

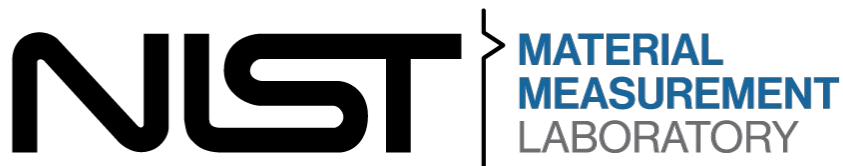


Lessons Learned Building a Modern Microscopy Data Ecosystem at NIST

Joshua A. Taillon

Presentation for Voyles Group

Monday, October 17, 2022



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 CHMaD

- Laura Bartolo
- Roberto dos Reis

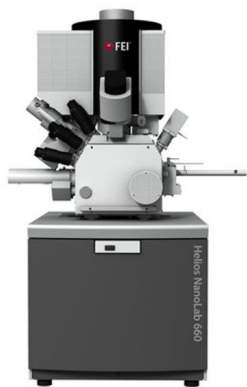


- Ao Liu
- Weinan Si

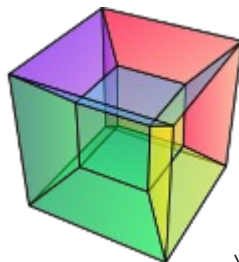
NIST MML LIMS Community of Interest

- 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



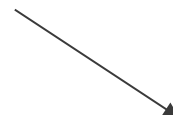
Hyperspy



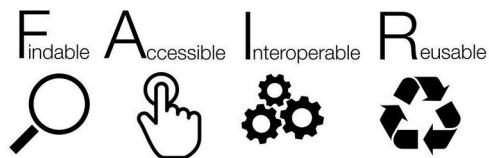
General scientific
programming



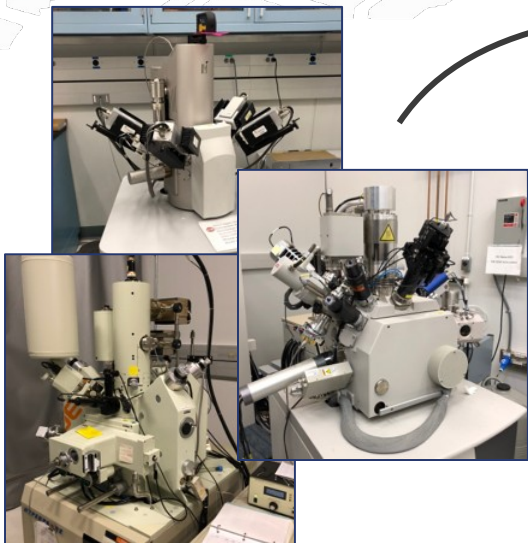
Databases (SQL)



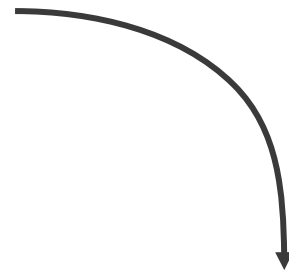
Dashboarding/app design



Our Data Problem...

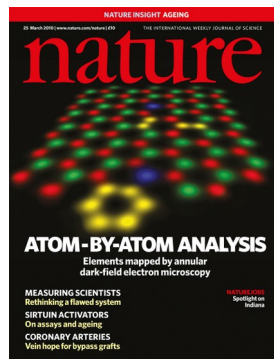


sophos.com



engadget.com

Can we see the data?



Vol 464 (7288), 2010. Nature.



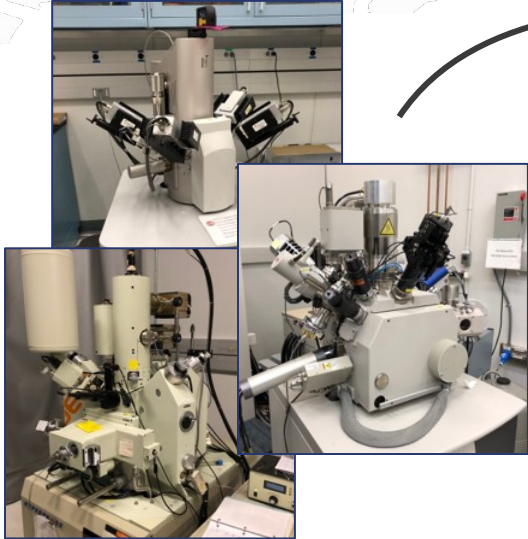
Our Data Problem...



sophos.com

1

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?

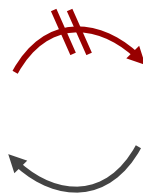
Can we see the data?



How do we associate that data with our great publications?



Vol 464 (7288), 2010. *Nature*.



3



engadget.com

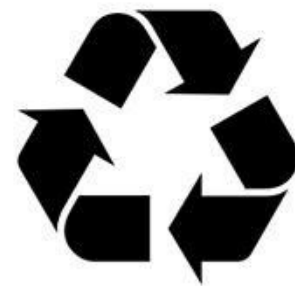
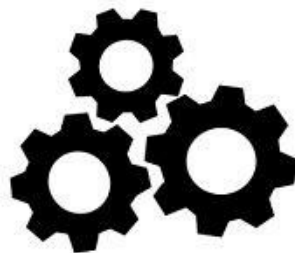
FAIR Data Principles

F
Findable

A
Accessible

I
Interoperable

R
Reusable

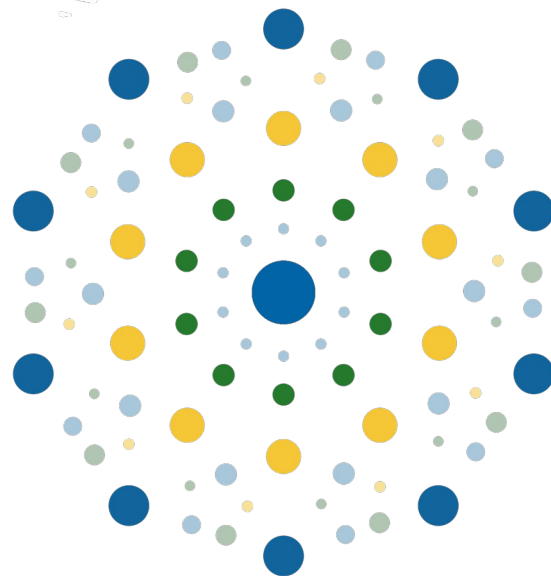


Wilkinson *et al.*, *Scientific Data*, 3, 160018, 2016 ([link](#))

Image: Sangya Pundir - [CC-BY-SA 4.0](#)

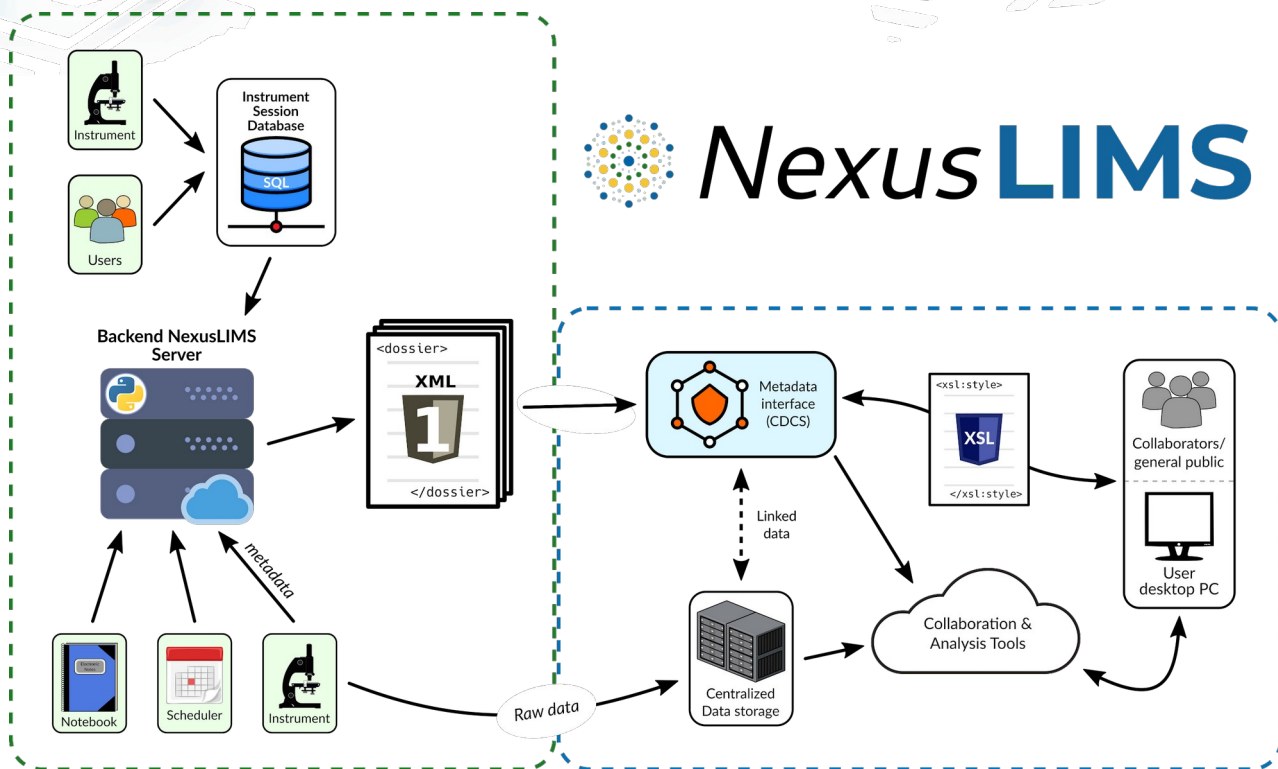
Let's solve it all! (or at least some...)

- 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](https://doi.org/10.18434/mds2-2355)
 - Described in detail in *Microscopy and Microanalysis*, 27 (3), 2021. pp. 511 - 527. [10.1017/S1431927621000222](https://doi.org/10.1017/S1431927621000222)



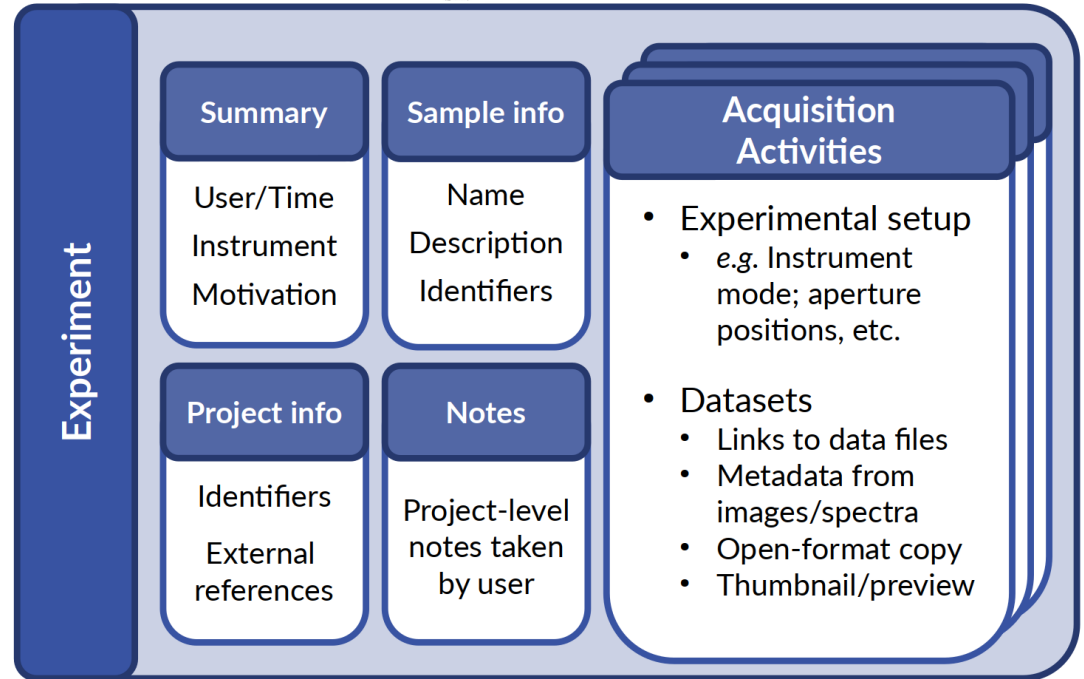
NexusLIMS

What does our LIMS for microscopy look like?



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

The screenshot displays the NexusLIMS web interface. At the top, the navigation bar includes the NexusLIMS logo, a search bar with the text "Browse and Search Records", and several utility links: MARLIN, NEMO, Tutorial, Help, and a user profile for "jat".

The search bar contains the query "EDS" and "david". A blue "Search" button is positioned to the right of the input fields. Below the search bar, the results section is titled "Found 4 Results:". To the right of this title are several interactive buttons: "Sort", "Share Query", "Share PIDs", "Download", and "Date".

The search results are presented as a list of four entries, each with a checkbox on the left and a pencil icon on the right. The entries are:

- EDS of Sn on Graphene** (FEI Quanta200) - 4 data files in 1 activity - 4 tif files. Date: June 17 2022, 1:00PM. Motivation: How much Sn is on Graphene!!!
- EDS W, Ag post-echem** (FEI Quanta200) - 25 data files in 2 activities - 25 tif files. Date: May 07 2021, 1:41PM. Motivation: Morphology and species identification
- EDS W, Ag post-echem** (FEI Quanta200) - 3 data files in 1 activity - 3 tif files. Date: May 07 2021, 9:20AM. Motivation: Morphology and species identification
- W Ref** (FEI Quanta200) - 6 data files in 2 activities - 6 tif files. Date: Dec. 10 2020, 10:47AM. Motivation: EDS

Blue brackets highlight the search bar area at the top and the bottom-most result entry.

Browsing and previewing (meta)data

Explore record:

Activity 1

SEM Imaging

EDS of Sn on Graphene

Complete filelisting for:

EDS of Sn on Graphene - June 17, 2022

Root path: /Quanta/SA20220617_Au patterned Graphene Sn02/

Search:

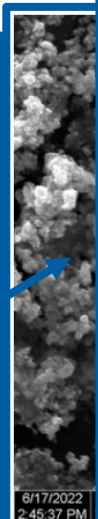
Select all Select none Download all as .zip Download selected as .zip

Copy CSV Excel Print

Total size of all datasets: 3.7 MiB.

Dataset Name	Path	Size	Type	Meta	D/L
<input type="checkbox"/> Region 1.tif	/	953.4 KiB	Image		
<input type="checkbox"/> Region 2.tif	/	953.4 KiB	Image		
<input type="checkbox"/> Region 3.tif	/	953.4 KiB	Image		
<input type="checkbox"/> Other chip region 1.tif	/	953.4 KiB	Image		

Showing 1 to 4 of 4 datasets



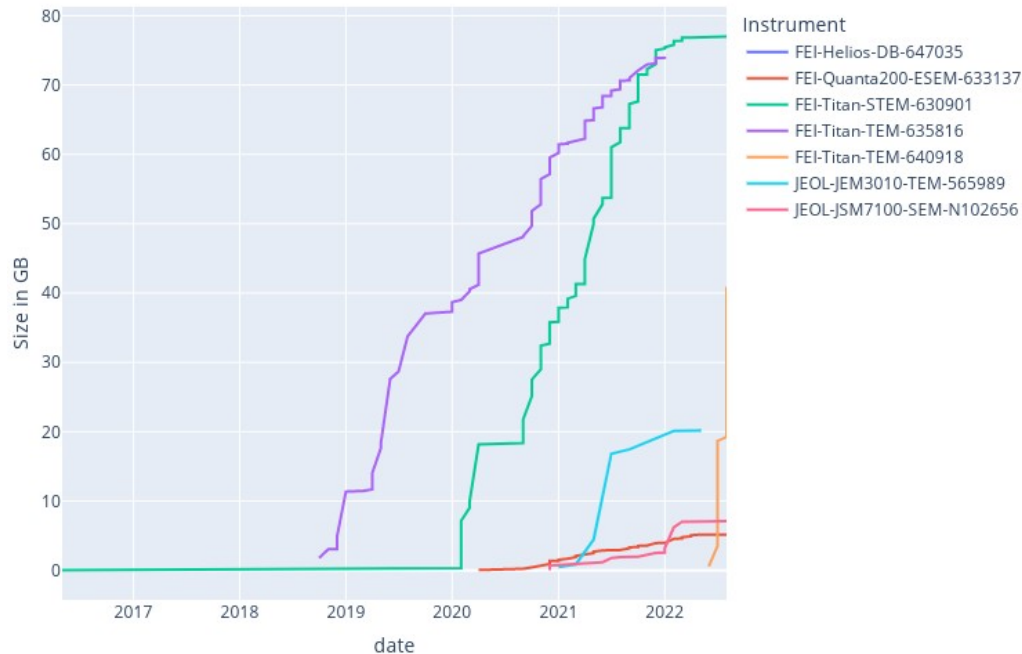
```
{
  "nx_meta": {
    "Acquisition Date": "06/17/2022",
    "Acquisition Time": "02:45:37 PM",
    "Beam Name": "EBeam",
    "Beam Tilt X": 0.0704791,
    "Beam Tilt Y": -0.0416927,
    "Chamber ID": "XL30SB",
    "Chamber Pressure (mPa)": 0.574172,
    "Column Type": "FEG SEM",
    "Creation Time": "2022-06-17T14:46:00.247976",
    "Data Dimensions": "(1024, 884)",
    "Data Type": "SEM_Imaging",
    "DatasetType": "Image",
    "Detector Brightness Setting": 40.3927,
    "Detector Contrast Setting": 38.8255,
    "Detector Grid Voltage (V)": 250,
    "Detector Name": "ETD",
    "Detector Signal": "SE",
    "Drift Correction Applied": true,
    "Emission Current (uA)": 130,
    "Horizontal Field Width (um)": 23.4009,
    "Instrument ID": "FEI-Quanta200-ESEM-633137",
    "Magnification Mode": 3,
    "Operator": "draciti",
    "Pixel Dwell Time (us)": 30,
    "Pixel Height (nm)": 22.8525,
    "Pixel Width (nm)": 22.8525,
    "Software Version": "4.1.15.2218 (build 2218)",
    "Spot Size": 4.5,
    "Stage Description": "50 x 50 manual tilt",
    "Stage Position": {
      "R": -0.0429575,
      "X": -0.000458187,
      "Y": -0.0128239,
      "Z": 0.0100184,
      "a": 0.000391165
    },
    "Stigmator X Value": 0.00128282,
    "Stigmator Y Value": 0.00154145,
    "System Type": "Quanta FEG",
    "Total Frame Time (s)": 29.5667,
    "Vacuum Mode": "High vacuum",
    "Vacuum Pump": "TMP",
    "Vertical Field Width (um)": 20.2016,
    "Voltage (kV)": 20,
    "warnings": [

```

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

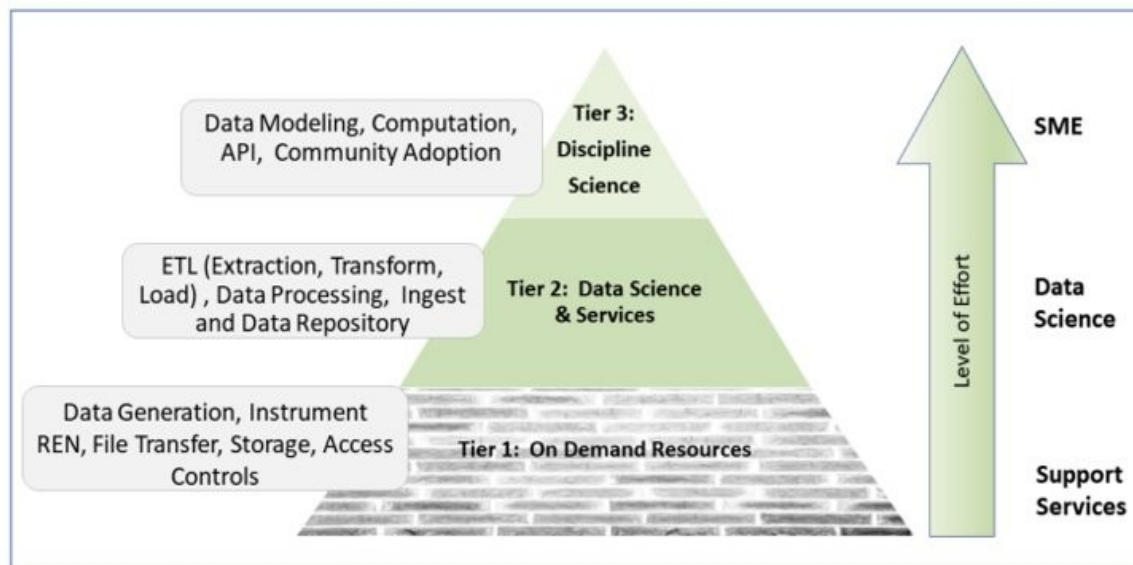


Fig. 1. LIMS three tiered model for implementation

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

- Networked instruments
- Centralized storage resources for working data
- Archival storage
- Networked computing

Software/Tools

- Data “plumbing”
- Microscopy specific LIMS (NexusLIMS) for working data
- Persistent identifiers
- Institutional data sources
- Public data repository

Culture

- Integrating with existing workflows
- Carefully changing user behavior
- Carrots vs. sticks

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

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”

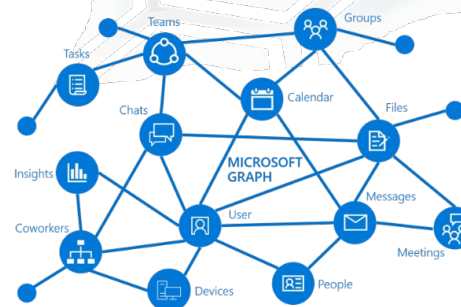


- Automates data flows from instruments across MML’s scientific laboratories into one or more centralized location(s)
- Each PC shares a read-only folder
 - This folder becomes the new “data” folder for users on the instrument
 - Users can use any folder hierarchy they wish - helpful to use usernames
- 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

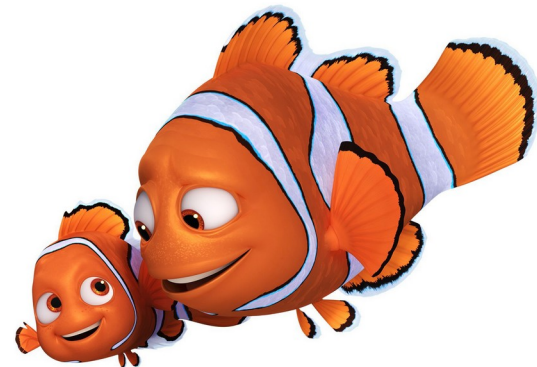
- Being able to programmatically access user information is very useful
 - Instrument PCs usually don't have user info
 - Associating files with users
 - Adding contact information into experimental records
 - Integrating organizational information (project, division, etc.) provides additional query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



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

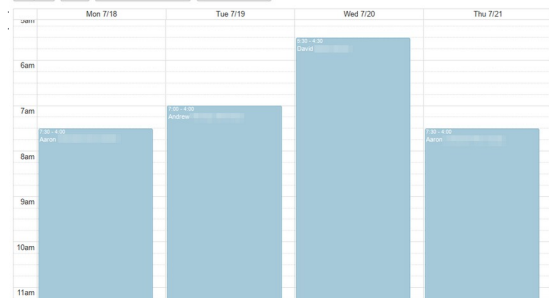
Tools

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
}
```



```
{
  "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 ]
}
```

#	sample_name	sample_or_pid	sample_details
1	CeO2	Sample Name	

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>

NIST Data Discovery
Explore data, tools, and resources for Science, Engineering, Technology and more

Search: Gallium

Categories: INFORMATION TECHNOLOGY, MATHEMATICS AND STATISTICS, MANUFACTURING, FORENSICS, MATERIALS, PHYSICS AND NEUTRON, ADVANCED COMMUNICATIONS, CHEMISTRY

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

NIST Science Data Portal

Search: electron microscopy

10 records found

Filters: Clear All

- Research Topics
 - Materials (2)
 - Standards (2)
 - Manufacturing (3)
 - Nanotechnology (1) Show More...
- Type of Resource
 - SRD (2)
 - Public Data Resource
 - Dataset (1)
 - Data Publication (2)
- Record has
 - Access Page (4)
 - Data File (6)

Authors and Contributors

Record 1: **NIST Electron Elastic-Scattering Cross-Section Database - SRD 64, Version 4.0**
Note that this SRD supersedes SRD 64 Version 3.2. The NIST Electron Elastic-Scattering Cross-Section Database provides values of differential elastic-scattering cross sections, total elastic-scattering cross sections, phase shifts, and transport cross sections in electr...
Subject Keywords: Auger electron spectroscopy, analytical electron microscopy, cross-section, elastic scattering, electron scatt...
[Visit Home Page](#)

Record 2: **NIST Electron Elastic-Scattering Cross-Section Database - SRD 64 Version 3.2**
Note that this SRD is superseded by SRD 64 Version 4.0. The NIST Electron Elastic-Scattering Cross-Section Database provides values of differential elastic-scattering cross sections, total elastic-scattering cross sections, phase shifts, and transport cross sections in ...
Subject Keywords: Auger electron spectroscopy, analytical electron microscopy, cross-section, elastic scattering, electron scatt...
[Visit Home Page](#)

Record 3: **Transmission electron microscope tomographic data of aligned carbon nanotubes in epoxy at volume fractions of 0.44%, 2.6%, 4%, and 6.9%.**
Transmission electron microscope tomographic data of aligned carbon nanotubes in epoxy at volume fractions of 0.44%, 2.6%, 4%, and 6.9%. Reduced data and analysis are available at <https://doi.org/10.1021/acsnano.5b01044>. This is the raw data used to generate the figu...
Subject Keywords: TEM, tomography, carbon nanotube composite, nanocomposite, CNT
[Visit Home Page](#)

Record 4: **NexusLIMS: a Python Package for EM Experiment Metadata Management**
This code repository contains the "back-end" of the Nexus Microscopy Facility Laboratory Information Management System (NexusLIMS), developed by the NIST Office of Data and Informatics. Its primary function is to build XML-formatted research experiment records by combin...
Subject Keywords: laboratory information management, materials microscopy, electron microscopy, data management, scientific data...
[Visit Home Page](#)

OAR – Public Data Repository

<https://data.nist.gov>



Public Data Resource

Transmission electron microscope tomographic data of aligned carbon nanotubes in epoxy at volume fractions of 0.44%, 2.6%, 4%, and 6.9%.

Contact: [James Alexander Liddle](#)

Identifier: [doi:10.18434/mds2-2344](https://doi.org/10.18434/mds2-2344)

Version: 1.1.0 Last modified: 2020-12-18

Description

Transmission electron microscope tomographic data of aligned carbon nanotubes in epoxy at volume fractions of 0.44%, 2.6%, 4%, and 6.9%. Reduced data and analysis are available at <https://doi.org/10.1021/acsnano.5b01044> ...

Research Topics: Materials: Composites , Materials: Materials characterization , Nanotechnology: Nanomaterials

Subject Keywords: TEM, tomography, carbon nanotube composite, nanocomposite, CNT

Data Access

These data are public.

Files Click on the file/row in the table below to view more details. Total No. files: 2

Name	Media Type	Size	Status
NIST_VACNT_3D_TEM.7z	application/x-zip-compressed	49.6 GB	
ReadMe 3D TEM data file organization_Final.docx	application/vnd.openxmlformats-officedocument.wordprocessingml.document	101.4 kB	

References

[Natarajan, B., Lachman, N., Lam, T., Jacobs, D., Long, C., Zhao, M., & Liddle, J. A. \(2015\). The Evolution of Carbon Nanotube Network Structure in Unidirectional Nanocomposites Resolved by Quantitative Electron Tomography. ACS Nano, 9\(6\), 6050&6058. doi:10.1021/acsnano.5b01044](#)

- Go To...
- Top
- Description
- Data Access
- References
- About This Dataset
- Use
- Citation
- Repository Metadata
- Fair Use Statement
- Data Cart
- Find
- Similar Resources
- Resources by Authors
- Dataset Metrics
- 20 dataset downloads
- 16 unique users
- 384.08 GB downloaded
- [More...](#)

Citable DOI for dataset

Citation

Copy the recommended text to cite this resource

James Alexander Liddle (2020), Transmission electron microscope tomographic data of aligned carbon nanotubes in epoxy at volume fractions of 0.44%, 2.6%, 4%, and 6.9%. , National Institute of Standards and Technology, <https://doi.org/10.18434/mds2-2344> (Accessed 2022-07-30)

See also the NIST Citation Recommendations.

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>