Data Infrastructure

Carelyn Campbell, Ben Blaiszik, Laura Bartolo

November 1, 2016
Data Landscape

Collaboration Tools
(e.g. Google Drive, DropBox, Sharepoint, Github, MatIN)

Data Sharing Communities
(e.g. Dryad, FigShare, NanoHub, Kaggle, NDS)

Data Repositories
(e.g. Aflow, MaterialsProject, OQMD, NIMS MaterialNavi, NoMaD, Materials Universe)

Data Curation

Data Analysis Tools

Software
Data curation is the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education.

http://ischool.illinois.edu/academics/degrees/specializations/data_curation
Representative list of tools focused on materials data, but not comprehensive
**Data Model Definition**

Defines the structure of metadata and data

<table>
<thead>
<tr>
<th>Measurement Data Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Metadata e.g.</strong></td>
</tr>
<tr>
<td>• Sample owner</td>
</tr>
<tr>
<td>• Date of measurement</td>
</tr>
<tr>
<td>• Sample stage position</td>
</tr>
<tr>
<td>• Apparatus temperature</td>
</tr>
<tr>
<td><strong>Data e.g.</strong></td>
</tr>
<tr>
<td>• As XML</td>
</tr>
<tr>
<td>• Raw data (text, ASCII, binary)</td>
</tr>
<tr>
<td>• Imported table</td>
</tr>
<tr>
<td>• *Link to image or raw data</td>
</tr>
</tbody>
</table>
Example APS Data (Large Data Example)
Workflows

• Large Data sets: Single Point Source (e.g. APS)
• Experimental data (small to medium size), multiple source generation
• Computational Data

• Infrastructure Selection Tool
Stress-Strain Measurement

Experimental Data Workflow
* Small → Medium
* Multiple modalities

Example: Stress-Strain Curve

- Sample Geometry Measurement
- Stress-Strain Measurement
- DIC
- LYTD

Measurement Data
* ICE
* Mechanical Commons

1. Measurement Analysis
2. Measurement Analysis

MTS - results
DIC
LVOF

Results

Final Stress-Strain
MTS

DIC sample

G, f/hub, Nono hub, Matlab

Mathlab

DIC sample

Calibration

Resin

Press

Clear

KAC-VP

Flexural
Sample Geometry

Stress-Strain Measurement

Analysis (Tools e.g. Github, Matlab)

Load Frame
DIC
LVDT

Load Frame Metadata
DIC Metadata
LVDT Metadata

Load Frame Data
DIC data
LVDT data

Curate with Materials Commons or ICE
Big Data Workflow

* APS, SNS, ... experimental or computational
  Leadership Computing Facilities

# Schema (for technique)

- Beamline
- Data Capture
- Metadata Capture
- Storage

- Beamline conditions
- Sample provenance
- Material composition

- Flow Long?
- Which data to keep?
- How to validate and value data
- What does it take to re-use?

- IoT, capture everything [2005]
- Sample mapping across instruments

- Share
  - Link w/ Papers
-

- MDF
  - Data
- MDCS

- Experts
  - Full w/ all metadata
- Non-experts (technique)
  - Segmented data, e.g.

- End US01
  - Aggregated
  - Dark Data search
Computational Data Workflow

• Lots of different techniques
• PhaseField modeling: no standards.
• Community standards needed

• Codes changes quickly
• Social change needed.
• FEM - more benchmark. -- more standardize
How do I select a Materials Data Infrastructure Tool?
Example: Workflow Tool Selection

### Table 2. Workflow design taxonomy mapping.

<table>
<thead>
<tr>
<th>Project name</th>
<th>Structure</th>
<th>Model</th>
<th>Composition systems</th>
<th>QoS constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAGMan</td>
<td>DAG</td>
<td>Abstract</td>
<td>User-directed</td>
<td>User specified rank</td>
</tr>
<tr>
<td>Pegasus</td>
<td>DAG</td>
<td>Abstract</td>
<td>User-directed Language-based</td>
<td>Expression for desired resources</td>
</tr>
<tr>
<td>Triana</td>
<td>Non-DAG</td>
<td>Abstract</td>
<td>User-directed Language-based</td>
<td>N/A</td>
</tr>
<tr>
<td>ICENI</td>
<td>Non-DAG</td>
<td>Abstract</td>
<td>User-directed Language-based</td>
<td>N/A</td>
</tr>
<tr>
<td>GridAnt</td>
<td>Non-DAG</td>
<td>Concrete</td>
<td>User-directed Language-based</td>
<td>N/A</td>
</tr>
<tr>
<td>GrADS</td>
<td>DAG</td>
<td>Abstract</td>
<td>User-directed Language-based</td>
<td>Estimated application execution time</td>
</tr>
<tr>
<td>Kepler</td>
<td>Non-DAG</td>
<td>Abstract</td>
<td>User-directed Language-based</td>
<td>N/A</td>
</tr>
<tr>
<td>Karajan</td>
<td>Non-DAG</td>
<td>Concrete</td>
<td>User-directed Language-based</td>
<td>N/A</td>
</tr>
<tr>
<td>GridFlow</td>
<td>DAG</td>
<td>Abstract</td>
<td>User-directed Language-based</td>
<td>N/A</td>
</tr>
</tbody>
</table>
### Example: Hardware Store Website

<table>
<thead>
<tr>
<th>Category</th>
<th>Option</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Brand</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Price</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Counter Depth (Yes/No)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Refrigerator Width (In.)</strong></td>
<td>28 - 29 (151)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29 - 29.9 (209)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>30 - 31.9 (64)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>32 - 32.9 (217)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>33 - 34.9 (23)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>35 - 36 (456)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>36 or Greater (138)</td>
<td></td>
</tr>
<tr>
<td><strong>Height to Top of Door Hinge</strong></td>
<td>66.25 in (21)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>66.31 in (1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>66.44 in (4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>66.5 (11)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>66.5 in (9)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>66.62 in (2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>66.625 (11)</td>
<td></td>
</tr>
</tbody>
</table>

Credit: homedepot.com

Any mention of commercial products is for information only; it does not imply recommendation or endorsement by NIST.
Example: Used Car Website

Credit: carmax.com

Any mention of commercial products is for information only; it does not imply recommendation or endorsement by NIST.
## Registry: Materials Data Infrastructure Tools

<table>
<thead>
<tr>
<th>Material Types</th>
<th>Metal</th>
<th>Semiconductor</th>
<th>Ceramic</th>
<th>Polymer</th>
<th>Biomaterial</th>
<th>Organic</th>
<th>Inorganic</th>
<th>Oxide</th>
<th>Composite</th>
<th>Nanomaterials</th>
<th>Superconductor</th>
<th>Non-Specific</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morphology/Structures</td>
<td>Crystalline</td>
<td>Amorphous</td>
<td>Fluid</td>
<td>Quasi-periodic</td>
<td>Bulk</td>
<td>2-Dimensional</td>
<td>1-Dimensional</td>
<td>Film</td>
<td>Nanotube</td>
<td>Fiber</td>
<td>Composite</td>
<td>Interphase</td>
<td>Line Defect</td>
</tr>
<tr>
<td>Material Property Classes</td>
<td>Optical</td>
<td>Mechanical</td>
<td>Thermodynamic</td>
<td>Structural</td>
<td>Simulation</td>
<td>Thermodynamic</td>
<td>Structural</td>
<td>Simulation</td>
<td>Dynamic</td>
<td>Thermal</td>
<td>Static</td>
<td>Indentation</td>
<td>Dilatometry</td>
</tr>
<tr>
<td>Sample Processing Methods</td>
<td>Casting</td>
<td>Annealing</td>
<td>Vapor Deposition</td>
<td>Milling</td>
<td>Extrusion</td>
<td>Pressing</td>
<td>Exfoliation</td>
<td>Melt Blending</td>
<td>Polymerization</td>
<td>Curing</td>
<td>Evaporation</td>
<td>Other</td>
<td></td>
</tr>
</tbody>
</table>

Need "checkboxes" for Materials Data Infrastructure Tools.
First Rough Draft Taxonomy

- Laboratory Information Management
  - Project Management
  - File/Directory Management
  - Database Creation
  - Collaboration Environment
  - Repository Creation
  - Registry Creation
  - Analysis
  - Workflow/Simulation Management
    - Data Movement
    - Materials Data Resource
  - Semiautomated
  - Automated
  - Data Harvesting
  - Functionality
First Rough Draft Taxonomy

Research Actions (w.r.t. data)
- Data Search
  - Scope
    - Global
    - Local
  - Type
    - Keyword
    - Parametric
- Data Mining
- Data Download
- Data Transformation
- Data Generation
- Data Exchange (collaboration)
- Data Analysis
- Data Publication
- Data Visualization
First Rough Draft Taxonomy

Data Format (in/out)

Unstructured
- Collection
  - Flat File
- Flat File
- XML
  - XML Schema
- JSON
  - JSON Schema
- JSON-LD
- HDF5
- MySQL
- NoSQL
- SciDB

Structured
- User-Defined
- Syntax
- Schema-Defined
- Software-Defined
Notes from Summit Wrap-up Session

• Integrate tools into undergraduate education
  – Tools need to more user friendly
• Embed data experts into experimental groups
  – Alternate: floating data experts available for experimental groups.
  – Need to define skills needed for these data experts
• Encourage more conference exchanges at Data Analytics and Materials communities
• Define data curation guidelines/code
  – Benefit to users
• Data Challenge (Student)
  – Prize for data set
  – Best paper/DOI/PID
• Develop implementation path
• Improve peer recognition
• Develop data cite profile

Interest in following up with small working groups on specific issues.
Data Cit. Profile

Tool Integ. into undergrad edu
- More user-friendly

Embedding data into exp.
on floating data experts. groups

Data Analytics/Materials conf. exchange

Guidelines for data curation/
(code

Data Challenge (student-band)
- prize
- best paper/DOC-PID

Peer Recognition
Role/Function Implementation Path