

Infrastructure for Nanoinformatics: Sharing & Reuse of Scientific Data

- Martin Fritts, SAIC-F, Frederick National Lab for Cancer and National Institute of Standards and Technology;
frittsmj@mail.nih.gov
- Do you see the reliability of database content as an issue? **Yes**
 - Insufficient detail on experimental methods, esp. sample preparation
 - Insufficient detail on models, esp. validation, range of validity, test suites
 - Reproducibility of data and sufficiency of metadata are both issues
- What are the infrastructure needs for databases?
 - Access to databases in different fields and disciplines
 - Evaluation of the quality of experimental and modeling data
 - Data curation, annotation and linking
- In which areas is the EU-US collaboration most likely to be successful?
 - Identifying challenge/pilot problems and use cases
 - Gathering of requirements and rapid prototyping
 - Collaboration among countries (BiLAT, CoRs) & programs (NNI, MGI)
- Should there be common databases? **No**

Develop a framework for assessing data quality (model and experiment) based on data reproducibility and sufficiency of the metadata



Data Readiness Levels (DRLs)

DRL 0. Invalid data

DRL 1. Raw or unscaled data

DRL 2. Scaled data

DRL 3. Data with defined precision or noise

DRL 4. Data with defined precision and noise

DRL 5. DRL 4 + data related to the larger body of scientific knowledge

DRL 6(X). Standards-quality data of X % measurement uncertainty



Metadata Levels

Poor: Insufficient information: data cannot be reproduced/ interpreted by others

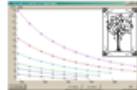
Acceptable: Others can reproduce and interpret the data; e.g., adequate descriptions of exp/comp methods used; descriptions of data formats.

Excellent: **Acceptable** + additional information; e.g., history/provenance, validation of the experimental methods and models.



Global Information Systems *Application to the Field of Thermodynamics*

- ❑ Software tools for mass-scale data capture



Guided Data Capture (GDC) s

Two levels of data curation

- ❑ Comprehensive Data Storage Facility



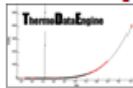
SOURCE Data Archival System

- ❑ Data Entry Facility



NIST/TRC Data Entry Facility

- ❑ Software Expert Systems



ThermoData Engine (TDE) software

- ❑ Data Communications Standard



ThermoML

- ❑ Data Reader Software



ThermoML Opener into

- ❑ Web Communication Portal



NIST Web-Oracle infrastr

Work with:

Journal of Chemical and Engineering Data,
Journal of Chemical Thermodynamics,
Fluid Phase Equilibria,
International Journal of Thermophysics,
and Thermochimica Acta
(~1000 articles per year)

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

MGI James Warren 28/6/2013

Tuesday, May 28, 13

[http://www.nitrd.gov/nitrdgroups/index.php?title=Data Sharing and Metadata Curation: Obstacles and Strategies](http://www.nitrd.gov/nitrdgroups/index.php?title=Data_Sharing_and_Metadata_Curation:_Obstacles_and_Strategies)

(Big Data Program , with agencies in Networking and IT R&D Program- NITRD)