The Materials Genome Initiative, Data, Open Science, and NIST

James A Warren
Director, Materials Genome Program
Material Measurement Laboratory
National Institute of Standards and Technology
Executive Secretary, NSTC Subcommittee on MGI

Science advances one funeral at a time - Max Planck
The Perfect is the Enemy of the Good - Voltaire
“To help businesses discover, develop, and deploy new materials twice as fast, we’re launching what we call the Materials Genome Initiative. The invention of silicon circuits and lithium ion batteries made computers and iPods and iPads possible, but it took years to get those technologies from the drawing board to the market place. We can do it faster.”

-President Obama (6/11)
There are two groups of people that don’t like the name Materials Genome Initiative
Develop a Materials Innovation Infrastructure

Achieve National goals in energy, security, and human welfare with advanced materials

Equip the next generation materials workforce
A Multi-Agency Effort
THE MGI SUBCOMMITTEE (SMGI)

who we are

- MGI Subcommittee, Committee on Technology, NSTC
- First meeting April, 2012
- Membership includes the Federal agencies: NIST, DOE, DOD, NSF, NASA, NIH, US Geological Survey, National Nuclear Security Administration, DARPA, and Office of Management and Budget
- Co-chairs: OSTP (Whitman), AFRL (Ward), DOE (Horton)
- Executive Secretary: NIST (Jim Warren)

what we do

- Coordinate across government
- Convene stakeholders to engage in strategy: Grand Challenge Summits
- Development of a National Strategy for MGI

SIGN UP FOR OUR STAKEHOLDER EMAIL LIST!

www.mgi.gov
MGI National Strategy: 4 Goals

- Enable a Paradigm Shift in Culture
- Integrate Experiments, Computation, Theory
- Facilitate Access to Materials Data
- Equip the Next-Generation Materials Workforce

https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/mgi_strategic_plan_-_dec_2014.pdf
The Relationship between Models, Measurement, Data, and Publication

A simple example to illustrate the key questions, gaps, obstacles to realizing materials by design
Better Models = Less Data

- Take the LHC as an example
- Data produced at 1PB/sec!
- Reduced data saved: 300 MB/s
- That’s a darn good model
- Other end: Biology?
- Materials: In the middle

- But you need to know the model to make sense of the data!
What should we be doing?

Science is characterized by the iteration of experiment and models, yielding higher fidelity with lower uncertainty.

The measurement or computation of a quantity (data) is generally meaningless without the associated quantifying model that defines both the data and its uncertainties.

Thus dissemination of data is ideally the dissemination of the following information:
1. Measured quantities,
2. Associated quantifying models, and
3. Raw data, including the protocols (and equipment) by which it was obtained

see R. Peng Science 2 December 2011: 1226-1227
Consequences of the traditional approach

• High barrier to adoption of methods and results
• Extra expense due to duplication of effort
• Lost opportunities to enable discovery & further science

So Now What?
Formulating the NIST Role in MGI

Some background and further thoughts
NIST Products and Services

Measurement Research
- ~ 2,200 publications per year

Standard Reference Data
- ~ 100 different types
- ~ 6,000 units sold per year
- ~ 226 million data downloads per year

Standard Reference Materials
- ~ 1,300 products available
- ~ 30,000 units sold per year

Calibration Tests
- ~ 18,000 tests per year

Laboratory Accreditation
- ~ 800 accreditations of testing and calibration laboratories
Data mandates in the Federal Government

OSTP “Public Access”
Memo Feb 22, 2013

Executive Order
May 9, 2013

OMB “Open Data”
Memo May 9, 2013

DATA Act, P.L. 113-101
May 9, 2014

For Immediate Release

Executive Order -- Making Open and Machine Readable the New Default for Government Information

EXECUTIVE ORDER

MAKING OPEN AND MACHINE READABLE THE NEW DEFAULT FOR GOVERNMENT INFORMATION

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. General Principles. Openness in government strengthens our democracy, promotes the delivery of efficient and effective services to the public, and contributes to economic growth. As one vital benefit of open government, making information resources easy to find, accessible, and usable can fuel entrepreneurship, innovation, and scientific discovery that improves Americans’ lives and contributes significantly to job creation.

Decades ago, the U.S. Government made both weather data and the Global Positioning System freely available. Since that time, American entrepreneurs and innovators have utilized these resources to create navigation systems, weather newscasts and warning systems, location-based applications, precision farming tools, and much more, improving Americans’ lives in countless ways and leading to economic growth and job creation. In recent years, thousands of Government data resources across fields such as health and medicine, education, energy, public safety, global development, and finance have been posted in machine-readable form for free public use on Data.gov. Entrepreneurs and innovators have continued to develop a vast range of useful new products and businesses using these public information resources, creating good jobs in the process.
Enable & Enhance Exchange
(repositories, disciplines, industries; standards)

Assess & Improve Quality
(Data & Models)

New Methods and Metrologies
(data driven analysis and models)

Materials w/ Targeted Properties
SCOPE: Goals of Initiative

To foster widespread adoption of the MGI Paradigm both across and within the multitude of materials development ecosystems

Goal 1: NIST establishes essential materials data and model exchange protocols

Goal 2: NIST establishes the means to ensure the quality of materials data and models

Goal 3: NIST establishes new methods, metrologies and capabilities necessary for accelerated materials development.
The NIST Portfolio and Wider Questions about Data
Computational Tools

SOFTWARE

DATA INFRASTRUCTURE

HARDWARE

Software Catalogs
Comp Repos
Workflow Tools

Computation
Quality Metrics
&
Data Driven
Materials Science

Best Practices
for HPC

Ingest &
Automated
Capture
Experimental Tools

- DATA INFRASTRUCTURE
  - Automated Capture
- DATA INFRASTRUCTURE
  - Expt. Quality Metrics
  - & Data Driven Materials Science
- METHODS AND MEASUREMENTS
  - Expt. Design
  - New Models
  - Best Practice Guides
- METHODS AND MEASUREMENTS
  - Machine Learning
Office of Data and Informatics

**SRD**
- continue existing SRD distribution
- Quality Framework
- SRD Modes
- assess external need
- new product ideas
- SRMDS
- data streams
- alternative delivery methods
- Open Data Initiative
- Open Govt Directive
- Data.gov

**Research Data**
- deal w/ data deluge
- provide advice to MML bench staff
- gather best practices
- interpret external rules & regulations
- reduce redundancy
- promote cooperation and coherent action
- manage changes in scholarly publishing
- coordinate with
  - WERB
  - Library
  - JResNIST

**Lead/Liaison**
- partner with ITL
- represent MML
- NIST committees
- NSTC & IWGs
- NIH, NSF, DOE
- other Fed Govt
- Research Data Alliance (RDA)
- data standards
- champion proposals
- budget initiatives
- IMS
- inter-agency, RDA

**Data Science**
*The 4th paradigm?*
- will it stand next to
  - theoretical
  - experimental
  - computational
- Cloud
- Statistical Learning
- Big Data
- Knowledge Discovery
- very fast growing
- many new jobs
- new degrees/depts
What’s NIST, the USG and International Community doing about “Data”

• This is really a helicopter view of the broad issues
• You can hopefully not understand much of this part of talk and it shouldn’t matter
• Many issues
  • getting the data “in”
  • getting the data “out”
  • Evaluating the data for quality
  • Using the data for maximum effect
• Who are we working with, who else is a “player”?
Google Materials

Search
Data Discovery, What do we Need?

1. Some sort of interface/api that the user interacts with to deposit their information such that (2) is enabled. This entails
   a. The establishment of **repositories** to store the data
   b. The reposed data should then be **marked up with sufficient metadata** to inform someone else how the data was created, including attribution or provenance information relevant to citation
   c. Assignment of a **persistent digital identifier** (like the DOI for journals) so the data can be cited and discovered by others
   d. Tools must be developed to
      i. enable the ingest of data from computation or experiment for deposition in a repository
      ii. simplify metadata collection, e.g. by
         1. Automatically assigning some basic metadata
         2. Implementing automatic extraction of metadata to the extent possible
      iii. Assign persistent digital identifiers

2. Some sort of interface/api that the user interacts with to find needed information. This requires (at least)
   a. the **registration of the availability of the data into some sort of “registry”** to enable discovery without prior knowledge of the existence of the repository/specific data features.
   b. Various types of **policy** enforcement
   c. Tools to evaluate the **“quality”** of the data
      i. Enabling determinations of the data’s uncertainty, validity, sensitivity, as well as other metadata qualities
      ii. Enabling assessments of the relevance of the data to the question at hand, such as quick-look plotting or imaging capabilities]
   d. Tools for data manipulation
   e. **Terminologies and ontologies** to enhance search
Repositories
Property Calculations on the Interatomic Potential Repository

Zachary Trautt, Materials Measurement Science Division
Chandler Becker, Materials Science and Engineering Division

http://www.ctcms.nist.gov/potentials/
WebFF-A Smart Force-Field Repository of Soft Materials

http://www.nist.gov/mml/msed/polymers/webff.cfm

Frederick R. Phelan Jr., Material Science & Engineering Division, Polymers & Complex Fluids Group
NIST, Gaithersburg, MD
Huai Sun, Aeon Technology Inc. and School of Chemistry and Chemical Engineering, Shanghai Jiao Tong University, Shanghai
Customized DSpace repository for materials

➢ Enables sharing of a variety of data types, including text, images, and video

https://materialsdata.nist.gov/
Digital Identifier
http://hdl.handle.net/11250/22

Related Work

Data files

Offer licenses with attribution 3.0
Metadata Curation

Markup
THERMODYNAMICS RESEARCH CENTER

- Expanding to metallic systems
- Initial focus on phase equilibria data and thermochemical property data.

http://trc.nist.gov/
Vision of Materials Data Curation System
A Successful Example

“We present a data model describing the structure of spectrophotometric datasets with spectral and temporal coordinates and associated metadata. This data model may be used to represent spectra, time series data, segments of SED (Spectral Energy Distributions) and other spectral or temporal associations.”

“In this document we present a proposed abstraction for spectral data and serializations in VOTABLE, FITS, and XML, for use as a standard method of spectral data interchange.”

“This is a Proposed Recommendation, developed with the intention to support the Simple Spectral Access Protocol.”

“The Simple Spectral Access (SSA) Protocol (SSAP) defines a uniform interface to remotely discover and access one-dimensional spectra.”
Astronomy vs. Material Measurement

Position in the sky

Measurement Type

Singular data models

Modular data models

Material synthesis and processing history

Measurement Type

e.g.

Additive Manufacturing

Thermomechanical Processing

Physical Vapor Deposition

Chemical Vapor Deposition

XRD, NIR, XPS, EBSD, AFM

Chemical Vapor Deposition
Written in python Django
Backed by MongoDB
SPARQL Query interface
XML-based Schema
Ability to store templates
Schema management tools
REST API interface
Features in progress: Schema composer and Links to Dspace repository
Data Curation: Tracer Diffusivity Test Schema
Development of Composer

User develops needed template from smaller schema

Material
- Metals
- Al Alloys
- Fe Alloys
- Ni Alloys
- …
- Polymers

Processing
- Heat Treatment
- Casting
- Thermomechanical processing
- Sintering
- Rapid Cooling
- …

Sample Preparation
- Polishing
- Etching
- Electro-polishing
- …

Measurement Method
- EPMA/EDS
- EMSD
- DSC/DTA
- Dilatometer
Application Programming Interface in a nutshell

**WWW** interface to the Curator:

Hand data entry

**API** interface to the Curator:

Automated data entry
Raw Data Curation

Curation of Raw Data from Computation

Curation of Raw Data From Instrument 2

Curation of Raw Data From Instrument 1

Send Data via API

Send Data via API

Send Data via API

Interface with workflow tools

Interface with workflow tools

Interface with workflow tools

User defined tools
Registries
The case for a Materials Resource Registry


Research Data Alliance
Collaborators

National Institute of Standards and Technology (NIST)

National Data Service (NDS)

Center for Hierarchical Materials Design (CHiMaD)

Research Data Alliance (RDA)
Data Policy

A NIST Approach (in progress)
**Working Data**
The digital equivalent of entering data in a laboratory notebook. Working data may be raw observational data that is acquired directly from an instrument or a measurement system, or digital values acquired or generated during experiments or simulations. In some cases the researcher responsible for generating the working data may determine that this data has immediate value and is worth preserving, or the researcher expects the data will have value after it has been manipulated or further evaluated, and the data has the potential to develop into a publication or will be used to draw conclusions. In other cases working data may be recognized as not appropriate for broader use in its present form. It may have value to the data producers and their collaborators, but it should be recognized that the data could be easily misinterpreted by people not closely involved in its production because some metadata and important facts about its status or acquisition are not readily available beyond the immediate research team (i.e. adequate metadata for re-purposing is not attached to the data itself, expending resources to codify needed metadata is not justified, etc.
Derived Data
Derived data are those that underpin the conclusions provided in a publication or report. Derived data come from working data that have been manipulated, analyzed, processed, or evaluated in some way. The data must have passed some minimal (perhaps ad hoc) evaluation and be considered by the responsible researcher (typically the data producer) to be ready for the next steps in the workflow or project/product development effort.

Publishable Results
All final or summary results that comply with relevant NIST policies (e.g., SI units, uncertainty statements), that have been reviewed internally and approved by an appropriate NIST authority, and that could be published either in a scientific publication or as a standalone data product.

Published Results
Results that are publishable and that are contained in a document that has been reviewed and approved for publication by the necessary NIST organizational authorities, submitted to its intended publisher, and made public. We see Reports of Analysis as fitting in this category but no longer call them out because they seem unique to MML.
• Preservation Levels
  1. No additional requirements
  2. Individual user responsibility
  3. Data backed up using a tested/automated process

• Discoverability Levels
  1. No additional requirements
  2. Persistent Identifier (PID) assigned
  3. Entered in NIST Enterprise Data Data Inventory\(^1\)
  4. Inventory record flagged for public access

\(^1\)PID assigned + NIST minimum metadata present
Data policy under discussion

mapping levels onto categories

Preserv = Preservation
Discov = Discoverability

<table>
<thead>
<tr>
<th>Preserv</th>
<th>Discov</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Collaborations
Discussions
Convening
Research Data Alliance Interest Group

RDA/CODATA Materials Data, Infrastructure & Interoperability IG

James A Warren and Laura Bartolo, Co-Chairs

Materials Resource Registry, Working Group, Robert Hanisch, Chair
Exploration of Opportunities
• The aim of the council (for the moment) is to do preparatory networking to establish activities necessary in the materials modelling field.
• The Council could build on existing activities happening in Europe and make use of the network to complement these activities.
THE NATIONAL DATA SERVICE

The National Data Service is an emerging vision of how scientists and researchers across all disciplines can find, reuse, and publish data. It is an international federation of data providers, data aggregators, community-specific federations, publishers, and cyberinfrastructure providers. It builds on the data archiving and sharing efforts under way within specific communities and links them together with a common set of tools.

VISION

It is widely believed that ubiquitous digital information will transform the very nature of research and education. The reasons for this excitement are clear: in essentially every field of science, simulations, experiments, instruments, observations, sensors, and/or surveys are generating exponentially growing data volumes. Information from different sources and fields can be combined to permit new modes of discovery. Data, including critical metadata and associated software models, can capture the precise scientific content of the processes that generated them, permitting analysis, reuse, and reproducibility. By digitizing communication among scientists and citizens, discoverable and shareable data can enable collaboration and support repurposing for new discoveries and cross-disciplinary research enabled by data sharing across communities. Open, shareable data also promise to transform education, society, and economic development.

However, while some communities are making progress in developing discipline-specific data services, the U.S. and international scientific communities lack a unified framework and supporting services for storing, sharing, and publishing data, for locating data, or for verifying data. More specifically, we are lacking standard means of
**Goal:** Establish well-pedigreed and curated demonstration datasets for non-proprietary metallic structural materials data over all length scales.

**NIST’s role**
- Provide data schemas and meta-data formats for diffusion and phase equilibria data.
- Provide sample diffusion and phase equilibria data for the Al-Mg-Si system.
- Use expanded TRC Guided Data Capture program with available binary and ternary phase equilibria literature
- Expand use and implementation of DSpace Repository
- Link with developing ontology and semantic web tools

March 2014: Phase 1 release.
June 2014: Phase 2 release.
Dec 2014: Project Completion
NIST Center of Excellence for Advanced Materials
First US-Japan MGI Workshop, June 2015
STOP