How the Virtual Observatory Has Influenced Data Discovery, Access, and Re-Use in Other Disciplines

Robert Hanisch
Director, Office of Data and Informatics
Material Measurement Laboratory
National Institute of Standards and Technology

Tuesday November 21, 2017

MATERIAL MEASUREMENT LABORATORY

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

Common Goals

Discover

- Materials Resource Registry (data, code)
- International Metrology Resource Registry
- NIST Enterprise Data Inventory
- data.gov
- NIST Public Data Repository and Search Portal

- Standard Reference Data
- Materials Data Repository
- Materials Data Facility
- Persistent identifiers (DOIs, handles)

Access

Interoperate

- Materials Data Curator
- Data type registry
- Schema repository
- Lab info mgmt systems

MATERIAL MEASUREMENT LABORATORY

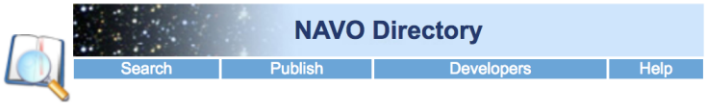
NIST National Institute of Standards and Technology U.S. Department of Commerce

Discover, Access, Interoperate

- Why?
  - Support FAIR* principles: Findable, Accessible, Interoperable, Re-usable
  - Assure maximum return on national investment in basic research
  - Demonstrate best practices
  - Address reproducibility "crisis"
- US OMB, OSTP directives; FASTR legislation
- But, Astronomy was here ~20 years ago!

*Wilkinson et al. 2016, *Nature Scientific Data*, DOI: 10.1038/sdata.2016.18

Registries



Find Astronomical Data Resources

Enter terms in the text box that describe the type of data you are looking for, or use the filters below to browse data by
will show catalogs and data collections that have these terms as part of their descriptions. Note that the search interfa
slashes by default, so when searching for full or partial IDs you may want to put the search string in quotes (e.g. "ivo:/
/ssap/befs").

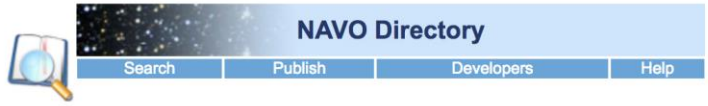
All Fields | Search... | Search

Examples: quasar, AGN, binary stars, Chandra, GALEX, far ultraviolet

Use the Data Discovery Tool to search and view data for a particular object or position.

Limit your search

- Subject
- Publisher
- Resource Type
- Content Level
- Instrument
- Facility
- Validation Level
- Rights
- Capability Type
- Result Type



Hosted By
Space Telescope Science Institute

Find Astronomical Data Resources

Enter terms in the text box that describe the type of data you are looking for, or use the filters below to browse data by facet. Results will show catalogs and data collections that have these terms as part of their descriptions. Note that the search interface ignores slashes by default, so when searching for full or partial IDs you may want to put the search string in quotes (e.g. "ivo://mast.stsci /ssap/befs").

All Fields | quasar | Search

Examples: quasar, AGN, binary stars, Chandra, GALEX, far ultraviolet

Use the Data Discovery Tool to search and view data for a particular object or position.

Limit your search

Subject

qsos	460
redshifts	243
spectroscopy	123
agn	109
photometry:wide-band	101
galaxies	91
photometry	70
equivalent_widths	57
survey source	38
bl_lac_objects	34

more »

- Publisher
- Resource Type
- Content Level
- Instrument

You searched for: quasar ✖ Start Over

« Previous | **1** - **10** of **671** | Next » Sort by relevance 10 per page

1. **Liverpool Quasar Lens Monitoring**

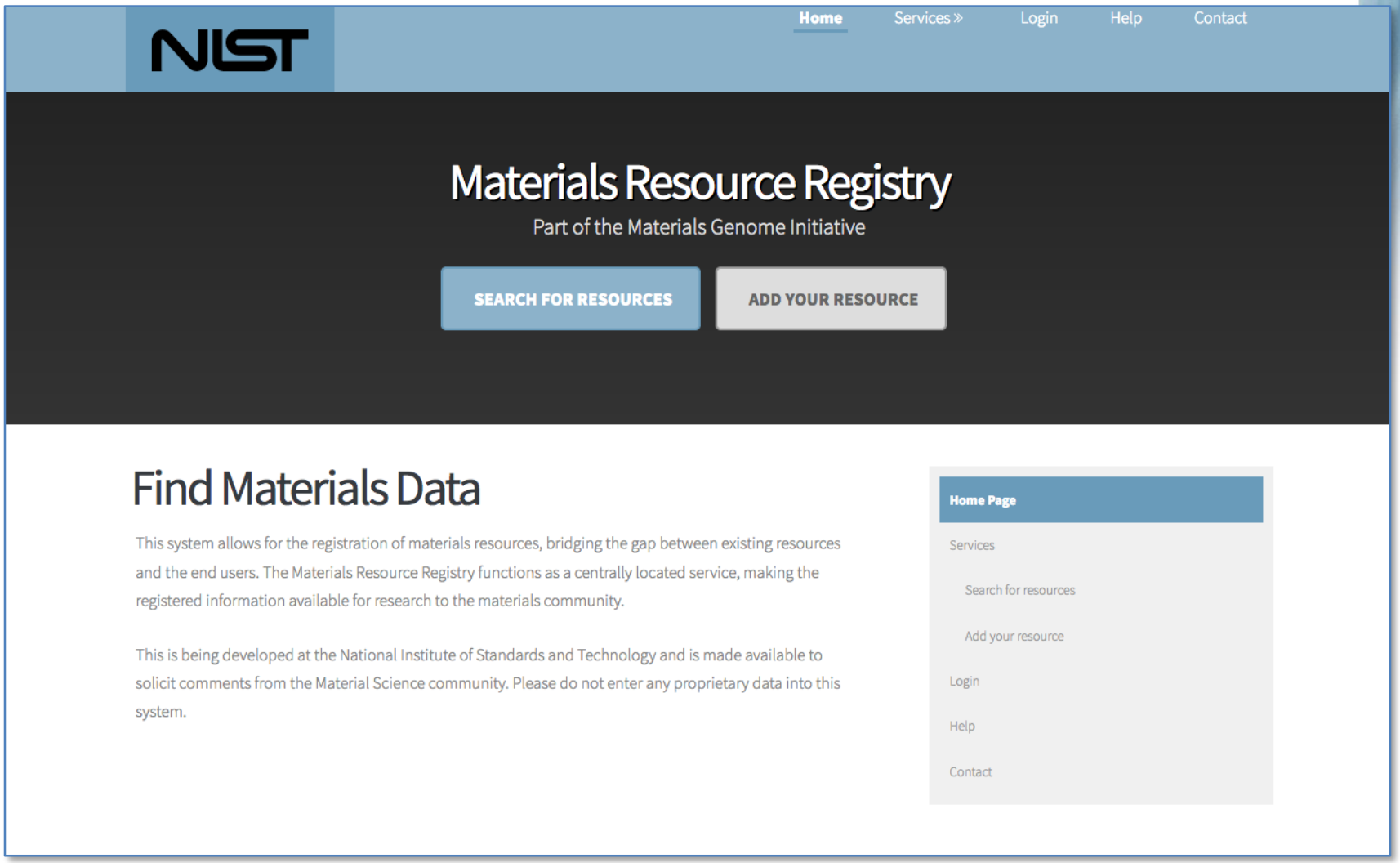
ID:	ivo://org.gavo.dc/liverpool/res/rawframes/q
Title:	Liverpool Quasar Lens Monitoring
Short Name:	LivQLensMon
Date:	08 May 2015 00:28:49
Publisher:	The GAVO DC team
Description:	This collection includes optical monitorings of gravitationally lensed quasars. The frames can be used to make light curves of quasar images and field objects. From quasar light curves, one may measure time delays and flux ratios, analyse variability and chromaticity, etc. These direct analyses/measurements are basic tools for different astrophysical studies, e.g., expansion rate of the Universe, mechanism of intrinsic variability in quasars, accretion disk structure, supermassive black holes, dark halos of galaxies (dust, collapsed dark matter, smoothly distributed dark matter,...)

2. Quasar polarization (Hutsemekers 1998)

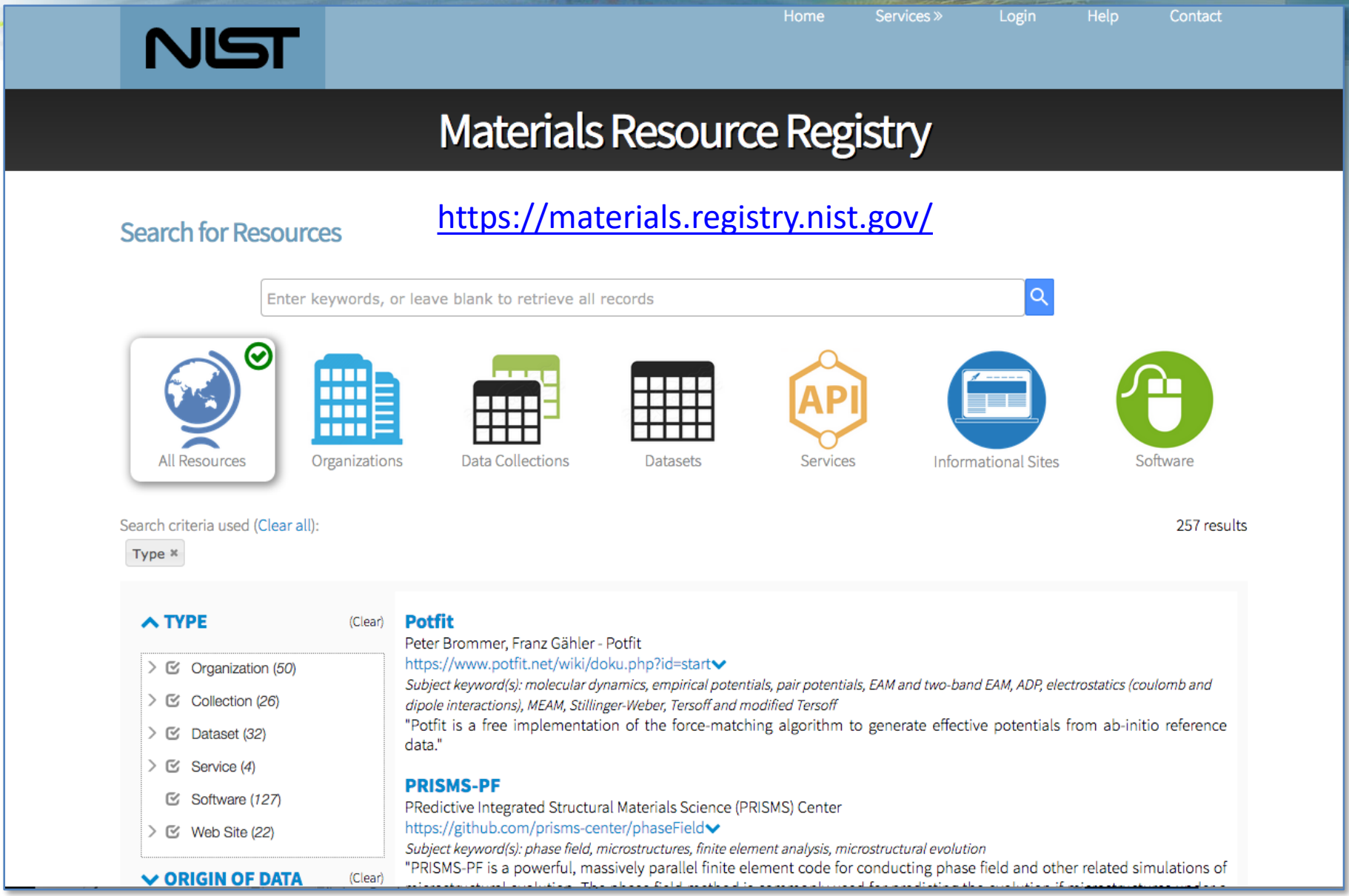
MATERIAL MEASUREMENT LABORATORY

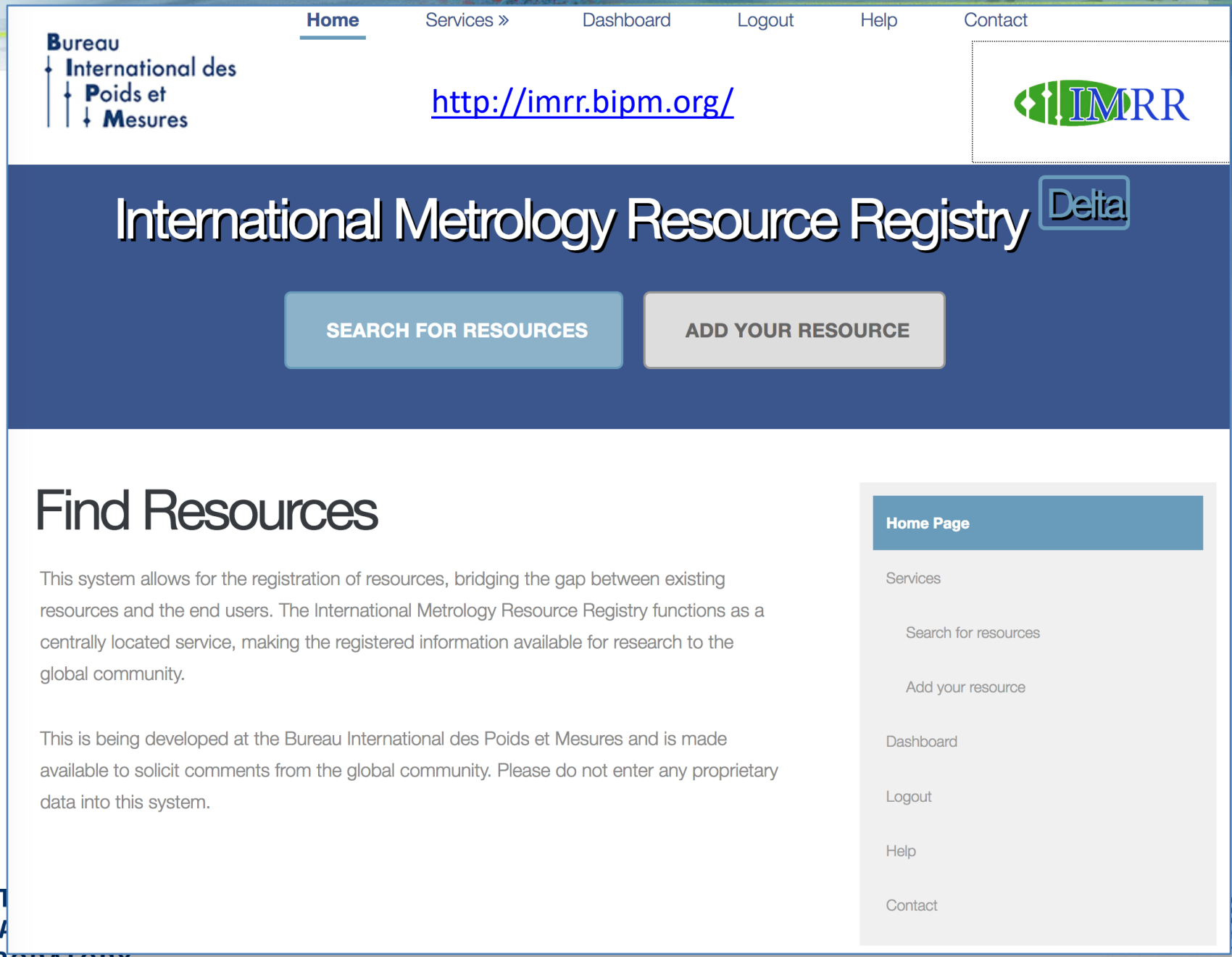
Standards and Technology
U.S. Department of Commerce

Materials Resource Registry



https://materials.registry.nist.gov/


NIST
Home
Services »
Login
Help
Contact
Materials Resource Registry
Search for Resources
https://materials.registry.nist.gov/
Enter keywords, or leave blank to retrieve all records
All Resources
Organizations
Data Collections
Datasets
API
Services
Informational Sites
Software
Search criteria used (Clear all):
257 results
Type ×
TYPE
(Clear)
Organization (50)
Collection (26)
Dataset (32)
Service (4)
Software (127)
Web Site (22)
ORIGIN OF DATA
(Clear)
Potfit
Peter Brommer, Franz Gähler - Potfit
https://www.potfit.net/wiki/doku.php?id=start
Subject keyword(s): molecular dynamics, empirical potentials, pair potentials, EAM and two-band EAM, ADP, electrostatics (coulomb and dipole interactions), MEAM, Stillinger-Weber, Tersoff and modified Tersoff
"Potfit is a free implementation of the force-matching algorithm to generate effective potentials from ab-initio reference data."
PRISMS-PF
PRedictive Integrated Structural Materials Science (PRISMS) Center
https://github.com/prisms-center/phaseField
Subject keyword(s): phase field, microstructures, finite element analysis, microstructural evolution
"PRISMS-PF is a powerful, massively parallel finite element code for conducting phase field and other related simulations of


Home
Services »
Dashboard
Logout
Help
Contact
Bureau International des Poids et Mesures
http://imrr.bipm.org/
IMRR
International Metrology Resource Registry
Delta
SEARCH FOR RESOURCES
ADD YOUR RESOURCE
Find Resources
This system allows for the registration of resources, bridging the gap between existing resources and the end users. The International Metrology Resource Registry functions as a centrally located service, making the registered information available for research to the global community.
This is being developed at the Bureau International des Poids et Mesures and is made available to solicit comments from the global community. Please do not enter any proprietary data into this system.
Home Page
Services
Search for resources
Add your resource
Dashboard
Logout
Help
Contact

International Metrology Resource Registry

Search for Resources

Enter keywords, or leave blank to retrieve all records

All Resources | Participating Institutes | Databases | Datasets | Standards Documents | Data Services | Data Portals | Metrology-related Software and Tools

Search criteria used (Clear all):

Resource Type: ×

58 results

RESOURCE TYPE: (Clear)

- Database (20)
- Dataset (7)
- Document (13)
- Metrology Institute (9)
- Portal (6)
- Service (1)
- Software (2)

STATUS OF THE DATA: (Clear)

STATUS OF THE DOCUMENT: (Clear)

APPLICATION AREA: (Clear)

TYPE OF PROPERTIES MEASURED: (Clear)

NIST/EPA Gas Phase Infrared Library (NIST-GPID)

US National Institute of Standards and Technology (NIST) (United States of America)

Home: 10.18434/T45K5N

Subject keyword(s): infrared spectra, vapor phase

This data collection contains 5,228 infrared spectra of different compounds along with chemical structures for most of them. Spectra are provided on a CD-ROM in the JCAMP-DX (Joint Committee for Atomic and Molecular Physical Data "Data Exchange") format. Chemical structures are provided in the MOL-file format. The IR data origin... show more

NIST Standard Reference Data Gateway

US National Institute of Standards and Technology (NIST) (United States of America)

Home: http://srdata.nist.gov/gateway/

Subject keyword(s): standards, standard reference data, metrology, materials, physics, computing, cybersecurity, law enforcement

The NIST Data Gateway is an online system designed to provide users with easy access to NIST scientific and technical data. The Gateway provides (1) direct links to free NIST online systems and full text online versions of many Journal of Physical and Chemical Reference Data (JPCRD) articles, (2) links to descriptions and orderi... show more

NIST Thermophysical Properties of Hydrocarbon Mixtures Database (SRD#4)

US National Institute of Standards and Technology (NIST) (United States of America)

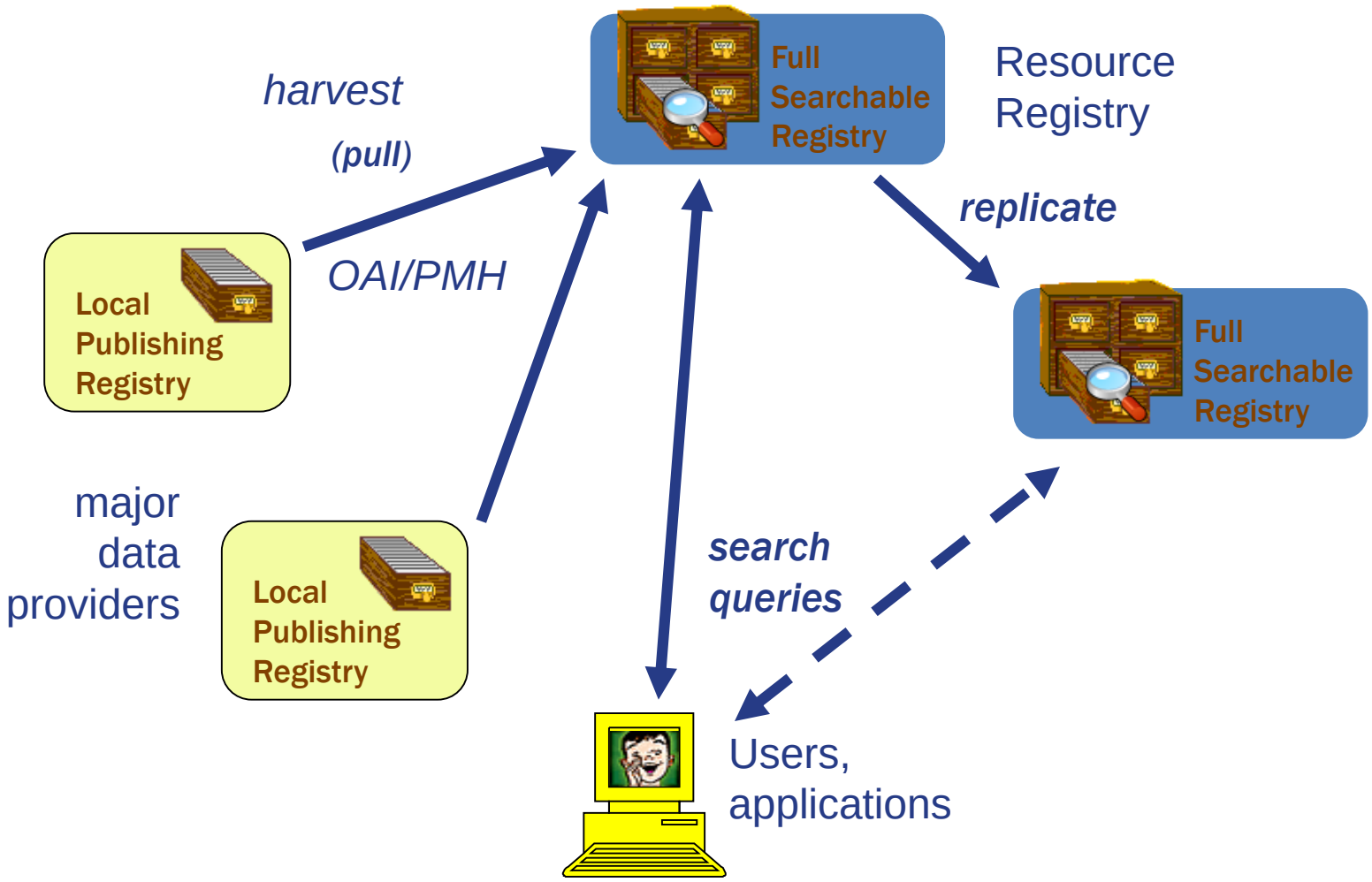
Home: 10.18434/T4CC76

Subject keyword(s): Chemical engineering, chemical manufacturing, chemical property, chemistry, dielectric constants, equation of state, fluids, heat capacity, hydrocarbons, LNG, mixtures, natural gas, petrochemical, phase equilibria, thermal conductivity, thermodynamics, thermophysics, transport property, viscosity

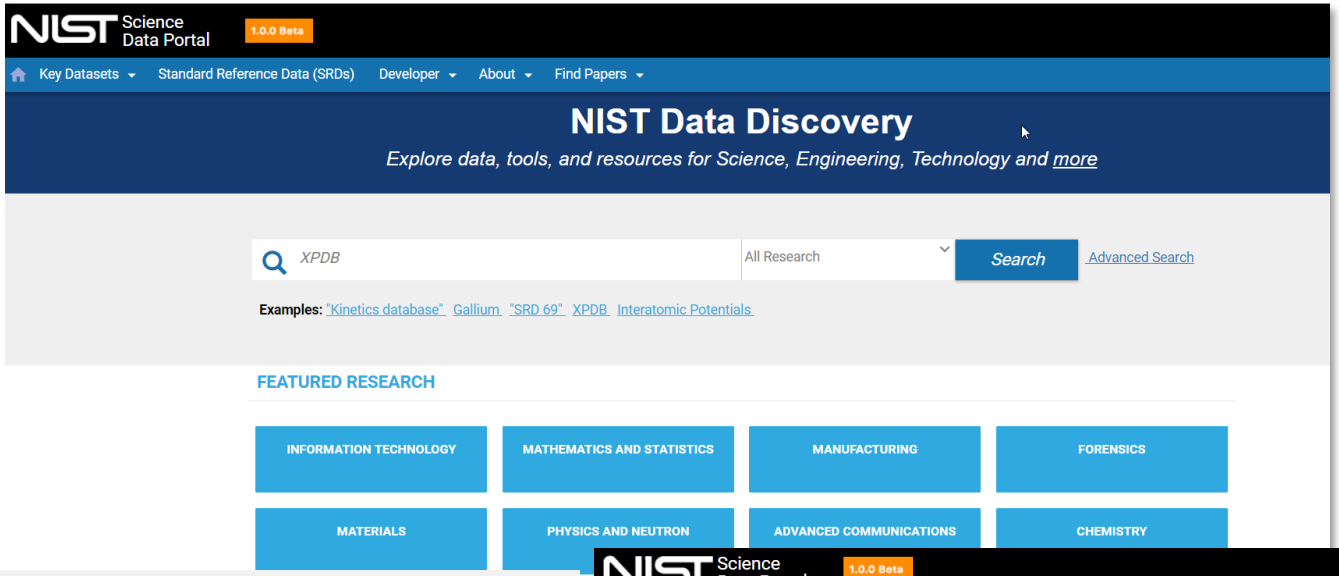
MAT
MEA
LABORATORY

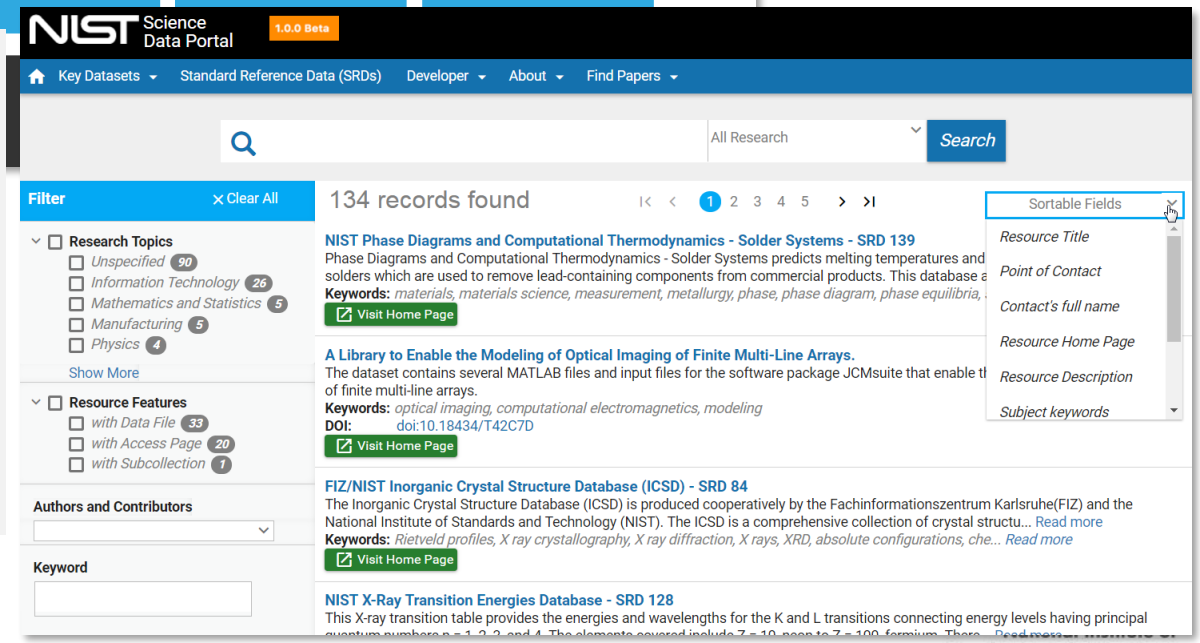
NIST National Institute of Standards and Technology U.S. Department of Commerce

Federated Architecture



Data Discovery for Public Research Data



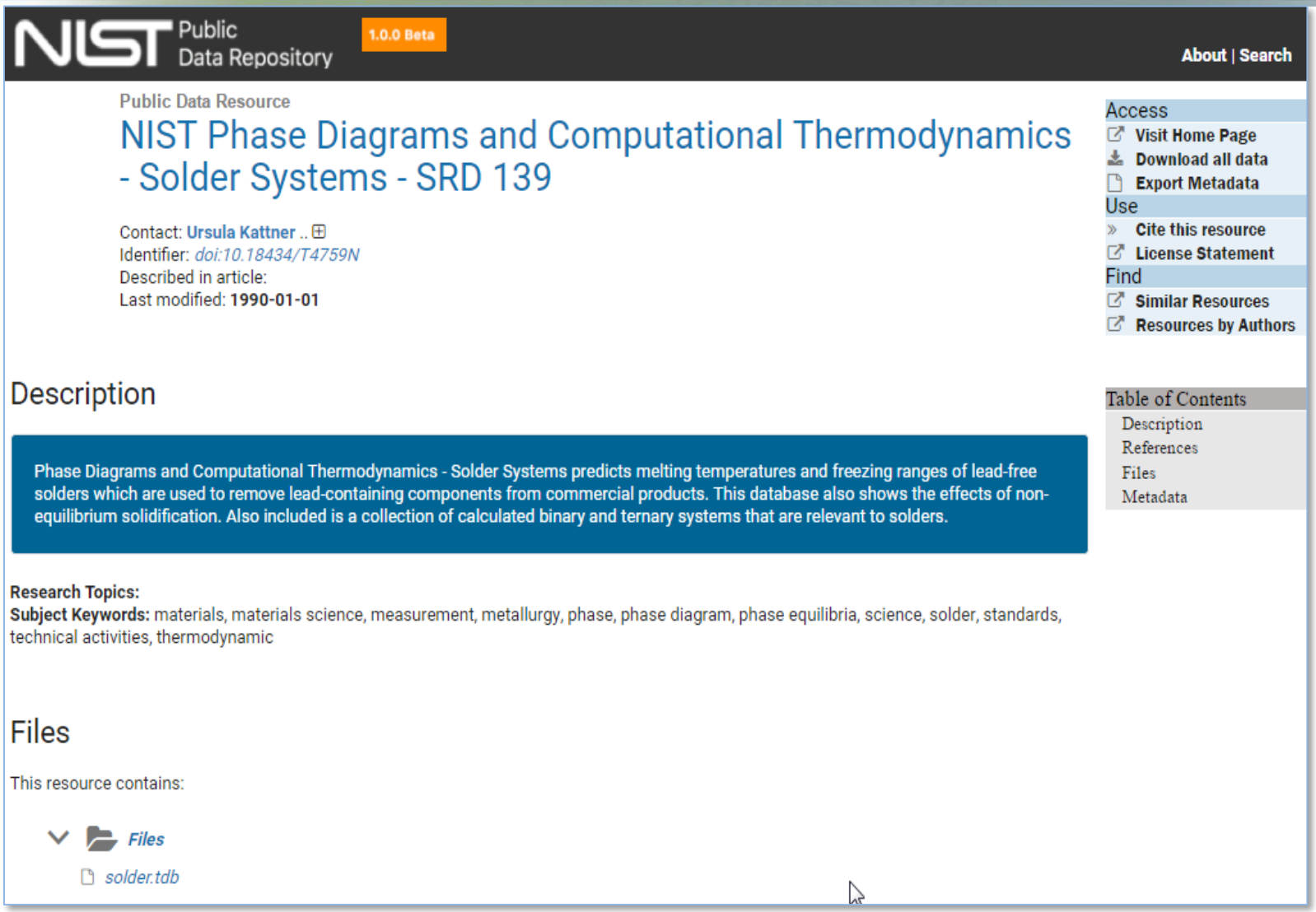


Search NIST public data records

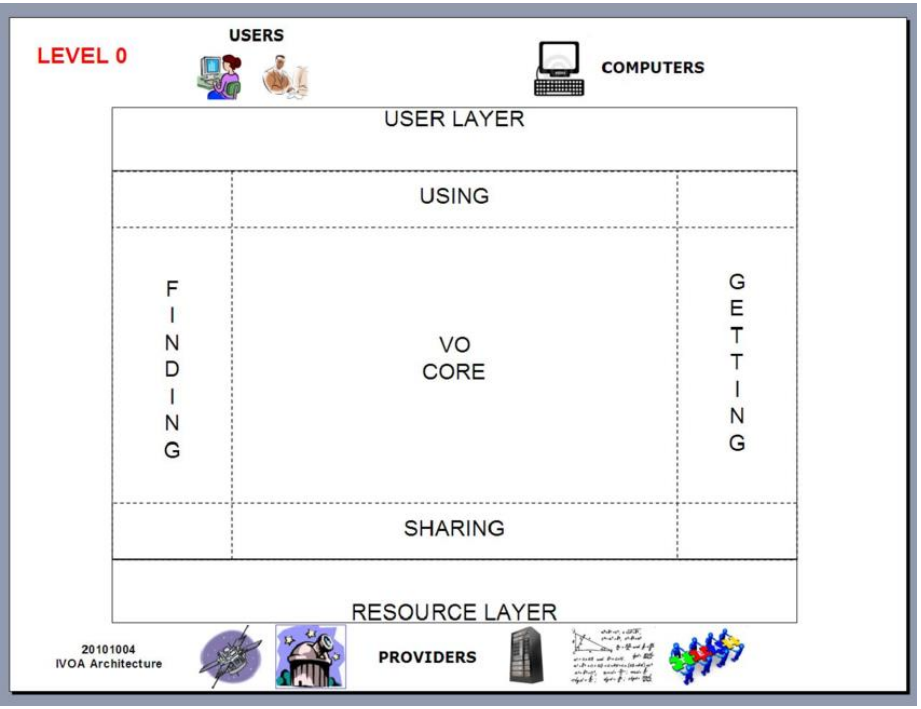
- View metadata
- Filter results
- Access data files, metadata
- APIs allow interoperability with client tools
- Records link to Public Data Repository

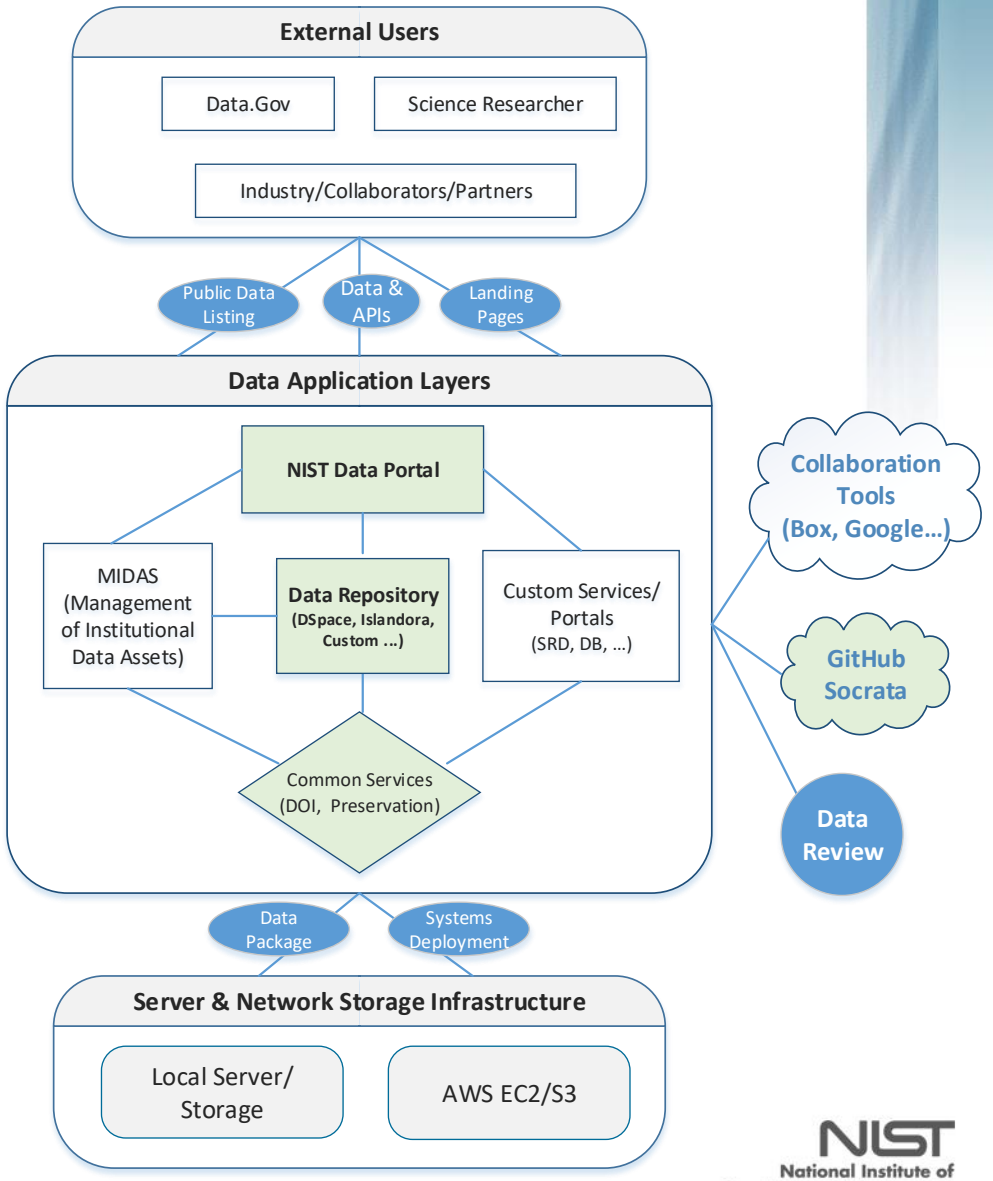
MATERIAL MEASUREMENT LABORATORY

NIST Public Data Repository – Basic Landing Page

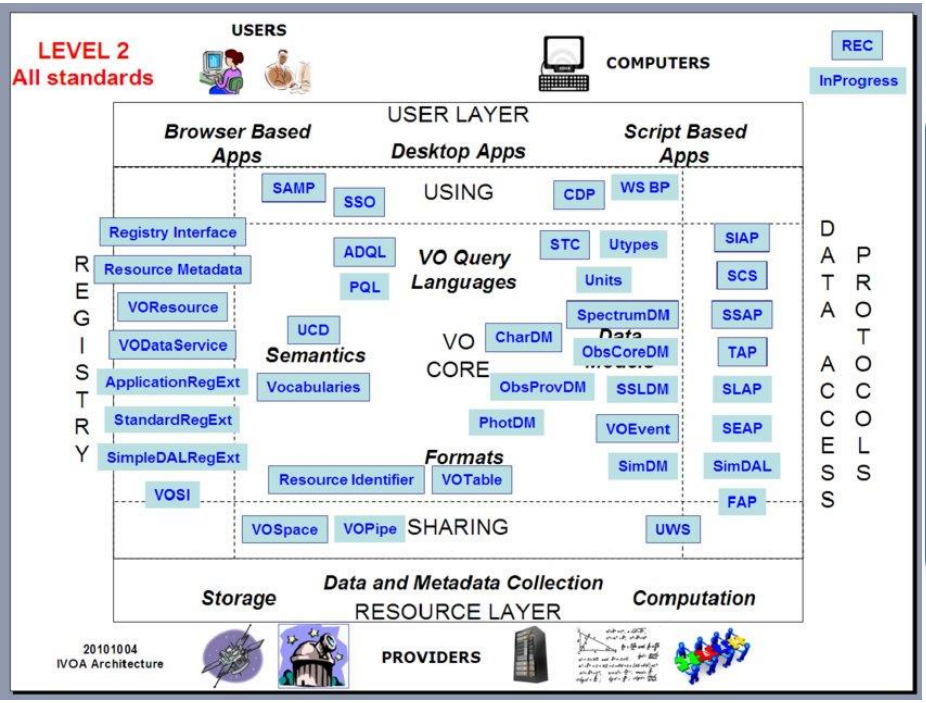


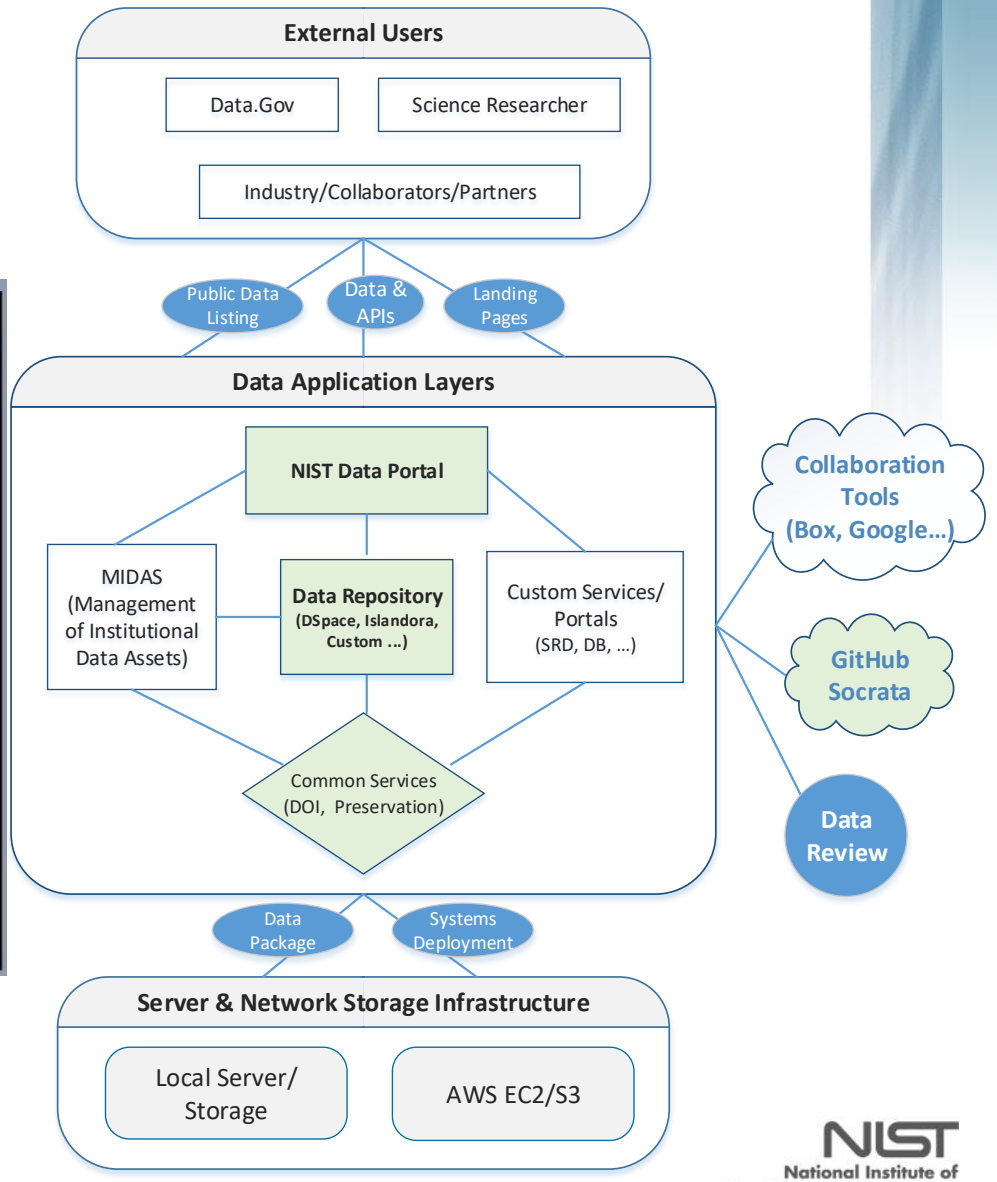
Research Data Infrastructure





Research Data Infrastructure

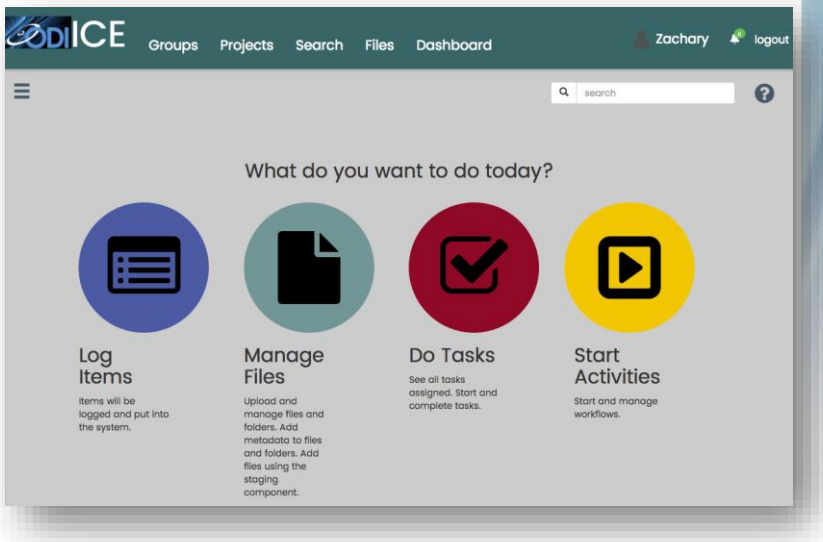


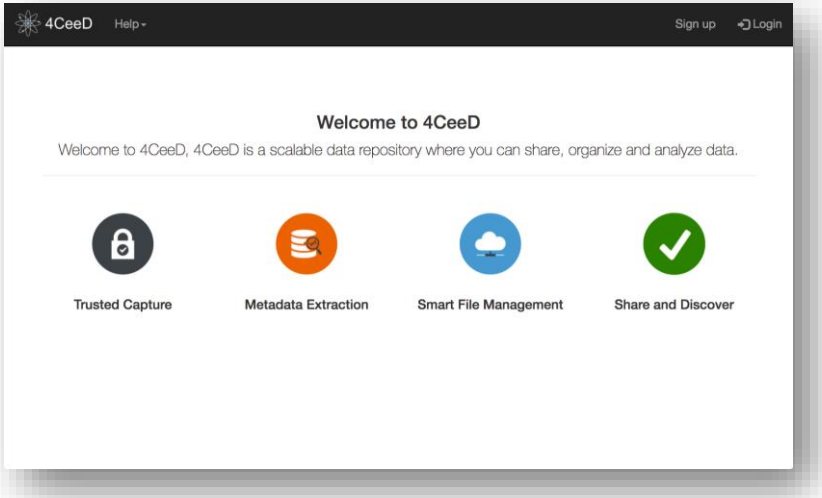


Interoperability

Laboratory Information Management Systems

- Integrated Collaborative Environment (ICE)
  - Running now at http://ice.nist.gov
  - Developed by Air Force Research Laboratory
- Timely and Trustworthy Curating and Coordinating Data Framework (T2C2) *4CeeD* system
  - Running now at http://t2c2.nist.gov:32500/
  - Developed by University of Illinois at Urbana-Champaign
- Also considering Discovery Environment for Relational Information and Versioned Assets (DERIVA) from USC

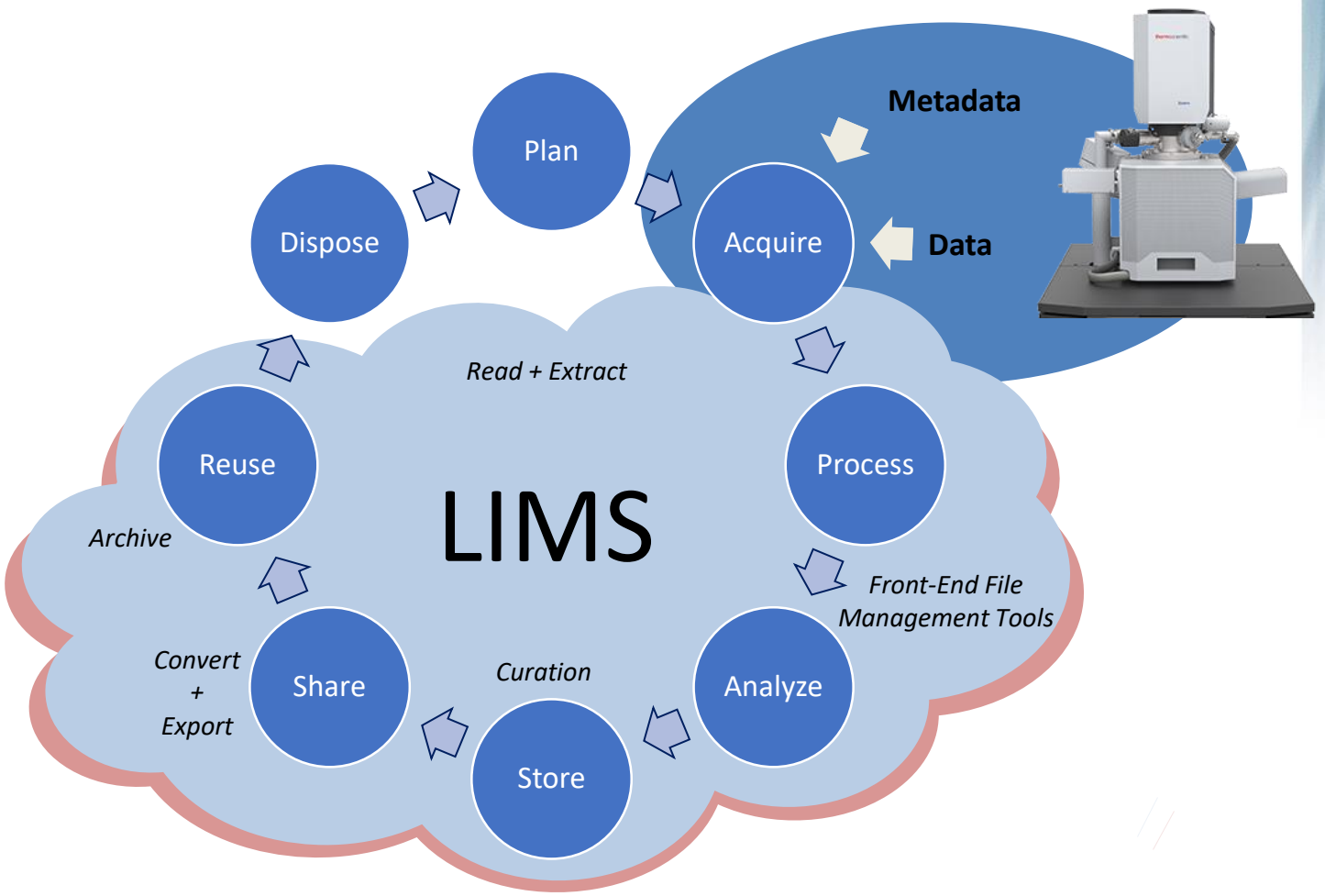




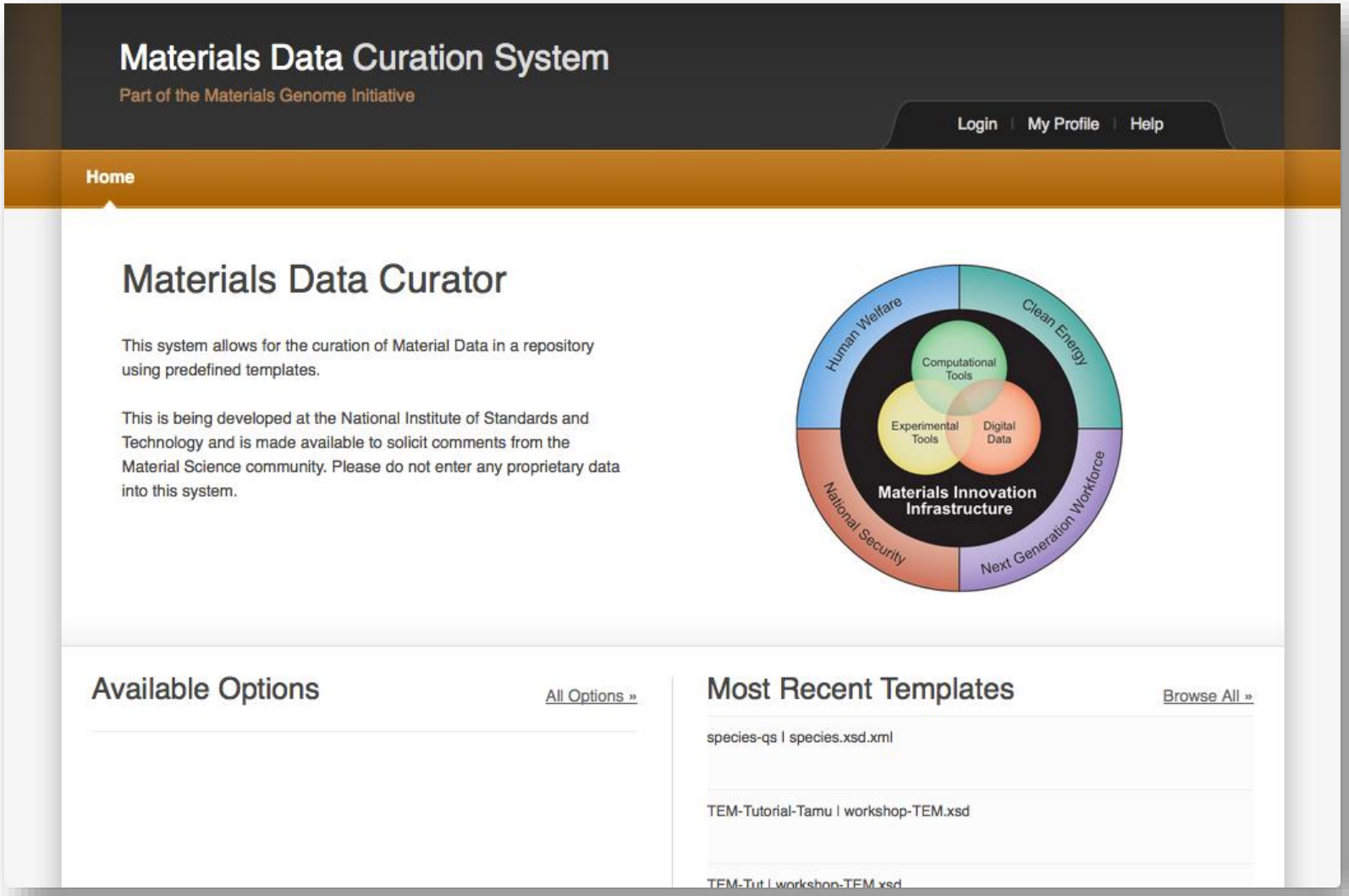
Laboratory Information Management Systems

- Capture instrument metadata at the source
  - Metadata extractors
  - Often must reverse engineer proprietary binary formats
- Move experiment metadata into database
  - Enable search across many experiments
  - Do not use filenames/file system for metadata storage
- Enable scripted data processing, calibration, feature extraction
- Support data management from acquisition to publication; improve reproducibility

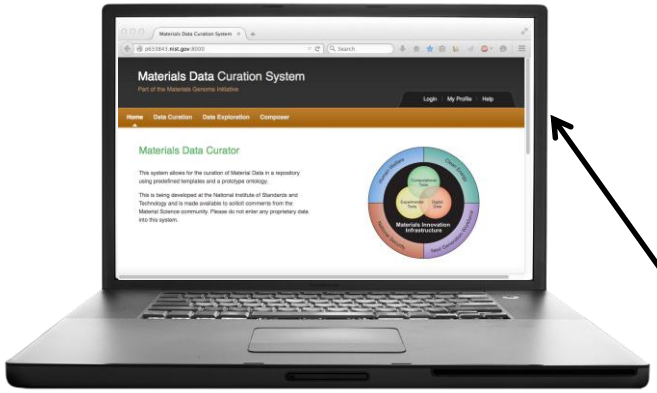
LIMS Help Manage the Data Lifecycle



Materials Data Curation System



Materials Data Curation System



Digital Data
& Metadata
(any format)

Data Analysis
Infrastructure

GUI

Web
Framework

Exporter

REST API

Data Provider

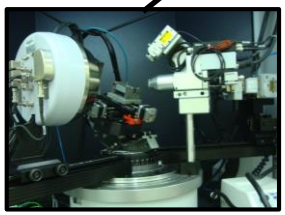
Harvester

Simulation

User Scripts

Data Management
& Search Engine

Harvester



Measurement

Data
Metadata

Images
Large Files
BLOBs

Database

Large Dataset
Repository

Data Provider

LABORATORY

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

Challenges with Experimental Data

Undefined Structure

Different Formats

```
SampleIdent CPD RR S
BANK    1       7251                                    0.000     0.000
 1   153 1   161 1                                   39 1   129 1   132
 1   129 1   129 1   151 1                                      46
 1   129 1   134 1   125 1   114 1   12                  27 1
```

```
SampleIdent CPD RR Sample 1B        DataFileNam
DiffrType PW3710   GeneratorVoltage 40   TubeCurr
Anode Cu    Alpha1  1.54056    Alpha2  1.54439    Ratio
MonochromatorUsed YES   DivergenceSlit 1      ReceivingSlit 0.3
   5.000   0.020 150.000    MeasureDateTime 20/12/1997   17:18   StepTime   3.00
     184     171     182     184     158     169     182     151     171     136
     171     163     146     158     158     169     182     151     171     136
     156     158     148     153     151     156     139     158     125     163
```

Only an expert human can understand this number.

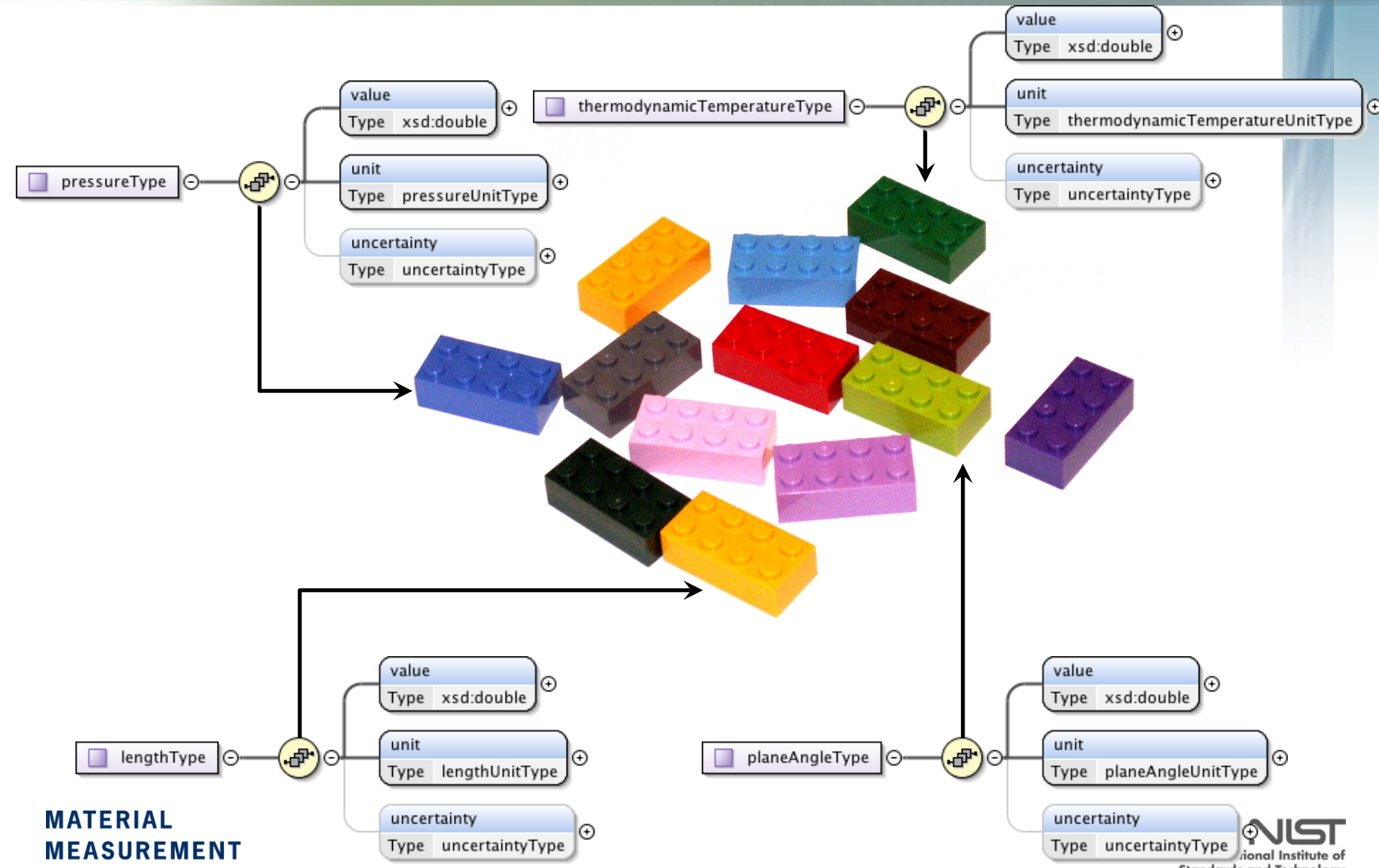
To a computer, this is a meaningless collection of numbers

```
 5.0000  150.0000    0.0200    1.0000
177.   182.   174.   154.   177.   156.   172.   161.   146.   169.
144.   154.   161.   156.   144.   166.   164.   119.   182.   135.
164.   128.   154.   114.   142.   121.   144.   154.   137.   137.
```

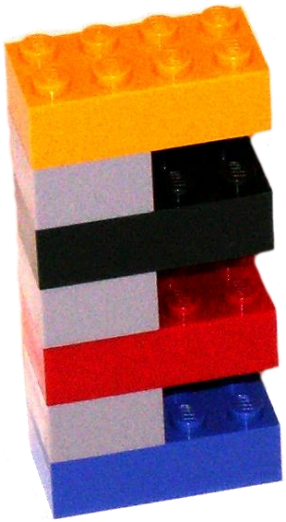
Structured Data Based on Data Model


```
{"diffractogram": {
    "xray-source": {
        "tube": {"anode-material": "Cu",
            "spectra": {"emission-line": [
                    {"Siegbahn": "Kalpha",
                     "wavelength": {"value": 1.54184,"unit": "angstrom"}},
                    {"Siegbahn": "Kalpha1",
                     "wavelength": {"value": 1.54056,"unit": "angstrom"}},
                    {"Siegbahn": "Kalpha2",
                     "wavelength": {"value": 1.54439,"unit": "angstrom"}}
                    ]}}},
    "pattern-data": {
        "angle-2-theta": {
            "value": [9.3,9.32,9.34, ... 75.16,75.18,75.2],
            "unit": "degree"},
        "intensity": {
            "value": [681.02,687.34,703.49, ... 127.52,124.29,118.32],
            "unit": "arbitrary"}}}}
```

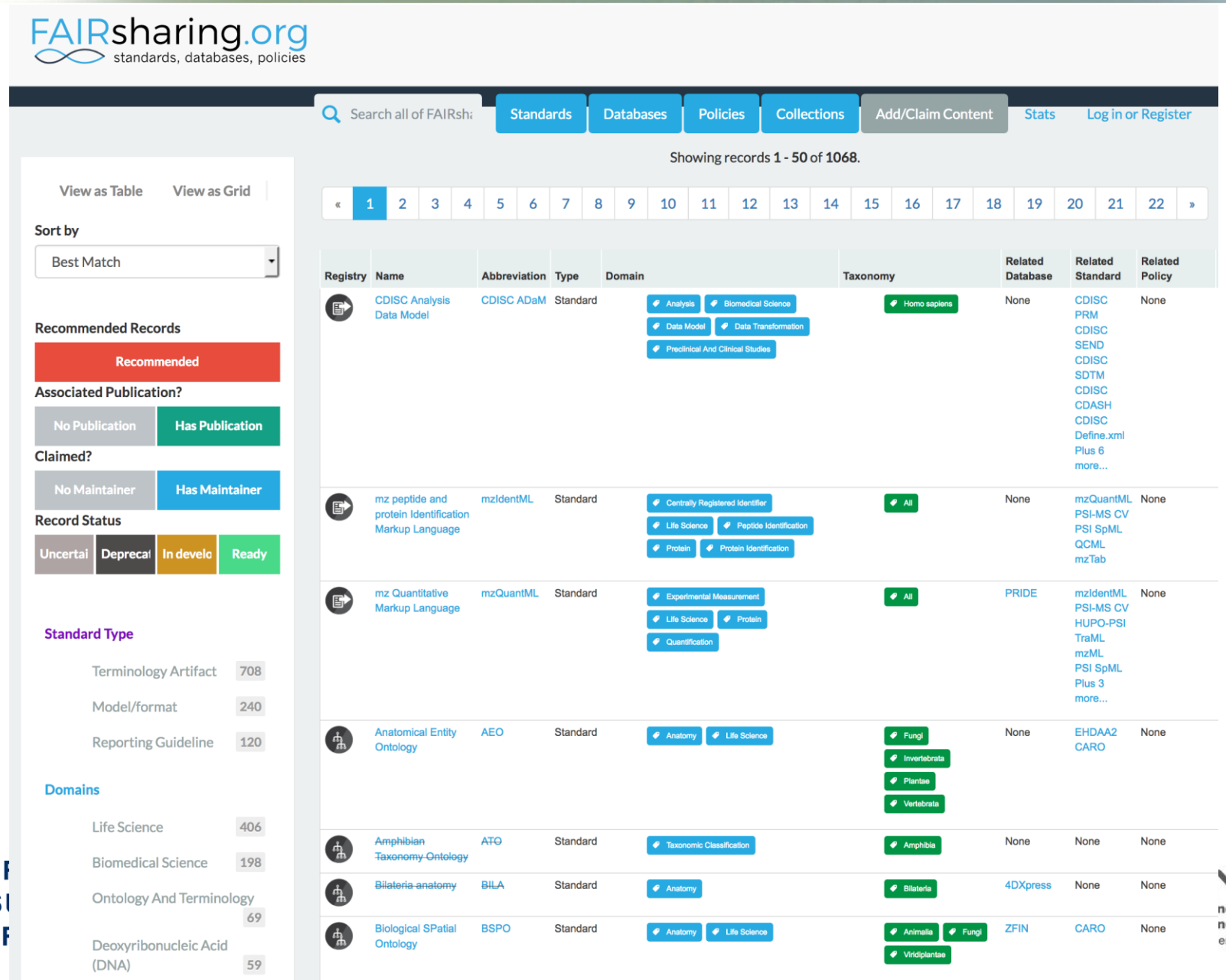
Modularity: Foundational Types



Data Models Re-Use Components



Metadata Standards



OME

Home | Products | Documentation | Community | About | Help | Blog

The Open Microscopy Environment

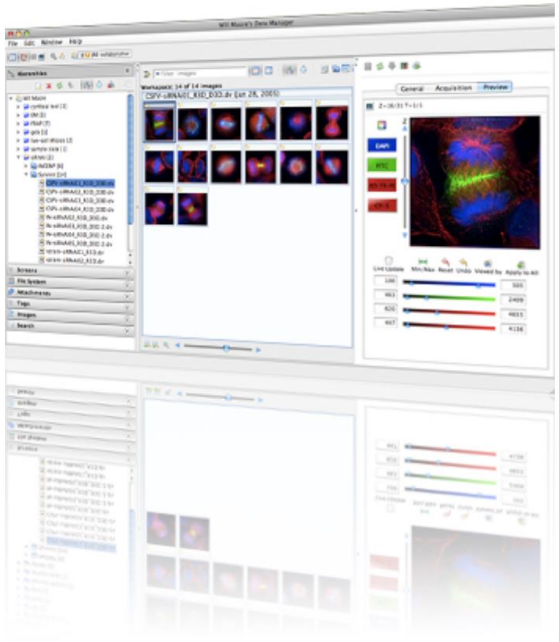
OME develops open-source software and data format standards for the storage and manipulation of biological microscopy data. It is a joint project between universities, research establishments, industry and the software development community.

OMERO

OMERO is client-server software for visualization, management and analysis of biological microscope images.

About OMERO | Download | New Users | Projects | Shared Scripts

Latest Release: OMERO 5.3.2
(May 2017)
Full details of this release can be found in the Release announcement.



Bio-Formats

Bio-Formats is a Java library for reading and writing biological image files. It can be used as an ImageJ plugin, Matlab toolbox, or in your own software.

About Bio-Formats | Download

Latest Release: Bio-Formats 5.5.1
(May 2017)
Full details of this release can be found in the Release announcement.

OME-TIFF Format

A TIFF-based image format that includes the OME-XML standard. Read more

OME Data Model

A common specification for storing details of microscope set-up and image acquisition. See

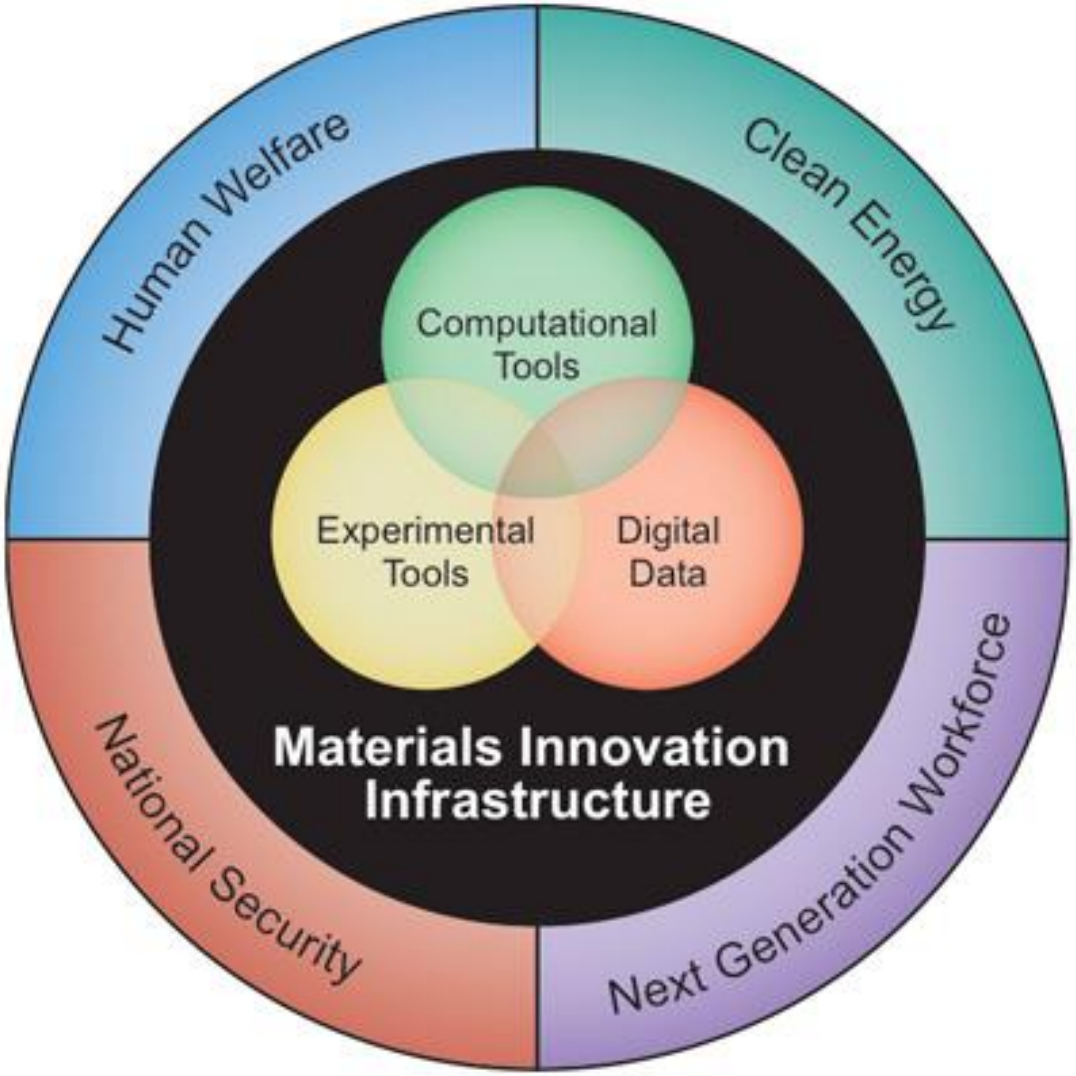
MATERIAL MEASUREMEI LABORATORY

NIST National Institute of ndards and Technology .S. Department of Commerce

Inspiration

- Data management practices in astronomy, and the VO in particular, have inspired similar efforts in other fields
  - Space science/space physics VxOs (VSO, VHO, VMO, VITMO)
  - Materials science
  - Metrology
  - Ecology/environmental science
  - Life/bioscience
  - Neuroscience
- Other communities are envious of astronomy's global data format, FITS
  - Other fields must contend with myriad of formats, many proprietary
- Also similar challenges, such as interoperability and semantic standards
- Independent development of similar architecture

Materials Genome Initiative



Environmental Science

What is DataONE?

Data Observation Network for Earth (DataONE) is the foundation of new innovative environmental science through a distributed framework and sustainable cyberinfrastructure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data.

Life Sciences

CYVERSE

VISION: TRANSFORMING SCIENCE THROUGH DATA-DRIVEN DISCOVERY.

MISSION: OUR MISSION IS TO DESIGN, DEPLOY, AND EXPAND A NATIONAL CYBERINFRASTRUCTURE FOR LIFE SCIENCES RESEARCH, AND TO TRAIN SCIENTISTS IN ITS USE.

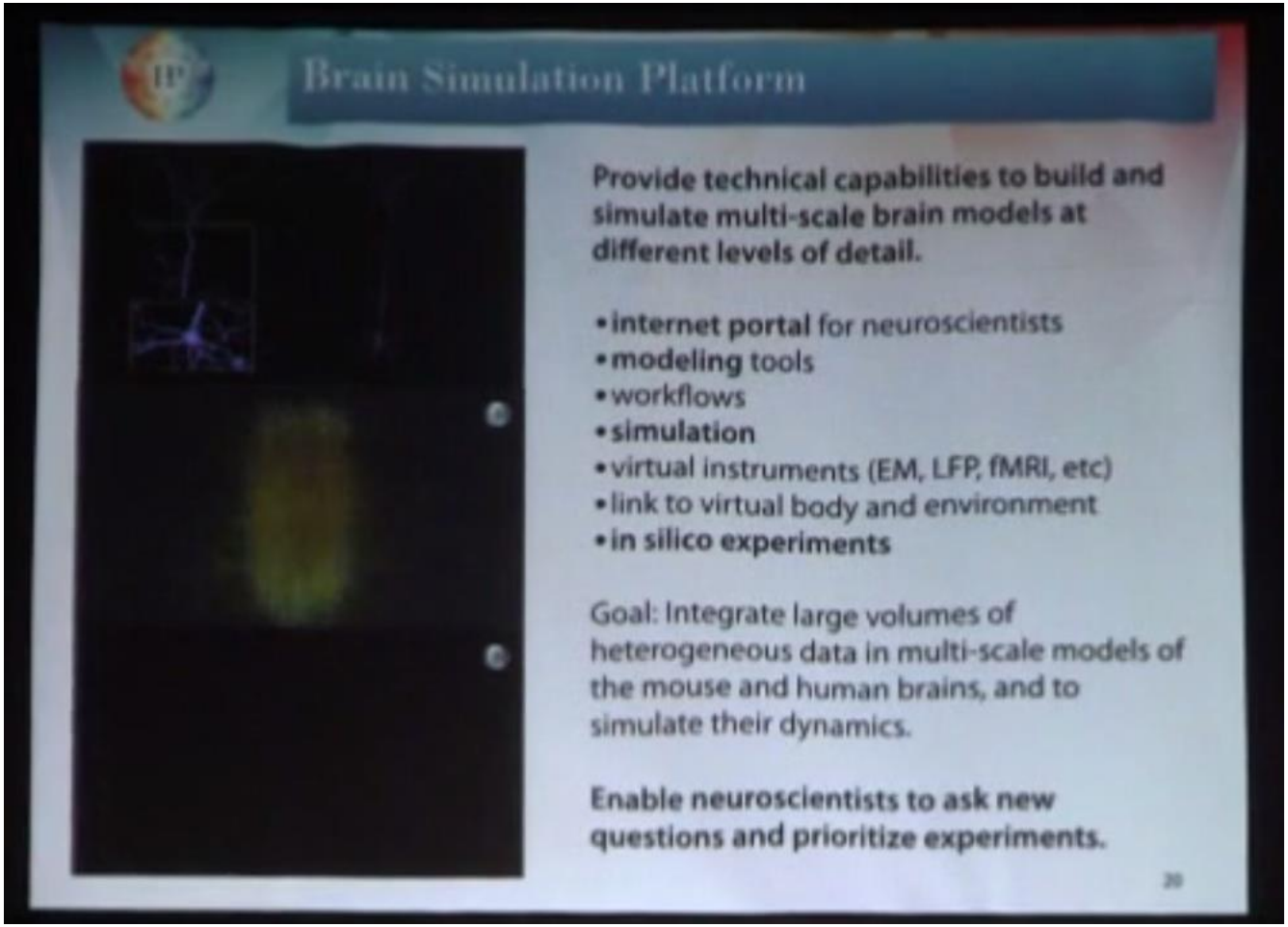
THE PROJECT

CyVerse is funded by the National Science Foundation's Directorate for Biological Sciences. We are a dynamic virtual organization led by the University of Arizona to fulfill a broad mission that spans our partner institutions: Texas Advanced Computing Center, Cold Spring Harbor Laboratory, and the University of North Carolina at Wilmington.

CyVerse fills a niche created by the computing epoch and a rapidly evolving world. Developing solutions to today's grand scientific challenges means that we must understand how the organisms that contribute to our food, fuels, and ecosystem are shaped by interactions with their environment. CyVerse provides life scientists with powerful computational infrastructure to handle huge datasets and complex analyses, thus enabling data-driven discovery. Our powerful extensible platforms provide data storage, bioinformatics tools, image analyses, cloud services, APIs, and more.

Neuroscience

- Sean Hill, European Brain Initiative…



Neuroscience

NSF 16-047

Dear Colleague Letter: National Brain Observatory: A Phased Approach for Developing a National Research Infrastructure for Neuroscience

February 19, 2016

Dear Colleagues:

With this Dear Colleague Letter (DCL), the National Science Foundation (NSF) is announcing the intention to foster the development of a national research infrastructure for neuroscience (National Brain Observatory) to support collaborative and team science for achieving a comprehensive understanding of the brain in action and context.

- Facilitates the emergence of a coherent national infrastructure comprising the above shared and accessible tools, resources and networks that will allow rapid integration, analysis, and modeling of brain data associated with behaviors from multi-disciplinary projects and enable large-scale collaborative research efforts nationally and internationally that will advance our understanding of brain structure and function.

Conclusions

- Core VO design/architecture is strong and has been adopted by major data providing organizations in astronomy; now informing data system design in other fields
- Federated, distributed systems have many benefits…
  - Scalability
  - Flexibility
  - Distributed curation by experts
  - Communities can build on core infrastructure
- …and challenges
  - Consensus must be reached on metadata standards and realistic goals for interoperability

https://www.nist.gov/mml/odi