Working Group 6: CALPHAD Proto Data

Working group members: Greg Olson, Peisheng Wang, Wei Xiong, Ken Kroenlein, Ursula Kattner, Alden Dima, Guillaume Sousa Amaral, James Saal, Matthew Peters, Zachary Trautt, Carrie Campbell
Significance of WG’s Focus

• Generation, curation and dissemination of CALPHAD files (e.g. TDB, POP, etc) and CALPHAD proto data.

• CALPHAD Proto Data: Phase-base property data (temperature, composition, and pressure dependent)
  – Data are diverse
  – Data semi-structured

Data and files are essential for the development of multicomponent databases that serve as a building block for materials design.

Basis of data tools and schemas develop are material class independent.
Focus (phase-based property data needed to build Composition, temperature, pressure dependent CALPHAD-base databases)

G. B. Olson, 2013
Summary of WG’s Goals

• Improve the dissemination/discovery of CALPHAD files, including functional descriptions, evaluated data, and macros/scripts (e.g. TDB, POP, TCM, etc)

• Improve the curation and dissemination of CALPHAD proto data
  – Developing curation tools and schemas
  – Establish best practices (e.g. ThermoML 5.0 for the curation of thermodynamic data)
Goal: Implement Co data for Spring Design Project

- Distribution existing work Feb 2016
- Assemble available experimental data in the MDCS as possible (thermodynamics, diffusion, DFT)
  - Refinement of data schemas
  - Add Re to existing Co-base descriptions (Co-Al-W-Ni-Ti-Ta-V) to Questek database (Co-C-Al-Cr-Fe-Mn-Ni-Ti-V-W)
  - Deadline: End of March
- Distribute available TDB/POP files (DSpace - CHiMaD only community)
  - During the design course
- Demonstrate search strategy for component selection using Granta software
  - Short term Goal: Shared data schemas and implement common search strategies in Granta and MDCS
  - Long term Goal: Indexing a federated system
Technical Requirements/Needs

• Dissemination of CALPHAD files
  – Ease of use (input and discovery)
  – Flexible inputs
  – Ability to link to additional resources (e.g. files)

• Curation of CALPHAD proto data requires a system able to
  – Handle diverse data sets: computational and experimental, from single data points to complex 3D atom probe data
    • Need modular data schemas
  – Transform data into new formats
  – Combine data from multiple sources
  – Find data and reuse it
  – Associate metadata with data values
  – Automated data curation and search (REST API)
Solutions/Actions

• Dissemination of CALPHAD files
  – Where
    • NIST Dspace Repository: CALPHAD Assessments community
    • Materials Data Facility
    • Journals (JPED, CALPHAD -subscribers only)
  – How: Need to encourage publishers to require files and link data resources

• Curation of CALPHAD Proto data
  – Evaluate available and developing tools
  – Evaluate data curation schemas
  – Engage the community in using developed tools and developing needed data schemas
Data Curation/Archival Tools

- **MDCS/ThermoML** ([https://github.com/usnistgov/MDCS](https://github.com/usnistgov/MDCS))
  - Commercial
- **Citrine** ([http://citrination.com/](http://citrination.com/))
  - Focused collection
- **Materials Commons**
  - ([http://www.prisms-center.org/#/mcommons/overview](http://www.prisms-center.org/#/mcommons/overview))

ESPEI-V2 - pre-CALPHAD data assessment tool (Penn State)
# Evaluation of Materials Data Curation Tools

<table>
<thead>
<tr>
<th></th>
<th>MDCS</th>
<th>Materials Commons</th>
<th>Granta MI</th>
<th>Citrine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data import formats</td>
<td>Anything</td>
<td>Anything</td>
<td>Excel-based, csv, tabular</td>
<td>Excel, cvs, tabular</td>
</tr>
<tr>
<td>Image/Large Data capability</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes (limited)</td>
<td>??</td>
</tr>
<tr>
<td>Data stored as</td>
<td>XML</td>
<td>JSON</td>
<td>Modified MatML</td>
<td>JSON</td>
</tr>
<tr>
<td>REST API</td>
<td>yes</td>
<td>yes</td>
<td></td>
<td>Yes?</td>
</tr>
<tr>
<td>Data Storage</td>
<td>MongoDB</td>
<td>ReThinkDB</td>
<td>Relational DB</td>
<td>??</td>
</tr>
<tr>
<td>Type of Code</td>
<td>Open</td>
<td>Open</td>
<td>Commercial</td>
<td>Partially??</td>
</tr>
<tr>
<td>Data output</td>
<td>Anything (future)</td>
<td>Anything (future)</td>
<td>Excel, cvs, mod-MatML, code specific</td>
<td>???</td>
</tr>
</tbody>
</table>
Solutions/Actions

• Evaluation of ThermoML V5 for the curation of thermodynamic data (applications to metals and alloys) test in terms of Co-base superalloys
  – Test draft schema with data from participants (computational and experimental)
  – Evaluate the how to extend or re-use parts of ThermoML for other phase-based properties (i.e. diffusion)
According to statistics gathered at TRC, 1 in 3 articles has a fundamental issue in its reported data.

What is NIST doing to help repeatability in science?

Since 2003, TRC has collaborated with 5 journals to stem the flow of erroneous data:
- Present workflow in place since 2013

We work with IUPAC and:
- Journal of Chemical and Engineering Data
- Fluid Phase Equilibria
- The Journal of Chemical Thermodynamics
- International Journal of Thermophysics
- Thermochimica Acta

Our experts review 1000 articles per year and fix 500 issues per year.
Collaborations/Synergies

• Inputs need from other working groups
  – Materials Resource Registry and Repositories to disseminate and find data (Working Group 1: MDCS, DSpace, MRR & MDF)
  – Experimental and Computational data need for CALPHAD assessments (Working Group 2: Experimental Data and Working Group 5: DFT)
  – Ability to ease search and data re-use (Working Group 4: NLP)

• Need for outputs: Use-Cases
  – Co-base Superalloys for additive manufacturing
  – Shape-memory alloys (PdTi-based)
  – In-Situ Si composites/Thermoelectrics
Collaborations/Synergies

- Others at NIST
  - Informatics
    - Working with ITL (Dima) to develop MDCS system
    - Working with TRC to develop ThermoML
    - Working NLP efforts by providing data to mine to develop terms
  - Use Cases
    - Co-base superalloys
    - Additive Manufacturing (Ni-base superalloys, steels, Ti6Al4V)
  - DFT Bench Marking
  - NIST Computational Tools
    - OpenCALPHAD
    - Materials Genome Toolkit
  - Development of CALPHAD assessments uncertainty quantification
    - Uncertainty of output (assessment)
    - Uncertainty of inputs (experimental data, computational data)
    - De-couple weighting of data for assessment process from the uncertainty of assessment
• Slides to help with discussion
Examples of Files for a CALPHAD Thermodynamic Assessment

Functional Description File
Thermodynamics

- Elements: e.g. A,B
- Phases: e.g. FCC_A1 and BCC_A2 (crystal structure)
- Reference states used
- Software/version used
- Reference info
  - Bibliographic info
  - DOI link
- Contributor info
  - Name
  - E-mail
- User Comments
- Links to other files

Note: Thermo-Calc based file extensions used, but information on file types should also be included
Examples of Files for a CALPHAD: Diffusion Mobility Assessment

**Database files**

- **Diffusion Mobility Description**
  - Elements: e.g. A,B
  - Phases: e.g. FCC_A1 and BCC_A2 (crystal structure)
  - Software/version used
  - Reference info
  - Bibliographic info
  - DOI link
  - Contributor info
  - Name
  - E-mail
  - User Comments
  - Links to other files

- **Thermodynamic Description**
  - Link to TDB file or reference

- **Evaluated Data Files (e.g. DOP)**
  - Link to associated TDB file
  - Elements: e.g. A,B
  - Phases: e.g. FCC_A1 and BCC_A2 (crystal structure)
  - Software/version used (e.g. TC-PARROT, ver S)
  - Reference info for data
    - Bibliographic info, DOI link
  - Author info
    - Name
    - Reference info
    - Bibliographic info, DOI link
  - User Comments
  - Links to other files

- **MACRO/Script Files (e.g. to run simulations, DCM)**
  - Elements
  - Phases
  - Reference
  - Author
  - Links to other files (grid data, start values)
  - User comments

- **Auxiliary Data Files (e.g. *.EXP)**
  - Link to POP and/or TDB
  - Elements
  - Phases
  - Reference info
  - Author
  - Links to other files
  - User comments

Note: Thermo-Calc based file extensions used, but information on file types should also be included.
Customized DSpace repository for materials

- Enables sharing of a variety of data types, including text, images, and video
Data files

Offer licenses with attribution 3.0

Related Work

Digital Identifier
Material Data Curation System (MDCS)

- **Need**
  - Difficult to
    - Combine data from multiple sources
    - Understand and reuse existing data
    - Find associated metadata
    - Transform data into new formats

- **Objectives**
  - Facilitate collection, use, and reuse of materials data
  - Provide needed informatics infrastructure to enable High Throughput Experimentation (HTE)

https://github.com/usnistgov/MDCS

Alden Dima, NIST
The MDCS Approach

- **Web-based:** Python/Django/MongoDB, REST API, XML-based, SPARQL queries
- **Store data in XML-based templates**
- **Store, manage, & compose templates**
- **Spreadsheet input**
All MDCS functions available via the User Interface are also available via the REST API. The MDCS can be fully automated.
Curation of Raw Data

Curated Data
(Data to share)

User defined tools

Other Data/ Users

Interface with workflow tools

Send Data via API

Send Data via AIK

Also see separate API demo slides
Raw Data from a dilatometer:
Length vs Temperature
Python script to convert ASCII format to XML format
Need: Modular Data Models

- **Data Model Definition**
  - Defines the structure of data and metadata associated with the measurement or synthesis

- **Modularity**
  - Via the MDCS composer assemble modules for your workflow
  - Modify as modules needed

- **Not a standard**
  - Flexible data structures
  - Common reusable types
  - Domain modules

**Foundation:**
- **Shared Types**
  - E.g. Chemical Substance
- **Physical Quantities**
  - E.g. Pressure

Zach Trautt, NIST
Reusable Data Types

**Base Types**

<table>
<thead>
<tr>
<th>Base Type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Length</strong></td>
<td>- <code>value</code>: <code>xsd:double</code> &lt;br&gt; - <code>unit</code>: <code>lengthUnitType</code> &lt;br&gt; - <code>uncertainty</code>: <code>uncertaintyType</code></td>
</tr>
<tr>
<td><strong>Mass</strong></td>
<td>- <code>value</code>: <code>xsd:double</code> &lt;br&gt; - <code>unit</code>: <code>massUnitType</code> &lt;br&gt; - <code>uncertainty</code>: <code>uncertaintyType</code></td>
</tr>
<tr>
<td><strong>Time</strong></td>
<td>- <code>value</code>: <code>xsd:double</code> &lt;br&gt; - <code>unit</code>: <code>timeUnitType</code> &lt;br&gt; - <code>uncertainty</code>: <code>uncertaintyType</code></td>
</tr>
<tr>
<td><strong>Wafer Diameter</strong></td>
<td>- <code>value</code> &lt;br&gt; - <code>unit</code> &lt;br&gt; - <code>uncertainty</code></td>
</tr>
<tr>
<td><strong>Grain Size</strong></td>
<td>- <code>value</code> &lt;br&gt; - <code>unit</code> &lt;br&gt; - <code>uncertainty</code></td>
</tr>
<tr>
<td><strong>Diffusion Length</strong></td>
<td>- <code>value</code> &lt;br&gt; - <code>unit</code> &lt;br&gt; - <code>uncertainty</code></td>
</tr>
</tbody>
</table>

Zach Trautt, NIST
4.1  SI base units

Table 1. SI base units

<table>
<thead>
<tr>
<th>Base quantity</th>
<th>Name</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>meter</td>
<td>m</td>
</tr>
<tr>
<td>mass</td>
<td>kilogram</td>
<td>kg</td>
</tr>
<tr>
<td>time</td>
<td>second</td>
<td>s</td>
</tr>
<tr>
<td>electric current</td>
<td>ampere</td>
<td>A</td>
</tr>
<tr>
<td>thermodynamic temperature</td>
<td>kelvin</td>
<td>K</td>
</tr>
<tr>
<td>amount of substance</td>
<td>mole</td>
<td>mol</td>
</tr>
<tr>
<td>luminous intensity</td>
<td>candela</td>
<td>cd</td>
</tr>
</tbody>
</table>

4.2  SI derived units

Derived units are expressed algebraically in terms of base units or other derived units. The symbols for derived units are obtained by means of the mathematical operations of multiplication and division. For example, the derived unit for the derived quantity mass divided by amount of substance is the kilogram per mole, symbol kg/mol. Additional examples of derived units expressed in terms of SI base units are given in Table 2. (The rules and style conventions for printing and using SI unit symbols are given in Secs. 6.1.1 to 6.1.8.)

Table 2. Examples of SI coherent derived units expressed in terms of SI base units

<table>
<thead>
<tr>
<th>SI coherent derived unit</th>
<th>Derived quantity Name</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>area</td>
<td>square meter</td>
<td>m(^2)</td>
</tr>
<tr>
<td>volume</td>
<td>cubic meter</td>
<td>m(^3)</td>
</tr>
<tr>
<td>speed, velocity</td>
<td>meter per second</td>
<td>m/s</td>
</tr>
<tr>
<td>acceleration</td>
<td>meter per second squared</td>
<td>m/s(^2)</td>
</tr>
<tr>
<td>wavenumber</td>
<td>reciprocal meter</td>
<td>m(^{-1})</td>
</tr>
<tr>
<td>density, mass density</td>
<td>kilogram per cubic meter</td>
<td>kg/m(^3)</td>
</tr>
<tr>
<td>specific volume</td>
<td>cubic meter per kilogram</td>
<td>m(^3)/kg</td>
</tr>
<tr>
<td>current density</td>
<td>ampere per square meter</td>
<td>A/m(^2)</td>
</tr>
<tr>
<td>magnetic field strength</td>
<td>ampere per meter</td>
<td>A/m</td>
</tr>
<tr>
<td>luminance</td>
<td>candela per square meter</td>
<td>cd/m(^2)</td>
</tr>
</tbody>
</table>

4.2.1  SI coherent derived units with special names and symbols

Certain SI coherent derived units have special names and symbols; these are given in Table 3. Consistent with the discussion in Sec. 4, the radian and steradian, which are the two former supplementary units, are included in Table 3. The last four units in Table 3 were introduced into the SI for reasons of safeguarding human health.
Modularity: Foundational Types
Modularity: Foundational Types

Zach Trautt, NIST
Modularity: Synthesis Example

Substance Module

Physical Quantity Types
Composing a Template in MDCS:

Easy plug-and-play!

Primary Goal:
- Reduce startup cost: time required to create Template

Warning: Side effects include standardization through popularity and ease of use
Examples of CALPHAD Data

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
- Tracer Diffusivities

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

3-D
- Crystal structures
- Micrographs/Morphologies
- 3-D Atom probe Tomography
Not a standard

Module not quite right?

Change it!

Share it.

Zach Trautt, NIST
- Expanding to metallic systems
- Initial focus on phase equilibria data and thermochemical property data.