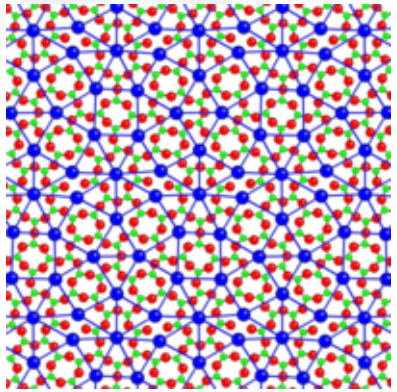
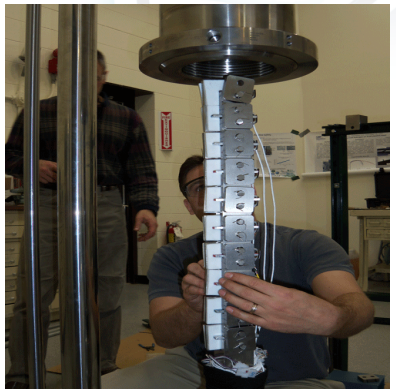
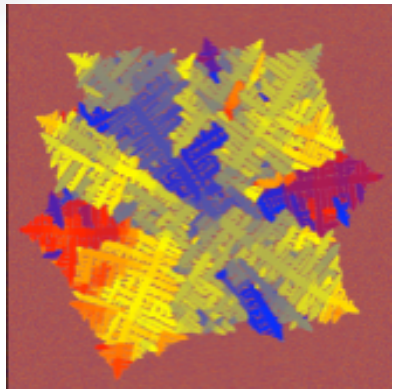






NIST Research Data Framework (RDaF)

Dr. Robert Hanisch
Director, Office of Data and Informatics (ODI)
Material Measurement Laboratory
US National Institute of Standards and Technology

About NIST and ODI

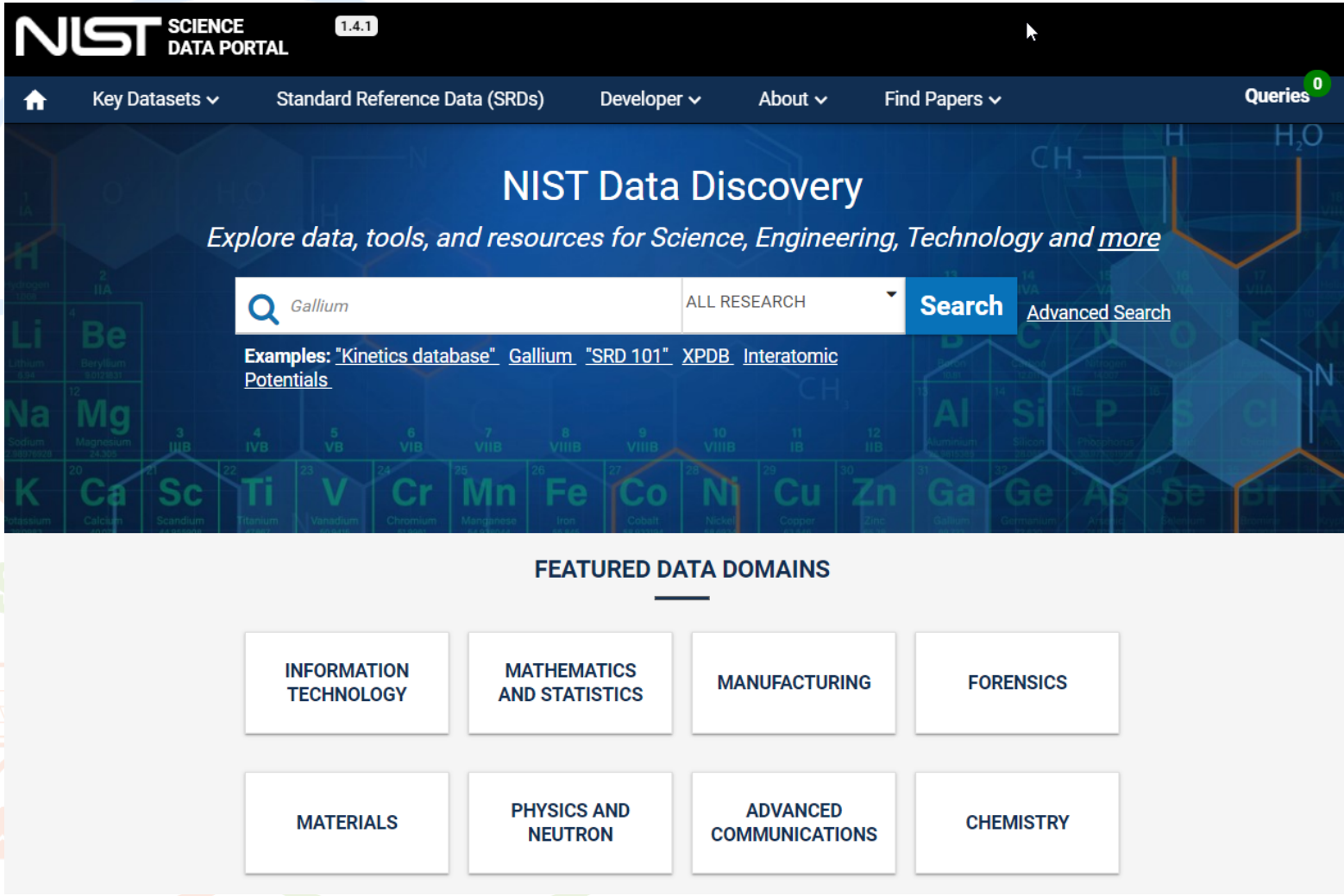
- The National Institute of Standards and Technology is a federal agency under the US Department of Commerce
  - Known as the National Bureau of Standards until 1988, originally founded in 1901
- Non-regulatory
- State of the art in measurement science and technology
- US National Metrology Institute, amongst network of 103 NMIs globally organized under the Bureau International des Poids et Mesures (BIPM, or International Bureau of Weights and Measures), Paris
- ~3,000 staff at NIST (Gaithersburg, Maryland headquarters; Boulder, Colorado; Charleston, South Carolina; Brookhaven National Laboratory)
- 6 major research laboratories
  - Material Measurement Laboratory
    - Office of Data and Informatics (15 people)

Primary ODI Activities

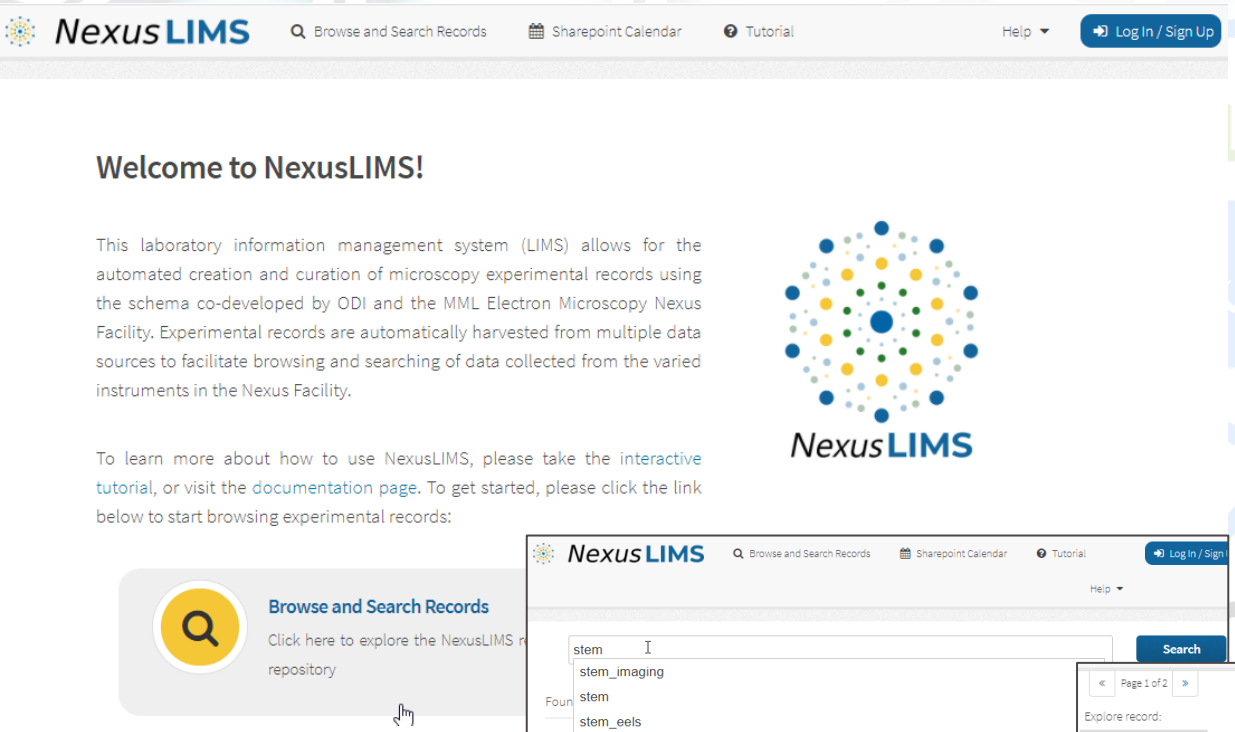
- Data management
  - Public Data Repository (PDR) and Science Data Portal (SDP), data.gov compliance
  - Laboratory Information Management Systems (LIMS)
  - Configurable Data Curation System (CDCS), Python-based metadata extractors (HyperSpy)
  - Data Management Plans (DMPs)
- Standard Reference Data (SRD)
- Informatics and analytics
- External engagements

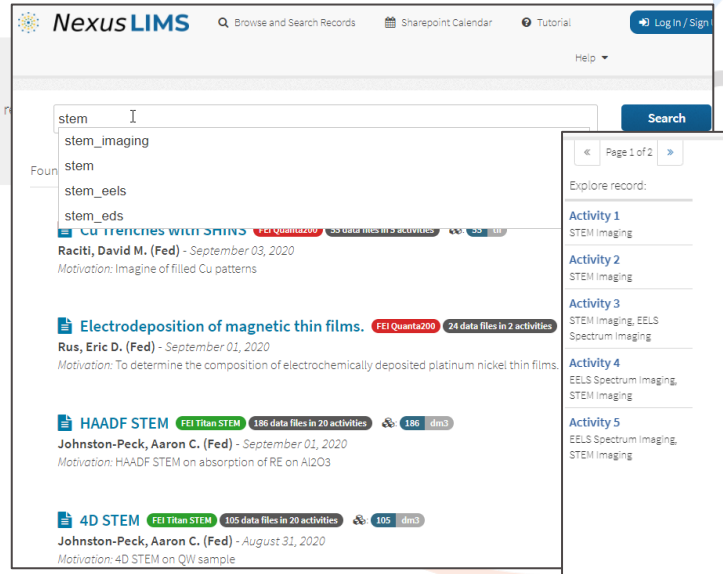
Science Data Portal and Public Data Repository

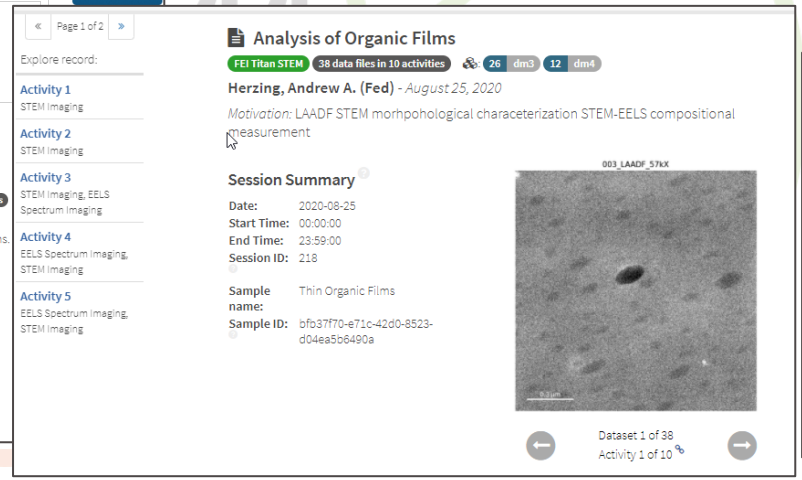
- Modern website for search and discovery of NIST public data sets
  - https://data.nist.gov
- Developed and operated by ODI for NIST
  - Front end to the NIST Public Data Repository
  - Implements the NIST taxonomy for research domains
- Open source code base – hosted on github/ USNISTGOV

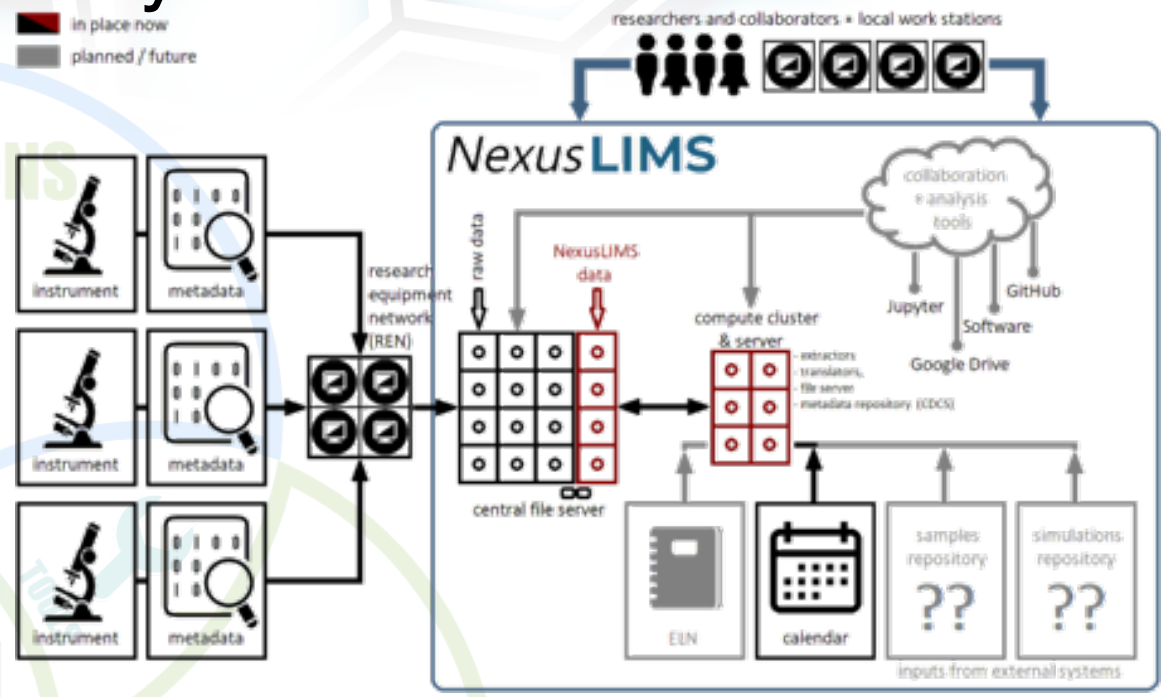


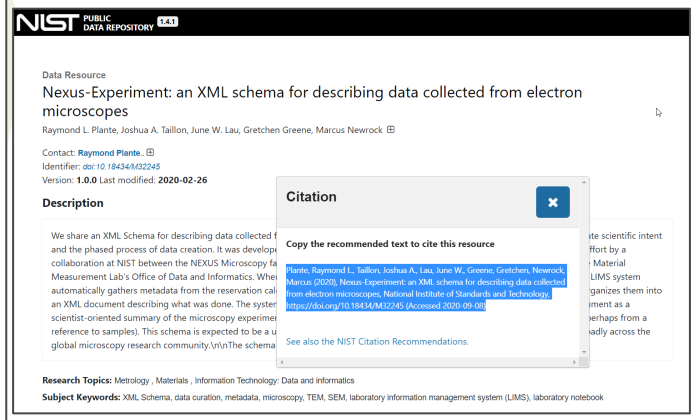
Laboratory Information Management Systems





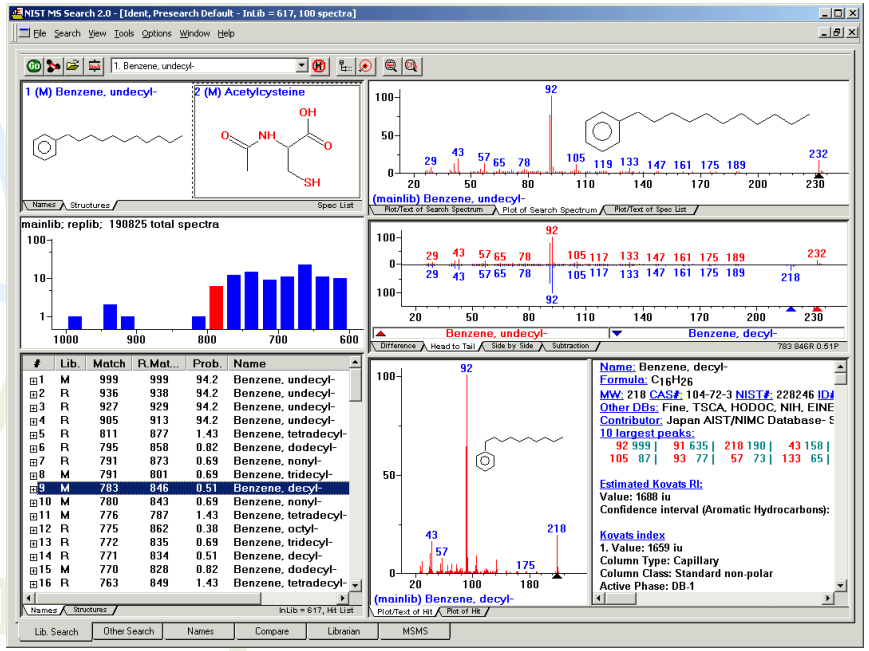






NIST Standard Reference Data

- Most highly vetted data products of NIST
  - SRD Act of 1968
- 65 databases, free and subscription based
- 6,000 units sold/year as downloads and agreements including royalties on instrument sales
- Online SRD Metrics
  - 2M views a month webbook.nist.gov
  - 300K views a month XPS – NIST X-Ray Photoelectron Spectroscopy Database



Stephen E. Stein (2014), NIST/EPA/NIH Mass Spectral Library with Search Program – SRD 1a, National Institute of Standards and Technology, https://doi.org/10.18434/T4H594 (Accessed 2020-09-08)

Informatics and Analytics Support

Data Informatics Resources

- Python and R
- AI and Machine Learning
- Data Analytics and Uncertainty Quantification
- Data Seminars and Training
- Scientific computing

Data Informatics Resources

A curated collection of data informatics references and learning materials relevant to NIST's mission in the materials, chemical, and biological sciences.

Data Seminars and Training

Information about Software/Data Carpentry, Python and R Slack channels, ODI seminars, and other events

Python and R

Programming languages used widely in science and engineering

Artificial Intelligence and Machine Learning

Instructional material related to applications of AI/ML in scientific research

Data analysis and uncertainty quantification

References and resources, including some maintained by the NIST Statistical Engineering Division

Scientific computing links

Includes high performance computing (Enki, Nisaba, etc.), scientific software, and data storage

External Engagement

- Commerce Data Governance Board, Data Inventory Working Group
- OSTP/NSTC subcommittees (Subcommittee on Open Science, Subcommittee on Rigor and Integrity of Research)
- Research Data Alliance, CODATA (Digital Representation of Units of Measure task group), GO-FAIR (FAIR Digital Object Framework), World Data Service (Technical Advisory Board)
- Digital SI (BIPM/CIPM), Digital Calibration Certificates
- Commerce, Energy, NASA, Defense Information (CENDI) network
- National Academies Roundtable, Incentives for Open Data
- Association of American Universities (AAU) / Association of Public and Land-Grant Universities (APLU) / Association of Research Libraries (ARL) workshops on improving public access to research data
- Materials Research Data Council (MaRDaC) / Materials Research Data Network (MaRDaN)

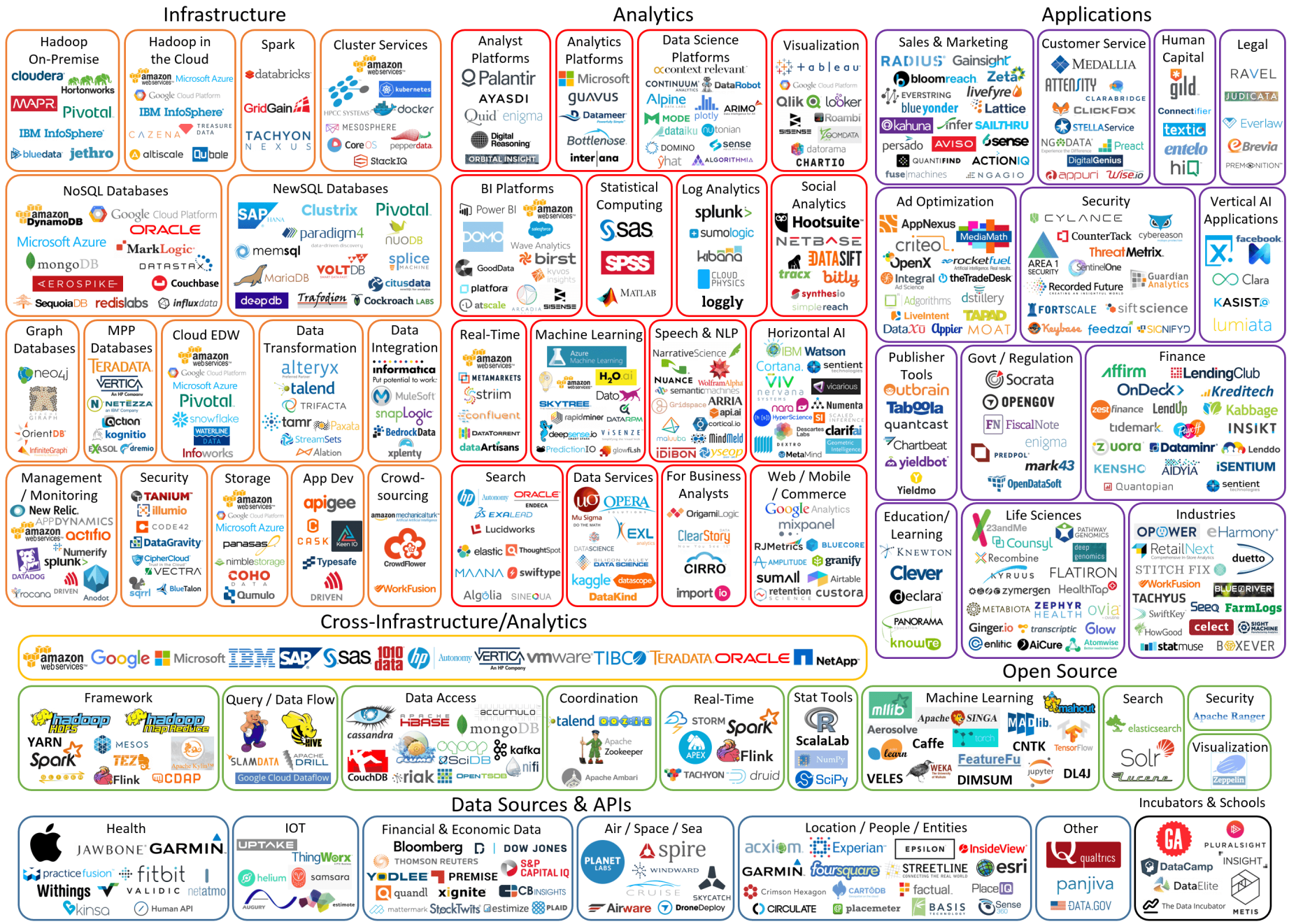
What is a Research Data Framework?

- A map of the research data space: who, what, where, why, when?
- A dynamic guide for the various stakeholders in research data to understand best practices for research data management and dissemination
- A resource for understanding costs, benefits, and risks associated with research data management
- A consensus document based on inputs and conversations amongst the stakeholders in research data

Why a Research Data Framework?

- Research data ecosystem is very complex!
  - Lots of players, various funding models and sustainability plans
  - How long should data be kept?
  - How should data quality be assessed?
  - How do we measure the value of research data?

Big Data Landscape 2016 (Version 3.0)



Last Updated 3/23/2016

FIRSTMARK

Stakeholders

- Government agencies
- National laboratories
- Universities and research libraries
- Data repositories
- Scholarly publishers
- Professional societies
- National and international collaboration organizations (e.g., CENDI, BRDI, CODATA, RDA, WDS, GO-FAIR)
- Standards bodies
- Funders (public and private)
- Industry and the private sector
- Researchers
- General public

Why a Research Data Framework?

- Leverage research data to address global challenges

United Nations Sustainable Development Goals (SDGs)

RDaF Benefits

- **Increase research integrity** with quality data and improved transparency of the research process
- **Reduce costs and maximize efficiency** by establishing best practices for data management
- **Guide risk management and reduction** through assessment of risk positions and roadmaps for improvement
- **Increase scientific discovery and innovation** with the FAIR principles (Findable, Accessible, Interoperable, Reusable) for better utilization of data

National and International Need

- Data is proliferating at an exponential rate
- Data management is complex and confusing
- Mismanaged data has dire social and economic consequences, including loss of global leadership in critical technical fields
- The U.S. needs a coordinated effort to establish a research data infrastructure, but research data are global in nature so international collaboration / coordination is necessary
- NIST is well-positioned to lead the project; our business is consensus building through being a neutral convener of diverse communities

Process

- Pilot program to provide an overall guide to the actors and stakeholders in the research data space
- NIST Cybersecurity Framework is the model
- Community consensus, not NIST imposition
- If I am a ____ , then I need to know ____ .
- Initial scoping workshop held in December 2019 at NIST
  - 50 invited participants representing stakeholders, both US and international

Research Data Framework

Robert Hanisch

Director, Office of Data and Informatics

Material Measurement Laboratory

National Institute of Standards and Technology

Bonnie Carroll

Founder & CDO, Information International Associates

Secretary General, CODATA

RDaF Steering Group

Laura Biven, NIH

Mercé Crosas, Harvard

Josh Greenberg, Sloan

Hilary

Hanahoe, RDA

Heather Joseph, SPARC

Barend Mons, CODATA and GO-FAIR

Beth Plale, NSF

Anita de Waard, Elsevier

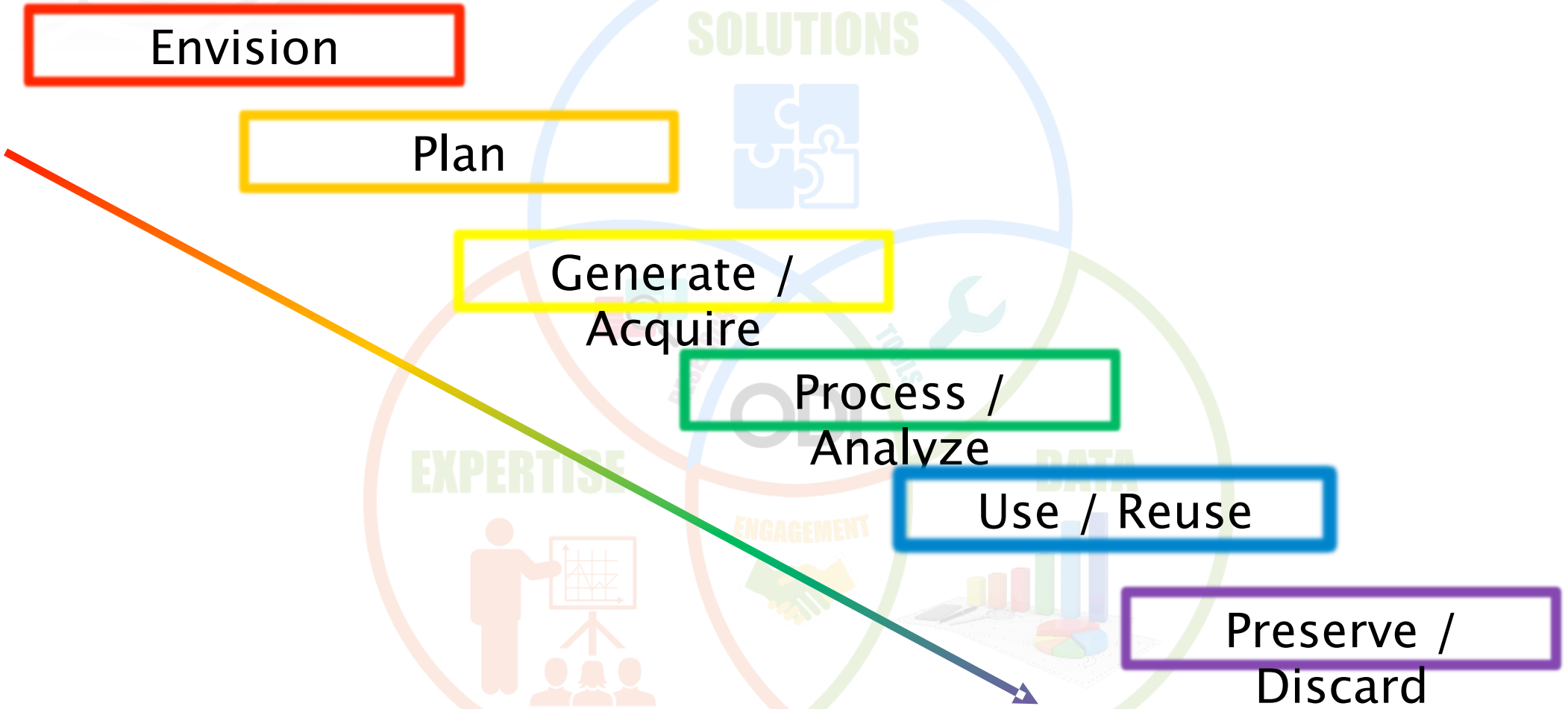
Mark Da

Workshop Summary

- **Status:** Confirmed support by government agencies, academic organizations, private sector companies, not-for-profit organizations, and international stakeholders.
- **Next Steps:** Management commitment to complete the scoping, pilot testing, and community building for the Framework.
  - Proposed pilots
    - Materials science
    - Universities and research libraries (AAU, APLU, ARL)

Will need cooperation across government to move fully forward with the Framework.

RDaF Structure Based on “Functions”



RDaF Structure

Function	*Category*	Subcategory
1) Envision	*Data governance*	Data vision, data policy
		Data management organization
		Data quality, privacy, ethics
	Community engagement	Communication, interactions
		Cross-domain
	Data culture	FAIR principles
		Value of data
		Roles and responsibilities
	Reward structure	Value of data professionals
		Incentives for sharing and re-use

RDaF Structure

Function	*Category*	Subcategory
2) Plan	*Costs*	Cost-benefit analysis Costs by data lifecycle stage
	Funding	Direct, overhead, mixed, other
	Data objects	Data (experimental, simulation) Software, instruments Publications, presentations
	Data management planning	DMPs (intent, update) Formats, standards

RDaF Structure

Function	*Category*	Subcategory
3) Generate / simulation Acquire	*Sources*	In-house, experiment or Collected from external sources
	Experiment	Instruments and their metadata Measurement protocols Data capture and recording
	Simulation	Commercial or custom software Metadata capture and recording
	External sources	Identification, provenance Metadata harvesting
	Data formats	Standards development and/or adoption

RDaF Structure

Function	*Category*	Subcategory
4) Process / Analyze data	*Provenance*	Origin, version, time-stamp Data copied or derived from other
management	*Data architecture*	Design, security, configuration
		Hosting and storage On-premise or Cloud
	Software	Commercial or custom software Versions
maintenance		Stability, resilience, adaptability, Workflows, ELNs, LIMs
	Publishing, curation	Processes, tools, stewardship Metadata

RDaF Structure

Function	*Category*	Subcategory
5) Use / Reuse	*Legal and licenses*	Ownership, IP, rights and restrictions
		Agreements, permissions
		Citation expectations
	Data access	Internal, external
		APIs
		Downloads vs. visiting
	Analysis tools	AI/ML
		Performance
	Impact	Usage tracking, citation

RDaF Structure

Function	*Category*	Subcategory
6) Preserve / Discard	*Sustainability*	Longevity requirements Who pays? Orphan data sets
	Preservation	Media and media migration Back-up Repositories (domain, institutional, general) Migration between organizations
	Retention and disposition	Decision processes End-of-life (dark archives, deaccession, gravestones)

Status

- Briefed OSTP Subcommittee on Open Science and OSTP Director Kelvin Droegemeier (03/26/2020)

- Developed roadmap and structure, vetted with Steering Group
- Seeking ~$500k to fund two pilots: materials science and research universities/libraries/scholarly publishers
  - NIST plus other agencies/laboratories, either $$ or in-kind support
  - Professional societies
  - Scholarly publishing community

https://www.nist.gov/programs-projects/research-data-

NIST Frameworks

Framework for Improving Critical Infrastructure Cybersecurity:

https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf

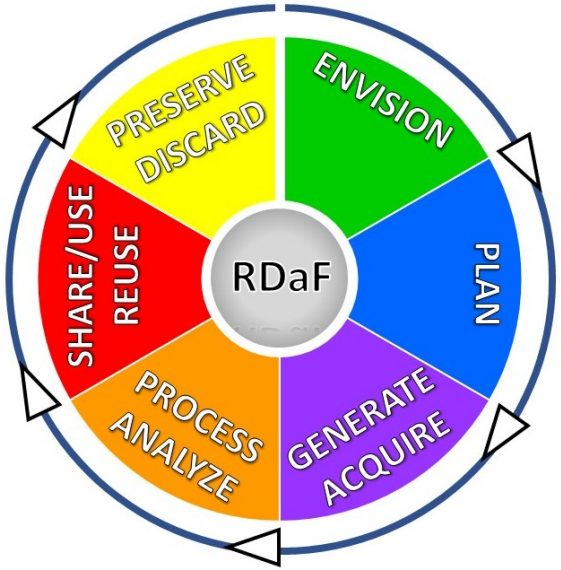
NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, September 6, 2019 (Preliminary Draft)

https://www.nist.gov/system/files/documents/2019/09/09/nist_privacy_framework_preliminary_draft.pdf

NIST Big Data Interoperability Framework: Volume 1, Definitions October 2019 Version 3

https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1r2.pdf

RDaF Summary



- Successful in building community interest and engage
  - Diverse stakeholders
  - National and international
- Challenges
  - Resources
  - Timeliness: the research data ecosystem is changing rapidly. How to keep pace and assure ongoing updates?
  - Controlling scope and scale
- Strategy for moving forward
  - Start with pilot projects in order to validate approach and re-tune as necessary
  - Collaborate with other federal agencies, professional societies, scholarly publishing community, etc., to garner the necessary resources and take advantage of work in progress

Contact

Robert Hanisch
robert.hanisch@nist.gov
https://nist.gov/people/robert-hanisch