The Materials Genome Initiative and the Data Infrastructure

James A Warren
Technical Program Director for Materials Genomics
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
Outline

• A very few slides on MGI
• The Needs
• The MGI Data Infrastructure: A Vision
• The intersection of Information and Materials Sciences
• What can your organizations do?
There are two groups of people that don’t like the name Materials Genome Initiative

Scientists Hate Metaphors
THE MATERIALS GENOME INITIATIVE

to decrease time-to-market by 50% while <$$

Develop a Materials Innovation Infrastructure

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

Equip the next generation materials workforce
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: 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
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
Data Sharing is Important Beyond MGI

OSTP “Public Access” Memo
Feb 22, 2013

OMB “Open Data” Memo
May 9, 2013

Executive Order
May 9, 2013

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.
Today's Approach to Computational Materials Design

Data Models

Experiment

Simulation

Materials w/ Targeted Properties

Quantum Nano Micro Macro
An infrastructure for Open Science

MGI

How do we do it?

Data Repositories

Models

Repositories

Simulations

Experiments

Metadata

Material w/ Targeted Properties

Quantum

Nano

Micro

Macro

Experiment

Simulation

Materials
Data

Community-based Curated Repositories

Models

MGI

Ecosystems

Materials w/ Targeted Properties
Enable & Enhance Exchange
(repositories, disciplines, industries; standards)

Data

Repositories

Models

Experiment

Simulation

Quantum

Nano

Micro

Macro

Materials w/ Targeted Properties
Data Repositories

Models

Simulations Experiments

Data Models Code Metadata

Assess & Improve Quality
(Data & Models)

Materials w/ Targeted Properties
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
Enable & Enhance Exchange

• Develop and deploy repositories
• Develop and disseminate materials informatics infrastructure
  – Enable data discovery through tools and standards
  – Capture data from scientific workflows and archival sources
  – Engage with stakeholders to determine needs and disseminate best practices
• Integrate across length and time scale
• Build and Test infrastructure through Pilots
Assess & Improve Quality

- Validate Experiments and Models
- Verify Model accuracy
- Quantify Uncertainty
- Quantify Sensitivity
- Define Next Generation of Experiments and Models
New **Methods and Metrologies**

- Develop Data Driven Materials Science
- Integrate with Modeling Expertise
- Enable Crowdsourced/Open Science
- Achieve targets in Materials by Design/ICME
Another way of telling the story

- Industry needs good data and models
  - How do they get them? (exchange)
  - Can they trust them? (quality)
  - How can they use them best and who has the talent? (new methods)
Objective 3.1: Identify Best Practices for Implementation of a Materials Data Infrastructure

- Milestone 3.1.1: Convene a series of multiagency workshops that engage stakeholders, including researchers from academia, industry, publishing, and government to establish the needs of the disparate materials communities, identify the barriers to creating a materials data infrastructure, and define potential methods of overcoming these obstacles. [DOD and NIST, Here we are!]

- Milestone 3.1.2: Foster ongoing discussion of best practices in data management plans used by participating agencies with the opportunity to leverage these for broader applications within the MGI community. [SMGI]
Objective 3.2: Support Creation of Accessible Materials Data Repositories

Milestone 3.2.1: Develop and implement at least three materials data repository pilot projects to assess a range of repository models and initiate the definition of a materials data infrastructure model. [DOD, DOE, and NIST]
MGI Data Flow

How do we Integrate this?
What Data Should We Save?

How do I decide?  Policy Questions
**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.
• 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 Inventory
  4. Inventory record flagged for public access

1PID assigned + NIST minimum metadata present
Data policy under discussion

mapping levels onto categories

<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>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

SRD
Reference Data
Resource Data
Published Results
Publishable Results
Derived Data
Working Data

STRAW MAN-DRAFT
NIST 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

Stolen from Hanisch, MGI Seminar, 1/12/2015
Data in astronomy

1-d, 2-d, 3-d: intensity/polarization vs. energy, time, position, velocity

tables: catalogs, x-ray event lists, radio visibility measurements

• various data processing levels, from raw to “science-ready”

Stolen from Hanisch, MGI Seminar, 1/12/2015
Quantity and distribution

- ~70 major data centers and observatories with substantial on-line data holdings
- ~10,000 data “resources” (catalogs, surveys, archives)
- Data centers host from a few to ~100 TB each, currently ~1+ PB total
- Current growth rate ~0.5 PB/yr, expected to increase soon
- Current request rate ~1 PB/yr
Data discovery

Stolen from Hanisch, MGI Seminar, 1/12/2015
NIST Data Efforts

Collaborations

- ASM International: Structural Data Demonstration Project
- DOE/EERE Kinetics of Cast Mg Alloys
- Journals collaboration
  - IMMI
  - Others under discussion

DATA CAPTURE
- TRC: Guide Data Capture
- Materials Data Curator

DATA PROCESSING
- Uncertainty Analysis
- Data Analytics
- Data Mining Tools
- Bench marking activities (DFT)

DATA DISSEMINATION
- DSpace
- Ontologies
  - Schemas (XML based)
  - NLP
- WebFF-Repository
- Interatomic Potential Repository
- Semantic Media Wiki Meta data standards

DATA CAPTURE
- TRC: ThermoML Archive
First Take-Away

• This Problem is huge

• National Center for Biotechnology Information (NCBI) costs about 200M/year, 500 staff including 300 software engineers.
  – That’s “just” genomics

• We need all the help and participation we can get
What can the Professional Societies Do?

- Test parts of the infrastructure
- Cross linking pubs and data
- Experiment with new “business models” for materials information sharing
- Provide pieces of the infrastructure for their members
- Develop de facto (to start) standards for data exchange
- ???
Some Initial Partnerships and Modalities
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/DOE Collaboration to Collect Data for Advanced Automotive Cast Mg Alloys - Kinetics

- **High-Throughput Study of Diffusion and Phase Transformation Kinetics of Magnesium-Based Systems for Automotive Cast Magnesium Alloys:** J-C Zhao and A. Luo, The Ohio State University

- **Phase Transformation Kinetics and Alloy Microsegregation in High Pressure Die Cast Mg Alloys:** John Allison, University of Michigan

- **In-situ Investigation of Microstructural Evolution During Solidification and Heat- Treatment in a Die-Cast Mg Alloy:** Aashish Rohatgi, Pacific Northwest Laboratory

- **A systematic multiscale modeling and experimental approach to protect grain boundaries in Mg alloys from corrosion:** Mark Horstemeyer at Mississippi State and Santanu Chaudhuri at Washington State

- **Corrosivity and Passivity of Metastable Mg Alloys ---An Introductory Study to Future Stainless Mg Alloys:** Guang-Ling Song, ORNL

- **Dealloying, Microstructure and the Corrosion/Protection of Cast Mg Alloys:** Karl Sieradzki, Arizona State U.
The Research Data Alliance (RDA) builds the social and technical bridges that enable open sharing of data Cross-border & cross-disciplinary challenges

The current global research data landscape is highly fragmented, by disciplines or by domains, from oceanography, life sciences and health, to agriculture, space and climate. When it comes to cross-disciplinary activities, the notions of “building blocks” of common data infrastructures and building specific “data bridges” are becoming accepted metaphors for approaching the data complexity and enable data sharing. The Research Data Alliance enables data to be shared across barriers through focused Working Groups and Interest Groups, formed of experts from around the world – from academia, industry and government. Participation in RDA is open to anyone who agrees to its guiding principles of openness, consensus, balance, harmonisation, community driven and non-profit approach. It was started in 2013 by a core group of interested agencies – the European Commission, the US National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation. Other agencies, countries, companies, associations and institutes are due to join. RDA also has a broad, committed

http://rd-alliance.org/
http://www.codata.org/

Stolen from Hanisch, MGI Seminar, 1/12/2015
National Data Service

http://www.nationaldataservice.org/

Stolen from Hanisch, MGI Seminar, 1/12/2015
Halt!