon JA, Bina TF, Plante RL, Newrock MW, Greene GR, Lau JW (2021) NexusLIMS: A Laboratory Information Management System for Shared-Use Electron Microscopy Facilities. *Microsc Microanal*. doi:10.1017/S1431927621000222

© National Institute of Standards and Technology 2021 outside of the United States of America. As a work owned by the United States Government, this Contribution is not subject to copyright within the United States. Outside of the United States, Cambridge University Press is the nonexclusively licensed publisher of the Contribution. Published by Cambridge University Press on behalf of the Microscopy Society of America

most directly applicable to those at facilities with stringent networking and firewall configurations. This article discusses the design philosophies, the LIMS components, deployment considerations, preliminary observations, and lessons learned from the implementation of NexusLIMS.

It should be noted that the concept of LIMS is a well-established one (Gibbon, 1996), and numerous mature commercial and open source projects exist to support both highly specific and general laboratory data workflows (Cheung et al., 2009; Blaiszik et al., 2016; Carey et al., 2016; Jacobsen et al., 2016; Arkilic et al., 2017; CARPi et al., 2017; Nguyen et al., 2017; Zakutayev et al., 2018; Abbott Laboratories, 2020; Bika Lab Systems, 2020; Dataworks Development, Inc., 2020),¹ many of which were marshaled as part of the materials data infrastructure efforts of the Materials Genome Initiative (MGI; Warren & Ward, 2018). As such, the implementation of a LIMS in general is not a novel effort. The unique contribution of this work lies in the development of an extensible LIMS framework focused on materials EM that is modular in nature and adapts to existing user practices without forcing users to modify their behaviors to gain the benefits provided by the system.

Designing a LIMS for EM at NIST

As the reproducibility and integrity of scientific research data has gained more prominence throughout the scientific community, the implementation of a LIMS within the EM Nexus (a small, user-focused community of researchers) quickly became apparent as a unique opportunity to encourage responsible data habits among users by building around existing behaviors, and without enforcing draconian operating procedures. Such an approach has resulted in consistent increases in the number of users and amount of data managed since the project launched because it does not require significant investment from the individual researchers beyond their extant workflows. What follows is a detailed description of NexusLIMS's features, fundamental concepts, design, required infrastructure, and user behavior requirements. Microscope users (i.e., researchers) will likely be most interested in the features of the system, described in this section, as well as the screenshots and discussion of how to use the web interface to find and view the research records generated by the system (the "Accessing Research Records Using the NexusLIMS Web Interface" section). The "Components of the NexusLIMS Backend" section focuses on the mechanisms of how data are harvested from various sources, the infrastructure required, and design considerations, and will be of greater interest to facility managers, developers, and those users with a penchant for the intricacies of data management.

What is NexusLIMS?

NexusLIMS is a data workflow engine that assists in the capture and management of research data and metadata. Separated into two parts, the *backend* automatically captures information about user experiments with few user inputs (i.e., no long forms to complete), while the *frontend* streamlines the ability of users to search,

explore, and access data produced by EM Nexus instruments from any networked device at NIST [including remotely via the virtual private network (VPN)], together with records of each microscopy session. All together, NexusLIMS is a suite of interconnected software platforms, storage systems, servers, and a custom Python package (to assemble the research records), and has become the data management centerpiece for the EM Nexus. NexusLIMS began with the modest goal of helping NIST researchers find and reuse EM data, particularly for data from postdocs and researchers who have moved on. The FAIR data principles (Wilkinson et al., 2016) were used as a guide during the development of NexusLIMS to promote data findability, accessibility, interoperability, and reusability. Furthermore, FAIR data enable scalable machine learning (ML) and artificial intelligence (AI), and NexusLIMS will serve as both a human-accessible and machine-readable clearinghouse for EM data from the Nexus Facility.

Summary of Features

From a user perspective, the most important features of NexusLIMS are those that solve the painful data management challenges commonly faced by electron microscopists. Through extensive engagement with researchers in the Nexus Facility, a set of highly desired features was identified and prioritized for implementation. Chief among these is the automatic backup and archiving of all raw research data collected by instruments in the Nexus EM Facility. As part of the data harvesting process, each file has any readable metadata extracted into a common format text-based file and a preview image is automatically generated. Any data (along with the extracted metadata) collected during a user's span of time on an instrument is bundled into a structured text document representing a snapshot of the experiment, which is stored in a curated database. A web-based portal provides access to all of the research records, enabling users to search for prior experiments by date, user, instrument, sample, or any other metadata parameter. They can also view a rich representation of those experiments (including previews and metadata from proprietary data formats) without needing to install any additional software. These research records are created automatically with nearly zero added effort from the researcher, and in this manner augment a user's existing workflow without requiring any modifications to it.

Design Philosophy

In a prior work, Lau et al. (2019) evaluated existing open source LIMS solutions for use at NIST and presented the key LIMS attributes that would serve the needs of the Nexus Facility. In that work, the circumstances leading to the specific implementation choices for NexusLIMS were discussed. The Nexus is an instrument co-op (as opposed to a typical user facility model); it is an at-will arrangement in which each microscope "owner" freely shares their microscope with other co-op members. Thus, the co-op management has limited top-down control of membership data compliance. Any requirement for annotating acquired datasets with detailed instrument configuration, specimen holder used, sample details, the purpose of the experiment, etc., is unenforceable. Some EM Nexus members take impeccable care to document metadata, but such behavior and the methods used are not uniform. When asked, however, most members agreed that an intuitive interface to help them find well-documented old data would be very useful. The core inspiration for the design

¹Certain commercial equipment, instruments, or materials are identified in this presentation 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.

of NexusLIMS thus became: how can this outcome be achieved without enforcing facility-wide behavior change? NexusLIMS must be able to curate and compile research records with little or no input from the researchers.

The importance of harmonizing with the established workflow patterns and behaviors of the research community cannot be overstated, and this has been critical to the success of NexusLIMS. Early software demonstrations of several LIMS-like systems produced user responses that ranged from tepid to hostile (see Lau et al., 2019) because the end benefit was seen as not worth the up-front and ongoing effort, often in the form of meticulous and laborious data entry by the researcher. Instead, this project has focused on devising a LIMS that leverages institutionally supported data infrastructure and preexisting experimental practices. For example, all staff at NIST have access to the extensive Microsoft Office software suite.¹ The Microsoft SharePoint¹ platform is used for overall management of the EM Nexus Facility, which includes spaces for shared documents, safety and operating procedures, and microscope scheduling/reservation. Some Nexus members additionally use Microsoft OneNote¹ as an electronic lab notebook (ELN). Other institutionally provided resources that are leveraged by NexusLIMS include a NIST-wide central file server (CFS) where Nexus raw experimental data and harvested metadata are deposited, and a division-level computational server that hosts the NexusLIMS backend services. A customized instance of the Configurable Data Curation System (CDCS; Dima et al., 2016) (also developed at NIST) is used as the primary front-end interface for NexusLIMS.

A key tenet of the design of NexusLIMS has been the importance of system modularity and scalability. As described in the “What is NexusLIMS?” section, NexusLIMS is comprised of multiple interconnected systems, each responsible for different functionality in the software. By design, any of these pieces can be exchanged for another with minimal effort to allow for use of the overall system design at all types and sizes of institutions. For example, the current storage implementation relies on the NIST-wide CFS for archiving raw data (from which metadata are extracted), but it would be simple to exchange this storage location for any other storage type, provided it is accessible over a network [e.g., a local mounted disk, a remote storage bucket, or a Globus endpoint (Allcock et al., 2005)]. Likewise, the current SharePoint-based instrument scheduling tool could be replaced by any scheduling system with an appropriate application programming interface (API) enabling automated harvesting of reservation information. Such modularity ensures NexusLIMS can remain robust to changes in the underlying infrastructure, promoting a long operational lifespan.

Comparing NexusLIMS to Existing Platforms

Before undertaking a substantial development effort, it was apposite to compare the approach of NexusLIMS with that of other existing LIMS solutions, and in particular to compare with freely available open source packages designed for academic/research use. In the work of Lau et al. (2019), a thorough evaluation of the 4CeeD platform (Nguyen et al., 2017) was performed via pilot deployment, together with more rapid trials and demonstrations of other open source packages including Hyperthought (Jacobsen et al., 2016) and a LIMS developed at the National Renewable Energy Laboratory (White & Munch, 2014). 4CeeD was found to be a powerful tool that addressed many, but not all of the needs of the Nexus Facility, and it was decided to

proceed with an implementation of the custom-built NexusLIMS instead, replicating a few attributes of 4CeeD while proceeding with a different fundamental design and incorporating novel features.

Briefly, NexusLIMS recreates those features of 4CeeD identified as most important for the Nexus Facility (Lau et al., 2019): a web-accessible interface to research data, the previewing of proprietary data formats, the extraction of metadata from those formats, and providing a search interface to find previously acquired data. NexusLIMS differs from 4CeeD, however, in a few key ways. First, the system is designed to require little to no user interaction beyond making reservations on a tool scheduling system, and saving their data in a particular network-accessible folder (two behaviors already in practice at the facility). There is no need for a manual upload of data to the NexusLIMS system, as there is with 4CeeD. Next, the underlying schema for datasets in NexusLIMS is more closely aligned with experimental behaviors compared with the nested collections utilized by 4CeeD (see the “Development of an Experimental Schema for EM” section for further discussion of the data model). Finally, NexusLIMS does not currently integrate analysis capabilities as tightly as 4CeeD [such as through the py4Ceed library (Coordinated Science Laboratory, UIUC, 2020)], although this is envisioned as a future development direction.²

Another point of differentiation is in the distribution and design of the software system. Rather than a monolithic software stack, NexusLIMS is a highly modular collection of infrastructure decisions and software tools with an inherently flexible design, meaning the various software services and design features can be easily swapped (or omitted) to meet an institution’s or research group’s individual needs. In fact, the entire frontend system described in this work (the “Accessing Research Records Using the NexusLIMS Web Interface” section) could be replaced by a similar repository software, if so desired (although the approach presented here confers a number of useful benefits). Likewise, the record building backend can be fully customized for different data formats, or specialized processing as needed during metadata extraction. As every institution’s needs will be different, NexusLIMS can be best considered as a platform (or framework) for LIMS implementation. As such, it is unlikely to be a “turnkey” solution for every research facility, but instead acts as a model reference implementation for a customizable LIMS.

Components of the NexusLIMS Backend

As described above, the NexusLIMS backend makes use of multiple components, each of which is readily exchangeable with a replacement, if desired. The configuration described in this work represents the implementation at the time of publication, although since NexusLIMS is an evolving project, the specifics may change in the future (see Taillon et al., 2020, for the latest implementation). The approaches described in this section may not be directly relevant to all institutions, but the considerations discussed are broadly applicable and should assist in attempts to create a similar system beyond the borders of NIST.

The experimental metadata harvesting, dataset metadata extraction, and experimental record building process are controlled by the NexusLIMS *backend*, as shown in Figure 1. Here, microscopy data are generated by users at the top of the figure. This data

²N.B.: Data in NexusLIMS is certainly machine-readable via API, but this is not a “first-class” feature as in 4CeeD, in that no dedicated library exists to facilitate access.

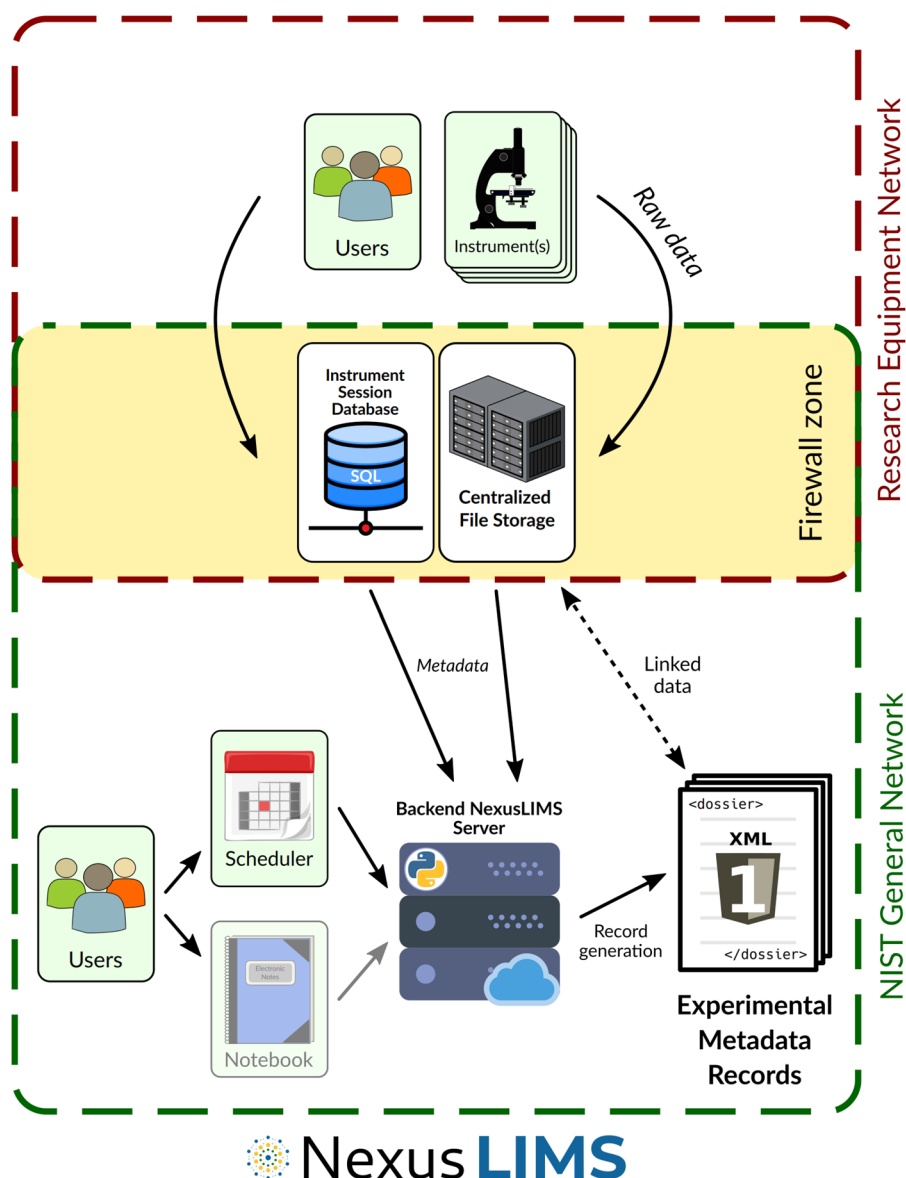


Fig. 1. Schematic representation of the NexusLIMS *backend* architecture. Sources of information are highlighted in green, and information flows are indicated by the arrows. The underlying network infrastructure is represented by the dashed boxes, with the highly restricted REN outlined in red, and the more widely available NIST general network in green. Firewall rules allow access to CFS and session database (colocated with the CFS) from the REN (indicated by the yellow background). The server represented in the bottom center of the diagram orchestrates the collection of information from all the data sources and assembles an experimental metadata record (in XML format) that conforms to the NexusLIMS Experimental Schema (see the “Development of an Experimental Schema for EM” section) (Plante et al., 2020). The ELN data source is faded to represent it has not yet been implemented, but is expected to be an important data source in the future. An important feature of the records generated by NexusLIMS is that the raw data are not included in the record itself. Rather, it is linked by storing the location of the data on the CFS instead.

could be one-dimensional (1D) spectra from X-ray, electron energy-loss, and cathodoluminescence, images of two or more dimensions (where the previous list of 1D signals may be used to compose higher-dimensional signals). Higher-dimensional signals are also found in grain maps of electron back-scatter diffraction (EBSD) and four-dimensional-scanning transmission electron diffraction (4D-STEM) datasets. Data can be held in open formats common for scanning electron microscope (SEM) images [typically stored as Tagged Image File Format (TIFF) images], or the various proprietary formats typical of transmission electron microscopy (TEM) and higher-dimensional data. This data together with information about individual experiments are stored in a centralized network-accessible file server and database. Once the experiment

has finished, the NexusLIMS server queries the session database, the instrument reservation system, and the individual files (saved by the instruments) to build a metadata record of the experiment in the eXtensible Markup Language (XML) format, which is uploaded to the user-facing frontend (not shown in Fig. 1). A number of components work together to enable this functionality, as detailed in the following sections.

Supporting Network Infrastructure

A critical component to any information management system is the ability to easily move data from one location to another. By networking the microscopes’ data acquisition (DAQ) control

computers, the data can be transferred or saved directly in a centralized location, enabling further processing (as described below). This further allows for intelligent controls on data access and prevents computer security and data integrity concerns as are common when simply accessing data via USB. It also provides a facility for users to access their data from their own workstation, or remotely via VPN, which has become critically important due to the recent growth in remote work.

In the NexusLIMS model, once microscopy data are collected by the DAQ computers, one-way outbound data are sent to a central file storage (CFS) server through a highly restricted and protected research equipment network (REN) (the red box in Fig. 1), separated from the primary organizational network by a dedicated enterprise firewall (Helu & Hedberg, 2015; Lau et al., 2019). Though a protected network such as the REN is not a requirement to build a LIMS, it confers a number of benefits worth discussion. Microscope DAQ PCs tend to be older, and frequently run legacy operating systems (OSs) possessing well-known cybersecurity vulnerabilities, meaning they must be treated as untrusted for connection to a wider network and are required to be white-listed prior to network connection [i.e., the REN aims to operate on a “Zero Trust” model (Rose et al., 2020)]. These protected instrument networks are meant to shield the organization’s other networks (green box in Fig. 1) by isolating vulnerable DAQ PCs onto highly restricted subnets and allowing very limited access to centralized resources on an as-needed basis. 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 that may be imposed by the organization.

At NIST, special access to certain network resources (such as the CFS) can be configured with appropriate approvals in place (yellow box in Fig. 1). To further enhance the cybersecurity stance of machines on the REN, all PCs connected to this network have their USB ports disabled, meaning the CFS (see below) is the only means by which users can access their data. Finally, the REN allows the DAQ PCs to access the *time.nist.gov* time synchronization servers, to ensure that the files written by the computer have accurate timestamp information. This is critical for the matching of individual files to a particular experiment (as described in the “Building Experimental Records” section) and allows for the metadata for files collected from different computers to be compared. The REN firewall also prevents general internet access to prevent the unintentional introduction of malware onto the DAQ systems. The REN confers many advantages and is an important component of the NexusLIMS architecture that other institutions may find useful as well. While the NIST system relies on enterprise-level dedicated firewalls, a similarly secure configuration can be recreated at any facility with commodity computing equipment and open source software tools at minimal (potentially zero) cost (Scott, 2015).

All data produced by instruments within the EM Nexus are stored on the NIST-managed CFS server. Using a centrally managed server confers a number of benefits, such as automatic data backups and recovery guarantees,³ as well as automatic expansion of available storage space, although it does incur a financial cost (similar to an external “cloud” storage facility). For the Nexus

Facility, this cost is borne by the sponsoring organization, although a cost-share model would be simple to implement as well. Having the storage on the local network provides acceptable performance with a reasonable cost structure. Local networked storage drives (NAS systems) could be used as well with essentially zero changes to the NexusLIMS system. To enable easy access to the data by users and machines, a simple folder structure is used, where all data from Nexus instruments are deposited into a folder called *mmfnexus* (blue outline in Fig. 2). Inside the *mmfnexus* folder are individual sub-folders for each Nexus microscope. Within each of these are sub-folders for each qualified user of that microscope. Beyond this structure, users are able to save their data in whatever folder structure/ilenames they prefer. This approach allows users to easily navigate through the (read-only) directory tree to locate and download their data post-experiment, while the automated tools of NexusLIMS can use the higher level structure and individual file metadata (such as file modification time) to locate the needed data.

To ensure the highest data reliability and security, the *mmfnexus* folder on the CFS is restricted to only be writable by each instrument’s DAQ PC (and even then, only that instrument’s individual folder within *mmfnexus*). The entire CFS is backed up daily by the NIST central IT services team, providing recovery capabilities in the event of unexpected data loss. In over a year of operation, this capability has yet to be needed due to the restrictive data access controls. Individual users as well as the backend NexusLIMS server (see Fig. 1) have read-only access to this folder, ensuring that users (or an errant piece of code) cannot delete, move, or overwrite any of the raw data collected by the instruments. Since the NexusLIMS code writes a number of accessory metadata files (described below), a parallel directory structure is used on the CFS within a folder named *nexusLIMS* (shown with red outline in Fig. 2). This provides a network-accessible location to store metadata, the session log database, and a backup of the XML metadata records, ensuring there is no danger to the original raw data.

Session Logging and User Workflow

One of the key features of NexusLIMS is the creation of experimental metadata records, which provide a “digital notebook”-like summary of a user’s individual experimental sessions on a microscope (see the “Development of an Experimental Schema for EM” section for more details on the structure of these records). A critical challenge with this process is correctly associating data files on the CFS with a particular experimental session (and the metadata that have already been collected about that session). Due to the wide variety of user behaviors when it comes to structuring and naming saved data, NexusLIMS cannot rely on folder structure or file names for grouping this data. Likewise, although users of the EM Nexus are required to reserve time on an instrument before using it, the reserved times do not always align with the actual time an instrument was in use (due to unexpected delays, instrument problems, etc.). This precludes using the instrument scheduler time ranges as authoritative information about when a user was on a microscope. To solve this problem, a “session logger” application was developed and deployed to the each microscope’s DAQ PC (see Fig. 3).

The session logging application is exceedingly simple (by design) and has been developed to run as a standalone (no installation required) application on Microsoft Windows XP and newer,¹ as well as on Linux, using the *PyInstaller* package (The

³While this is the case for the NIST CFS, data integrity and backup policies may differ at other institutions, so please check local policies prior to relying on a third-party service for data storage.

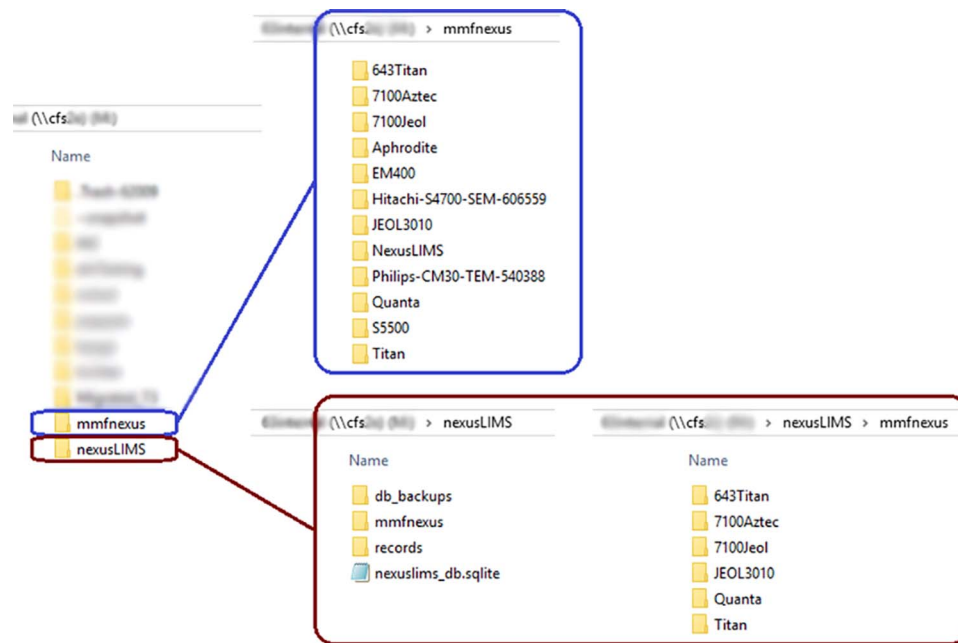


Fig. 2. Structure of the CFS used by NexusLIMS. The instruments write their data into the *mmfnexus* (blue outline), which has highly restrictive permissions to prevent data loss (both NexusLIMS and individual users have read-only access to this folder). NexusLIMS writes its dataset previews, metadata, and experimental records into the *nexusLIMS* folder (red outline), which contains a parallel directory structure to *mmfnexus* (in a subdirectory) to allow for predictable data paths for extracted metadata and previews.

PyInstaller Development Team, 2020). Upon starting the application, a connection is initiated with the session database (colocated on the CFS). This relational database [implemented using SQLite (Hipp, 2020)] has only two tables, as shown in Figure 3c. The instruments table contains information about each instrument in the EM Nexus and its DAQ PC configuration, such as the host-name, static IP address, and where on the CFS that instrument's data are stored.

Each instrument is also assigned a unique identifier, following the pattern *manufacturer-instrument type-NIST property number*. Although including semantic information (i.e., values with an inherent meaning) in a database primary key is not typically recommended (due to the possibility for confusion about the meaning), organizations such as the Research Data Alliance (RDA) do not explicitly discourage it (Wittenburg et al., 2017). A concatenated key (rather than a procedurally generated natural key) was used in this simple database to aid in human recognition of the database entries and is constructed from values that are guaranteed never to change over the course of an instrument's lifecycle. These keys are also easier to use for humans in systems not immediately connected to the database, such as the SharePoint calendar reservation system. Due to the limited number of rows (instruments) in the database, a further abstraction with the use of natural keys was not deemed necessary, but could certainly be used instead, if desired. The instrument table is also used by the NexusLIMS Python package to access information about the individual microscopes, such as the API URL for that instrument's reservation calendar. Holding this information in one place satisfies the DRY principle (Thomas & Hunt, 2019) ("Don't repeat yourself") and prevents errors and data inconsistencies from arising within the system. The session_log table contains timestamped logs for each session, with unique entries for the start and end of a session (as specified by the event_type column). The session logging application creates these logs

with a unique identifier for each session, associating a session with a particular instrument by looking up the DAQ PC's host-name in the instruments table.

From a user perspective, the addition of a session logger represents only a small change from their existing workflow and most critically, does not require the completion of complex forms or other disincentivizing tasks. This approach has been key in driving user adoption of the tool since the facility's management has few tools to change user behavior. As shown in Figure 4, the workflow begins prior to the actual experiment when a user creates a reservation using the scheduling system. This process not only indicates to other users when a microscope is in-use but also allows for the collection of basic metadata about the experiment (see the "Basic Experiment Metadata Collection" section for details). When it is time to start the experiment, the user double-clicks on the session logger application's desktop icon (Fig. 3a), which creates a START log in the database. They then collect data as normal, saving data to the appropriate location on the CFS, which is mounted on the DAQ PC as a Windows network drive (the users can also save locally and copy data to the network drive at the end of their session, if they prefer). Upon completion of data collection, the user closes the session logger application by clicking the "End session" button (Fig. 3b), which inserts a corresponding END log into the session database. This process adds only two new clicks to the existing workflow prior to NexusLIMS, but provides all the information needed for NexusLIMS to build a metadata record of the experiment.

Basic Experiment Metadata Collection

As mentioned previously, the facility management system for the EM Nexus is built using Microsoft SharePoint,¹ to allow for easy sharing of documents, microscope status, announcements, etc. This portal is connected the NIST *Active Directory*, meaning

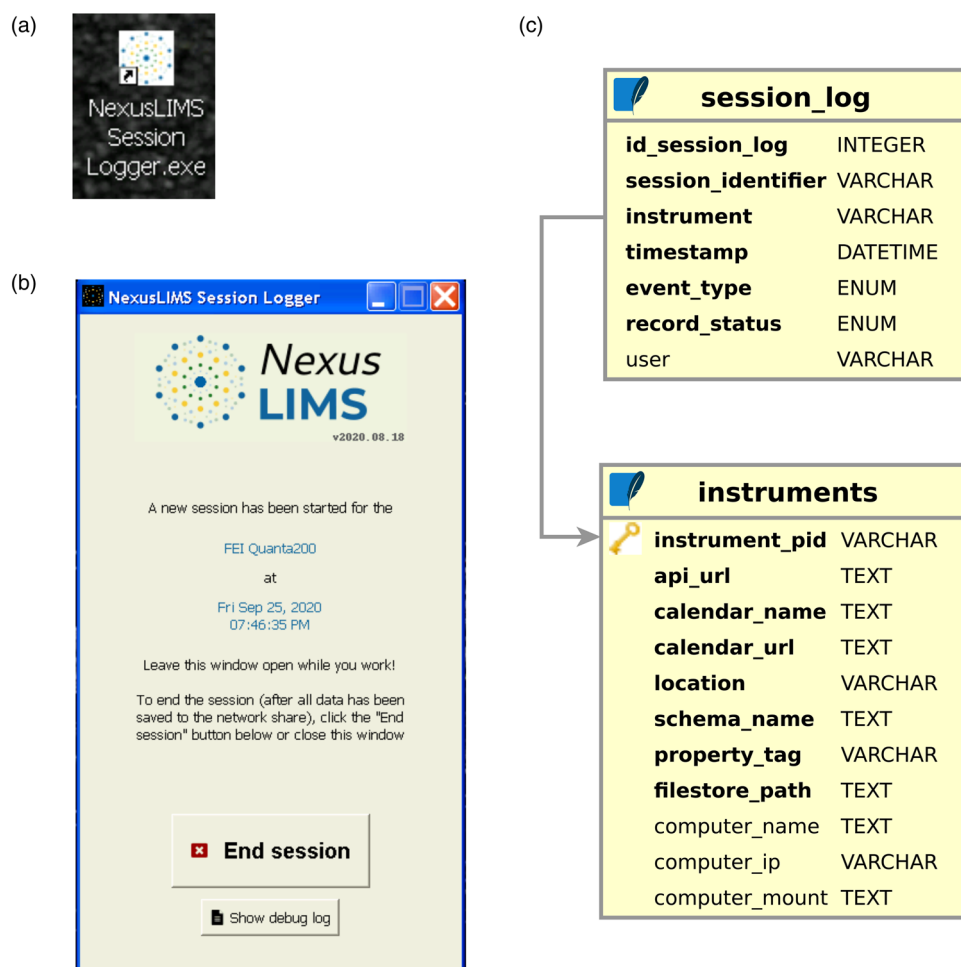


Fig. 3. To start a session, users simply click on (a) the desktop icon for the session logger. A session start is automatically logged to the session database (or if an interrupted session was found, the user is asked if they would like to continue). The user leaves the session dialog box (b) open while they are working (it consumes near-zero system resources), and simply clicks the “End session” button when they are finished, which logs a corresponding entry into the database. The session log SQLite database is likewise very simple and has only two tables (c) containing information about the individual session logs as well as the instruments supported by NexusLIMS. With this information, NexusLIMS has all the information it needs to build a metadata record.

every user automatically receives an account, and their user information stays up-to-date without input from the facility managers. Another resource provided by SharePoint is the concept of shared calendars, which have been utilized to implement a reservation system to prevent conflicts between users. This system, shown in Figure 5, is useful for both the humans that use the facility and for automated machine processes, such as NexusLIMS. The SharePoint calendar provides a central location for the current status and utilization of the microscopes (Fig. 5a), as well as a place to collect basic metadata about each experiment, including who is performing the experiment, the experiment’s title, details about the sample(s) being examined, and a general purpose of the experiment (Fig. 5b). Collecting this information at the time of tool reservation makes it available via a machine-readable API (Fig. 5c), meaning it can be harvested in an automated fashion and the values mapped into the experimental record built by NexusLIMS. Providing a clear link between the values collected in this form and those displayed in the resulting record (see the “Accessing Research Records using the NexusLIMS Web Interface” section) encourages users to thoughtfully complete this form, once they realize the metadata they enter can be queried. In this way, users connect their effort (form entry) to an outcome (searchable record of their work),

which has led to measurable improvements in data management practices throughout the organization, without management having to rule by fiat.

Development of an Experimental Schema for EM

Prior to the implementation of NexusLIMS, one of the most critical processes was the development of a schema to represent an *Experiment*, that is, a unit of time spent by a user on a microscope collecting data (Taillon et al., 2019). This is crucial because data (and metadata especially) are most useful when intelligently structured, allowing for browsing, querying, transforming, and validating the data. A schema is a mapping of real-life notions into a conceptual framework. In a technical sense, a schema is a formal representation of the allowable structure of a document. It can be thought of as similar to a template, in that the schema defines a set of rules, specifying what content is allowed (or often, required) in a document, as well as the values that content can have and the overall structure of the document. Defining the structure of an *Experiment* in this manner allows for the creation of metadata records (Fig. 1) that conform to the schema definition, which in turn allows for targeted queries on certain portions of the

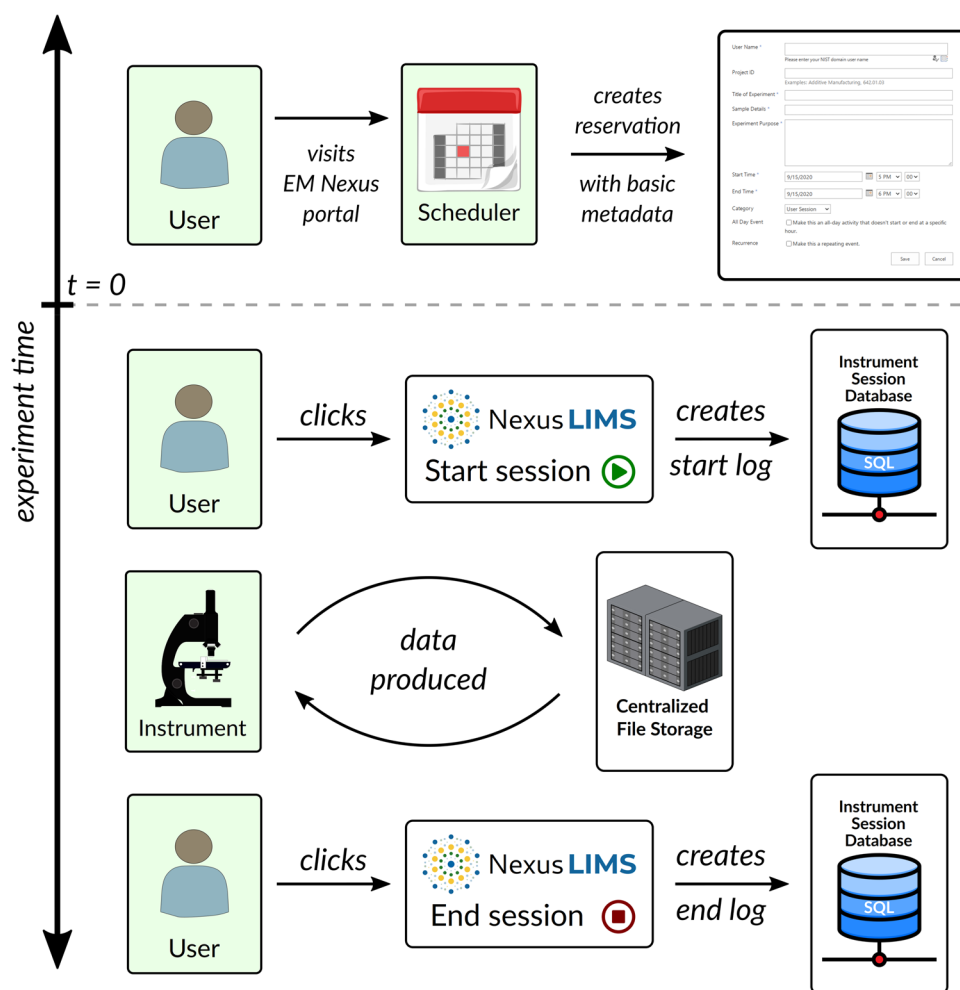


Fig. 4. A diagram of the user workflow prior to and during an experiment. From the user perspective, there is little change compared with the preNexusLIMS workflow. Prior to the experiment, a user visits the online scheduling portal and creates a reservation for the instrument with basic metadata about the experiment (time, date, instrument, purpose, etc.). Once sitting at the instrument, the only additional step required is to click on the NexusLIMS “Start session” desktop icon, followed by the “End session” button after they have finished collecting data. Otherwise, data are saved directly from the instrument to the CFS, where it can be accessed over the network. By the user clicking the button to mark the beginning and end of their session, timestamped logs are entered into the Instrument Session Database, allowing for accurate assignment of data files to a particular experiment.

schema to find all records that match a given set of parameters (this would not possible without a formal schema definition). It also allows for automated processing of the records since users of the system can assume a format for the documents and build workflows using those assumptions (such as through the use of transformation pipelines—see the “Displaying Records” section).

The *Nexus Experiment* schema (Plante et al., 2020) was developed in consultation with EM Nexus researchers at NIST through an iterative process where participants decided on the most important information to record from an experiment. Further efforts resulted in refinements after consultation with the wider materials microscopy community at the 2019 NIST/CHiMaD Materials Microscopy Data Conference (Center for Hierarchical Materials Design, 2020). The schema is defined in the XML Schema language (Vlist, 2002) for compatibility with the frontend CDCS, though this is a specific implementation decision and the concepts could be formalized in any schema language.

An overview of the schema design is shown in Figure 6. It is a hierarchical model, with the *Experiment* as the root-level node. *Experiments* have a number of high-level descriptive metadata nodes, such as *Summary*, *Project*, *Sample*, and *Notes*. Each of

these nodes consists of additional details. For example, the *Summary* node contains information about the *Experimenter*, the *Instrument* used, the declared *Motivation* for the work, and the start/end times. Please refer to Plante et al. (2020) for further details.

Besides the high-level metadata, the primary content of a record defined by the *Nexus Experiment* schema is contained within *Datasets*, which in turn are grouped into *Acquisition Activities*. A *Dataset* represents a file (or group of files) acquired during the *Experiment*, together with metadata about that dataset, including its name, a link to the raw data file location, its format, an optional description, and a link to a preview location. Since these documents are *metadata* records, the raw data are not stored within them and rely on the data being accessible in a linked location (such as the CFS). Each *Dataset* can also have one or more *Meta* nodes, which represent arbitrary metadata values about the dataset. Most often, these nodes contain metadata extracted from the proprietary data format as saved by the microscope or data collection instrument.

While the NexusLIMS system as a whole is modular and flexible, the use of a formally defined schema ensures consistency of the data model throughout the system. On the backend, the

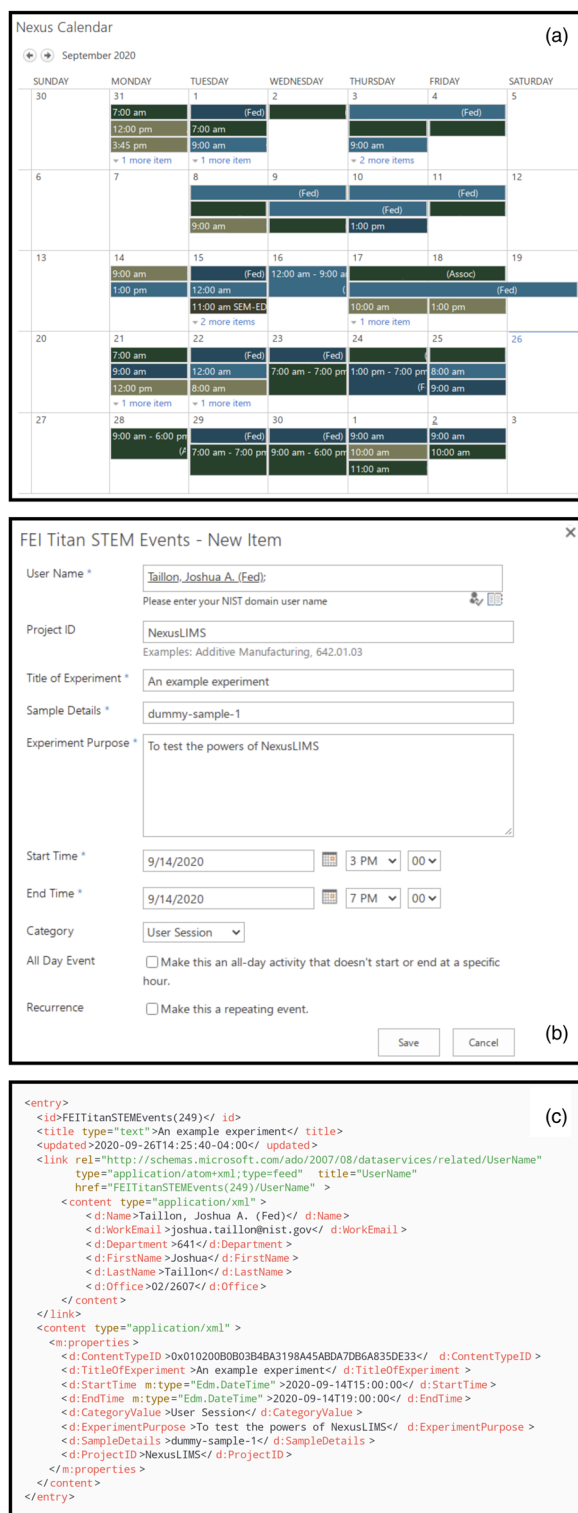


Fig. 5. The scheduling system provides human-accessible and machine-readable access to reservation information. (a) The overall calendar view provides a quick overview of current microscope utilization. Each instrument is represented by a unique color, and the name of the person for each reservation is displayed (removed for privacy in this figure). (b) Creating a reservation brings up a form with the expectation of basic metadata entry, such as the user (linked to the Active Directory³), a project identifier, a title and purpose of the experiment, as well as the date and time for the experiment. This information is then displayed on the overview calendar (a), but is also accessible in a structured XML format (c) that provides for machine-readable access to the information.

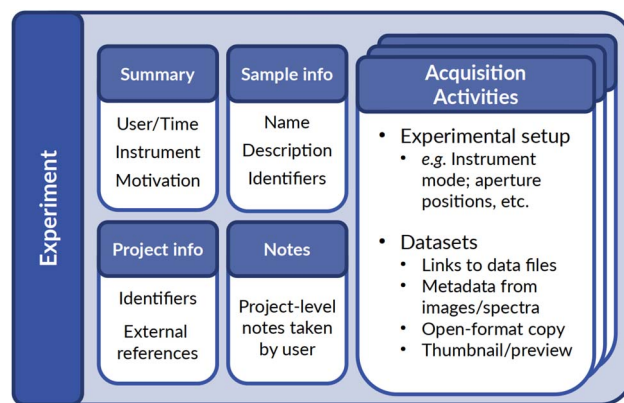


Fig. 6. A high-level overview of the *Nexus Experiment* schema, illustrating the hierarchical levels of an *Experiment*. At the top level is summary information about the experiment (the “who, what, when, where, why”), as well as details about the sample and any project of which the experiment is a part, together with any *Experiment*-level notes. The bulk of the information specified by the schema is contained in *datasets*, which are grouped into one or more *Acquisition Activities*, which represent a collection of *datasets* with common properties. Each *dataset* can contain additional details, such as instrumental metadata, a preview image, a link to where the raw data are stored, as well as a nonproprietary formatted version of the file, if desired.

schema informs what values are collected at the time of a user’s reservation, what sort of information (and in what format) is extracted from the individual data files, and how research records are structured by the record builder (see the “Building Experimental Records” section). The schema provides an expected structure against which records can be validated to ensure that the different software pieces are generating valid records, and to issue a warning or error if this is not the case. On the user-facing frontend, the schema provides a dependable data model on which to build the display of information (see the “Displaying Records” section), as well as a basis for reliable free text and faceted searching of metadata. In this way, the schema defines a set of expectations for the system that must be met in order for the various modular components to work together correctly.

It is important to note that the *Nexus Experiment* schema does not enforce a specific experimental metadata vocabulary or structure for the individual metadata values associated with each *Dataset*. For example, it does not specify that TEM images must have a value for accelerating voltage, magnification, dwell time, or any other experimental parameter. Defining a schema for this type of domain- and instrument-specific metadata was beyond the immediate scope of this work and will require obtaining a wider consensus among EM researchers and instrument vendors for standard vocabulary and metadata format definitions. Such an effort however, is urgently needed within the community to promote data interoperability, and a solution is likely to grow through the evolution of existing community efforts to standardize materials metadata (Cheung et al., 2009; Blaiszik et al., 2016) and file formats—including the HMSA HyperDimensional Spectral Data File Format (Microscopy Society of America : Standards Committee, 2019).

Dataset Metadata Extraction and Conversion to Open Formats

A key benefit of building a networked infrastructure for a LIMS is improved data utility, in part by automating interactions with the data produced in the laboratory. In pursuit of the FAIR data

principles, the data collected by the NexusLIMS microscopes are made more *findable* through the extraction of metadata from the individual data files (and its storage in a queryable system). Additionally, it is made more *interoperable* through the transformation from proprietary data formats into open formats that can be read without commercial software license(s). Finally, it is made more *accessible* through the generation of open-format preview images, such that a cursory examination of the data is possible without any specialized software beyond a web browser.

As part of the record building process (see the “Building Experimental Records” section), each file associated with a given experiment is analyzed and where possible, has its metadata extracted. An open-format preview image is also generated for the file as part of this process. To perform these operations, NexusLIMS makes extensive use of the community-developed open source library HyperSpy (de la Pena et al., 2017; de la Peña et al., 2020) due to its wide-ranging support for reading of proprietary EM data formats. Wherever possible, every metadata parameter stored by the data collection software is included in the resulting metadata record. Certain formats, however (such as Gatan’s *DigitalMicrograph*¹), include a great deal of software-specific metadata that are not of experimental or scientific interest, and are thus excluded from the final record. Once all the available metadata have been extracted, these are then filtered to ensure consistent formatting within NexusLIMS. A few additional values are added, such as a determination of the data type (SEM imaging, TEM diffraction, EELS Spectrum Imaging, etc.) based off the information available at the time of processing. Once complete, the extracted metadata are written into the XML record and also saved to the CFS in a JavaScript Object Notation

(JSON)-formatted file (Ecma International, 2020) for additional processing if a user desires.

For certain instruments and file formats, the metadata extraction process offers an opportunity to additionally perform basic data processing steps automatically, which have previously been carried out manually whenever the researcher needed to access their data. For example, one EM Nexus instrument collects 4D STEM data using a direct electron detector. The data are saved pixel-by-pixel as the sample is scanned, resulting in tens or hundreds of thousands of individual data files that must be combined and reduced prior to any sort of useful analysis. In collaboration with EM Nexus researchers, a pipeline was developed to allow NexusLIMS to perform this data preprocessing step noninteractively, meaning the data are in a format immediately useful to the researchers when they access it on the CFS (or through the NexusLIMS frontend), saving time and effort as well as ensuring complete reproducibility. Similar pipelines could likewise be developed for other data formats and processes with relative ease.

Building Experimental Records

With the constituent pieces (network infrastructure, session logging and user workflow, metadata collection, a detailed schema, and proprietary file format extraction) in place, NexusLIMS has all the required information to build detailed experimental metadata records. This functionality has been implemented using a custom Python package named `nexuslims` (Taillon et al., 2020), which runs unattended on a server connected to the general NIST network (recall Fig. 1’s schematic overview of the architecture). The `nexusLIMS` package comprises a number of modules

Table 1. Summary of the `nexusLIMS` Python Package Structure (in Alphabetical Order).¹

Subpackage	Module	Description
	<code>cdcs</code>	Code for interacting with the CDCS frontend via API (record uploading, deletion, and other helper methods)
	<code>instruments</code>	Pulls up-to-date instrument information from the NexusLIMS database and supplies a Python-object representation for instruments
	<code>utils</code>	Various project-level utility functions, including authentication for web requests, finding files by modification time, parsing XML, etc.
	<code>version</code>	A module to keep track of the current software version
<code>builder</code>	<code>record_builder</code>	Orchestrates the creation of metadata records using the other <code>nexusLIMS</code> modules. This module is the main entry-point that runs regularly to automatically create new records within NexusLIMS
<code>db</code>		Contains all functionality related to the NexusLIMS database and provides Python wrapper methods for common database operations
	<code>migrate_db</code>	Used to migrate session log records after any change to the NexusLIMS database SQL schema
	<code>session_handler</code>	Uses the NexusLIMS database to provide a Python-object representation of SessionLog (rows in the database) and Session (blocks of time from which to build a record)
<code>extractors</code>		Each extractor module contains the code necessary to extract metadata from a file format generated by one or more instruments in the EM Nexus
	<code>digital_micrograph</code>	Handles files saved by Gatan’s DigitalMicrograph software (.dm3 and .dm4 files)
	<code>fei_emi</code>	Handles files from FEI/Thermo Fisher’s Tecnai Imaging and Analysis software (.ser and .emi files)
	<code>quanta_tif</code>	Handles .tif images saved by FEI/Thermo Fisher’s SEM and FIB instruments based on the Quanta platform
	<code>thumbnail_generator</code>	Generates preview images for all types of files
<code>harvestors</code>	<code>sharepoint_calendar</code>	Communicates with the SharePoint calendar resource to obtain summary experimental metadata and performs XML response processing
<code>schemas</code>	<code>activity</code>	Provides a Python-object representation of an Acquisition Activity from the <i>Nexus Experiment</i> schema, as well methods to define activity time boundaries and metadata parameters

and subpackages, each responsible for a different component of the record building process. The structure of the entire package is summarized in Table I, which lists all the modules and their purposes.

Generally, the codebase can be understood as four primary components: *Extractors*, which are used to pull metadata out of the raw data files saved on the CFS; *Harvestors*, which collect metadata from external sources (currently only the reservation calendar, but other sources—such as ELNs—will be implemented in the future); *Database tools*, which interact with the NexusLIMS instrument and session logging database to determine what actions need to be performed; and the *Record builder*, which orchestrates the creation of the XML-formatted experimental metadata records. Other modules provide an assortment of additional functionality, such as interacting with the frontend system, or various utility functions that are shared throughout the codebase.

The workflow implemented to build these records from the various data sources is illustrated in Figure 7. This process happens asynchronously with the user workflow (Fig. 4) and is automated to run without user or administrator intervention [using the cron daemon (Vixie et al., 2013)], alerting the NexusLIMS administrators if any problems are encountered. Each “run” of the record builder begins with a check to the NexusLIMS database to determine if any new sessions have finished since the previous run (these sessions are logged by users using the Session Logging application—see Fig. 3). If not, no further actions are taken until the next iteration of the record builder (typically run every 30 min).

If one or more sessions are detected, the pipeline of steps 2 through 5 in Figure 7 is executed. For each session, the SharePoint calendar for that session’s instrument is queried for any reservations matching the time span of the session, and any user-entered metadata are stored. If no reservation is found, the record will still be built, but with only generic metadata (time, date, instrument, etc.). After this, the CFS is searched for any files created within the session’s time boundaries with file extensions that have a corresponding extractor implemented. At this point, the files are grouped into Acquisition Activities based on their creation time, using an adaptive clustering process that separates the groups of files at any points where there was a relatively large span of time between dataset acquisitions. Each file is then processed by the appropriate extractor, which causes a JSON-formatted metadata file and a preview image to be written to the CFS. The extracted metadata are added to one or more meta nodes within each dataset, as specified by the *Nexus Experiment* schema (see the “Development of an Experimental Schema for EM” section). Finally, all this information is written into an XML-formatted record that is saved to the CFS, and uploaded via API to the NexusLIMS frontend using the *nexusLIMS.cdcs* Python module. This process is repeated for all new sessions detected, and if any errors are detected in the output, the administrators are notified via email.

Accessing Research Records Using the NexusLIMS Web Interface

To this point, all the infrastructure and tooling described above has supported the *backend* capabilities of NexusLIMS to create experimental metadata records. Typical users in the EM Nexus facility, however, have little interest in or need for insight into this process and instead interact with NexusLIMS only through

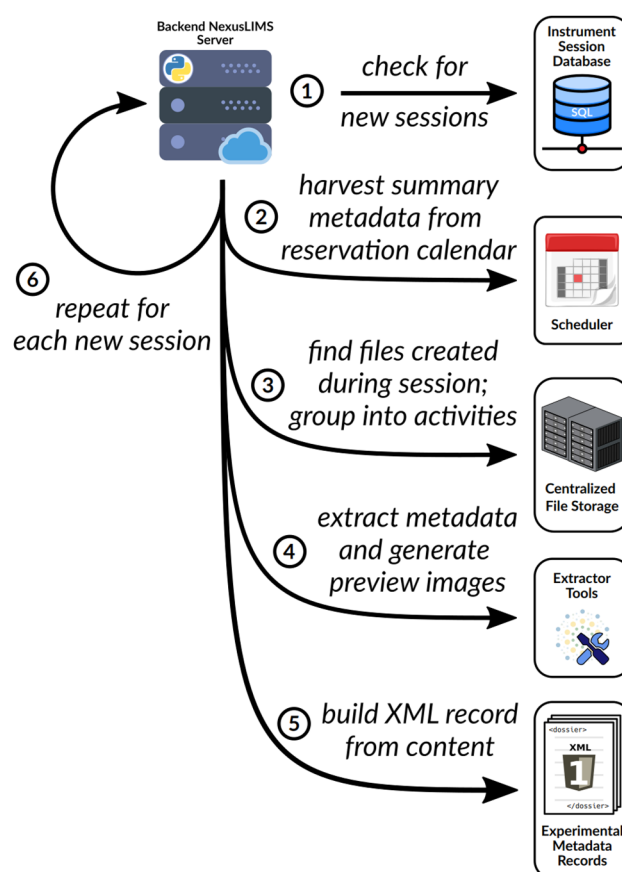


Fig. 7. The workflow used by NexusLIMS to create experimental research metadata records from the various available data sources. See the text for a full description of each step.

the web-based *frontend*, which provides all the searching, browsing, and data downloading capabilities they expect. This frontend has been built by customizing an instance of the CDCS, a project developed at NIST that grew out of the MGI (Dima et al., 2016). The CDCS allows for the collection, curation, dissemination, and display of XML-formatted structured documents. The system provides built-in querying capabilities, both via freeform text search and more detailed schema-based queries.

Displaying Records

A key feature of the CDCS is the use of stylesheets to display the structured XML records that are curated by the system. An XML document cannot be displayed by a standard web browser (besides a text-based view) since the nodes (such as <Experiment>, <summary>, <meta> in the case of NexusLIMS) have no meaning to a web browser. A *translation* is required to convert the XML format to an HTML document that can then be rendered by any web browser. In CDCS, these translations are performed using documents written using the eXtensible Stylesheet Language Transformation (XSLT) language (Tidwell, 2008). An XSLT document defines a roadmap to convert from a structured XML (regardless of the specific content of that XML) into a web page, allowing for precise control of the display based off the known structure of the input document (which is controlled by an XML Schema definition).

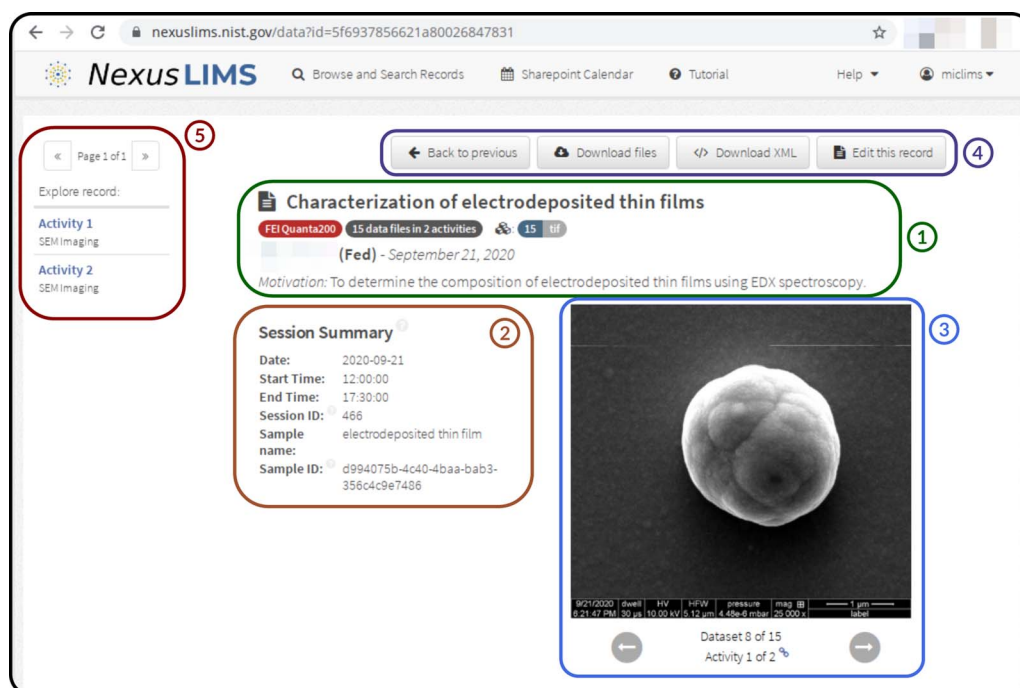


Fig. 8. A screenshot of a research metadata record as displayed when first loading the page. NexusLIMS creates this view by using the XSLT to transform the underlying XML into a standard HTML web page. ① (green box) The header information of the record contains the experiment title (as entered by the user for the reservation), the instrument (“FEI Quanta200” here), the number of datasets and their file types, together with the number of activities detected, the experimenter’s name (censored for privacy here), the date of the experiment, and the motivation as entered by the user; ② (brown box) the session summary contains a few more details about the reservation, information about the sample, and any associated project information (not shown in this example); ③ (blue box) the interactive preview gallery shows a preview image for every dataset contained in the record and can be quickly tabbed through using the provided buttons or the keyboard arrows; ④ (purple box) controls at the top of the record provide access to a dataset/metadata downloader tool, a link to download the XML record, and a way to edit the contents of a record (e.g., to correct any errors); and ⑤ (red box) the sidebar navigation allows the user to quickly jump to the detailed view for a specific activity (see Fig. 9) and displays the data type of the contents of each activity (SEM Imaging, in this example).

In NexusLIMS, a great deal of effort has been placed on generating an XSLT definition that results in a high density of immediately useful information presented to the user, with additional detail available close at hand wherever desired. Using XSLT to generate an HTML document means the full suite of modern web tools [JavaScript—including external libraries and cascading style sheets (CSS) (The World Wide Web Consortium (W3C), 2020)] can be used to enhance the display of the record content. Inspiration for the resulting output has been taken from *Wikipedia* (Wikimedia Foundation, Inc, 2020), where summary information and visual feedback (i.e., a gallery of previews) are available at first view, with additional detail available by scrolling down the page. This initial view is detailed in Figure 8.

For many users, the simple presence of a gallery of preview images is one of the most powerful features of NexusLIMS. With many data formats (such as .dm3 or .ser/.emi TEM images), the data cannot be easily previewed using the built-in operating system tools, meaning to find a particular file of interest, a user may need to open dozens (or more) of files in the proprietary software just to figure out which file is the one they wanted to share with a colleague. NexusLIMS enables efficient browsing of a large number of datasets with nothing more than a standard web browser. In addition to the gallery of preview images, the record view also includes high-level information such as the experiment title, the instrument used, the number of files contained within (and their types), the person who ran the experiment, and the motivation they noted when making a reservation. Additional details about the session and sample

(entered by the user at the time of reservation) are displayed next to the preview gallery. The navigation bar on the left of the page shows a list of the Acquisition Activities determined in the record building process and provides quick links to view the details of each one. The buttons at the top of the record provide a few various functionalities, such as opening the file downloading tool, which allows a user to download all their data (and the extracted metadata) as an archive (.zip) file. A user can also download the entire record as an XML file for their own processing, or click the “Edit this record” button to perform simple edits or corrections on the content of the record.

Scrolling down the page (or clicking one of the links on the left navigation bar) reveals the acquisition activity detail sections (Fig. 9). There will be one of these sections for each activity detected in the experiment. The header displays the types of data found in this activity and the number of files it contains, while the table on the right side lists every dataset included in the activity. Hovering over a line in the table with the mouse will reveal the preview of that dataset in the area next to the table on the left. Within the table, the name, creation time, data type, and role (i.e., experimental data or something else) for each dataset are listed. Links are also provided to view or download the extracted metadata or the individual data file itself (see Fig. 9 for more details).

A distinction is made within the *Nexus Experiment* schema (and is reflected in the display of the record) between a “setup parameter” for an acquisition activity and a “metadata value” for an individual dataset. The difference is defined by determining

Experiment activity 1
Activity contents: SEM Imaging
9 data files

Dataset Listing:

Dataset Name	Creation Time	Type	Role	Meta	D/L
20200914_001.tif	2020-09-21 18:11	Image	Experimental		
20200914_002.tif	2020-09-21 18:13	Image	Experimental		
20200914_003.tif	2020-09-21 18:14	Image	Experimental		
20200914_004.tif	2020-09-21 18:15	Image	Experimental		
20200914_005.tif	2020-09-21 18:17	Image	Experimental		

Experiment activity 1 Metadata:

Setup Parameter	Value
Start time	18:11:53
Acquisition Date	09/21/2020
Beam Name	EBeam
Beam Tilt X	-0.218728
Beam Tilt Y	0.302053
Chamber ID	XL305B
Column Type	FEI SEM
Data Dimensions	(1024, 884)
Data Type	SEM Imaging
Detector Grid Voltage (V)	0.0

20200914_005.tif Metadata:

Metadata Parameter	Value
Acquisition Time	06:16:50 PM
Chamber Pressure (mPa)	0.561393
Creation Time	2020-09-21 18:17
Detector Brightness Setting	38.1672
Detector Contrast Setting	63.2559
Horizontal Field Width (μm)	128.0
Pixel Height (nm)	125.0
Pixel Width (nm)	125.0
Stage Position X	-0.000205062
Stage Position Y	0.00197334

Fig. 9. A screenshot of activity details visible after scrolling down the page from Figure 8. There is one detail section for each activity detected in the record. ① (green box) The header for each activity contains a listing of the data types included (here just “SEM Imaging”) and the number of datasets contained in the activity. ② (brown box) The dataset listing for each activity provides a full listing of all the files associated with this activity, including their names, when they were collected, the type of data, and links to view or download both the extracted metadata and the raw data file. Hovering over a given row displays the preview of that dataset in ③ (blue box) the area on the left of the table. ④ (purple box) The metadata link at the top of the activity brings up a modal dialog showing a searchable list of metadata common to all files within the activity (e.g., values such as the electron column type do not change from dataset to dataset). The metadata link in each row of the table brings up ⑤ (red box) a modal box with the metadata unique to that specific dataset (e.g., values like stage position or magnification, which change from file to file).

which of the extracted metadata values are common among all datasets in the activity as opposed to those that might vary between individual files. Those that are uniform for all files are specified as *setup parameters* for the parent activity, while those that change from file to file are associated with each individual dataset as *metadata values*. These values are visible within the displayed record in two different locations (see boxes 4 and 5 in Fig. 9).

With both summary information and detailed metadata view for each dataset, the NexusLIMS frontend provides a full-spectrum view into an experimental record. It facilitates not only the browsing of data, but the examination of metadata and data access itself through the file downloading capabilities. Users can immediately recognize the value of the metadata they

enter at the time of reservation, which encourages more thoughtful completion of these forms over time, especially once the users realize they can search on these values as well (see the next section). It is the hope of the authors that the NexusLIMS record pages will become the first place researchers visit after their experiment to review their work and share their results with NIST colleagues.

Searching for Records

Besides the display of experiment records, one of the key features provided by CDCS for the NexusLIMS frontend is the ability to perform detailed queries on the repository of records, as shown in Figure 10. This feature is provided “out-of-the-box” by

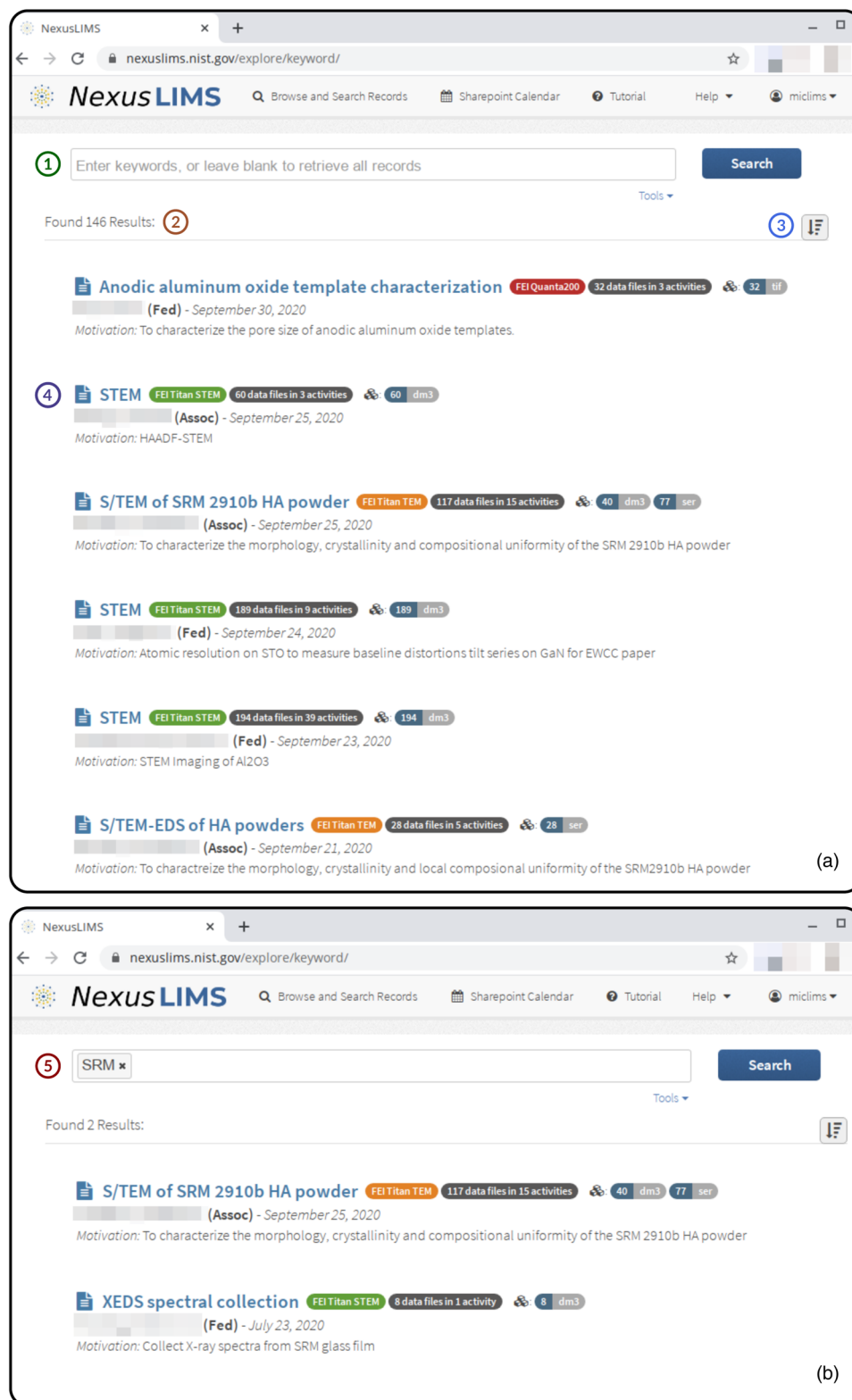


Fig. 10. Screenshots of the “explore” page that allows for querying the record repository (users’ names are obscured for privacy). (a) By default, when the search box ① is empty, this page shows all records in the system and reports the number of records found ② to the user. A sorting option ③ allows the records to be sorted by date (default) or alphabetically. A brief summary of each record ④ is displayed below, which can be clicked by the user to view the full record. The search bar performs a free-text search of the each metadata record (b) ⑤, returning records that match the query. This allows users to search for relevant terms from the information they entered at the time of reservation, by instrument, by date, by filetype, or any other query. In this example, two records in the repository were found matching experiments performed on Standard Reference Materials (SRMs), from two different users on two different instruments. By combining search terms, arbitrarily complex queries can be built that allow users to find exactly the data they had in mind.

CDCS and does not require any significant customization or configuration to enable. The basic text search powers the “Browse and Search” page. With an empty query (Fig. 10a), this page will by default display a small preview of each record found in the repository (paginated by 10 records per page). The preview is controlled by a separate XSLT document and has been customized to display only the most basic information contained in each record, such as the title, instrument, user, date, number of datasets, and motivation.

The power of this page comes from the search bar at the top, which accepts freeform text queries and searches the entire content of every record in the repository for matches. Using the box, a user can quickly find all their experiments by simply entering their username. In addition (or in alternative) to this, a specific instrument identifier can be entered to return experiments from only that microscope, or a sample identifier to match what was entered by the user at the time of reservation. The experiment titles and motivations are included in the search as well, so if a user has inputted useful information into these fields, they will be able to query on them using this page. This tool makes it easy for users to swiftly and effortlessly pare down a large repository of experiment records to just those that interest them, and is a vast improvement over browsing through a hierarchical folder structure using standard operating system tools.

Although not discussed to this point, NexusLIMS supports granular data access controls to limit what records are findable and viewable by which users. Within the Nexus Facility, the default data access model is that all users of the facility have rights to read (but not write or edit) all raw research data produced by any of the instruments (with a few exceptions). This model is reproduced in the current implementation of NexusLIMS, meaning any logged in user will be able to view or search the records of any other user. Obviously, this model does not translate to every research environment, where users may work on sensitive or proprietary samples, and the data must be protected. Thankfully, finer control over access levels is simple to implement (if desired) in the NexusLIMS frontend by using the CDCS concept of *workspaces*. In the CDCS platform, records are “owned” by individual users; this user is assigned by the NexusLIMS backend when the record is built using contextual information such as instrument reservation details and any username information contained in the filepaths of the harvested data files. Once uploaded to CDCS, a record can be assigned to one or more arbitrary workspaces, and users can be members of any number of workspaces. In this manner, any level of access control is possible, from global access (the current implementation) to highly restricted access, or shared workspaces (e.g., for a particular research group and project) that allow many, but not all, users to view a set of research records.

Future Development Directions

While the existing set of features in NexusLIMS provides many novel capabilities for users of the EM Nexus Facility, there are many future improvements and feature additions currently planned. Feedback is regularly solicited from users, and many of the ideas described here have been sourced from active users of the system.

Of primary importance is expanding extractor support for all filetypes produced by microscopes in the EM Nexus. The current extractor suite handles approximately 90% of the existing files produced (as calculated by file size), but the remaining 10% is

especially important to users that focus their efforts on those tools. The formats that are currently unsupported are those that do not have existing readers in HyperSpy, due to closed binary formats or poor vendor support for third-party tools. Enabling extraction from these types of files [particularly an issue with the files produced in analytical techniques such as energy-dispersive X-ray spectroscopy (EDS) and EBSD] will require significant reverse-engineering efforts by the community, or additional support from vendors.

Another feature in active development is the linking of NexusLIMS to other repositories at NIST that maintain information about samples and their histories. Because these repositories (and CDCS) implement persistent identifiers (PIDs; Borgman, 2010) for each record/sample, the use of a handle server (Corporation for National Research Initiatives, 2020) will allow NexusLIMS to resolve these PIDs to their originating repository, and vice versa. With this infrastructure in place, users will be able to easily jump from a research record to sample details and back through an interconnected system of repositories, confident that their digital data are FAIR and are being managed to modern best practices. Related to this effort is the implementation of PIDs for detectors and specimen holders that are used with the instruments in the EM Nexus, which will allow for another layer of rich metadata in the experimental records.

Additionally, while the *Nexus Experiment* schema has support throughout for note contents, these capabilities have yet to be fully utilized. This is partially due to existing user behavior, but also due to the breadth of research activities performed at NIST; not all users make use of ELNs, and those that do have a disparate range of practices, making uniform support difficult to implement. Many EM Nexus researchers do, however, make use of Microsoft OneNote¹ for their digital notetaking needs, and so, basic support is planned through the attachment of searchable digital copies of these notes, making use of Microsoft 365's OneNote API. Over time, users will be encouraged to use specific templates for their notes if they wish for that content to be integrated more fully into the research record, but this behavior will be completely optional.

Longer term, a few features are being considered, but have not been formally placed on the feature roadmap. Chief among these is the implementation and enforcement of instrument- or technique-specific schemas for the metadata (and their units) extracted from the datasets, most likely building off existing community efforts (Blaiszik et al., 2016). This will require coordination with the EM Nexus users, instrument vendors, and the larger materials microscopy community to reach a consensus arrangement that is satisfactory to all the involved stakeholders. Also, NexusLIMS currently handles only raw data as produced by the instruments and does not support processed/analyzed data (although the *Nexus Experiment* schema offers a place for this type of data). The developers plan to eventually incorporate support for analyzed data, including a Python API for data access using colocated services such as Jupyter (Kluyver et al., 2016). The *Nexus Experiment* schema will also continue to evolve alongside NexusLIMS as further enhancements are needed.

Summary and Lessons Learned

This work has presented NexusLIMS, a fully featured research data management system implemented by the Office of Data and Informatics and the Materials Science Engineering Division at NIST for a multi-user EM co-op. NexusLIMS is built from a collection of existing resources, as well as custom Python code

to handle the harvesting of information, extraction of metadata, and creation of research experiment metadata records. It relies on instruments being networked to a centralized data storage location, as well as user behavior to enter summary metadata when making a reservation on one of the tools, and to initiate the “Session Logger” through a single click when they are on the microscope. Once a user has completed their session, the backend server automatically harvests relevant metadata, finds and extracts information from the associated data files, and builds and uploads a research metadata record to a web frontend built using the CDCS. Users can then view, search, and download their data using an intuitive web interface through their favorite browser.

NexusLIMS was deployed shortly before the mandatory work-from-home orders of 2020, but even still, has built hundreds of experimental records in its first 6 months, representing the work of dozens of active users across three different NIST divisions. Over 7,000 datasets totalling hundreds of GB of data (and consistently increasing) have been ingested and collated using the system. Because usage of NexusLIMS cannot be mandated by the facility’s management, the authors have been pleased to observe that at the time of publication, nearly 90% of TEM data files are being captured within the system, while the harvested proportion of other instruments’ data is steadily growing. The goal for NexusLIMS is capture of 100% of research data produced by instruments in the EM Nexus, which will be obtainable through further development of data format extractors and extensive user outreach.

While NexusLIMS is highly tailored to the NIST/EM Nexus infrastructure, much of the underlying nexusLIMS code and general considerations presented in this work should be extendable to other organizations. By building tools and capabilities such as these, effective data management within of an organization can be built up with minimal effort from individual users, whose change in behavior is often the most difficult part of the challenge. Tools such as NexusLIMS encourage users to improve their practices through linking simple first steps with immediately visible benefits, and through these individual changes, NIST maintains its reputation for scientific data integrity.

Acknowledgments. The authors acknowledge the efforts of the Summer Undergraduate Research Fellowship (SURF) students and interns that have worked on the NexusLIMS project: Rachel Devers and Sarita Upreti. The authors also thank the early supporters of the NexusLIMS efforts in the EM Nexus, and especially those researchers who took the time to provide feedback and testing to ensure the resulting system was the most useful it could be for research scientists. Those users include (in alphabetical order): Drs. Andrew Herzing, Megan Holtz, Michael Katz, and Vladimir Oleshko. Finally, the authors thank Drs. Chandler Becker and J. Alexander Liddle for their insightful comments during review of this manuscript.

References

Abbott Laboratories (2020). Laboratory information management system (LIMS)—STARLIMS. Available at <https://web.archive.org/web/20200803100114/https://www.informatics.abbott/us/en/offerings/lims> (retrieved August 3, 2020).

Allcock W, Bresnahan J, Kettimuthu R, Link M, Dumitrescu C, Raicu I & Foster I (2005). The globus striped GridFTP framework and server. In *ACM/IEEE SC 2005 Conference (SC'05)*, vol. 2005, pp. 1–11. Seattle, WA: IEEE. Available at <http://ieeexplore.ieee.org/document/1560006/>.

Arkilic A, Allan DB, Caswell TA, Li L, Lauer K & Abeykoon S (2017). Towards integrated facility-wide data acquisition and analysis at NSLS-II. *Synchrotron Radiat News* 30, 44–45.

Bika Lab Systems (2020). Bika open source LIMS. Available at <https://web.archive.org/web/20200811104846/https://www.bikalims.org/> (retrieved August 11, 2020).

Blaiszik B, Chard K, Pruyn J, Ananthakrishnan R, Tuecke S & Foster I (2016). The materials data facility: Data services to advance materials science research. *JOM* 68, 2045–2052.

Borgman CL (2010). *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Carey NS, Budavári T, Daphalapurkar N & Ramesh KT (2016). Data integration for materials research. *Integr Mater Manuf Innov* 5, 143–153.

CARPi N, Minges A & Piel M (2017). eLabFTW: An open source laboratory notebook for research labs. *J Open Source Softw* 2, 146.

Center for Hierarchical Materials Design (2020). Data and database efforts. Available at https://web.archive.org/web/20200710235929/https://chimad.northwestern.edu/news-events/CHiMaD_Data_Database_Efforts.html (retrieved September 27, 2020).

Cheung K, Hunter J & Drennan J (2009). MatSeek: An ontology-based federated search interface for materials scientists. *IEEE Intell Syst* 24, 47–56.

Coordinated Science Laboratory, UIUC (2020). 4CeeD+Jupyter. Available at <https://web.archive.org/web/20201230005737/https://t2c2.csl.illinois.edu/4ceedjupyter/> (retrieved December 20, 2020).

Corporation for National Research Initiatives (2020). Handle.Net Registry. Available at <https://web.archive.org/web/20200901232239/https://handle.net/index.html> (retrieved September 30, 2020).

Dataworks Development, Inc (2020). Freezerworks—Laboratory software for freezer and biorepository tracking. Available at <https://web.archive.org/web/20200919013601/https://freezerworks.com/> (retrieved September 19, 2020).

de la Peña F, Ostasevicius T, Tonaas Fauske V, Burdet P, Jokubauskas P, Nord M, Sarahan M, Prestat E, Johnstone DN, Taillon J, Caron J, Furnival T, MacArthur KE, Eljarrat A, Mazzucco S, Migunov V, Aarholt T, Walls M, Winkler F, Donval G, Martineau B, Garmannslund A, Zagonel LF & Iyengar I (2017). Electron microscopy (big and small) data analysis with the open source software package HyperSpy. *Microsc Microanal* 23, 214–215.

de la Peña F, Prestat E, Fauske VT, Burdet P, Jokubauskas P, Nord M, Furnival T, Ostasevicius T, MacArthur KE, Johnstone DN, Sarahan M, Lähnemann J, Taillon JA, Migunov V, Eljarrat A, Aarholt T, Caron J, Mazzucco S, Martineau B, Somnath S, Poon T, Walls M, Slater T, Winkler F, Tappy N, Donval G, Myers JC, McLeod R & Hoglund ER (2020). HyperSpy 1.6.0. Available at <https://github.com/hyperspy/hyperspy>.

Dima A, Bhaskarla S, Becker C, Brady M, Campbell C, Dessauw P, Hanisch R, Kattner U, Kroenlein K, Newrock M, Peskin A, Plante R, Li SY, Rigodiat PF, Amaral GS, Trautt Z, Schmitt X, Warren J & Youssef S (2016). Informatics infrastructure for the materials genome initiative. *JOM* 68, 2053–2064.

Ecma International (2020). Introducing JavaScript object notation. Available at <https://web.archive.org/web/20200927011530/https://www.json.org/json-en.html> (retrieved September 28, 2020).

Gibson GA (1996). A brief history of LIMS. *Lab Autom Inf Manag* 32, 1–5.

Helu M & Hedberg Jr. T (2015). Enabling smart manufacturing research and development using a product lifecycle test bed. *Procedia Manuf* 1, 86–97.

Hipp RD (2020). SQLite. Available at <https://web.archive.org/web/20200927012340/https://www.sqlite.org/index.html>.

Jacobsen MD, Fourman JR, Porter KM, Worrig EA, Benedict MD, Foster BJ & Ward CH (2016). Creating an integrated collaborative environment for materials research. *Integr Mater Manuf Innov* 5, 232–244.

Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S & Willing C (2016). Jupyter notebooks—A publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, Loizides F & Schmidt B (Eds.), pp. 87–90. Amsterdam, Netherlands: IOS Press.

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 Stand Technol* 124, 124034.

Microscopy Society of America: Standards Committee (2019). MSA / MAS / AMAS Hyper-Dimensional Data File Specification- Version 1.02. <https://web.archive.org/web/20200921123650/https://www.microscopy.org/resources/>

- [scientific_data/HMSA_Specification-20191120.pdf](#) (accessed September 21, 2020).
- Nguyen P, Chan M, Mchenry K, Paquin N, Konstanty S, Nicholson T, O'Brien T, Schwartz-Duval A, Spila T, Nahrstedt K, Campbell RH & Gupta I (2017). 4CeeD: Real-time data acquisition and analysis framework for material-related cyber-physical environments. In *2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, pp. 11–20. IEEE. Available at <https://ieeexplore.ieee.org/document/7973684/>.
- Plante RL, Taillon JA, Lau JW, Greene GR & Newrock MW (2020). Nexus-Experiment: An XML schema for describing data collected from electron microscopes. *NIST Public Data Repository*. Available at <https://data.nist.gov/od/id/mds2-2245>.
- Rose S, Borchert O, Mitchell S & Connelly S (2020). Zero trust architecture. Tech. rep., National Institute of Standards and Technology, Gaithersburg, MD. Available at <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-207.pdf>.
- Scott JHJ (2015). Strategies for managing information technology (IT) in microscopy facilities. *Microsc Microanal* **21**, 373–374.
- Taillon JA, Devers RF, Plante RL, Newrock MW, Lau JW & Greene G (2019). Harvesting microscopy experimental context with a configurable laboratory information management system. *Microsc Microanal* **25**, 140–141.
- Taillon JA, Plante RL, Newrock MW, Greene GR & Lau JW (2020). NexusLIMS: A python package for EM experiment metadata management. *NIST Public Data Repository*. <https://doi.org/10.18434/mds2-2355>
- The PyInstaller Development Team (2020). PyInstaller. Available at <https://web.archive.org/web/20200926012000/http://www.pyinstaller.org/>.
- The World Wide Web Consortium (W3C) (2020). Cascading style sheets. Available at <https://web.archive.org/web/20201006201653/https://www.w3.org/Style/CSS/Overview.en.html> (retrieved October 6, 2020).
- Thomas D & Hunt A (2019). DRY—The evils of duplication. In *The Pragmatic Programmer, The Pragmatic Bookshelf*, 2nd ed. Boston, MA: Addison-Wesley. Available at <https://pragprog.com/titles/tpp20/the-pragmatic-programmer-20th-anniversary-edition/>.
- Tidwell D (2008). *XSLT*. Sebastopol, CA: O'Reilly.
- Vixie P, Mašláňová M, Dean C & Mráz T (2013). cron. Available at <https://web.archive.org/web/20200707115220/https://man7.org/linux/man-pages/man8/cron.8.html> (retrieved September 29, 2020).
- Vlist E (2002). *XML Schema*. Sebastopol, CA: O'Reilly.
- Warren JA & Ward CH (2018). Evolution of a materials data infrastructure. *JOM* **70**, 1652–1658.
- White RR & Munch K (2014). Handling large and complex data in a photovoltaic research institution using a custom laboratory information management system. *MRS Proceedings*, vol. 1654, mrsf13–1654–nn11–04. Available at https://www.cambridge.org/core/product/identifier/S1946427414000311/type/journal_article.
- Wikimedia Foundation, Inc (2020). Wikipedia: The free encyclopedia. Available at <https://web.archive.org/web/20200930035725/https://www.wikipedia.org/> (retrieved September 30, 2020).
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J & Mons B (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018.
- Wittenburg P, Hellström M, Zwölf CM, Abroshan H, Asmi A, Di Bernardo G, Couvreur D, Gaizer T, Holub P, Hooft R, Häggström I, Koureas D, Kuchinke W, Milanesi L, Rosato A, Padfield J, Staiger C, van Uytvanck D & Tobias W (2017). Persistent identifiers: Consolidated assertions. Tech. rep. GEDE: Group of European Data Experts. Available at https://www.rd-alliance.org/system/files/PID-report_v6.1_2017-12-13_final.pdf.
- Zakutayev A, Wunder N, Schwarting M, Perkins JD, White R, Munch K, Tumas W & Phillips C (2018). An open experimental database for exploring inorganic materials. *Sci Data* **5**, 180053.