

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
- ~5,000 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

The screenshot displays the NIST Science Data Portal interface. At the top, the logo reads "NIST SCIENCE DATA PORTAL" with a version number "1.4.1". A navigation bar includes links for "Key Datasets", "Standard Reference Data (SRDs)", "Developer", "About", and "Find Papers", along with a "Queries" counter showing "0". The main heading is "NIST Data Discovery" with the tagline "Explore data, tools, and resources for Science, Engineering, Technology and more". A search bar contains the text "Gallium" and a dropdown menu set to "ALL RESEARCH". A "Search" button and a link to "Advanced Search" are also present. Below the search bar, examples of search results are listed: "Kinetics database", "Gallium", "SRD 101", "XPDB", and "Interatomic Potentials". The background features a periodic table of elements. A section titled "FEATURED DATA DOMAINS" lists eight categories in a grid: INFORMATION TECHNOLOGY, MATHEMATICS AND STATISTICS, MANUFACTURING, FORENSICS, MATERIALS, PHYSICS AND NEUTRON, ADVANCED COMMUNICATIONS, and CHEMISTRY.

Laboratory Information Management Systems

Welcome to NexusLIMS!

This laboratory information management system (LIMS) allows for the automated creation and curation of microscopy experimental records using the schema co-developed by ODI and the MML Electron Microscopy Nexus Facility. Experimental records are automatically harvested from multiple data sources to facilitate browsing and searching of data collected from the varied instruments in the Nexus Facility.



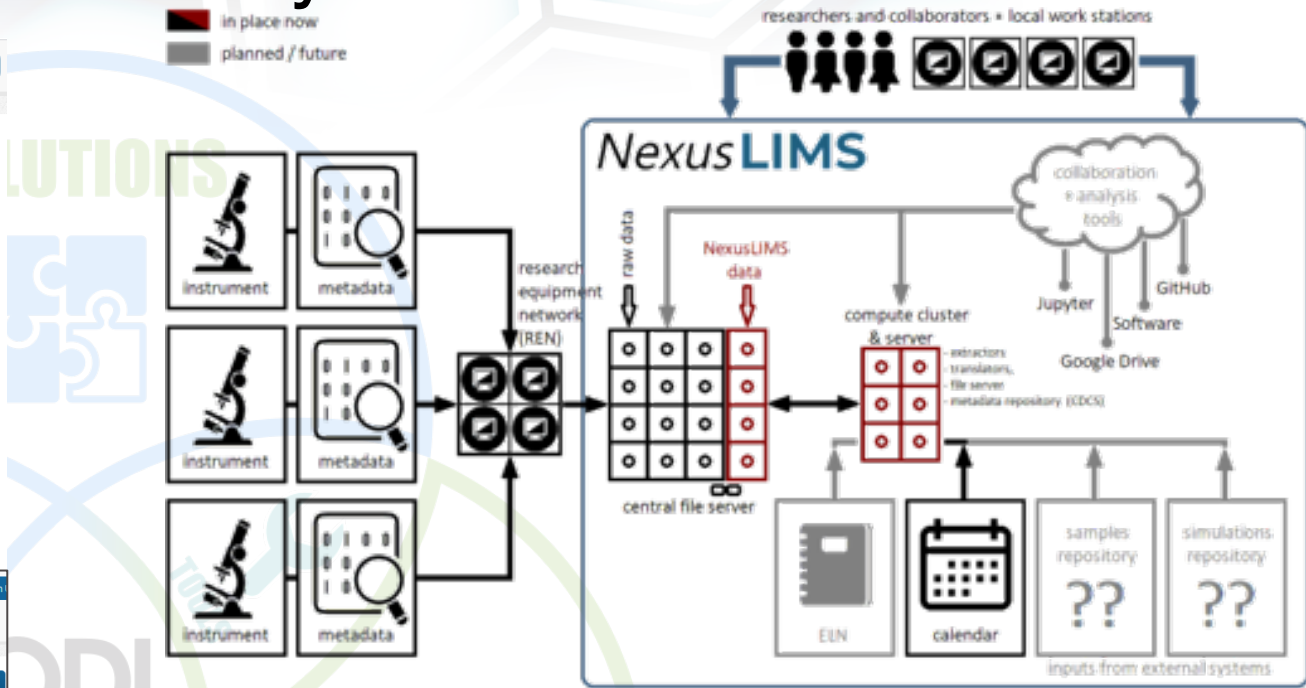
To learn more about how to use NexusLIMS, please take the [interactive tutorial](#), or visit the [documentation page](#). To get started, please click the link below to start browsing experimental records:

Browse and Search Records
Click here to explore the NexusLIMS repository

Analysis of Organic Films
Herzing, Andrew A. (Fed) - August 25, 2020
Motivation: LAADF STEM morphological characterization STEM-EELS compositional measurement

Session Summary
Date: 2020-08-25
Start Time: 00:00:00
End Time: 23:59:00
Session ID: 218

Sample name: Thin Organic Films
Sample ID: bfb37f70-e71c-4200-8523-d04ea306490a



NIST PUBLIC DATA REPOSITORY
Data Resource
Nexus-Experiment: an XML schema for describing data collected from electron microscopes
Raymond L. Plante, Joshua A. Tallon, June W. Lau, Gretchen Greene, Marcus Newrock

Contact: Raymond Plante
Identifier: doi:10.18434/M32245
Version: 1.0.0 Last modified: 2020-02-26

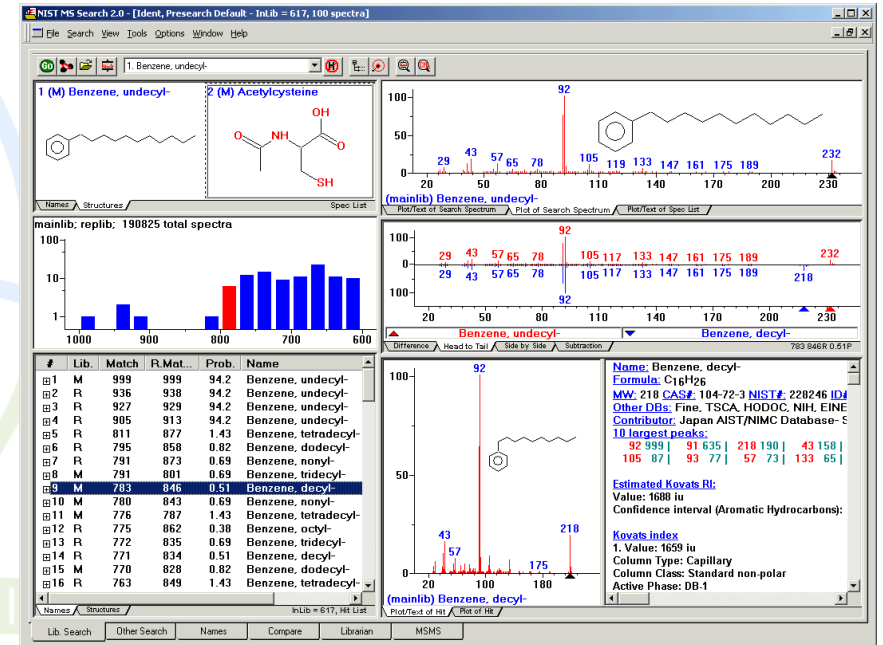
Description
We share an XML Schema for describing data collected and the phased process of data creation. It was developed in collaboration at NIST between the NEXUS Microscopy Facility Measurement Lab's Office of Data and Informatics. When automatically gathers metadata from the reservation call an XML document describing what was done. The system scientist-oriented summary of the microscopy experimenter reference to samples). This schema is expected to be a global microscopy research community.

Citation
Copy the recommended text to cite this resource
Plante, Raymond L., Tallon, Joshua A., Lau, June W., Greene, Gretchen, Newrock, Marcus (2020). Nexus-Experiment: an XML schema for describing data collected from electron microscopes. National Institute of Standards and Technology. <https://doi.org/10.18434/M32245> (Accessed 2020-09-08)

Research Topics: Metrology, Materials, Information Technology, Data and Informatics
Subject Keywords: XML Schema, data curation, metadata, microscopy, TEM, SEM, laboratory information management system (LIMS), laboratory notebook

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)

Infrastructure

Hadoop On-Premise
 cloudera, Hortonworks, Pivotal, IBM InfoSphere, bluedata, jethro

Hadoop in the Cloud
 Amazon Web Services, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, du bale

Spark
 databricks, GridGain, TACHYON NEXUS

Cluster Services
 Amazon Web Services, Kubernetes, Docker, HPCC SYSTEMS, MESOSPHERE, Core OS, pepperdata, StackIQ

Analytics

Analyst Platforms
 Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITAL INSIGHT

Analytics Platforms
 Microsoft, GUAVUS, Datameer, Bottlenose, interlana

Data Science Platforms
 context relevant, DataRobot, CONTINUUM, Alpine, MODE, dataiku, DOMINO, yhat, ALGORITHMIA

Visualization
 Tableau, Google Cloud Platform, Qlik, looker, Roambi, SISENSE, YOODATA, datarama, CHARTIO

Applications

Sales & Marketing
 RADIUS, Gainsight, bloomreach, Zeta, EVERSTRING, livefyre, blueyonder, Lattice, kahuna, inifer, SAILTHRU, persado, AVISO, sense, QUANTIFIND, ACTIONIQ, fuse|machines, ENAGGIO

Customer Service
 MEDALLIA, ATTENTIFY, CLARABRIDGE, CLICKFOX, STELLAService, NGDATA, Preact, DigitalGenius, appurfi, Wiseio

Human Capital
 gild, Connectifier, textic, entelo, hiQ

Legal
 RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

NoSQL Databases
 Amazon DynamoDB, Google Cloud Platform, Microsoft Azure, ORACLE, mongoDB, DATASTAX, KEROPIKE, Couchbase, SequoiaDB, redislabs, influxdata

NewsQL Databases
 SAP HANA, Clustrix, Pivotal, paradigm4, nuODB, memsql, splice MACHINE, MariaDB, VOLTDB, citusdata, deep db, Trafodion, Cockroach LABS

BI Platforms
 Power BI, Amazon Web Services, DOMO, Wave Analytics, GoodData, birst, kyvos insights, platforma, atscale, ARCADIA, SISENSE

Statistical Computing
 SAS, SPSS, MATLAB

Log Analytics
 Splunk, sumologic, kibana, CLOUD PHYSICS, loggly

Social Analytics
 Hootsuite, NETBASE, DATASIFT, track, bitly, synthetio, simplereach

Ad Optimization
 AppNexus, MediaMath, Criteo, OpenX, rocketfuel, Integral, theTradeDesk, Ad Science, Algorithms, dstillery, LiveIntent, TAFAD, DataXu, Appier, MOAT

Security
 Cylance, CounterTack, cybereason, ThreatMetrix, AREA 1 SECURITY, SentinelOne, Recorded Future, Guardian Analytics, FORTSCALE, sift science, Keybase, feedzai, SIGNIFYD

Vertical AI Applications
 Facebook, Clara, KASIST@, lumiata

Graph Databases
 neo4j, GIAPH, OrientDB, InfiniteGraph

MPP Databases
 TERADATA, VERTICA, NETEZZA, Qcton, Kognitio, EXASOL, dremio

Cloud EDW
 Amazon Web Services, Microsoft Azure, Pivotal, snowflake, WATERLINE DATA, Infoworks

Data Transformation
 Alteryx, talend, TRIFACTA, tamr, StreamSets, Alation

Data Integration
 Informatica, MuleSoft, snapLogic, Bedrock Data, xplenty

Real-Time
 Amazon Web Services, METAMARKETS, Streamium, Confluent, DATATORRENT, dataArtisans

Machine Learning
 Azure Machine Learning, H2O, Amazon Web Services, SKYTREE, rapidminer, DATARPM, deeppearl, VISENZE, PredictionIO, glowfish

Speech & NLP
 NarrativeScience, NUANCE, semantic machines, ARRIA, apiai, Gridspace, maluba, MindMeld, IDIBON, yseop

Horizontal AI
 IBM Watson, Cortana, sentient, viv, vicarious, nara, Numenta, HyperScience, SI, Datascales, clarifai, MetaMind

Publisher Tools
 Outbrain, Taboola, quantcast, Chartbeat, yieldbot, Yieldmo

Govt / Regulation
 Socrata, OPENGOV, FN FiscalNote, enigma, PREDPOL, mark43, OpenDataSoft

Finance
 Affirm, LendingClub, OnDeck, Kreditech, Crest Finance, LendUp, Kabbage, tidemark, Insikt, ZUORA, Dataminr, Lenddo, KENSHC, AIDYA, ISENTIUM, Quantopian, Sentient Technologies

Management / Monitoring
 New Relic, APPDYNAMICS, Amazon Web Services, actifio, NUMERIFY, splunk, DATADOG, DRIVEN, Anodot

Security
 TANIUM, illumio, CODE42, DataGravity, CIPHERCLOUD, VECTRA, sqrl, BlueTalon

Storage
 Amazon Web Services, Microsoft Azure, panasas, nimblestorage, COHO DATA, Qumulo

App Dev
 Apigee, CASK, KEEN IO, Typesafe, DRIVEN

Crowd-sourcing
 Amazon Mechanical Turk, CrowdFlower, WorkFusion

Search
 HP, Autonomy, ORACLE, ENDECA, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swifttype, Algolia, SINEQUA

Data Services
 UC OPERA, MU SIGMA, EXL, DATASCIENCE, DATA SCIENCE, kaggle, datascopie, DataKind

For Business Analysts
 OrigamiLogic, ClearStory, CIRRO, import io

Web / Mobile / Commerce
 Google Analytics, mixpanel, RJMetrics, BLUECORE, AMPLITUDE, granify, sumall, Airtable, retention, custora

Education / Learning
 KNEWTON, Clever, Declara, PANORAMA, knowre

Life Sciences
 23andMe, Counsyl, RECOMBINE, KYRUS, FLATIRON, zymogen, HealthTap, METABIOTA, ZEPHYR HEALTH, ovia, Ginger.io, transcriptic, Glow, enlithic, AiCure, Atomwise

Industries
 OPOWER, Harmony, RetailNext, STITCH FIX, WorkFusion, BLUE RIVER, TACHYUS, SwiftKey, Seeq, FarmLogs, HowGood, select, SIGHT MACHINE, statmuse, BOEVER

Cross-Infrastructure/Analytics

Amazon Web Services, Google, Microsoft, IBM, SAP, SAS, HP, Autonomy, VERTICA, vmware, TIBCO, TERADATA, ORACLE, NetApp

Open Source

Framework
 Hadoop, HADOOP HDFS, YARN, Spark, MESOS, TEZ, Apache Kylin, Flink, CDAP

Query / Data Flow
 SLAMDATA, HIVE, DRILL, Google Cloud Dataflow

Data Access
 ACCUMULO, HBASE, mongoDB, cassandra, kafka, CouchDB, riak, OPENTDB, nifi

Coordination
 talend, Apache Ambari

Real-Time
 STORM, Spark, APEX, Flink, TACHYON, druid

Stat Tools
 Scalalab, NumPy, SciPy

Machine Learning
 mlilb, Aerosolve, Apache SINGA, MADlib, caffe, CNTK, TensorFlow, WEKA, FeatureFu, jupyter, DL4J, DIMSUM

Search
 elasticsearch, Solr, Lucene

Security
 Apache Ranger, Zeppelin

Data Sources & APIs

Health
 Apple, JAWBONE, GARMIN, practicefusion, fitbit, Withings, VALIDIC, netatmo, kinsco, Human API

IOT
 UPTAKE, ThingWorx, helium, samsara, AUGURY, estimate

Financial & Economic Data
 Bloomberg, THOMSON REUTERS, DOW JONES, YODLEE, PREMISE, S&P CAPITAL IQ, quandl, xignite, CB INSIGHTS, mattermark, Stocktwits, estimate, PLAID

Air / Space / Sea
 PLANET LABS, spire, WINDWARD, CRUISE, SKYCATCH, Airware, DroneDeploy

Location / People / Entities
 acxiom, Experian, EPSILON, InsideView, GARMIN, foursquare, STREETLINE, Crism Hexagon, CARTODB, factual, PlaceIQ, CIRCLATE, placemeter, BASIS, Sense

Other
 qualtrics, panjiva, DATA.GOV

Incubators & Schools
 GA, PLURALSIGHT, DataCamp, INSIGHT, DataElite, The Data Incubator, METIS

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

NIST



RDaF Workshop

December 5-6, 2019

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
Harvard, RDA



Mercé Crosas, Harvard



Josh Greenberg, Sloan



Hilary



Heather Joseph, SPARC



Barend Mons, CODATA
and GO-FAIR



Beth Plale, NSF



Anita de Waard, Elsevier



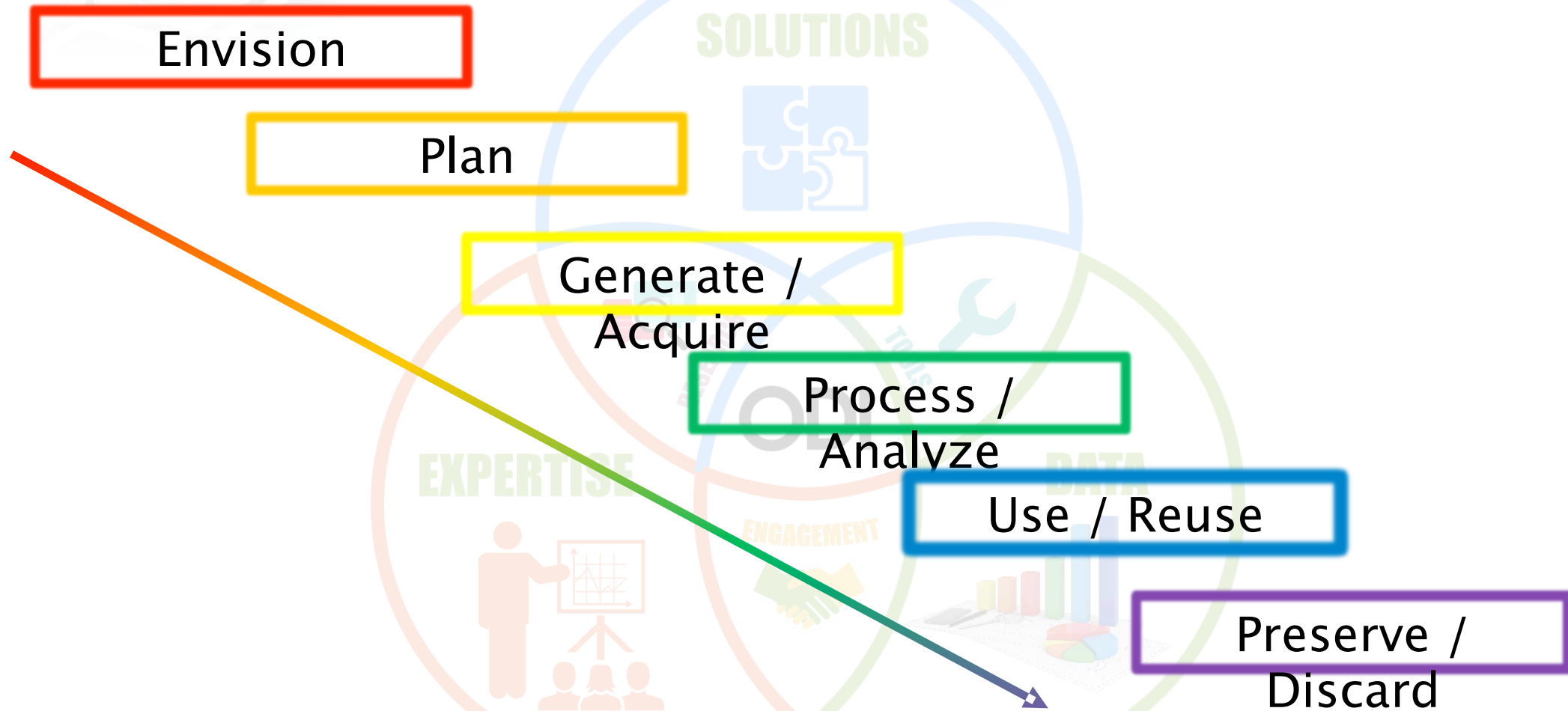
Mark
Da

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

2) Plan

Category

Costs

Funding

Data objects

*Data management
planning*

Subcategory

Cost-benefit analysis
Costs by data lifecycle stage

Direct, overhead, mixed, other

Data (experimental, simulation)
Software, instruments
Publications, presentations

DMPs (intent, update)
Formats, standards

RDaF Structure

Function

Category

Subcategory

3) **Generate** / simulation
Acquire

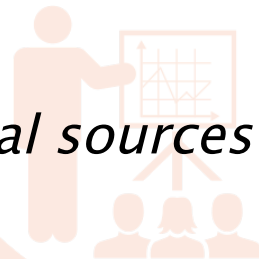
Experiment



In-house, experiment or
Collected from external sources

Simulation

EXPERTISE



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

maintenance

Software

Hosting and storage
On-premise or Cloud

Commercial or custom software
Versions
Stability, resilience, adaptability,

Workflows, ELNs, LIMs

Publishing, curation

Processes, tools, stewardship

Metadata

RDaF Structure

Function

Category

Subcategory

5) Use / Reuse
restrictions

Legal and licenses

Ownership, IP, rights and

Agreements, permissions
Citation expectations

Data access



Internal, external
APIs
Downloads vs. visiting

Analysis tools

EXPERTISE



AI/ML
Performance

Impact

Usage tracking, citation

DATA



RDaF Structure

Function

6) Preserve / Discard

Category

Sustainability

Subcategory

Longevity requirements

Who pays?

Orphan data sets

Preservation

Media and media migration

Back-up

Repositories (domain, institutional, general)

Migration between organizations

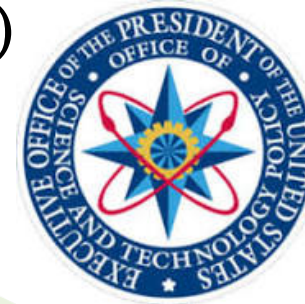
Retention and disposition

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

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)

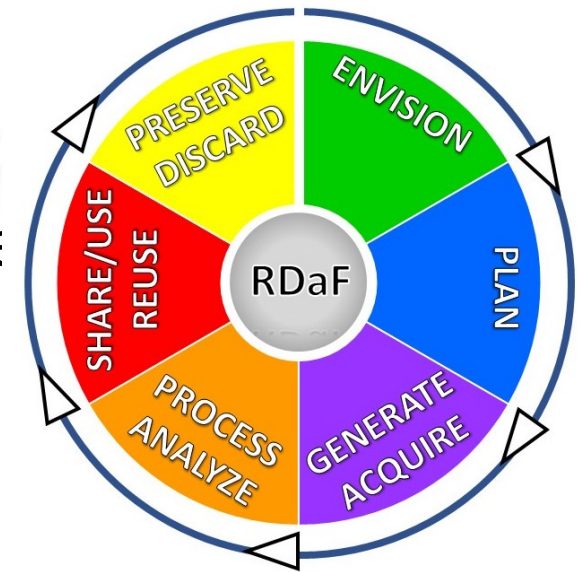
https://www.nist.gov/system/files/documents/2019/09/09/nist_privacy_framework_preliminary_draft.pdf

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 engagement
 - 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

SOLUTIONS

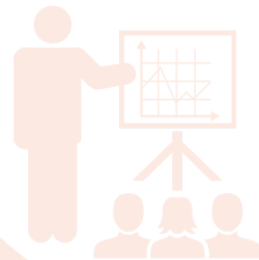
Robert Hanisch

robert.hanisch@nist.gov

<https://nist.gov/people/robert-hanisch>



EXPERTISE



DATA

