Lessons Learned Building a Modern Microscopy Data Ecosystem at NIST

Joshua A. Taillon

Presentation for USGS Brown Bag Lunch Group

Friday, October 28, 2022

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.

Personal Disclaimer

Lessons “learned” does not mean
we’re not still learning….

We are still in the process of building
(and probably always will be)

Efforts like these involve huge teams of people

Acknowledgements

NIST Office of Data and Informatics

- June Lau
- Gretchen Greene
- Marcus Newrock
- Ray Plante
- Ryan White (detail)
- Mike Katz (detail)

NIST MML IT Team

- Gary Hardin
- Ann Leith
- Michael LaRue
- Sergiy Domalevskyy

MML Microscopy Users

- Mike Katz (again)
- Andy Herzing
- Will Osborn

Northwestern CHiMaD

- Laura Bartolo
- Roberto dos Reis

euclid TECHLABS

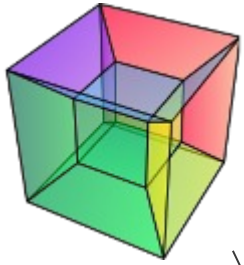
- Ao Liu
- Weinan Si

NIST MML LIMS Community of Interest

- Jared Ragland
- Zachary Trautt
- Adam Creuziger
- Chandler Becker
- Joseph Bennett
- Niksa Blonder
- Lisa Borsuk
- Carelyn Campbell
- Adam Friss
- Lucas Hale
- Michael Halter
- Robert Hanisch
- Lyle Levine
- Samantha Maragh
- Sierra Miller
- Christopher Muzny
- John Perkins
- Anne Plant
- Bruce Ravel
- David Ross
- John Henry Scott
- Chris Szakal
- Alessandro Tona
- Peter Vallone

About Me

Materials Characterization



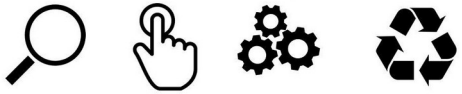
Hyperspy

General scientific programming

Databases (SQL)

Dashboarding/app design

Findable Accessible Interoperable Reusable



NIST MML LIMS Community of Interest (COI)

LIMS:

Laboratory Information Management System

COI brings together interested researchers from across the laboratory to share knowledge and resources

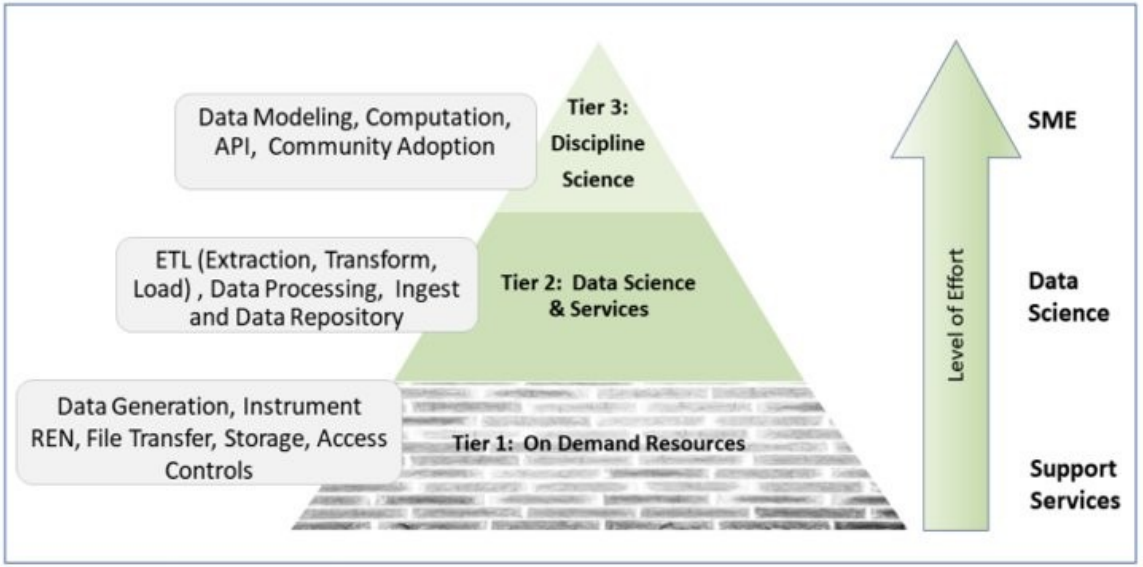
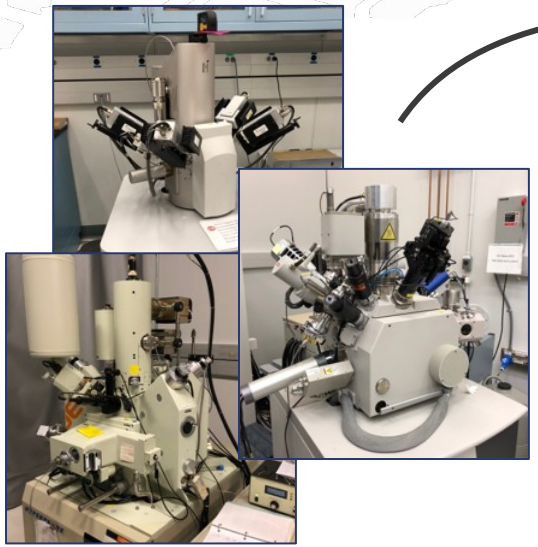


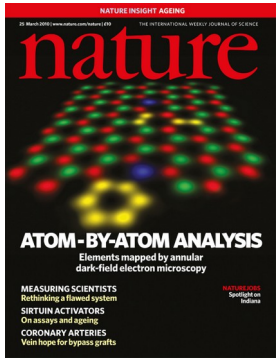
Fig. 1. LIMS three tiered model for implementation

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

Our Data Problem…

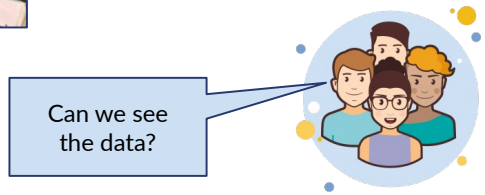


sophos.com

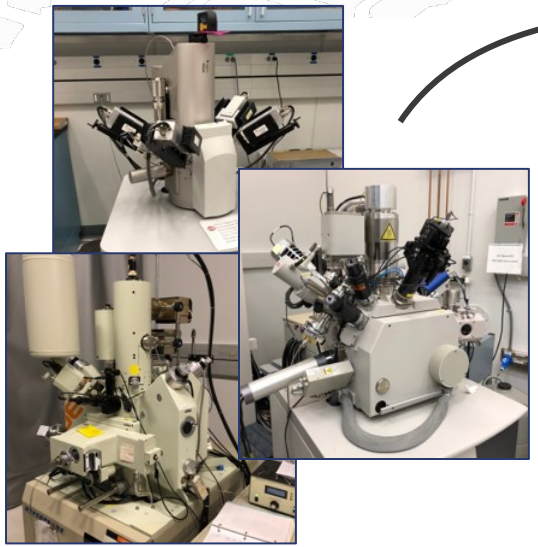


Vol 464 (7288), 2010. Nature.

engadget.com



Our Data Problem…



1

sophos.com

How do we get the data off the microscopes to a place where we can work with it?

Our Data Problem…

sophos.com

2

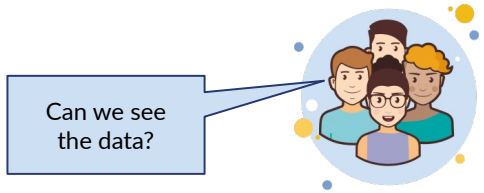
Once we're "done" with it, how do we store it long term?
(and how long is that?)

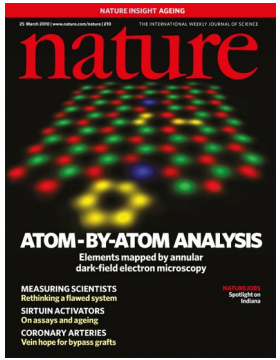
engadget.com

Our Data Problem…

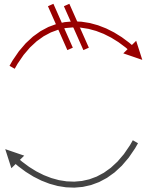
What do we do with requests for data? How do we find data?

How do we associate that data with our great publications?





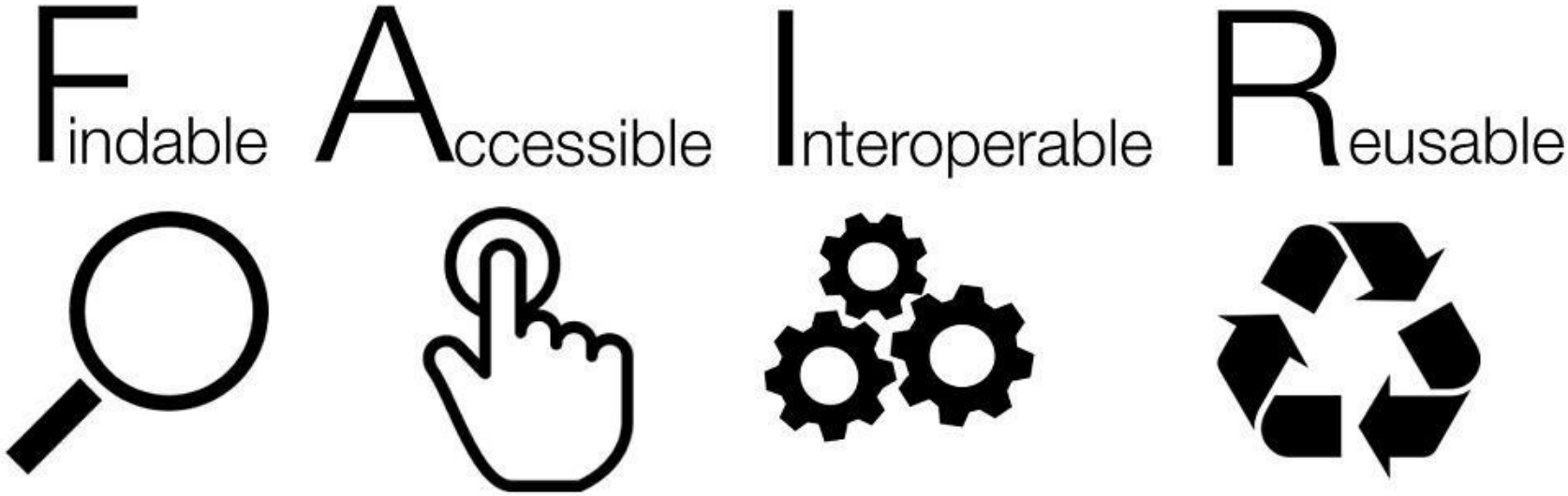
Vol 464 (7288), 2010. Nature.



3

engadget.com

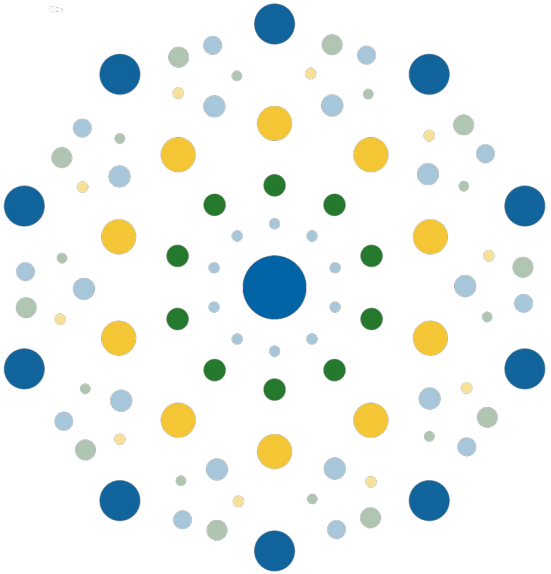
FAIR Data Principles



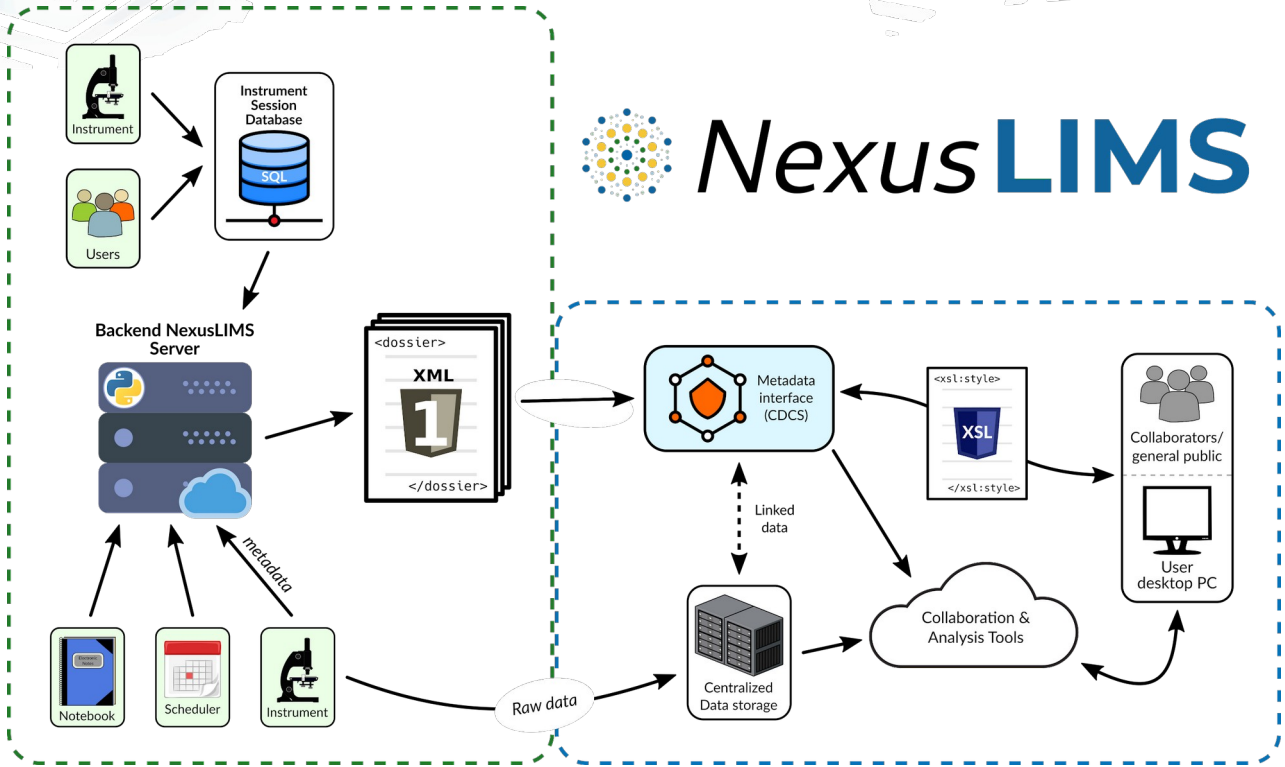
Wilkinson *et al.*, *Scientific Data*, **3**, 160018, 2016 (link)
Image: Sangya Pundir - CC-BY-SA 4.0

Our EM-focused effort

- Prior to community efforts (ca. 2018), we wanted to solve these issues for our shared microscopy facility
- Built 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



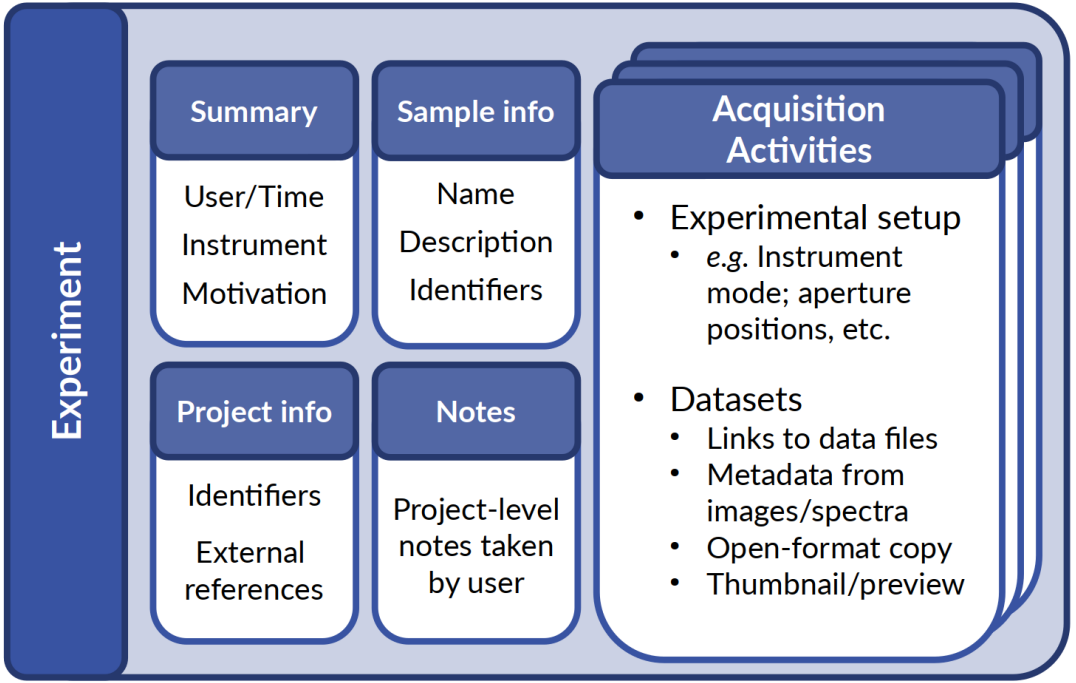
What does our LIMS for microscopy look like?



NexusLIMS

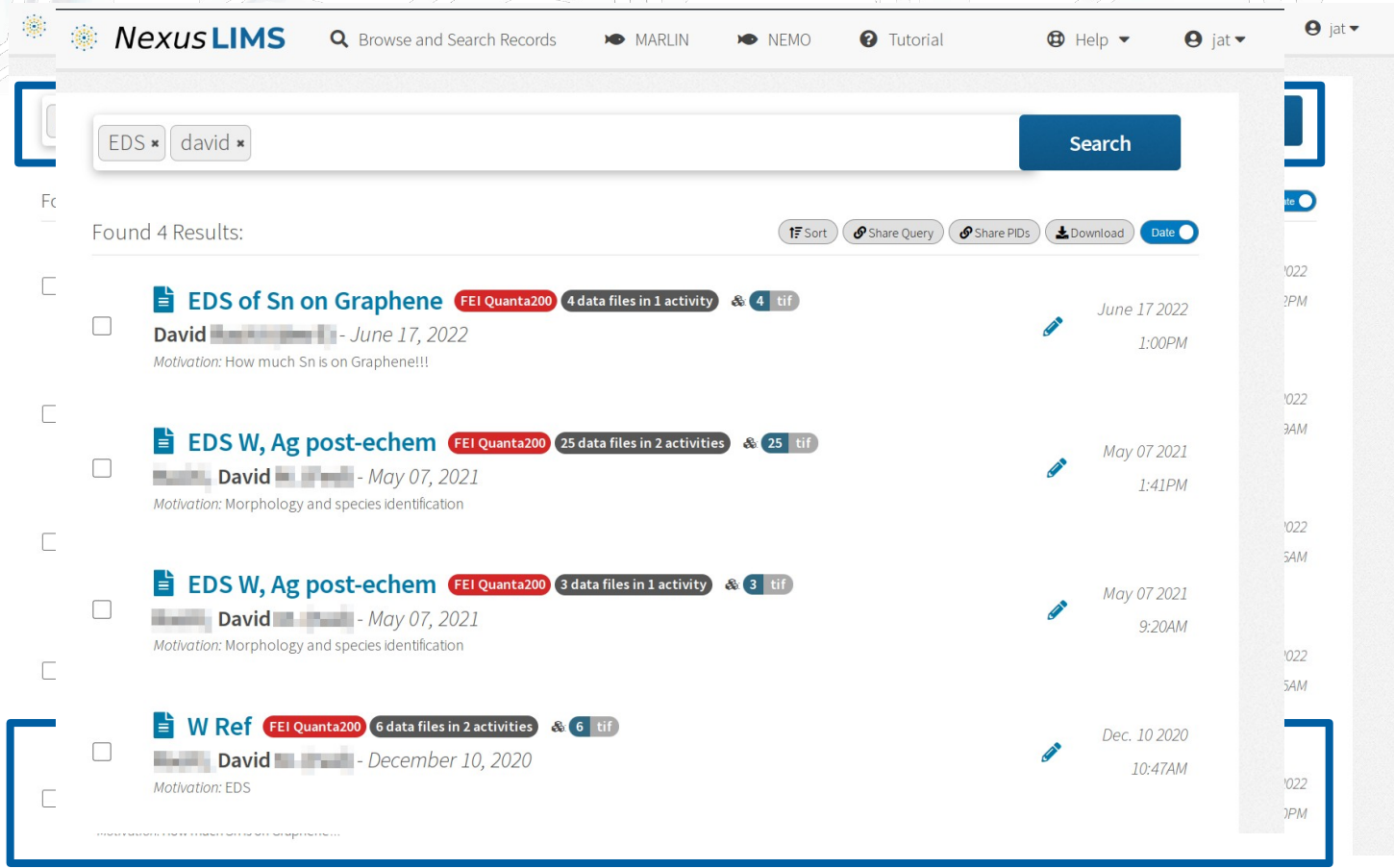
Mapping EM workflows into a data model

- 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

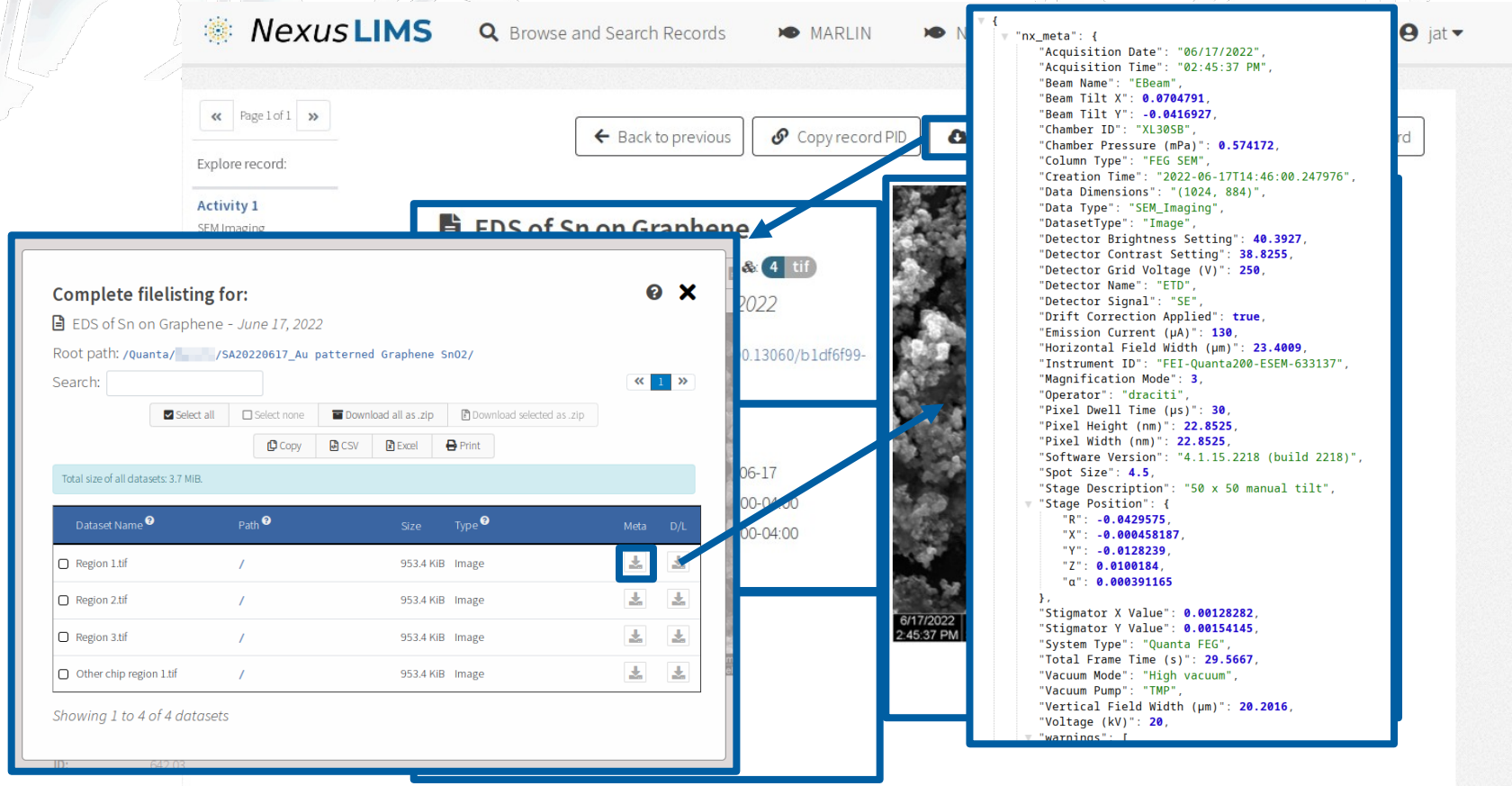


J.. Taillon, *et al.*, *Microscopy and Microanalysis*, vol. 25, no. S2, pp. 140–141, 2019.

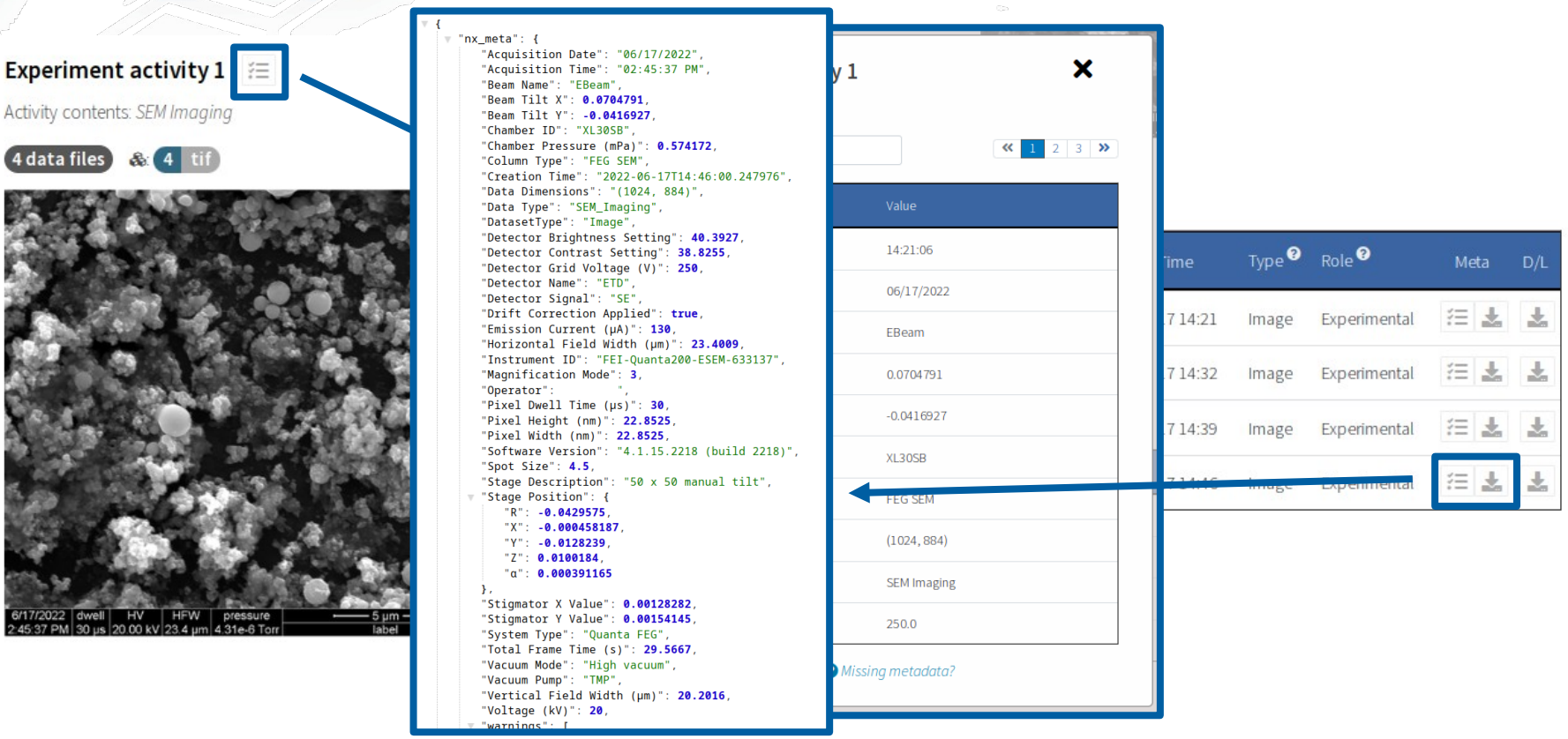
Querying the database



Browsing and previewing (meta)data



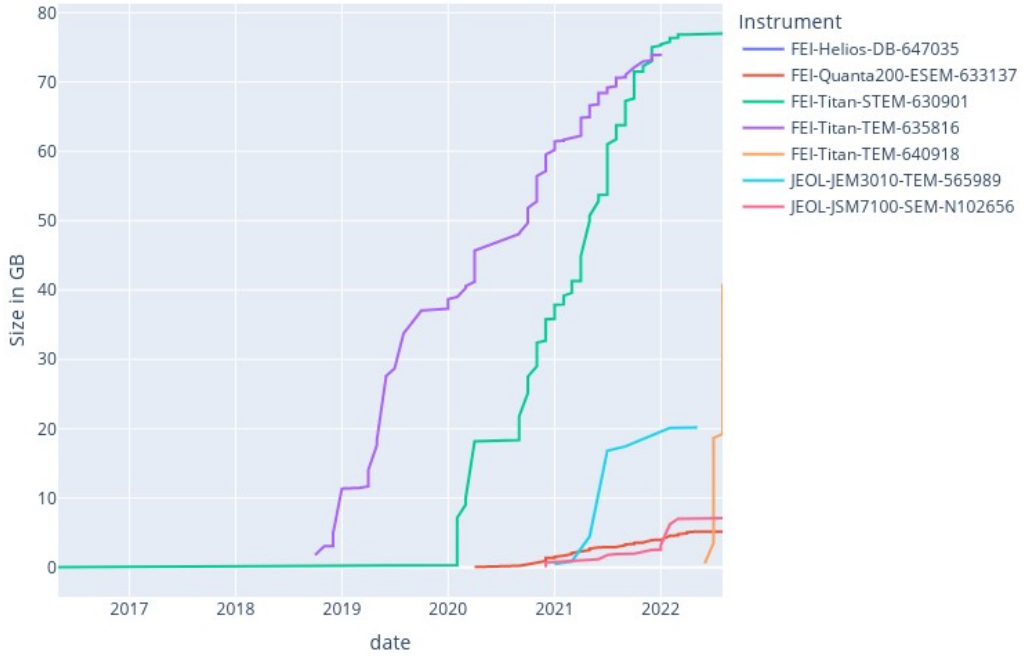
Browsing and previewing (meta)data



How's it going?

As of July 2022:

- 10 instruments "under management"
- ~ 600 individual "records" from ~ 40 users
- ~ 240 GB of files processed (mostly `.dm3/4` and `.tif`)
- New instruments being added regularly



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.

The LIMS “pyramid”

With NexusLIMS, we built most of the pyramid

Now, a focus on building out common infrastructure that all research can benefit from

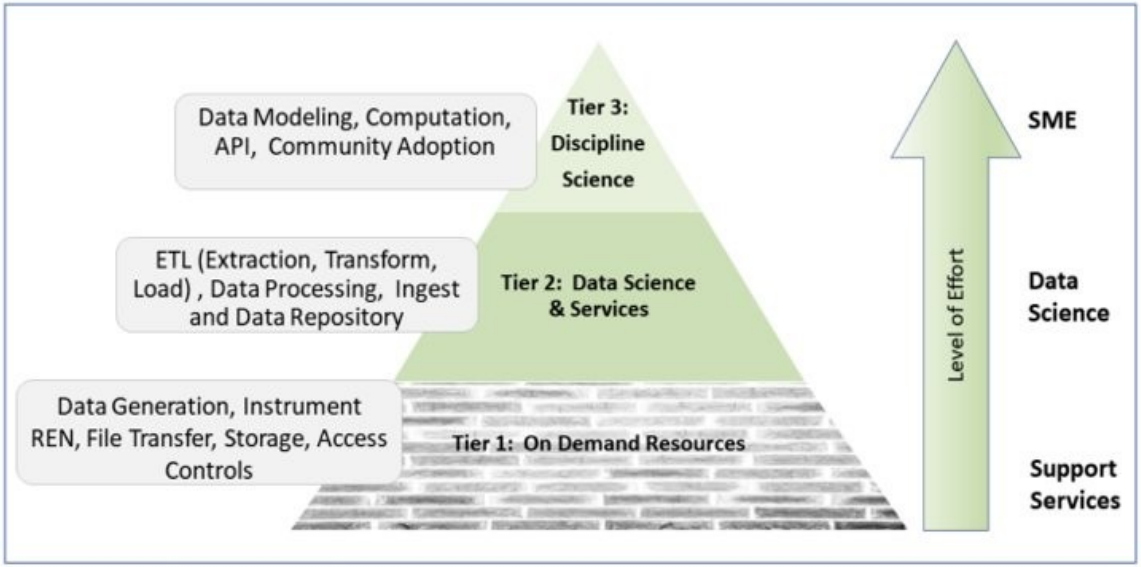


Fig. 1. LIMS three tiered model for implementation

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

An analogy…

Building “off the grid”

Septic, solar panels, battery storage, well water, etc.

Building in city limits

City provides electric, gas, water, trash, etc.

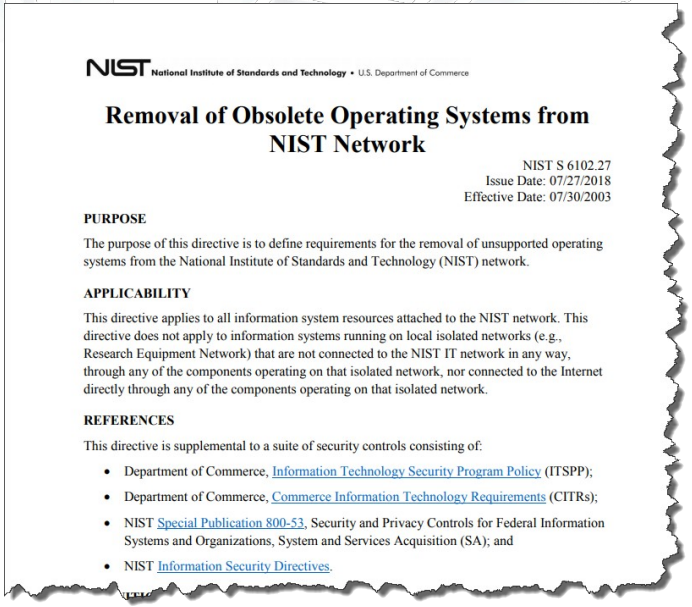
Parts of the more general solution

Infrastructure	Software/Tools	Culture
• Networked instruments • Centralized storage resources for working data • Archival storage • Networked computing	• Data “plumbing” • Microscopy specific LIMS (NexusLIMS) for working data • Persistent identifiers • Institutional data sources • Public data repository	• Integrating with existing workflows • Carefully changing user behavior • Carrots vs. sticks

Slide content courtesy of John Henry J. Scott

Networking instruments

- Can't we just plug in an ethernet cable?
- Are you sure you can trust your instrument control PC (or the ones it connects to)?
- Requirements on PC can come from organization, vendor, or often both
- How do we give these tools network capabilities while keeping everyone safe?



NIST National Institute of Standards and Technology • U.S. Department of Commerce

Removal of Obsolete Operating Systems from NIST Network

NIST S 6102.27
Issue Date: 07/27/2018
Effective Date: 07/30/2003

PURPOSE

The purpose of this directive is to define requirements for the removal of unsupported operating systems from the National Institute of Standards and Technology (NIST) network.

APPLICABILITY

This directive applies to all information system resources attached to the NIST network. This directive does not apply to information systems running on local isolated networks (e.g., Research Equipment Network) that are not connected to the NIST IT network in any way, through any of the components operating on that isolated network, nor connected to the Internet directly through any of the components operating on that isolated network.

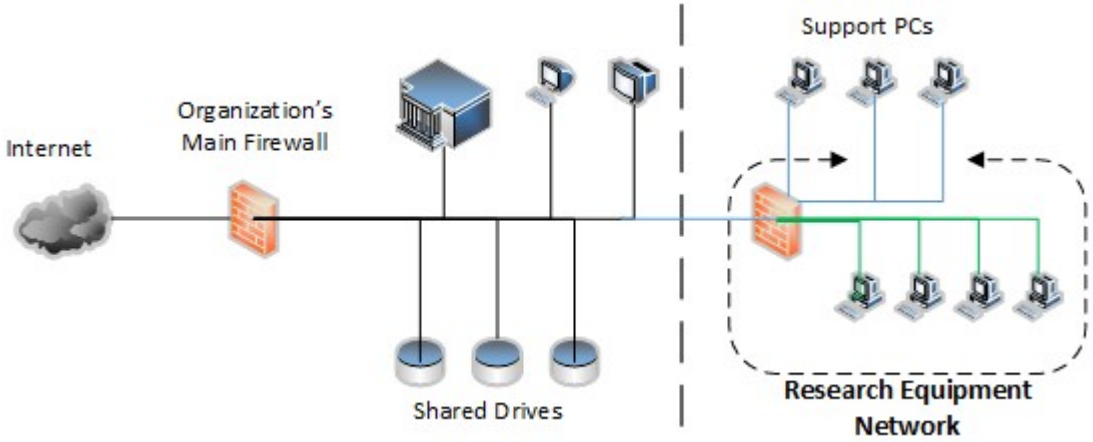
REFERENCES

This directive is supplemental to a suite of security controls consisting of:

- Department of Commerce, Information Technology Security Program Policy (ITSPP);
- Department of Commerce, Commerce Information Technology Requirements (CITRs);
- NIST Special Publication 800-53, Security and Privacy Controls for Federal Information Systems and Organizations, System and Services Acquisition (SA); and
- NIST Information Security Directives.

Networking instruments with a REN

Research Equipment Network



Segregates computers via firewall between REN and general organization network

Pinholes for OS updates and critical network resources

The REN at NIST

- Introduced late 2013 NIST-wide
- For digital tools, equipment, and computers that cannot meet federal IT security requirements
- Provides additional network security for both equipment and NIST network
- Effectively provides private virtual local area networks (PVLANs) for each instrument connected to the REN

Instruments *can*:

- Run any OS or hardware platform
- Access NIST central resources, like file or license servers (with limitations)

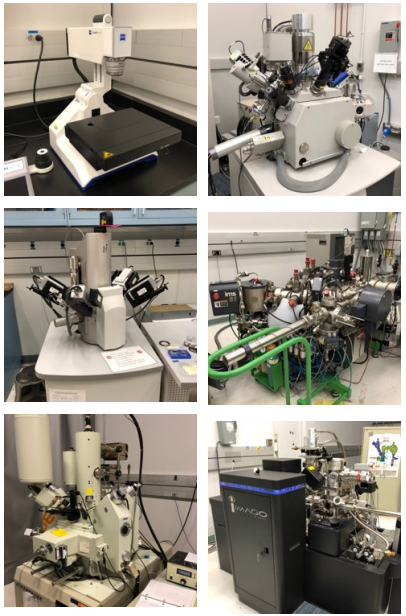
Instruments *cannot*:

- Access the internet
- Receive email
- Communicate with other REN computers (by default)

Centralized file storage

- Most institutions have some sort of “central” storage that is network accessible
- Often targeted for “business” uses, not scientific ones (NIST’s was)
- Many are being replaced by “cloud” offerings (NIST’s is)
- Given the size and bandwidth requirements, onsite “scientific” file storage is generally a requirement
- For a group or department, could be a commercial NAS system
- Larger institutions may benefit from enterprise-level storage
  - Backup, redundancy, storage sizes, etc.

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

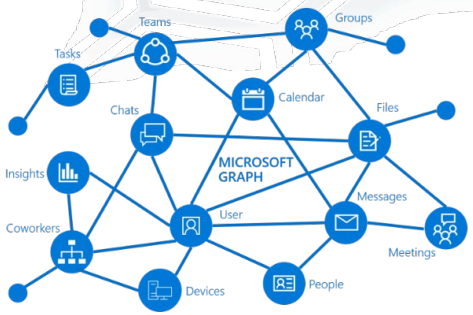
Data “Plumbing”

- 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

Institutional Data Sources

Information about people

- Being able to programmatically access user information is very useful
  - Instrument PCs usually don't have user info
  - Associating files with users
  - Adding contact information into experimental records
  - Integrating organizational information (project, division, etc.) provides additional query facets
- Looks different at every institution, but API access is key…



Microsoft Active Directory

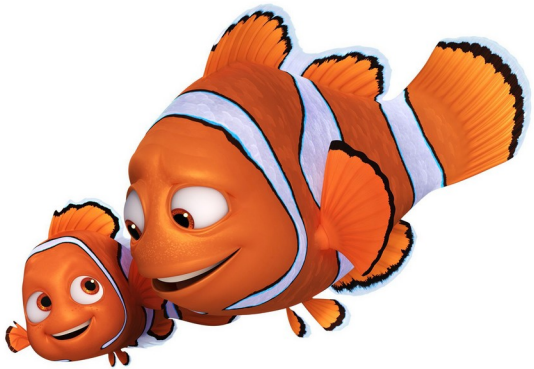
NIST Org

National Institute of Standards and Technology

Institutional Data Sources

Information about instruments and usage

- Interactive and programmatic information about instruments, who's using them, and when
  - Shared calendars can work (Google, Outlook, SharePoint, etc.)
  - A dedicated laboratory management system is better
- NEMO (https://github.com/usnistgov/NEMO) (NanoFab Equipment Management & Operations) is an open-source web application designed to manage the shared instrumentation facilities
- MML runs its own installation, named MARLIN



Institutional Data Sources

Information about instruments and usage

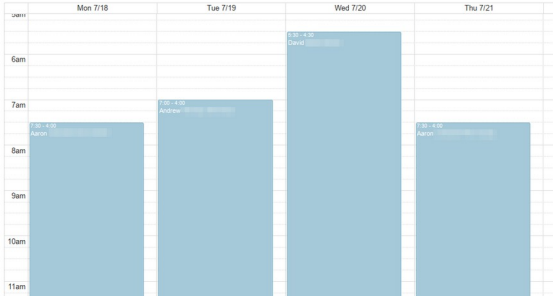
Reservations


```
{
  "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
```

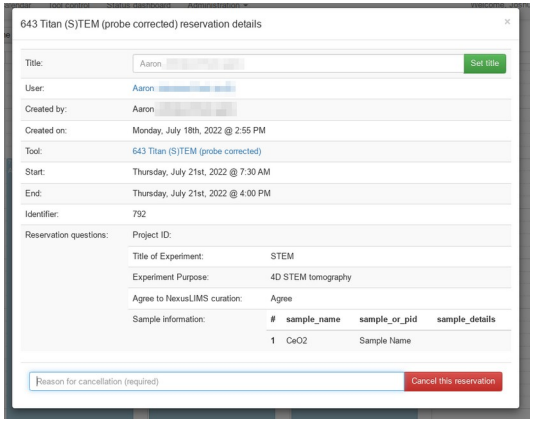

Usage Events


```
{
  "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 ]
}
```

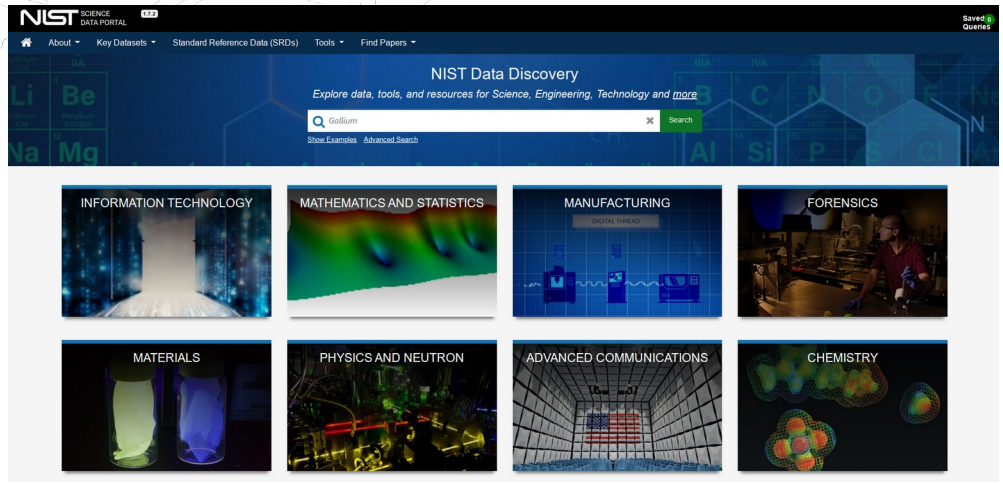



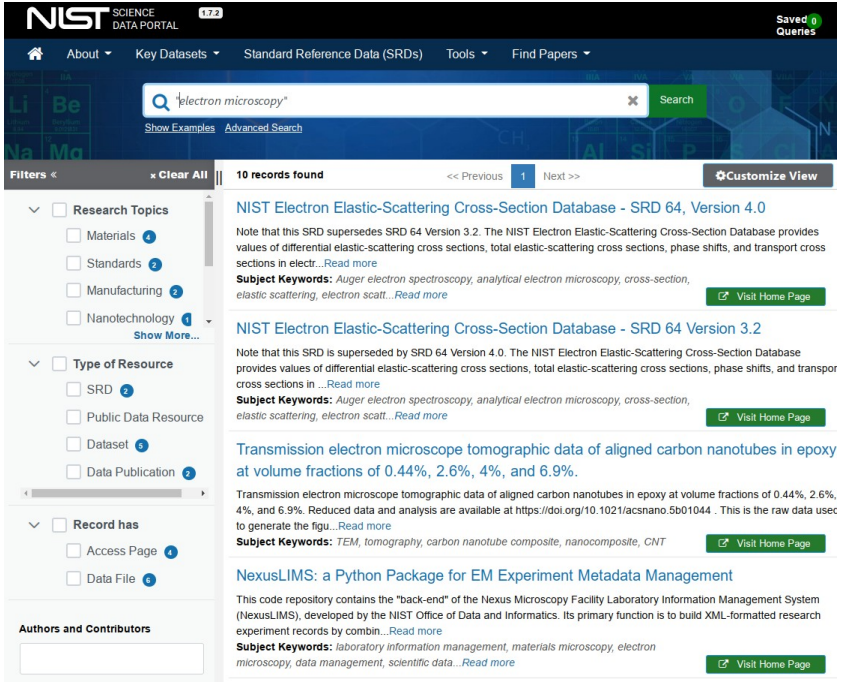
Open Access to Research (OAR)

- Since 2013, a variety of governmental memos, Executive Orders, and laws passed to require open access to government data (also, a good idea for science!)
- Published papers increasingly require (or at least allow) published data
  - How to publish data? What data gets published? Where does it get published?
- NIST OAR project has provided a framework for data publishing at NIST, making it easy for researchers to publish to https://data.nist.gov, which further populates https://data.gov
  - https://github.com/usnistgov/?q=OAR

OAR – Public Data Repository

https://data.nist.gov

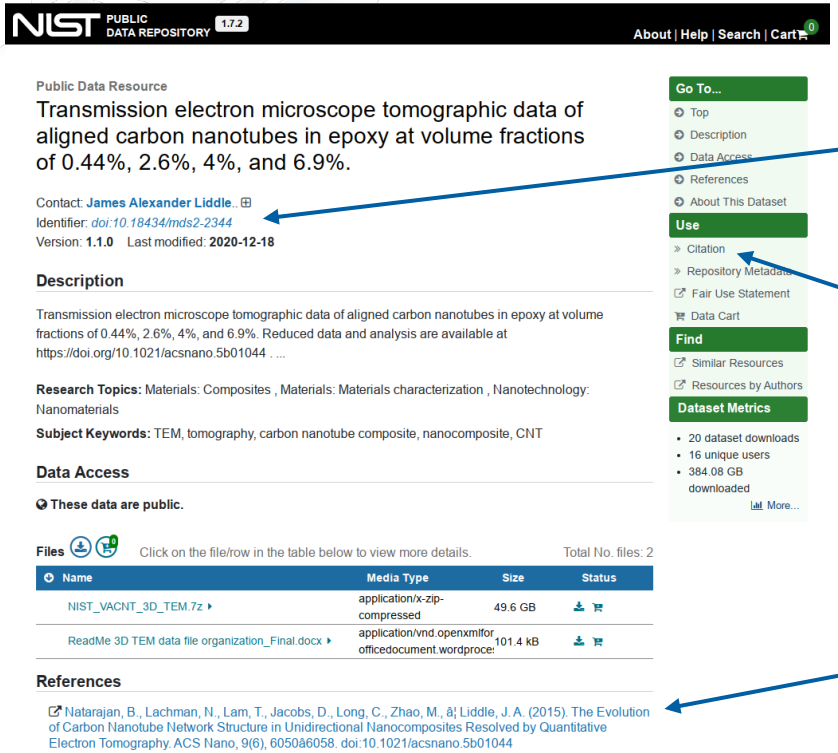




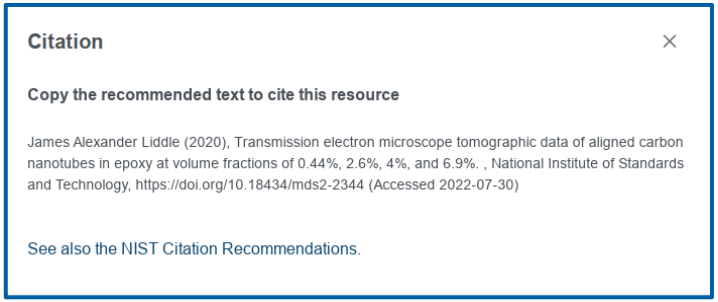
Faceted Browsing and free-text search of NIST **Public Data Repository** resources

OAR – Public Data Repository

https://data.nist.gov



Citable DOI for dataset



Published article associated with this data

Working with your organizational culture

- People like the way they already do things, so a real benefit has to be demonstrated
- Identify your “champions” – those who have a desire and motivation to change their data handling practices
- Need to build to be as inclusive of various workflows as possible – include inputs from across all the research areas, if possible
- Carrots generally work better than sticks, but sometimes sticks are necessary

What else can we do?

- Automated metadata extraction from *all* research files, not just in 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

Final takeaways

- These efforts take a lot of work; let's provide a better starting point
  - "Rising tides…" as the saying goes
- Improvements can be made from group- to organization-level
- Much of the work will be consensus-finding and workflow analysis
- Keep your eye on the scientific benefits
  - What *new thing* is possible or what *old thing* is much easier?

Thank you for your attention!
Questions?

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

NIST MML Central Storage

Storage / Connections	• 1.26 PB total storage space • SMB and NFS protocols supported
Data Protection	• Daily snapshots retained monthly • Striping across disks for protection against disk failure • Archival to Amazon Glacier for critical files
Use Cases	• Read-only target for all harvested instrument data • Read-write shared project spaces • Individual user workspaces (as needed)

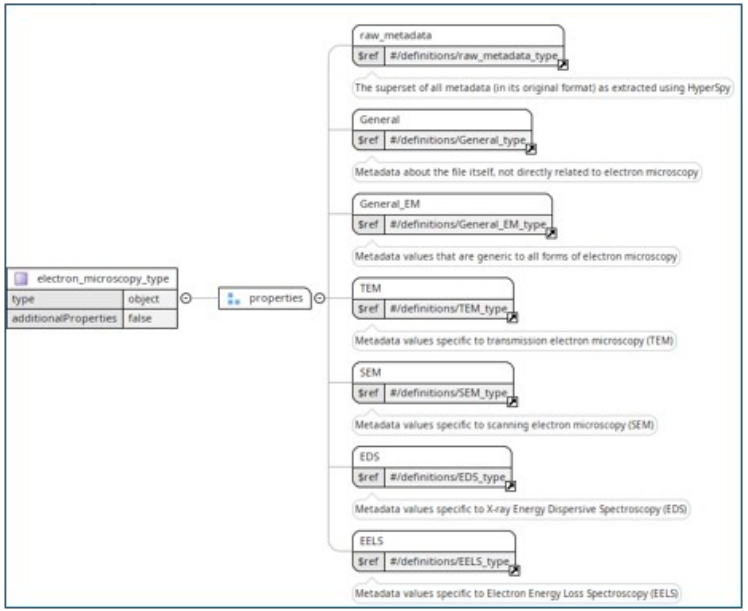
Extracting metadata with Materials-IO

https://github.com/materials-data-facility/MaterialsIO

Materials-IO is an example of an open-source tool to extract metadata into schema-controlled JSON representations with arbitrary “extractors as plugins”

Can be as simple as:

```
import materials_io
materials_io.execute_parser('em', [file_name])
```



Electron microscopy JSON Schema

Access control (if you want)

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