Overview of NexusLIMS Usage and Development

Joshua A. Taillon

Northwestern University Meeting

Monday, May 1, 2023

NIST MATERIAL MEASUREMENT LABORATORY

NIST Disclaimer

Certain commercial equipment, instruments, materials, vendors, and software are identified in this talk for example purposes and to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Any opinions expressed are my own, and not a statement on behalf of the U.S. Government.

What do we mean by LIMS?

LIMS:

Laboratory Information Management System

Ideally start at the bottom of the pyramid, but scientific value comes at the top

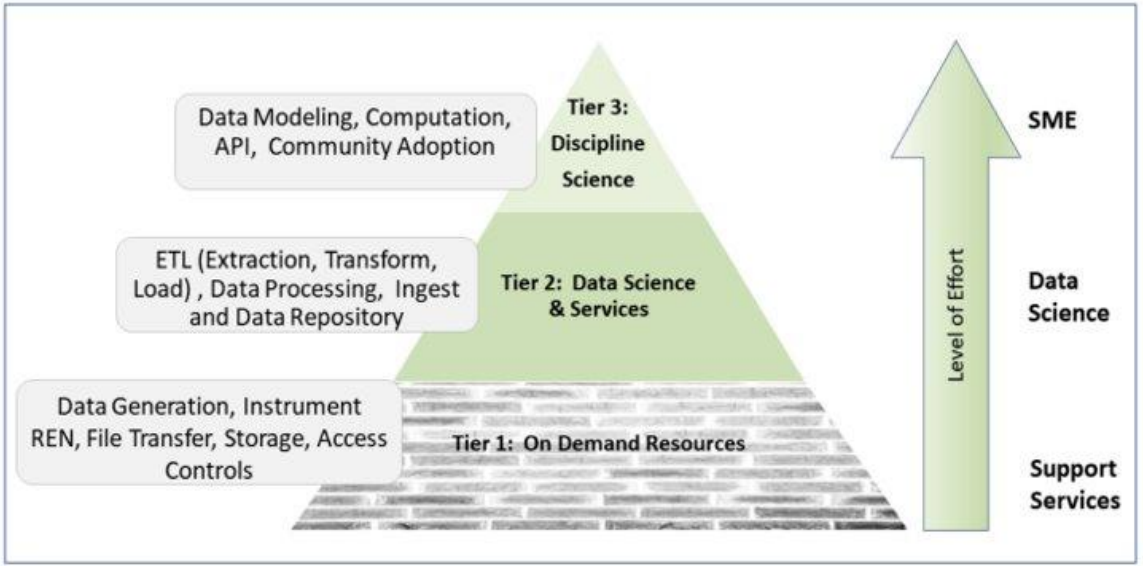


Fig. 1. LIMS three tiered model for implementation

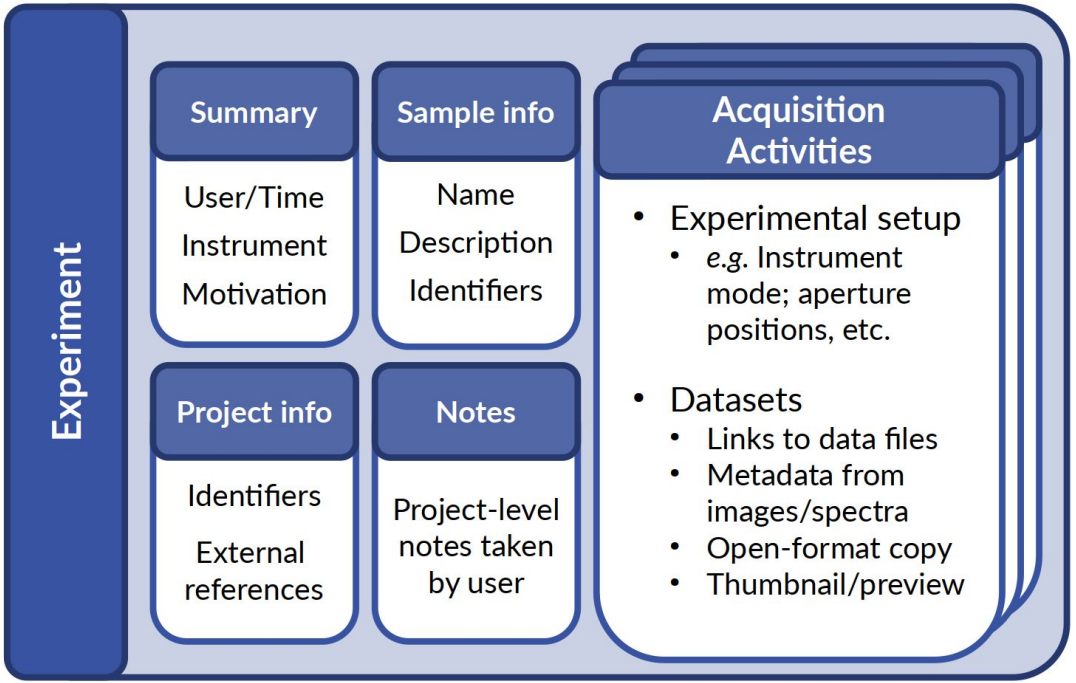
NIST Technical Note 2216 - https://doi.org/10.6028/NIST.TN.2216

NexusLIMS attempted to build from scratch

- Prior to community efforts (ca. 2018), we wanted to solve these issues for our shared microscopy facility
- Built NexusLIMS, a microscopy LIMS mostly from scratch
  - Open-sourced at https://github.com/usnistgov/NexusLIMS
  - DOI: 10.18434/mds2-2355
  - Described in detail in *Microscopy and Microanalysis*, 27 (3), 2021. pp. 511 - 527. 10.1017/S1431927621000222

Mapping EM workflows into a data model

- Data is most useful when intelligently structured
  - Allows browsing, querying, transforming, validating, etc.
- Structure should be tailored to context
  - What information could a researcher/manager/auditor want to see?
- A "record" represents an individual experimental session on microscope
- Schema published at https://doi.org/10.18434/M32245

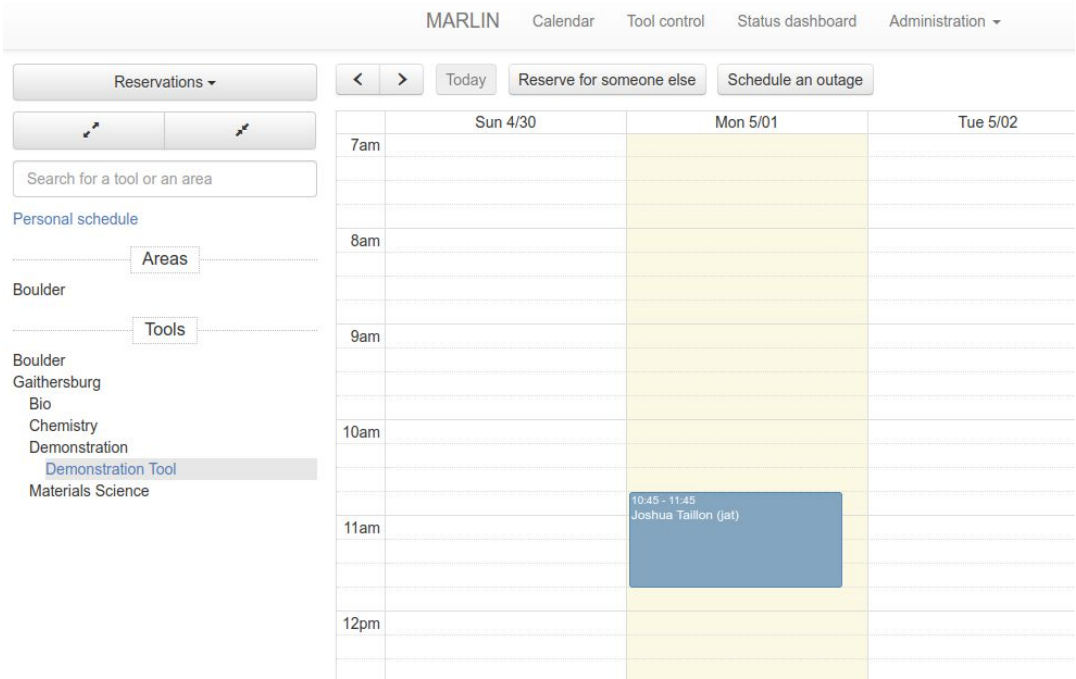


J.. Taillon, *et al.*, *Microscopy and Microanalysis*, vol. 25, no. S2, pp. 140–141, 2019.

What does it look like for users?

These slides (6-10) show features of NEMO/MARLIN, our facility management system

When a user wants to use a tool, they log in to "MARLIN" and make a reservation:



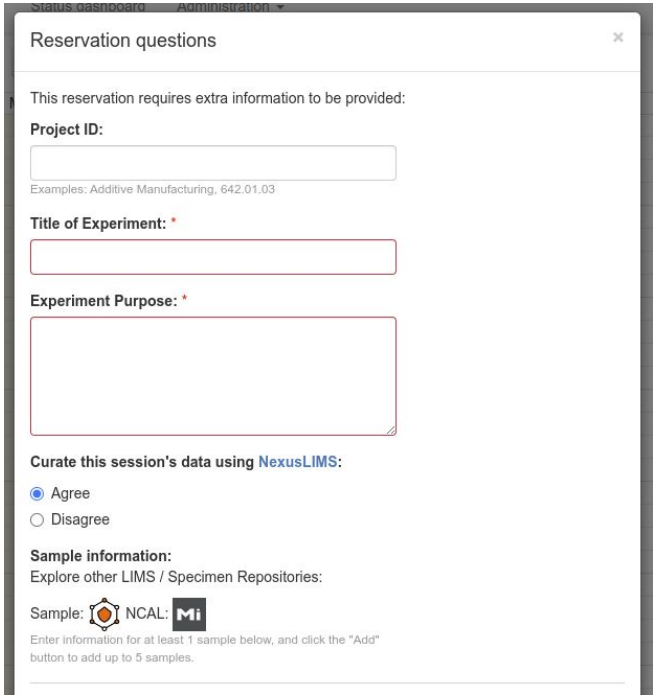
Select a tool, then click and drag to reserve a time

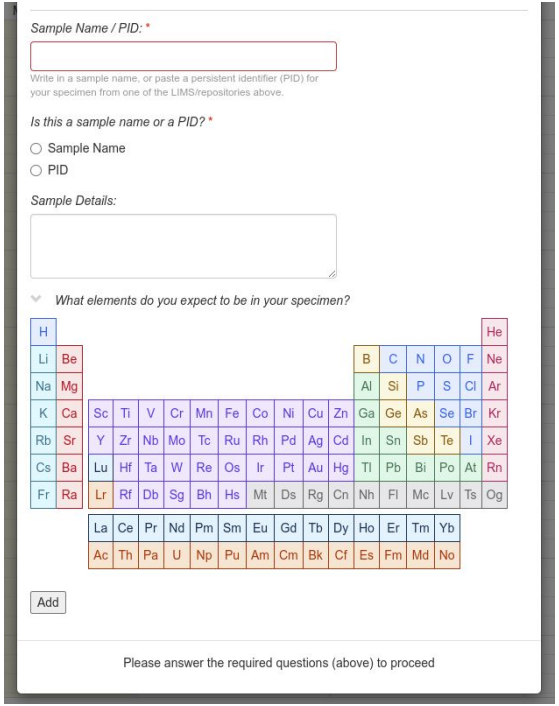
See https://github.com/usnistgov/NEMO

What does it look like for users?

These slides (6-10) show features of NEMO/MARLIN, our facility management system

When a user wants to use a tool, they log in to "MARLIN" and make a reservation:



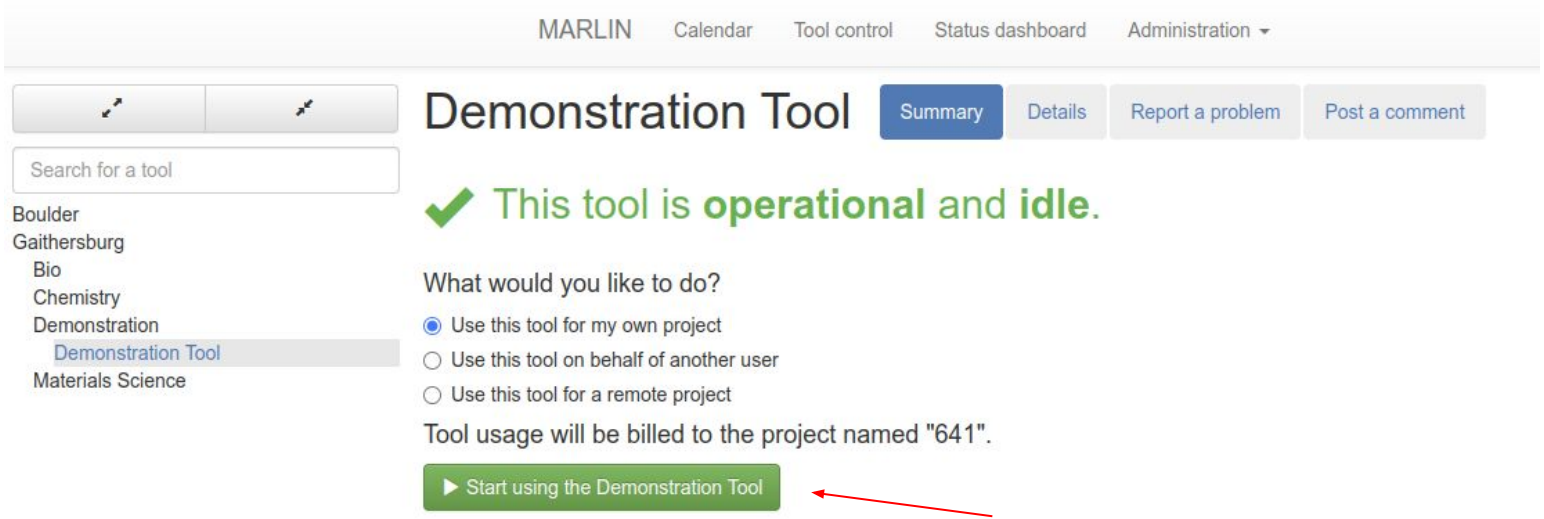


User fills out basic metadata about their experiment at reservation time

What does it look like for users?

These slides (6-10) show features of NEMO/MARLIN, our facility management system

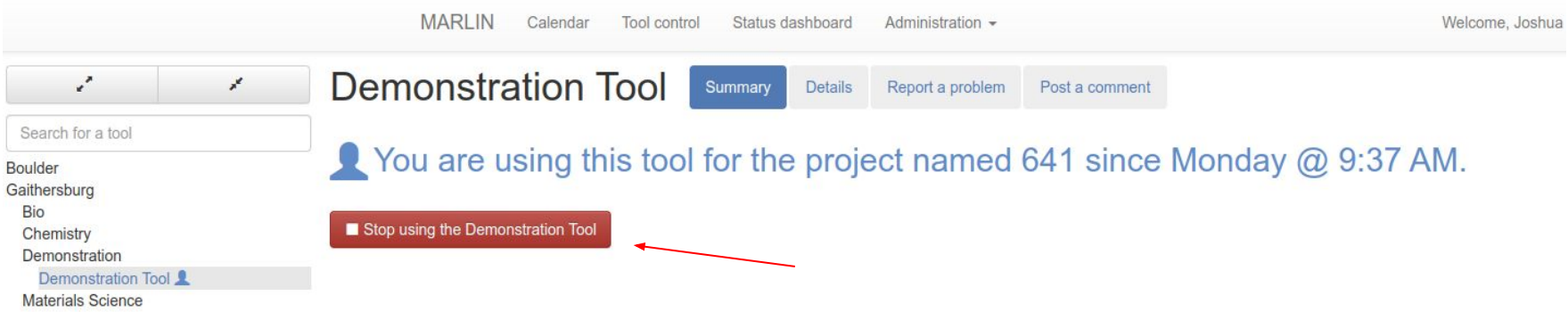
When it comes time to use a tool, they log in to “MARLIN” and enable it:



What does it look like for users?

These slides (6-10) show features of NEMO/MARLIN, our facility management system

When they are done, they log in to “MARLIN” and disable it:



Access control (if you want)

These slides (6-10) show features of NEMO/MARLIN, our facility management system

- Depending on desired level of control, NEMO/MARLIN can physically lock-out tools that are not enabled
- This can be done for billing
  - Could be used to ensure metadata entries are collected prior to tool use
  - May be otherwise useful for group/division management

Building of research records

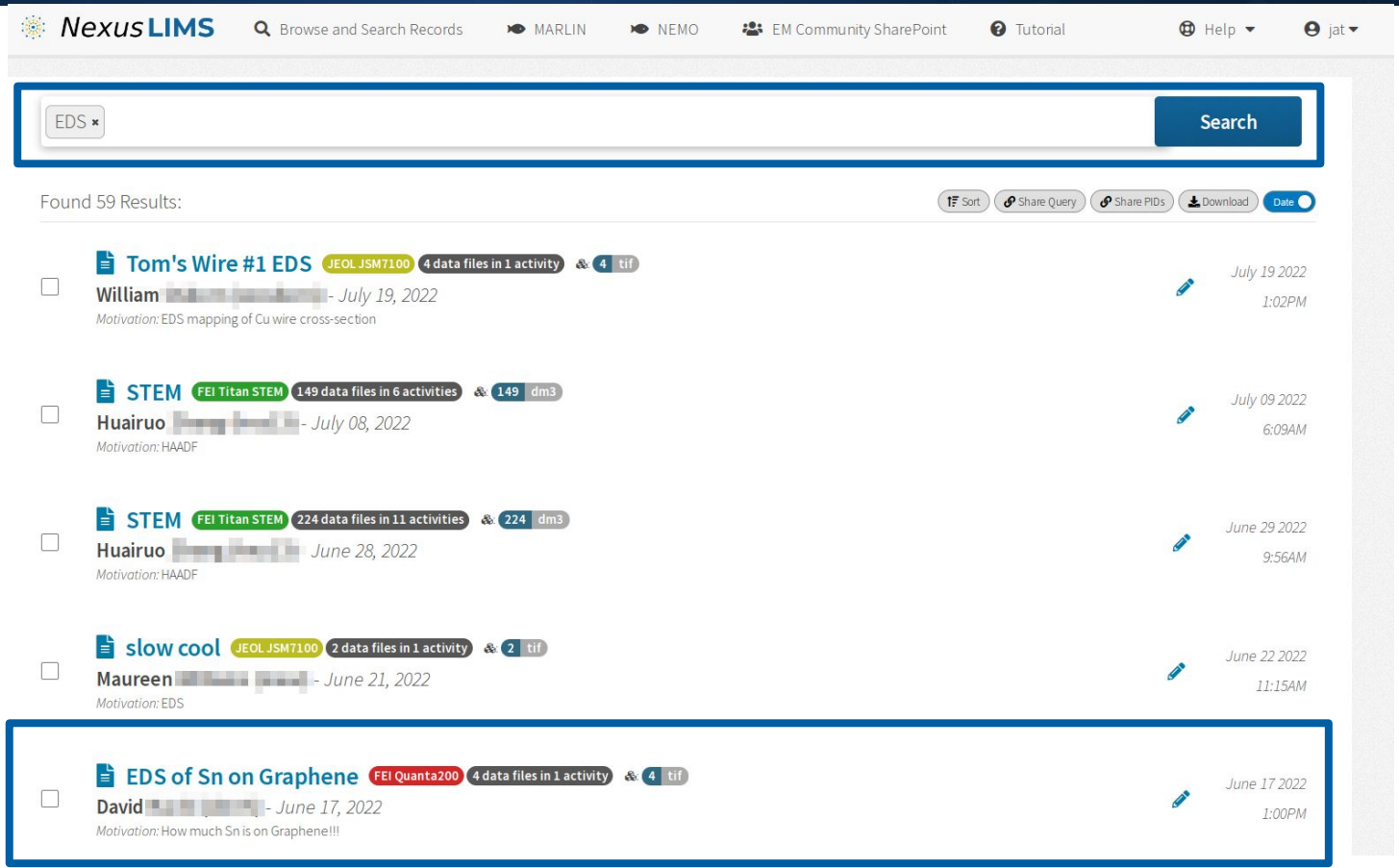
More details in later slides (slide 21); but once a user finishes their session, data is copied automatically to centralized file storage and a “record” of that session (matching the schema from slide 5) is built automatically and loaded into the NexusLIMS web interface (next slide)

Web-based user interface is based off open-source CDCS platform

Querying the database - Searching for records

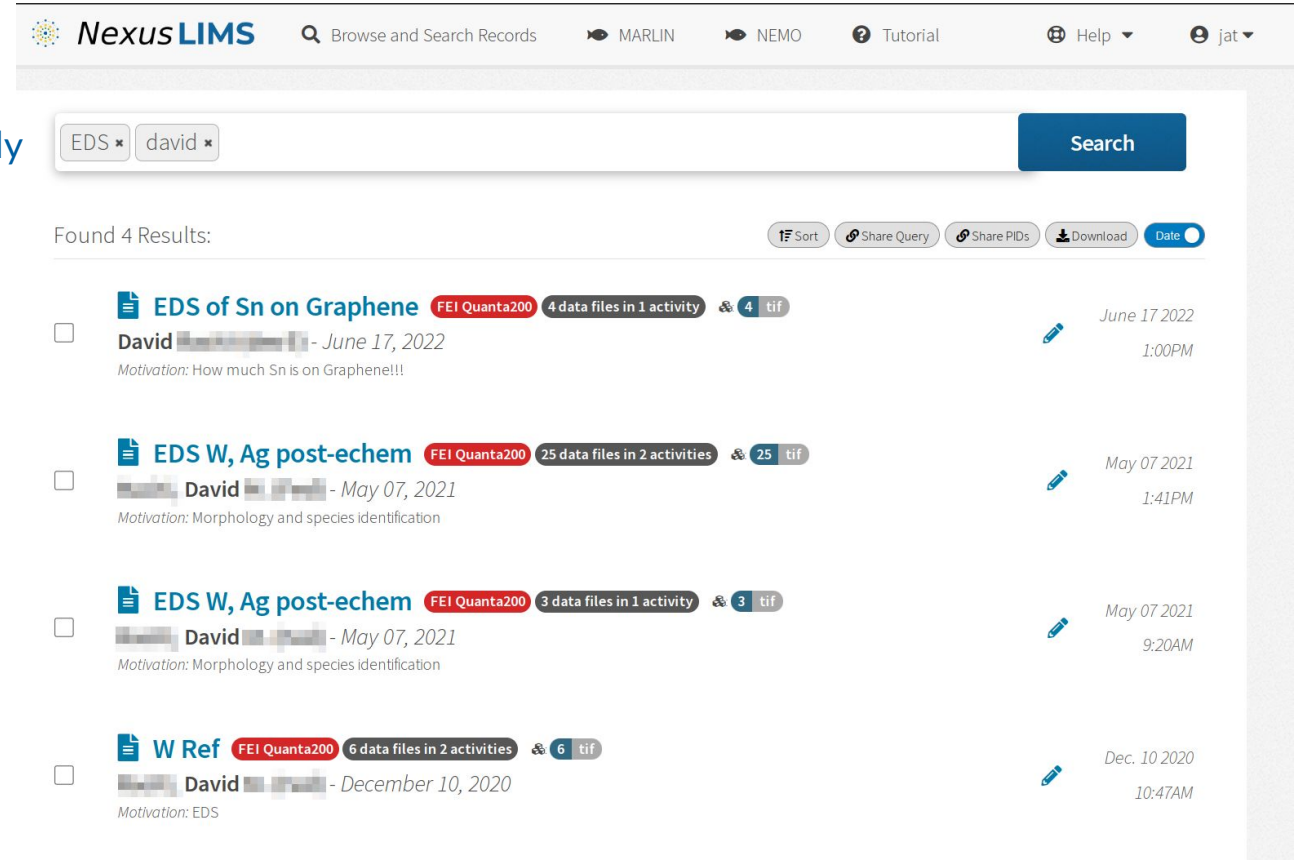
Free-text search for metadata terms

Summary display of record contents



Querying the database - Searching for records

Further refinement allows for quickly finding record of interest



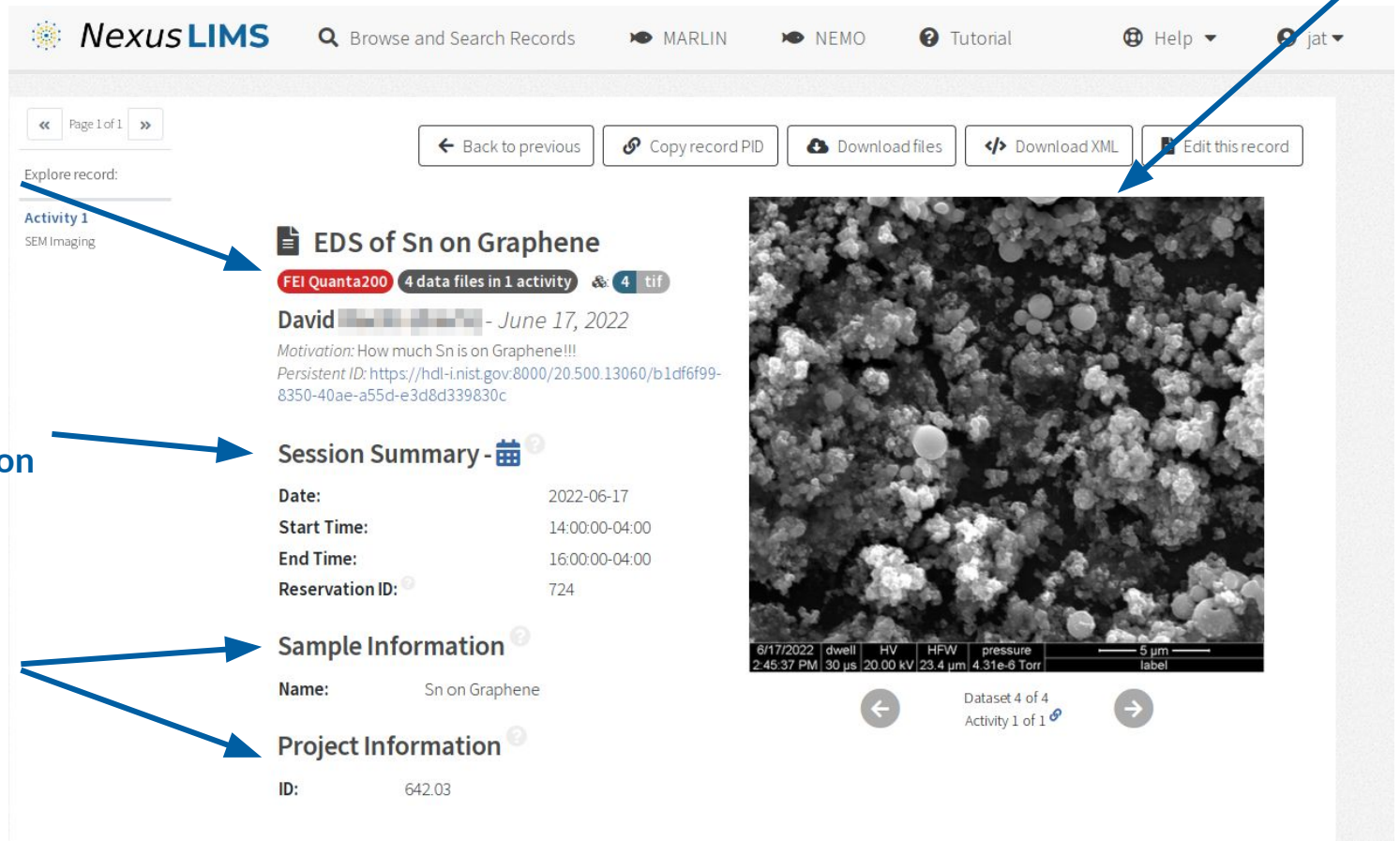
Viewing a record

Dataset preview gallery

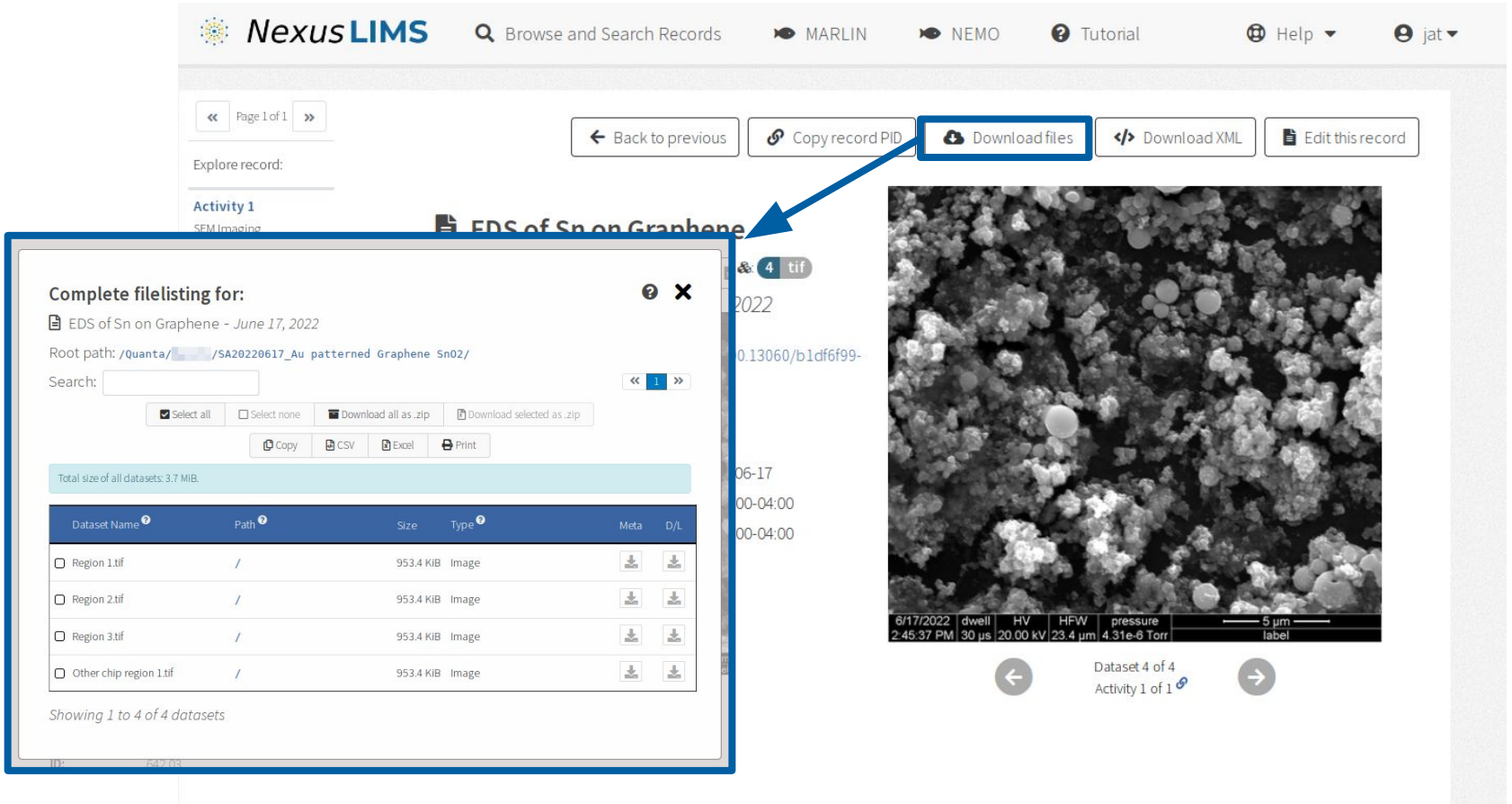
Summary info

Session/ reservation info

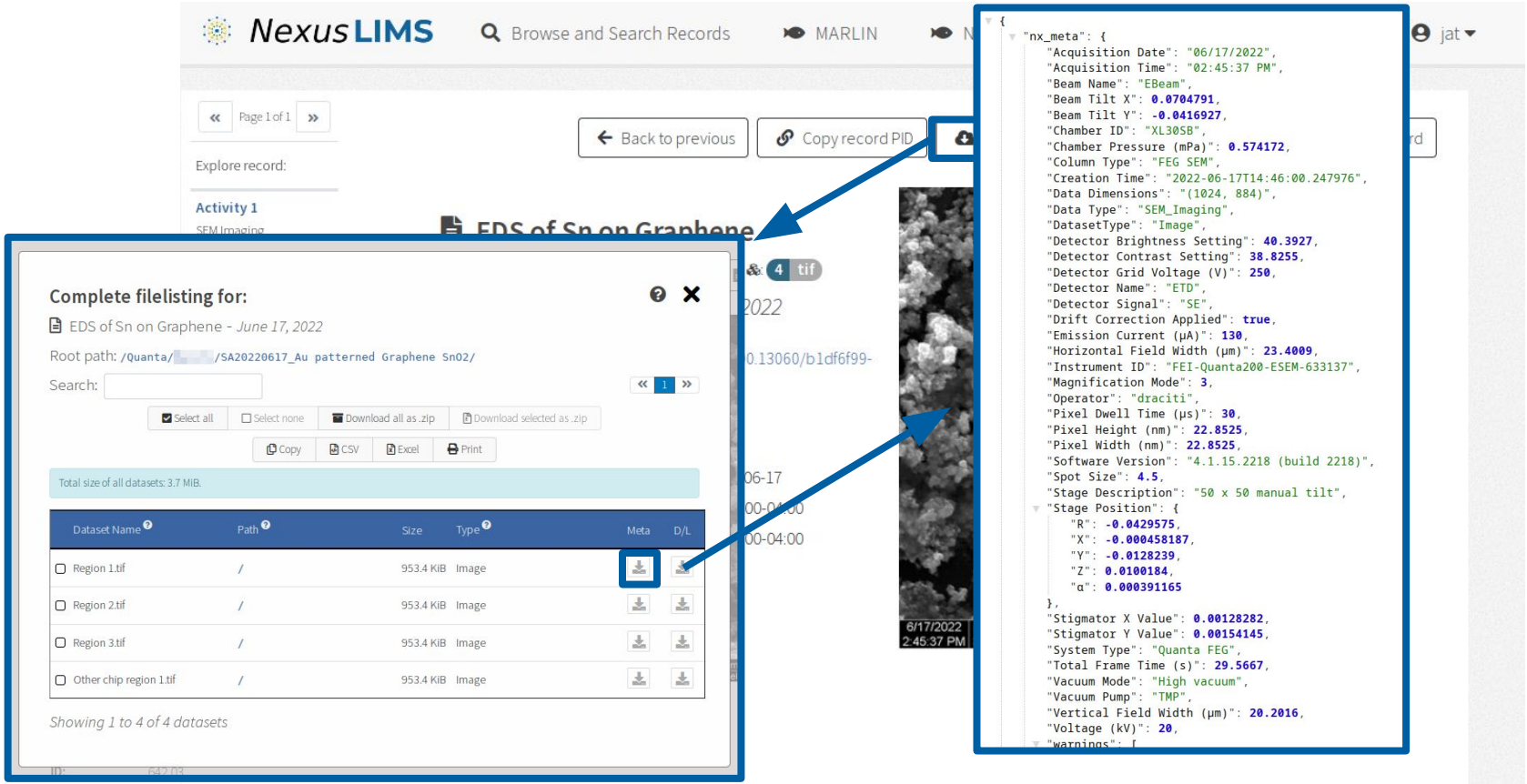
Sample and project



File downloading via Web UI



Viewing individual file extracted metadata

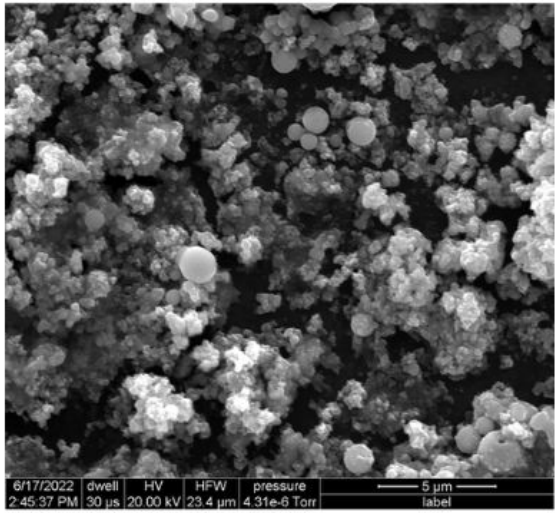


Viewing metadata specific to individual “activities”

Experiment activity 1

Activity contents: *SEM Imaging*

4 data files & 4 tif



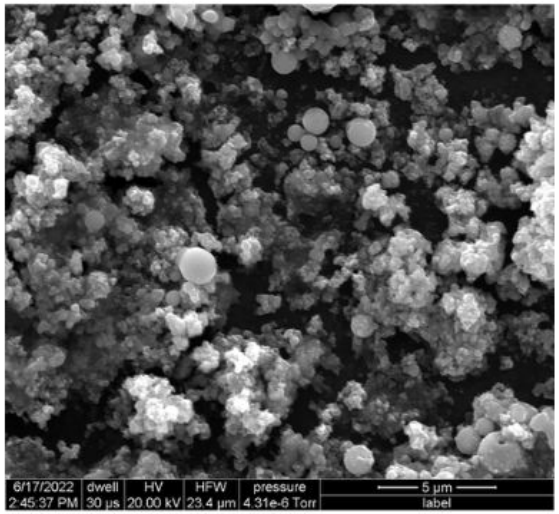
Dataset Name	Creation Time	Type	Role	Meta	D/L
Region 1.tif	2022-06-17 14:21	Image	Experimental		
Region 2.tif	2022-06-17 14:32	Image	Experimental		
Region 3.tif	2022-06-17 14:39	Image	Experimental		
Other chip region 1.tif	2022-06-17 14:46	Image	Experimental		

Viewing metadata specific to individual “activities”

Experiment activity 1

Activity contents: *SEM Imaging*

4 data files & 4 tif



Experiment activity 1

Activity contents: *SEM Imaging*

Search:

« 1 2 3 »

Setup Parameter	Value
Start time	14:21:06
Acquisition Date	06/17/2022
Beam Name	EBeam
Beam Tilt X	0.0704791
Beam Tilt Y	-0.0416927
Chamber ID	XL30SB
Column Type	FEG SEM
Data Dimensions	(1024, 884)
Data Type	SEM Imaging
Detector Grid Voltage (V)	250.0

Missing metadata?

Dataset N	ime	Type	Role	Meta	D/L
Region 1.1	7 14:21	Image	Experimental		
Region 2.1	7 14:32	Image	Experimental		
Region 3.1	7 14:39	Image	Experimental		
Other chi	7 14:46	Image	Experimental		

How's it going?

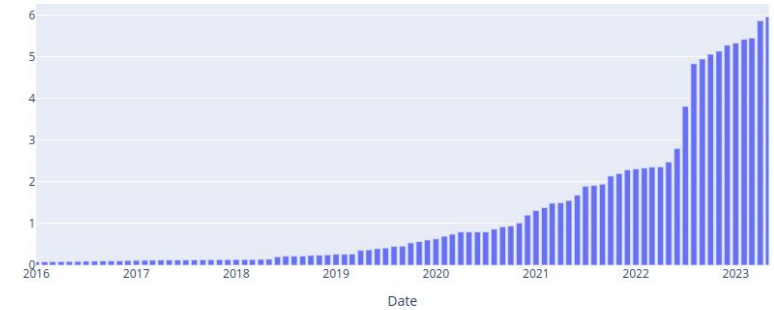
As of May 2023:

- 16 instruments "under management"
- ~ 800 individual "records" from ~ 90 users
- ~ 500 GB of files processed (mostly `.dm3/4` and `.tif`)
- Lots of types of files (EBSD, 4D-STEM, etc.) we're still not "capturing"

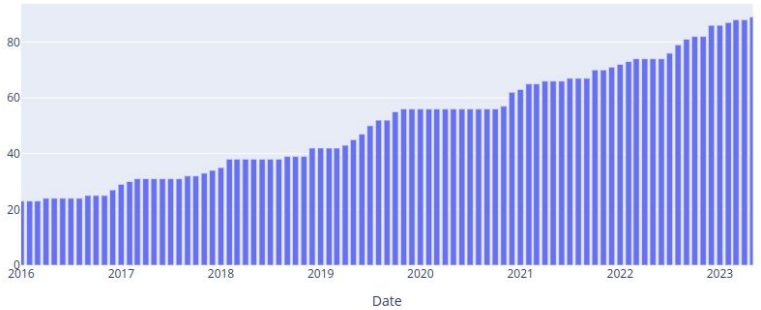
Nexus Data Explorer

Data Explorer Summary Stats

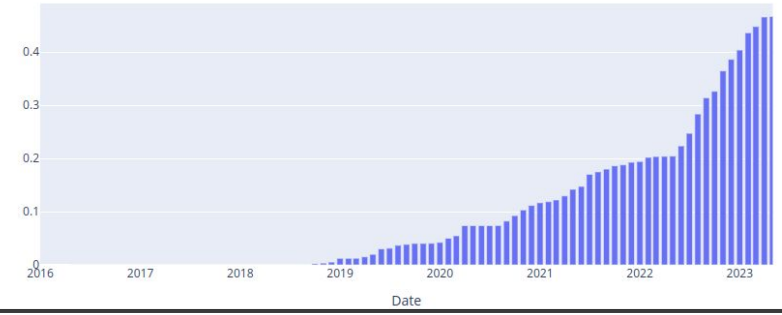
Total CFS data usage (TB)



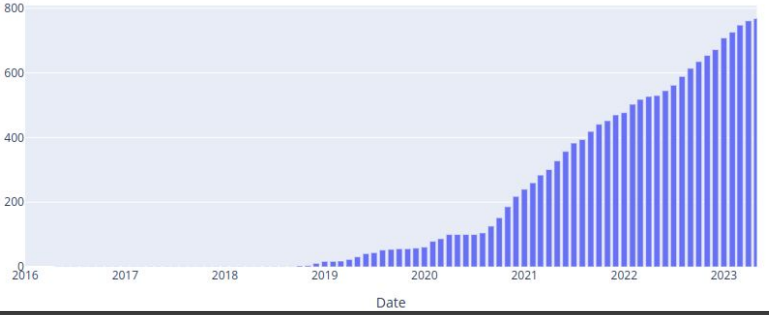
Users with files



Total Nexus data usage (TB)



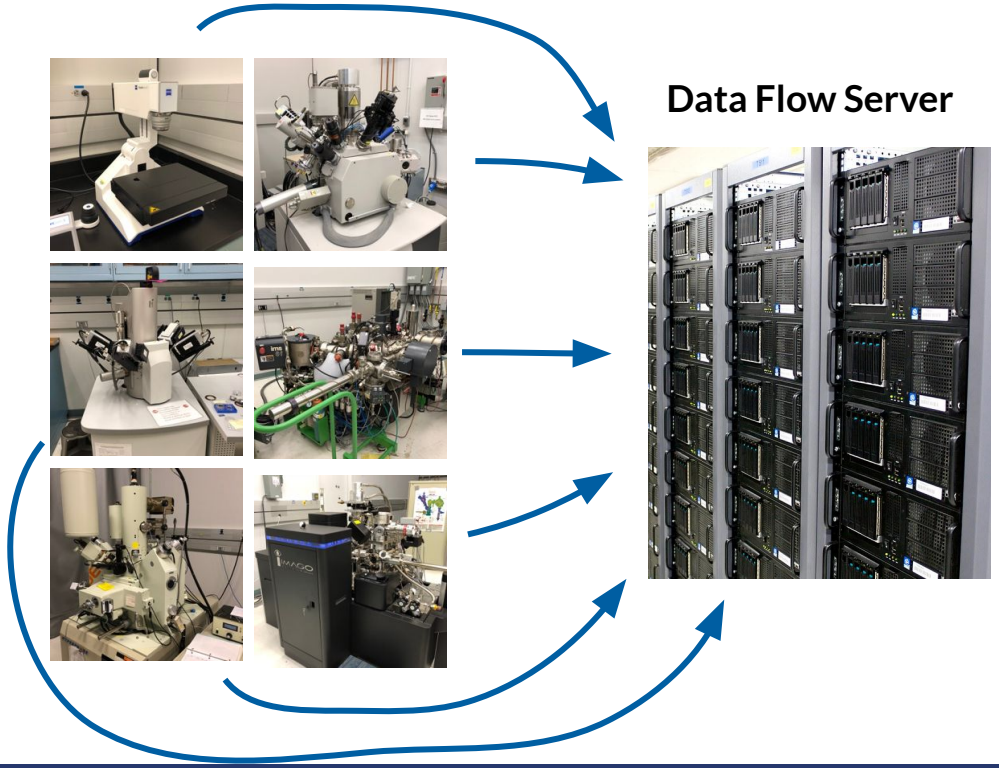
Number of Nexus records



How the sausage gets made…

- Make sure to check out documentation: https://pages.nist.gov/NexusLIMS
  - This should stay up-to-date
- In particular:
  - The record building docs: https://pages.nist.gov/NexusLIMS/record_building.html
  - The development docs: https://pages.nist.gov/NexusLIMS/development.html
- Assumption: instrument data somehow ends up in one central place
  - Diversion on next two slides describing what this looks like at NIST

Data “Plumbing”



Data Flow Server

Centralized storage; one folder per instrument PC with persistent names

InstrumentData

Name	Size	Modified
ABSciex-QTrap_MS-G000019	8 items	3/8/22 10:12 AM
Dell-servohydraulic_imaging_computer-G000003	4 items	1/4/22 10:46 AM
EDAX-Gemini_300_EBS-000025	1 item	4/11/22 4:40 PM
EDAX-LEO_1525_EDAX-000022	1 item	4/11/22 3:53 PM
FEI-Helios_FIB_SEM-G000025	63 items	7/28/22 2:57 PM
FEI-Quanta_200F_SEM-G000007	57 items	7/15/22 12:17 PM
FEI-Quanta_400_SEM-000023	1 item	4/7/22 3:29 PM
FEI-Quanta_Bruker-G000008	70 items	5/19/22 9:03 PM
FEI-Titan_80_300_STEM-G000020	18 items	7/15/22 4:42 PM
FEI-Titan_TEM-G000021	26 items	4/15/22 6:05 PM
Gatan-K2_IS-G000022	5 items	7/7/22 8:12 AM
Hitachi-S4700-SEM-606559	2 items	3/5/21 9:35 AM
Illumina-MiSeq_FGx_DNA_Sequencer_Server-G000023	2 items	7/27/22 4:40 PM
Illumina-MiSeq_FGx_DNA_Sequencer-G000023	8 items	7/5/22 10:39 PM
JAWoollam-A330_glove_box_ellipsometer-G000001	81 items	6/21/22 12:07 PM
JAWoollam-A330_insitu_ellipsometer-G000002	10 items	3/3/22 11:00 AM
JEOL-3010_Gatan_S_TEM-G000012	4 items	3/30/22 4:37 PM
JEOL-3010_Strobo_S_TEM-G000013	7 items	3/30/22 5:08 PM

As of July 2022:

- 36.7 TB of data harvested from 66 instruments on 2 campuses

Data “Plumbing”

- Automates data flows from instruments across MML’s scientific laboratories into one or more centralized location(s)
- Each PC shares a read-only folder
  - This folder becomes the new “data” folder for users on the instrument
  - Users can use any folder hierarchy they wish - helpful to use usernames
- Networked server periodically copies all data (rsync) to centralized storage
- Instruments are added via user-submitted form and automated script

The core pieces:

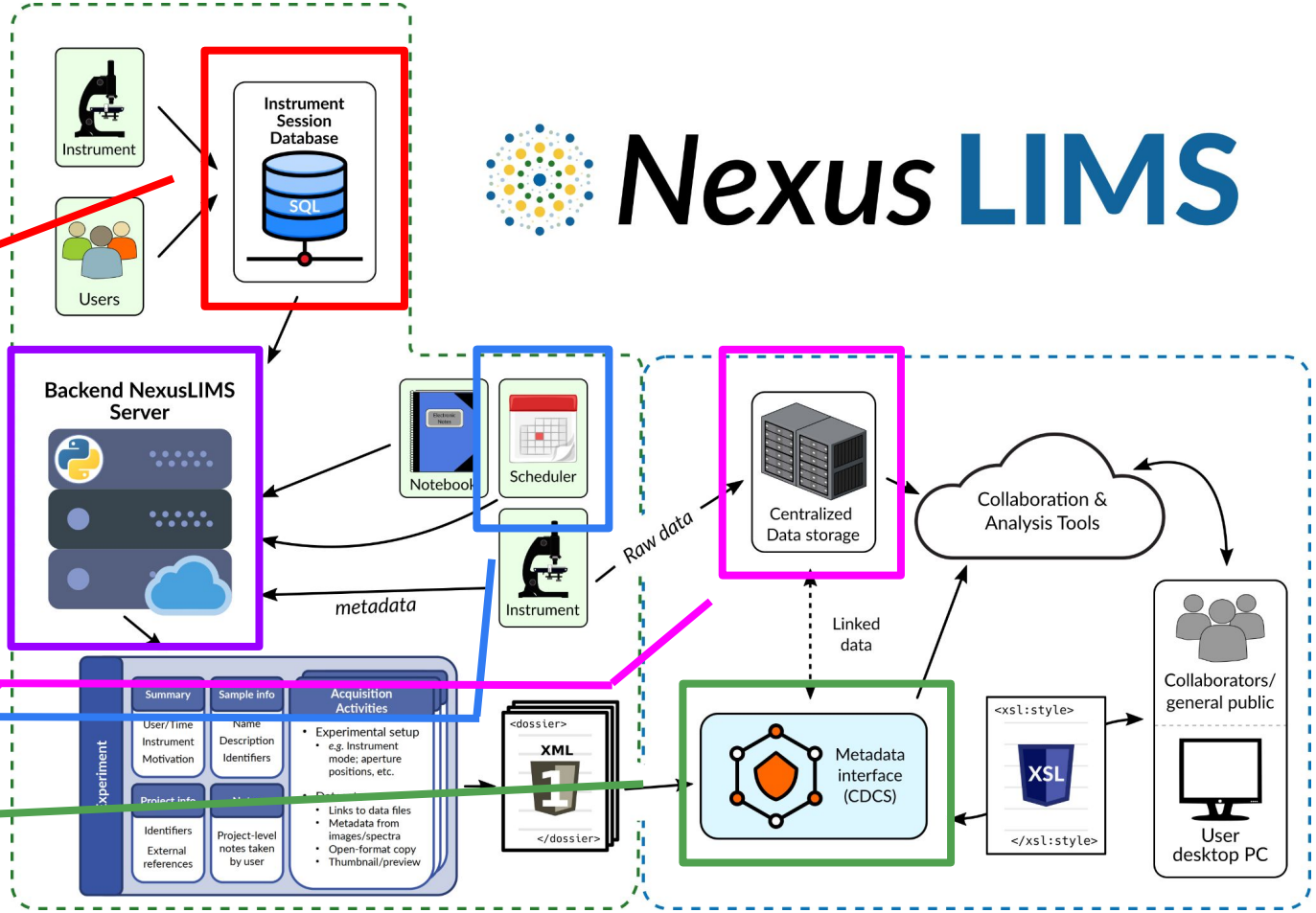
Backend Database

Record building server

Data storage

Equipment scheduler

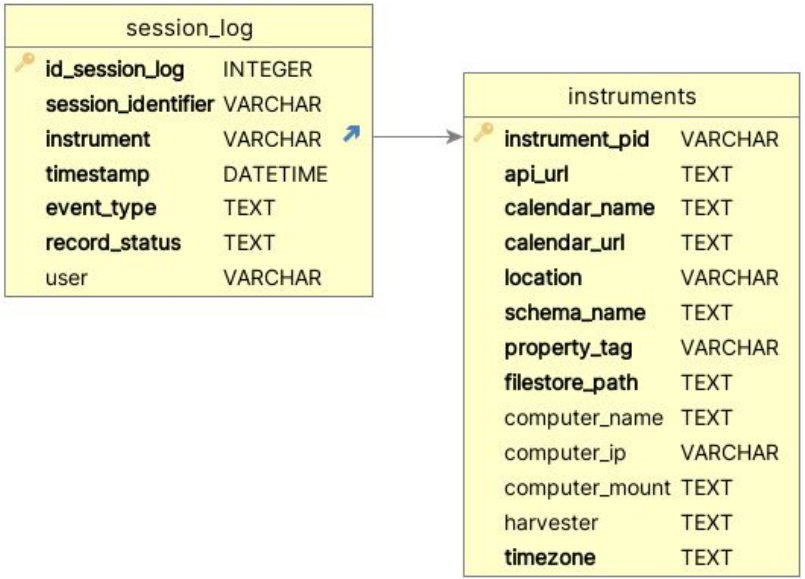
Front-end Web UI



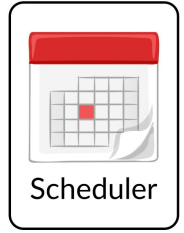
Nexus LIMS

NexusLIMS Backend Database

- Very simple model implemented In SQLite:



Scheduler Data Access



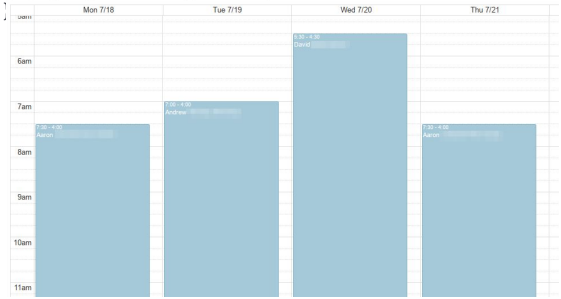
API access is critical for non-interactive access (NEMO has this baked in)

Reservations


```
{
  "id": 246,
  "question_data": {
    "project_id": "Hydrogen",
    "experiment_title": "Deformation evolution",
    "experiment_purpose": "Compare
microstructures
                          after various ...",
    "data_consent": "Agree",
    "sample_group": {
      // could have additional samples defined
      "0": {
        "sample_name": "4130-no strain",
        "sample_or_pid": "Sample Name",
        "sample_details": ""
      }
    }
  },
  "creation_time":
"2022-01-18T15:48:10.987314-07:00",
  "start": "2022-02-02T08:00:00-07:00",
  "end": "2022-02-03T16:00:00-07:00",
  "user": 2,
  "tool": 15,
  "project": 14
}
```

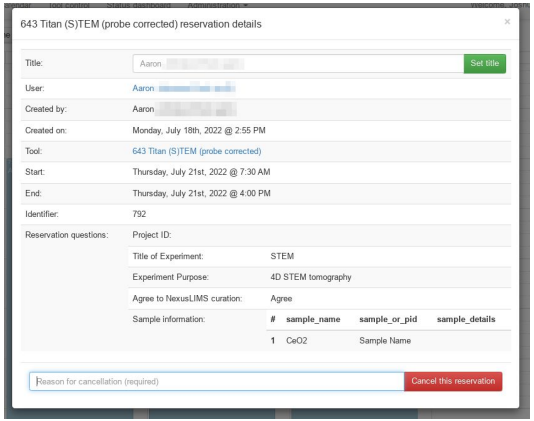

Usage Events


```
{
  "id": 51,
  "start":
"2022-01-21T08:20:53.879161-07:00",
  "end":
"2022-01-24T06:45:55.363185-07:00",
  "run_data": "",
  "user": 2,
  "operator": 2,
  "project": 13,
  "tool": 15
```

Tools


```
{
  "id": 15,
  "timezone": "America/New_York",
  "name": "642 JEOL 3010",
  "_description": "Stroboscopic TEM,
Thermionic
                   LaB6 emitter, 300 keV",
  "_image": "http://******.nist.gov/media/
            tool_images/642-jeol-3010.png",
  "_tool_calendar_color": "#33ad33",
  "_category": "Gaithersburg/(S)TEM",
  "_location": "223 A132",
  "_phone_number": "301-975-2000, x12345",
  "_notification_email_address":
"xyz.abc@nist.gov",
  "_superusers": [ 2 ]
}
```

Record building process

```
python -m nexusLIMS.record_builder
```

Find new sessions

Check for new `usage_events` for our instruments using NEMO API

Add each new usage event to NexusLIMS `session_log` table

For each, return `session_handler.Session` object with instrument, timestamps, and user info

For each `Session`...

Find matching reservation for this instrument and timespan (with reservation question answers)

Find files created on by this instrument during the usage event timespan

Cluster files by creation time, and extract metadata/create preview for each

Generate XML record from calendar information and files

Upload to front-end

Using web interface API, upload XML directly to CDCS

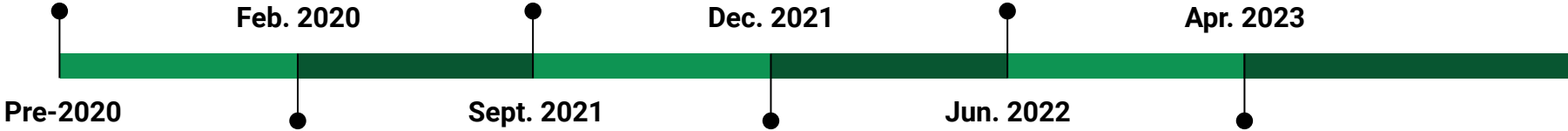
CDCS assigns a persistent identifier (PID) to each record via a local handle server deployment

Entire process runs without user interaction on a configurable 15 minute interval

Development Process

- Development takes place on using internal Gitlab project with issues, merge requests, CI/CD pipelines, etc.
- Releases periodically pushed to public repo at https://github.com/usnistgov/NexusLIMS
- Backend is a Python 3.9+ library with 100% test coverage (*caveat*) and auto-built documentation
- Most development has been done by me, with a few student contributions
- Recently added another part-time resource to work on new features (mostly extractors, at this point)

Development timeline



Pre-release development

Research on alternative data management systems; initial schema design; initial extractors for FEI FIB/SEM and .dm3 files

First official release

Automated record building using SharePoint calendar and "Session Logger App"; Implemented "acquisition activity" bundling; Improvements in data display and backend configuration

v 1.0.1

Added extractor for FEI .ser/.emi TEM files; Provided file downloading tool in front-end; Improved documentation; Bug fixes

v 1.1.0

Support for NEMO calendar and "reservation questions"; Moved project to poetry for package and dependency management

v 1.1.1

Bug fixes; Improved logging and display of NEMO reservation information in front-end; Deprecated SharePoint harvester

v 1.2.0 & 1.3.0

Added minimal extractor for otherwise non-supported files; non-supported files can now be included in built records; Added extractor for EDAX EDS spectra; Bug fixes and documentation enhancements

What else can/should we do?

- Automated metadata extraction from *all* research files, not just those managed by NexusLIMS
- Tools to query and find data by user, instrument, or any other arbitrary metadata
- Additional institutional data sources:
  - Organization-wide instrument database with persistent identifiers
  - Project database; Sample database
- Generalizing capabilities across MML and lowering barrier to entry

What have we learned from NexusLIMS?

- It's extremely hard to do everything yourself!
- If you want to use it, data must be centralized and accessible
- Our problems (mostly) are not particularly unique to microscopy
- As an organization, we need to invest in data-first infrastructure
    - Infeasible to repeat NexusLIMS process for every project, group, etc.

Thank you for your attention!
Questions?

joshua.taillon@nist.gov
https://orcid.org/0000-0002-5185-4503