Standards, Platforms, and IT

Ram D. Sriram
National Institute of Standards and Technology, U.S.A
sriram@nist.gov

Slides Courtesy:
Mary Brady, Alden Dima, Eric Lin, E. (Sub) Subrahmanian
Session 7 Questions

• Q1. How can we evolve unified formats for microstructure representation?
• Q2. How can the flow of information through the chain of simulations be managed both effectively and accurately?
• Q3. What are the foundational principles for platformization of approaches for the integrated realization of engineered materials and products?
• Q4. How does one build an enhanced materials innovation infrastructure to enable academic and industrial researchers to effectively exploit the capabilities of advanced computers and networks?
• Q5. What information models can be flexible in the development of ontologies and inter-operability of disparate systems? Specifically for applications in evolution of design and manufacturing information and design processes?
• Q6. Computational Tools: What is the best approach to provide scalable and maintainable computational infrastructures for ICME?
• Q7. How can user-friendly computational platforms be developed to manage, store and retrieve information which is useful for various types of users – engineers, designers and researchers?
• Q8. Security and privacy issues: How can security and privacy related barriers to collaboration be reduced?
MGI @ NIST
Meeting Societal Needs
Advanced materials are at the heart of innovation, economic opportunities, and global competitiveness. They are the foundation for new capabilities, tools, and technologies that meet urgent societal needs including clean energy, human welfare, and national security.

Accelerating Our Pace
The U.S. Materials Genome Initiative (MGI) challenges researchers, policymakers, and business leaders to reduce the time and resources needed to bring new materials to market—a process that today can take 20 years or more.

Building Infrastructure for Success
The MGI is a multi-agency initiative to renew investments in infrastructure designed for performance, and to foster a more open, collaborative approach to developing advanced materials, helping U.S. Institutions accelerate their time-to-market.

- Computational tools
- Experimental tools
- Collaborative networks
- Digital data

---

"...to discover, develop, and deploy new materials twice as fast, we’re launching what we call the Materials Genome Initiative”
—President Obama, 2011
Communities are diverse, and efforts are organic rather than top-down.

Data, metadata, and toolsets may be unique to each community.

Need new methods for rapid definition, discovery and automatic integration.
Scope: Goals of the Initiative

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

- Digital Data
- Collaborative Networks
- Computational Tools
MATERIALS DATA AND INFORMATICS

- Designed New Material
- Material Optimizer
  - Microstructure Prediction Tools
  - Processing Modeling Tools
  - Data Informatics & Tools
- Error propagation/Uncertainty Analysis
- MATERIALS DATA AND INFORMATICS
- Data Informatics and Tools
  - DATA (Experimental and Calculated)
    - Tools: e.g. CMS
  - Reference Data
  - First Principles
  - Atomistic Simulations
- CALPHAD
Data Tools and Informatics for Materials Data

Need

- Improved efficiency and reproducibility of thermodynamic and kinetic simulations (e.g. CALPHAD, first principles, atomistic methods).
- Accessible phase-based data described as functions of composition, temperature and pressure.

Objectives

- Develop file repositories that enable links to data files and the ability to identify key metadata.
- Develop informatics tools to enable data capture and retrieval of phase-based data.
- Develop tools that are available for the community to contribute towards and use.

Achievements and Impact

- NIST D-Space File Repository established with Kent State (http://nist.matdl.org/)
- NIST Interatomic Potentials Repository (www.ctcms.nist.gov/potentials) : 100+ downloadable referenced interatomic potentials
- Materials Data Curator system is under development
MODELING TOOLS FOR ADV COMPOSITES

↑

s

ms

μs

ns

↓

Atomistic

Coarse-Grained

Particle Properties

Suspension Rheology

Macroscopic Properties

Continuum

Molecular Modeling

R_h / R_g = 1.661
Topics Covered

• Data Representation and Challenges
  – Types of data and most efficient representation
  – Meta data
  – Standards

• Data Management
  – Organization and maintenance of digital data repositories
  – Methods for access and use that respect intellectual property rights
  – Long time sustainability

• Data Quality
  – Uncertainty properly quantified both from experiments and simulations

• Data Usability
  – Easy of use
Two Primary Breakout Groups

• **Data Challenges at Different Length Scales**
  – Atomic, Nano, Micro, and Macro – What kind of data infrastructure is needed

• **Specific Domains**
  – Electrochemical Storage
  – High Temperature Alloys
  – Catalysis
  – Lightweight Structural Materials
## Table 2.3 Nano and Molecular Length Scale Challenges

(*) one vote for potential short-term impact/≤5 years
(•) one vote for potential long-term impact/5 years and beyond

### Data Representation and Interoperability

<table>
<thead>
<tr>
<th>High Priority</th>
<th>Creating a taxonomy to systematically map out materials spaces down to the nanoscale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Developing stringent standards for what is considered &quot;data&quot; (4,8)</td>
</tr>
<tr>
<td></td>
<td>Limited ability to move from data to credible information and decision making (3,4)</td>
</tr>
<tr>
<td></td>
<td>Informing the community about web standards for data representation (6)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Medium Priority</th>
<th>Creating representations of particle distributions in nanocomposites (beyond atomic force microscopy pictures) (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Transitioning between models (lack of a single model representation) (1,3)</td>
</tr>
<tr>
<td></td>
<td>Lack of pre-competitive data for model development (4)</td>
</tr>
<tr>
<td></td>
<td>Consistently using proper definitions (e.g., &quot;a property that emerges under certain conditions relative to a material that exists at those conditions&quot;) (3)</td>
</tr>
<tr>
<td></td>
<td>Overcoming data model dependency at the nanoscale when models are not currently known (i.e., complexity at a scale larger than nanoscale may lead to emergent behavior) (1,2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lower Priority</th>
<th>Lack of the following: (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- Potential functions to model all species of interest through bond breakage</td>
</tr>
<tr>
<td></td>
<td>- Standards for reporting validation of potential functions</td>
</tr>
<tr>
<td></td>
<td>- True multi-scale simulation capability</td>
</tr>
<tr>
<td></td>
<td>- Nanoscale data for validation</td>
</tr>
<tr>
<td></td>
<td>- Continuously improving timescale data</td>
</tr>
<tr>
<td></td>
<td>- Relating the product to the materials property data</td>
</tr>
</tbody>
</table>

### Data Management

<table>
<thead>
<tr>
<th>High Priority</th>
<th>Lack of incentives to disseminate data (10,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Constructing a metadata interface between disparate databases (10)</td>
</tr>
<tr>
<td></td>
<td>Establishing a flexible framework to allow new forms of data and storage (8,2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Medium Priority</th>
<th>Accessing proprietary experimental data and using it with other databases and simulation data (1,4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ensuring data flexibility in the MII (4)</td>
</tr>
</tbody>
</table>

### Data Quality

<table>
<thead>
<tr>
<th>High Priority</th>
<th>Establishing/improving the provenance of experimental data and models (2,8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avoiding a strict focus on specific accuracy targets in simulations (1,7)</td>
</tr>
<tr>
<td></td>
<td>Systematically comparing and validating density function theories against experiments and higher-order theories (6)</td>
</tr>
</tbody>
</table>
# Table 3.4 Lightweight Structural Material Challenges

(● = one vote)

## Specific to Lightweight Structural Materials

| High Priority | • Establishing a data repository(s) for aluminum, titanium, and magnesium ●●●●●●●●●● (12)  
                     • Developing models of structures or materials comprising multiple materials, such as fiber plus matrix in a composite; two metals plus weld; or single crystal plus precipitate in a polycrystalline structure ●●●●●●●● (10)  
                     • Developing phase and interface properties, homogenization theories/models ●●●●● (6)  
                     • Modeling joints ●●●●● (6) |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium Priority</td>
<td>• Lack of good data on elastic coefficients ●●● (3)</td>
</tr>
</tbody>
</table>
| Lower Priority | • Understanding how constituent material properties relate to a specific (laminate) orientation (i.e., in a composite structure) ●● (2)  
                     • Establishing preferred orientation for texture ●● (2)  
                     • Modeling processes to convert material to usable goods |

## Generic to Lightweight Structural Materials

<table>
<thead>
<tr>
<th>High Priority</th>
<th>• Developing standards for defining data quality “data readiness level” ●●●●●●●●●● (9)</th>
</tr>
</thead>
</table>
| Medium Priority | • Drive data cost to zero; create, manage, explore ●●●●● (5)  
                      • Creating data “social networks” ●●●●● (5)  
                      – Data and models: what’s available and who’s willing to share?  
                      • Establishing an expert rating system for data and data reproducibility ●●●● (4)  
                      • Establishing ownership and maintenance of the MGI database ●●●● (4)  
                      • Lack of taxonomy, e.g., a rules-based classification system for items across disciplines, length scales, experiments, and analyses ●●● (3) |
| Lower Priority | • High cost behind data – unique or expensive data not available to potential users ●● (2)  
                      • Filtering data (i.e., one kind of information); combining targeted data across domains ●● (2)  
                      • Creating a large data network that is transparent, self-organizing, and flexible ● (1)  
                      • Understanding model readiness level ● (1) |
NIST Announces New Center for Materials Research to Advance Manufacturing and Innovation

From NIST Tech Beat: December 3, 2013

Contact: Michael Baum
301-975-2763

The National Institute of Standards and Technology (NIST) announced today that it has selected a consortium led by Northwestern University to establish a new NIST-sponsored center of excellence for advanced materials research. The new Center for Hierarchical Materials Design (ChiMaD) will be funded in part by a $25 million award from NIST over five years.

Other members of the ChiMaD consortium include the University of Chicago, the Northwestern-Argonne Institute of Science and Engineering (a partnership between Northwestern and the Department of Energy’s Argonne National Laboratory) and the Computation Institute (a partnership between the University of Chicago and Argonne.) The consortium also plans to work closely with QuesTek Innovations, a small business spin-off of Northwestern; ASM International, a well-known professional society of materials scientists; and Fayetteville State University.

“I’m particularly excited to announce this new alliance between NIST and two prominent research universities to drive innovation in the development of advanced materials,” said Patrick Gallagher, Under Secretary of Commerce for Standards and Technology and NIST Director. “This new Center for Hierarchical Materials Design is a natural fit for NIST, which has a long tradition of serving as a nexus with academia and industry to advance research and innovation for the nation’s benefit.”

“The launch of this new center represents a major milestone in support of the President’s Materials Genome Initiative and our national goal of doubling the pace of discovery and development of novel materials,” said Cyrus Wadia, assistant director for Clean Energy and Materials R&D at the White House Office of Science and Technology Policy. “By integrating the complementary strengths of computation, instrumentation, and creative modeling, this center promises to help keep America at the forefront of the materials revolution and a leader in the economically important domain of advanced manufacturing.”
Figure 4. Hierarchy of present and future Materials Genome methods, tools and databases.
REPRESENTATION: ROLE OF ONTOLOGIES
PHASE-BASED MATERIALS DATA

- Advanced materials often consist of several components \((n > 5)\) and multiple phases.
- The material properties are dependent on the microstructure.
- The microstructures changes as a function of processing and service conditions.
- The properties of the multiphase microstructures can be predicted from the properties of the individual phases and their interactions in the microstructure.

Examples of phase based properties:
- Molar Volume and Thermal Expansion
- Specific heat
- Heat capacity
- Enthalpy
- Entropy
- Intrinsic diffusivity
- Electrical conductivity and resistivity
- Thermal conductivity
- Bulk modulus
- Magnetic properties (e.g. Curie Temperature)
- Optical properties

Example: \(\gamma/\gamma'\) Superalloys (Ni or Co based

\(\gamma/\gamma'\) structure in a Co-Al-W at 900 °C for 1000 h
Examples of CALPHAD Data Types

For each assessment: Evaluated data file (e.g., POP, DOP)
Functional descriptions for phase quantity (e.g., TDB)
- Emphasis on binary and ternary data to predict multicomponent properties
- Data can be experimental or computational.

1-D (Points)
- Melting Temperatures
- Critical Temperatures (Phase Changes)
- Lattice Parameters
- Heat of Formations
- Phase Fractions and Compositions
- Activation Energies

2-D (Lines)
- Composition Profiles
- Heat Capacities
- Enthalpies of Mixing

3-D
- Crystal Structures
- Micrographs/Morphologies
- 3-D Atom Probe Tomography
Informatics Approach

Ontologies
XML Schemas
Low-Level Domain Language
Transform
XSLT Reasoners
NoSQL Databases
MongoDB, Neo4J

MatML
RDBMS
Thermo ML
Tool Formats
Other Formats

Materials Ontology currently being developed
Note this is a work in progress
What Is An Ontology

• An ontology is an explicit description of a domain:
  – concepts
  – properties and attributes of concepts
  – constraints on properties and attributes
  – Individuals *(often, but not always)*

• An ontology defines
  – a common vocabulary
  – a shared understanding
Example: A biological ontology is:

- A machine interpretable representation of some aspect of biological reality

  - what *kinds* of things exist?

  ![Diagram](image)

  - **eye disc**
    - develops from **eye**
    - is_a **sense organ**
  - **ommatidium**
    - part_of **eye**
The Foundational Model of Anatomy

- Anatomical entity
  - Physical anatomical entity
    - Material physical anatomical entity
      - Anatomical structure
        - Body
        - Organ
          - Solid organ
          - Cavitated organ
          - Organ with organ cavity
            - Esophagus
            - Stomach
            - Small intestine
            - Large intestine
            - Appendix
            - Anal canal (viewed anal to rectal opening)
            - Gallbladder
            - Vagina
            - Uterus
            - Urinary bladder
            - Duct (organ)
            - Hollow tree organ
  - Template Slots
    - continuous with
    - contained in
    - member of
    - arterial supply
    - venous drainage
    - lymphatic drainage
    - nerve supply
    - has boundary
    - bounded by
    - inherent 3-D shape
    - Has inherent 3-D shape
    - attributed part
    - adjacency
    - orientation
    - physical state
    - has mass
    - dimension
    - has dimension
    - shape
Engineering Ontology

Upper Ontology

- Collection
- Temporal Thing
- Event
- Pumping

Domain Ontology

- Individual
- Spatial Thing
- Mechanical Device
- Engine
- Supplies-fuel-to
- Part-of
- Connected-to
- Generalization
- Other Relationships

Hydraulic System
- Has-part
- Hydraulic Pump
- Aircraft Engine Driven Pump

Fuel System
- Fuel Filter

Fuel Pump
- Fuel Filter

Jet Engine
Other Ontologies

- Process Specification Language (PSL)
- Product Specification Representation Language (PSRL)
- Gene Ontology (GO)
- DAML ontologies
- Cyc
- Standard Upper Ontology
- Ontolingua
- TOVE
- Enterprise Ontology
- STEP
Ontology Spectrum

From less to more expressive

Conceptual Model

RDF/S

Extended ER

Thesaurus

ER

DB Schemas, XML Schema

Taxonomy

Relational Model, XML

Weak semantics

Strong semantics

First Order Logic

Modal Logic

Logical Theory

Description Logic

DAML+OIL, OWL

UML

Syntactic Interoperability

Structural Interoperability

Semantic Interoperability

Is Disjoint Subclass of with transitivity property

Is Subclass of

Has Narrower Meaning Than

Is Sub-Classification of

Courtesy: Leo Obrst, MITRE
Ontology Spectrum: Application

<table>
<thead>
<tr>
<th>Expressivity</th>
<th>Concept (referent category) based</th>
<th>Ontology</th>
<th>Logical Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term-based</td>
<td>Thesaurus</td>
<td>weak</td>
<td>strong</td>
</tr>
<tr>
<td>Taxonomy</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

More Expressive Semantic Models Enable More Complex Applications

| Application | Categorization, Simple Search & Navigation, Simple Indexing | Synonyms, Enhanced Search (Improved Recall) & Navigation, Cross Indexing | Enterprise Modeling (system, service, data), Question-Answering (Improved Precision), Querying, SW Services | Real World Domain Modeling, Semantic Search (using concepts, properties, relations, rules), Machine Interpretability (M2M, M2H semantic interoperability), Automated Reasoning, SW Services |

Courtesy: Leo Obrst, MITRE
Kinds of Ontologies: Alternate View

Terms
- ad hoc Hierarchies (Yahoo!)
- Thesauri
- structured Glossaries
- XML DTDs
- Principled, informal hierarchies
- XML Schema
- formal Taxonomies
- Data Models (UML, STEP)
- Frames (OKBC)
- General Logic

Glossaries & Data Dictionaries
- ‘ordinary’ Glossaries
- Data Dictionaries (EDI)

Thesauri, Taxonomies
- XML Schemas, & Data Models

MetaData, Inference
- Formal Ontologies & Inference
Benefits from an Ontological Approach

• Semantic Unification
  – The **unification of** lexically **different representations** that have the same semantics
    • Example: fcc phase in steels can be referred to as fcc, austenite or $\gamma$.

• Ontology-based Data Integration
  – Using ontologies to **unify data** that share some common semantics but originate **from unrelated sources**
    • Example: Are property data from two experiments consistent enough to be combined?
Sources
- Prototype MGI Ontology
- ThermoML
- MatML
- MatSeek
- UnitsML
- ChemML

Tools
- UML (Unified Modeling Language)
- Semantic Web (RDF, OWL)

Note: This is a generalized model depicting overall structure.
Broad concepts covered in materials data files (data have many types)
- Objects, Materials, and Events
- Physical Properties
- Documents
- Data Objects & Types
- People & Organizations
- Software
- Relations among these

An ontology renders shared vocabulary and taxonomy which models a domain with the definition of objects and/or concepts and their properties and relations.
Slight Digression:

Methodology for Image to Diagnosis Through Disease Ontology

Input Image

Image segmentation

Feature Vectors Computation

Mapping calculated feature vectors into disease ontology

Highlighting regions with deviation from normal conditions

Output:
Suggestion of the potential diagnosis

Suggested diagnosis: Ulcer
Fragment of Disease Ontology in Protege’-Ontoviz
UML Representation of Inflammatory Bowel Disease

**Idiopathic inflammatory bowel disease**
- endoscopic feature 1: <unspecified> = ulcers
- location: string(idl) = distal small intestine / proximal colon / distal colon / rectum

**Crohn's disease**
- endoscopic feature 2: <unspecified> = strictures
- endoscopic feature 3: <unspecified> = cobblestonning

**Ulcerative colitis**
- endoscopic feature 2: <unspecified> = pseudopolyps
Improving the Process...
NLP and Text Mining
Materials Science Informatics

NIST Materials Science Domain

Advanced Materials Group → CALPHAD Assessments → Standard Reference Data & Materials

Research Corpora

- Self-Diffusion Research Corpus > 760+ PDFs
- CALPHAD Research Corpus > 200+ PDFs
- ... and more ...

Heterogeneous data sources & types

- Tool-specific code, inputs, outputs
- Database files
- Research literature
- Related data & calculations
- Assessment-specific artifacts

Applications

- Semantic search
- Automatic summarization
- Semantic visualization
- Knowledge Representation
- Knowledge Discovery
- Thematic Clustering

NIST Materials Science Informatics

Tools

- Semantic Medline
- Carrot2
- And more...
Semantic Medline (SM)

Quick Facts:
- NIH semantic engine for PubMed
- Retargeting from biomedical domain to materials science

Natural Language Processing (NLP):
- Stochastic Part-of-speech Tagging
- Word Sense Disambiguation
- Semantic Predication

Categories
- Relations
- Groups

Search Results
- Auto Summarization
- Visualization

Semantic Search
- Thematic
- Iterative

Knowledge Representation
- Meta Thesaurus
- Semantic Network

Text\textsubscript{raw} → Text\textsubscript{clean} → SM-DB

Corpus\textsubscript{(PDFs)} → Text\textsubscript{raw} → Text\textsubscript{clean} → SM-DB
Informatics Task (discovery): Investigate Aluminum Alloy

Traditional Approach

- Google Search (aluminum alloy)
  - PageRank criteria
  - Spurious search results
  - Out-of-context browsing

Our Approach

- Semantic Medline Search (aluminum alloy)
  - Domain-specific representation
  - Graph & knowledge-based browsing
  - In-context browsing
Informatics Task (research): Research Ag Alloys for CALPHAD Assessment

Web of Science Search (Ag alloy, CALPHAD assessment)
- Term-only criteria
- Relies on pre-processed, structured corpora

Carrot2 Search (Ag alloy, CALPHAD assessment)
- Domain-specific representation
- Tunable clustering
- Can process less structured, noisy corpora
- Thematic visualization & browsing
## CALPHAD Corpus Results (125 files)

**CALPHAD Corpus Statistics**

<table>
<thead>
<tr>
<th>Property</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Files</td>
<td>125</td>
</tr>
<tr>
<td>Total Sentences</td>
<td>57,414</td>
</tr>
<tr>
<td>Min Sentence Length</td>
<td>1</td>
</tr>
<tr>
<td>Max Sentence Length</td>
<td>7,565</td>
</tr>
<tr>
<td>Avg Sentence Length</td>
<td>98</td>
</tr>
<tr>
<td>Total Concepts</td>
<td>195,053</td>
</tr>
<tr>
<td>Total Unique Concepts</td>
<td>5,732</td>
</tr>
<tr>
<td>Total Words</td>
<td>195,053</td>
</tr>
<tr>
<td>Total Unique Words</td>
<td>10,907</td>
</tr>
<tr>
<td>Total Relations</td>
<td>1,245</td>
</tr>
</tbody>
</table>

### Log Scale

<table>
<thead>
<tr>
<th>Property</th>
<th>Log Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Files</td>
<td>2.096910013</td>
</tr>
<tr>
<td>Total Sentences</td>
<td>4.759017805</td>
</tr>
<tr>
<td>Min Sentence Length</td>
<td>0</td>
</tr>
<tr>
<td>Max Sentence Length</td>
<td>3.878808932</td>
</tr>
<tr>
<td>Avg Sentence Length</td>
<td>1.989147886</td>
</tr>
<tr>
<td>Total Concepts</td>
<td>5.290152634</td>
</tr>
<tr>
<td>Total Unique Concepts</td>
<td>3.758306182</td>
</tr>
<tr>
<td>Total Words</td>
<td>5.290152634</td>
</tr>
<tr>
<td>Total Unique Words</td>
<td>4.037705313</td>
</tr>
<tr>
<td>Total Relations</td>
<td>3.095169351</td>
</tr>
</tbody>
</table>
Improving the Process...
Semantic Tools
Property Knowledge Discovery

Data Generation and Informatics

- Estimated diffusivity in Cu-Ga
- New Property Knowledge
- Ontologies XML Schemas
- CALPHAD Domain Language
- Other Data Sources

Material Optimizer

- Found information for Cu-Al; use to estimate diffusivity in Cu-Ga
- Search for diffusion data in phases with similar structures

- Machine Learning
- Data Mining
- Clustering
- Reasoning

Analysis

- Processing/Structure/Property Data
- Transformations
- XSLT Reasoners

Search

- Query: Data for the Cu-Ga phases the Cu-In-Ga-Se system
- Find: Estimated diffusivity in Cu-Ga with Provenance

Missing data for Cu-In-Ga-Se (photovoltaics)

C. Campbell and A. Dima, MGI and Big Data
Reference Data Workflow

- Literature
- Property Databases

Search

- Relevant Data

Curation

- A'
- A
- B'
- B
- C'
- C

Curated Data

Analysis & Selection

- Metadata
- Data

Reference Data

Generate

Computation

Experiment

Reference Data Workflow
Reference Data Workflow + Informatics

Semantic Search

Semi-Automated Curation System

Curation

Analysis & Selection

Reference Data

Search

Metadata

Data

Curated Data

Data Repository

Workflow System

Reference Data

Generate

Computation

Experiment

Property Databases

Literature

Relevant Data
BUILDING MATERIALS INNOVATION
INFRASTRUCTURE:
COLLABORATIVE NETWORKS
HubZero

“Facebook for scientists -- but built to facilitate serious research rather than socializing”

Web Application

Open source software platform
Create websites supporting
• Scientific research
• Education

Workspaces

In-browser Linux desktop
Can share screens
Upload/download files
Use development tools
Launch jobs on the grid
Some Existing Hubs

- **nanoHUB.org**: granddaddy of all hubs focused on nanotechnology
- **nees.org**: Network for Earthquake Engineering Simulation (NEES)
- **IndianaCTSI.org**: accelerating clinical and translational research in healthcare
- **pharmaHUB.org**: pharmaceutical product development and manufacturing
- **GlobalHUB.org**: global engineering education
- **cceHUB.org**: cancer care engineering
- **memsHUB.org**: microelectromechanical systems
- **CLEERhub.org**: Collaboratory for Engineering Education Research
iRODS

- **Integrated Rule Oriented Data System**
- “Intelligent Cloud” for sharing data
- Features include:
  - High-performance network data transfer
  - A unified view of disparate data
  - Support for a wide range of physical storage
  - Easy back up and replication
  - Manages metadata
  - Controlled access.
  - Policies, Rules and Micro-services
  - Workflows
  - Management of large collections

https://www.irods.org
How much data for Mg commercial alloy system

- Commercial Mg alloys including
  - Mg-Al based
  - Mg-Zn-Zr based
  - Mg-RE (rare earth) base alloys

- Thermodynamic database
  - 23 elements: Ag, Al, Ca, Ce, Cu, Fe, Gd, K, La, Li, Mg, Mn, Na, Nd, Ni, Pr, Si, Sn, Sr, Th, Y, Zn, Zr
  - 396 Solution and Intermetallic Phases
  - Current database contains 149 binary and 59 ternary systems
    - 59 ternaries = 59 x 10 GBs = 590 GBs

Full Thermodynamic Assessment

23 Components ➔ 1771 ternary systems

≈ 20 TB

Include data to describe diffusion, molar volume, elastic properties, nucleation, interfacial energies

⇒ 100 TB of data
Explosion of Database Systems
Popular Open Source NoSQL Databases

- **MongoDB**
  - Written in: C++
  - Main point: Retains some friendly properties of SQL. (Query, index)
  - Best used: If you **need dynamic queries**. If you prefer to define indexes, not map/reduce functions. If you **need good performance** on a big DB

- **Riak**
  - Written in: Erlang & C, some JavaScript
  - Main point: Fault tolerance
  - Best used: If you **need very good single-site scalability, availability and fault-tolerance**

- **HBase**
  - Written in: Java
  - Main point: Billions of rows X millions of columns
  - Best used:
    - Best if you use the Hadoop/HDFS stack already
    - Any place **where scanning huge, two-dimensional join-less tables are a requirement**

- **Redis**
  - Written in: C/C++
  - Main point: Blazing fast
  - Best used: For **rapidly changing data with a foreseeable database size** (should fit mostly in memory).

- **CouchDB**
  - Written in: Erlang
  - Main point: DB consistency, ease of use
  - Best used:
    - For accumulating, occasionally changing data, on which pre-defined queries are to be run
    - Places **where versioning is important**

- **Neo4j**
  - Written in: Java
  - Main point: Graph database - connected data
  - Best used: For **graph-style, rich or complex, interconnected data**

http://kkovacs.eu/cassandra-vs-mongodb-vs-couchdb-vs-redis
ASTERIX Project @ UCI

• Build a new Big Data Management System (BDMS)
  – Run on large commodity clusters
  – Handle mass quantities of semistructured data
  – Openly *layered*, for selective reuse by others
  – Share with the community via open source

• Conduct scalable information systems research
  – Large-scale query processing and workload management
  – Highly scalable storage and index management
  – Fuzzy matching, spatial data, date/time data (all in parallel)
  – Novel support for “fast data” (both in and out)

Train next generation of “Big Data” graduates
PUTTING THINGS TOGETHER
FUTURE MATERIALS INFORMATICS

Ontology

XML Schema

Data Capture

JSON

Various Database Platforms

Data Tools: Statistics; Machine Learning

XML

Semantic Web

UNIFIED MODELING LANGUAGE

php

JAVA

RDF
Toward a Multilevel Representation

• Form (Micro, Meso, Macro)
• Function
• Behavior (Qualitative, Quantitative)
• Meta Information (Rationale, etc..)
• .........
A Note About Standards

• Standards in general
  – Enhance compatibility
  – Facilitate interoperability
  – Reduce technology risk
  – Spur creativity and innovation ("The chromatic scale [of music], and its formal notation system, spawned two centuries of the most prolific and original composition" -- Buckingham and Coffman)

Source: Information Rules, Shapiro and Harian
Imagine!

- It took about 30,000 people to build the Taj Mahal
- It took about 100,000 people to build the Great Pyramid
- About 300-400,000 people were involved in putting a man on the moon
- Now, imagine what can the combined intelligence of millions of people on the Internet can achieve!!
Session 7 Questions

• Q1. How can we evolve unified formats for microstructure representation?
• Q2. How can the flow of information through the chain of simulations be managed both effectively and accurately?
• Q3. What are the foundational principles for platformization of approaches for the integrated realization of engineered materials and products?
• Q4. How does one build an enhanced materials innovation infrastructure to enable academic and industrial researchers to effectively exploit the capabilities of advanced computers and networks?
• Q5. What information models can be flexible in the development of ontologies and inter-operability of disparate systems? Specifically for applications in evolution of design and manufacturing information and design processes?
• Q6. Computational Tools: What is the best approach to provide scalable and maintainable computational infrastructures for ICME?
• Q7. How can user-friendly computational platforms be developed to manage, store and retrieve information which is useful for various types of users – engineers, designers and researchers?
• Q8. Security and privacy issues: How can security and privacy related barriers to collaboration be reduced?