Scientific Data at NIST: Current Practices and Vision

Dan Friend, Chief
Thermophysical Properties Division
Chemical Science and Technology Laboratory
National Institute of Standards and Technology
Boulder, CO (USA)

Scientific Data Management (SDM) for Government Agencies
Workshop to Improve SDM
29 June 2010
Washington, DC (USA)

NIST Mission

To promote U.S. innovation and industrial competitiveness by advancing

measurement science, standards, and technology

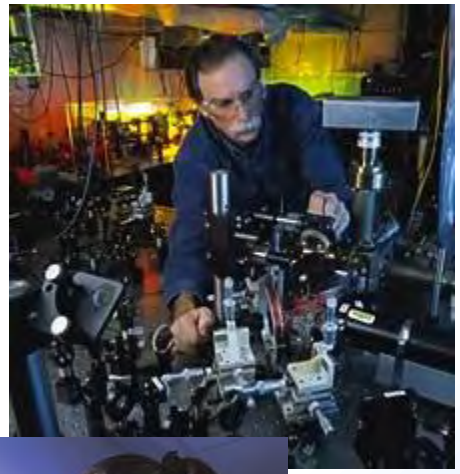
in ways that enhance economic security and improve our quality of life

©Robert Rathe

NIST at a Glance

Major Assets

- ~ 2,900 employees
- ~ 2600 associates and facilities users
- ~ 1,600 field staff in partner organizations
- ~ 400 NIST staff serving on 1,000 national and international standards committees



©Robert Rathe

Major Programs

- NIST Laboratories
- Baldrige National Quality Program
- Manufacturing Extension Partnership
- Technology Innovation Program

Data Climate at NIST

- Scientific Data Ubiquitous at NIST

- Some Unifying Themes
  - Requirement for Compliance with *Guide to the Expression of Uncertainty in Measurement* (ISO GUM, NIST TN 1297)
  - Standard Reference Data Act (P.L. 90-396)
  - Quality System Policies for NIST Measurement Services
  - Data Programs Must Support Agency Mission
  - Data Programs Prioritized According to Agency Priorities

- Agency Digital Data Policy
  - Evolving

- Data Management Plans
  - Evolving
  - Largely De-centralized
    - Key decisions/recommendations at the Division level
    - Coordination at the level of the Operating Unit (Laboratory)
    - Coordination through SRD (Measurement Services Division)

Standard Reference Data Act (P.L. 90-396)

"The Congress hereby finds and declares that reliable standardized scientific and technical reference data are of vital importance to the progress of the Nation's science and technology. It is therefore the policy of the Congress to make critically evaluated reference data readily available to scientists, engineers, and the general public. It is the purpose of this Act to strengthen and enhance this policy."

"Standard reference data … may be made available and sold… "

"…the Secretary may secure copyright … "

Thermophysical Properties Division

NIST Standard Reference Data Program

- **140 Scientific and Technical Databases**
  - 91 Available free on-line
  - 46 PC databases available for purchase
  - 3 On-line databases available by subscription

- **Journal of Physical and Chemical Reference Data**

- **Examples**

  (free) NIST Chemistry WebBook is the most widely used NIST data product and is used by scientists, engineers, educators and students worldwide for applications in the areas of chemical engineering, physical chemistry, analytical chemistry, and chemical informatics.

  (fee) NIST/EPA/NIH Mass Spectral Database is used by environmental, toxicology, forensic, and biomedical laboratories throughout the world and is distributed as an option by mass spectrometer manufacturers.

Thermophysical Properties DIVISION

Data Effort in Thermodynamics

Critically evaluated thermophysical property data are vital for numerous industrial sectors. Thousands of chemicals are of current interest and millions of potential future interest. Historical data may be of equal importance to current research.

•1942 NIST (then NBS) established Thermodynamics Research Center

•Data effort has been supported continuously since that time—within and outside of NIST

•Scope and focus have evolved

•Full range of data management plans have been developed at the programmatic level

Global Information Systems

Information Systems Designed to…

➢ Collect, Process, Integrate, Evaluate, and Communicate the *Entire* "Body of Knowledge" Pertaining to a Field and

➢ Support *Any* Application Requiring This Knowledge in an "on-Demand" Mode with Definitive Information Quality Assessments

Critical Components:

- ❑ Software tools for mass-scale data capture
- ❑ Comprehensive Data Storage Facility
- ❑ Data Processing Facility
- ❑ Data Communications Standard
- ❑ Data Reader Software
- ❑ Software Expert Systems
- ❑ Web Communication Portal

"That only a tiny fraction of all possible chemical compounds have been prepared and studied suggests that great discoveries and technological payoffs will come from further advances in knowledge, if researchers can efficiently find or make the compounds they want."

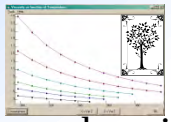
Graham Fleming, UC Berkeley
Mark Ratner, Northwestern Univ.

Physics Today, July 2008

Global Information Systems

Application to the Field of Thermodynamics

- Software tools for mass-scale data capture

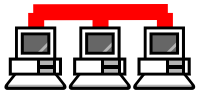


Guided Data Capture (GDC) software

- Comprehensive Data Storage Facility

SOURCE Data Archival System

- Data Entry Facility



NIST/TRC Data Entry Facility

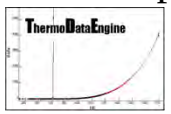
- Data Communications Standard

ThermoML

- Data Reader Software

ThermoML Opener into Microsoft Excel

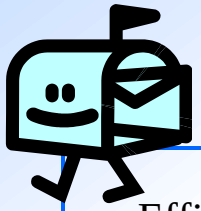
- Software Expert Systems

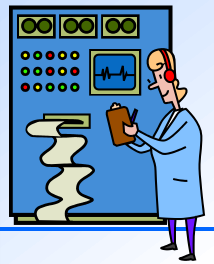


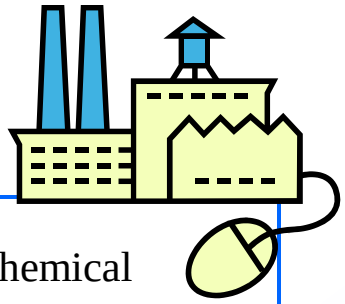
ThermoData Engine (TDE) software

- Web Communication Portal

NIST Web-Oracle infrastructure







Efficiency of Information Delivery

Journal Publication Quality

Instrument Calibration & Validation

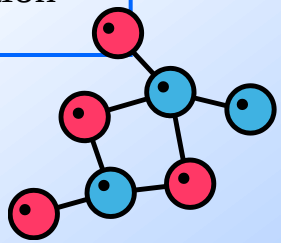
Chemical Process Design

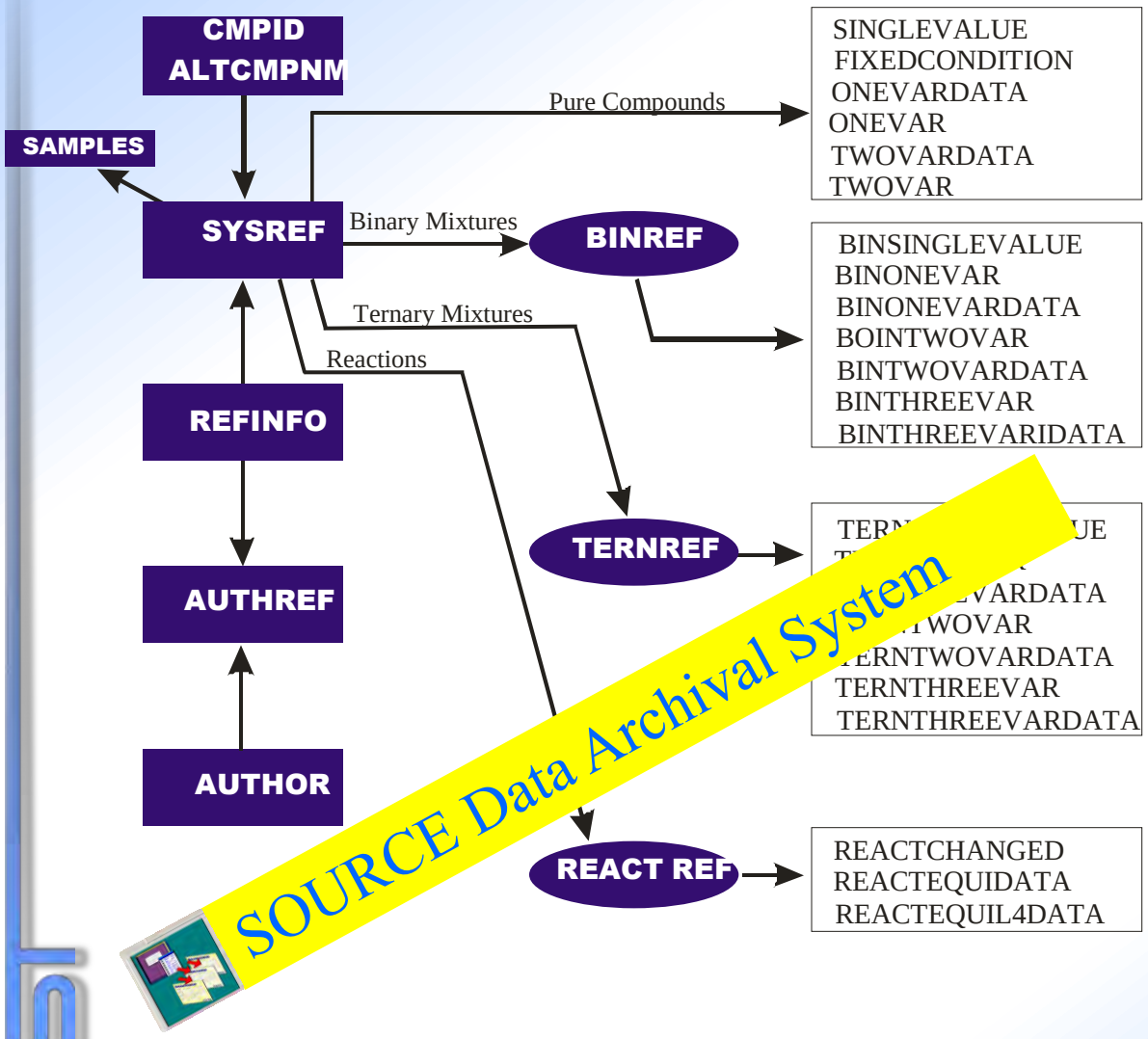
Global Information System In Thermodynamics

Scientific Discovery Process

Strategic Experiment Planning

Molecular Modeling & Property Prediction





- Structure is based on the *Gibbs Phase Rule*
- Accommodates a wide variety of reported data representations (*absolute, ratio, difference, a variety of composition measures, etc.*)

Thermophysical Properties DIVISION

Data Processing at NIST

Thermodynamics Research Center

Incoming Data Flow
500,000 data points/year

In-House Data Processing

Major journal cooperation

NIST/TRC Data Entry Facility

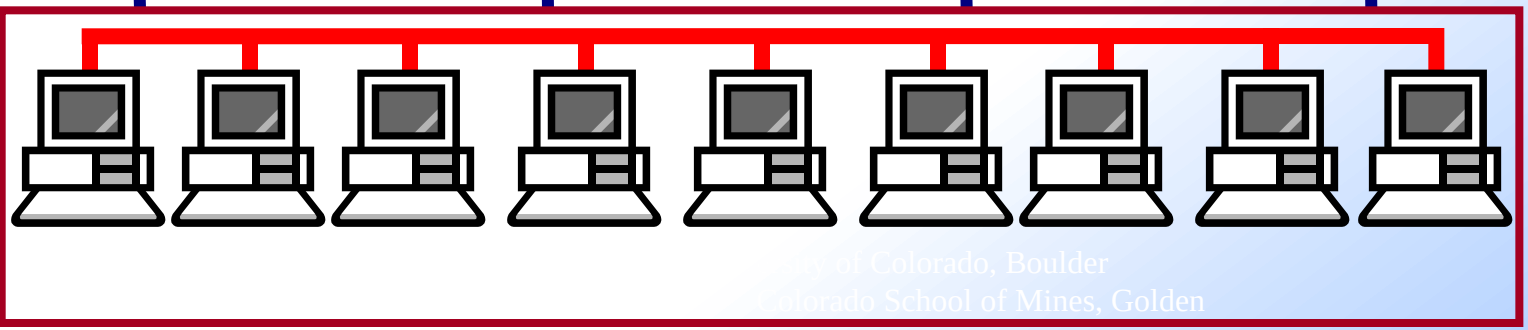
Communications

- Access to NIST SOURCE/Oracle Database
- Communications *Windows Network*

Tools

- Hardcopy digitizing
- Guided Data Capture: *GDC* software

- Data Capture Software Support
- Data Capture from international data repository resources (Russia/China)
- Managerial support



Pure Appl. Chem., Vol. 78, No. 3, pp. 541–612, 2006.
doi:10.1351/pac200678030541

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

COMMITTEE ON PRINTED AND ELECTRONIC PUBLICATIONS*

XML-BASED IUPAC STANDARD FOR EXPERIMENTAL, PREDICTED, AND CRITICALLY EVALUATED THERMODYNAMIC PROPERTY DATA STORAGE AND CAPTURE (ThermoML)**

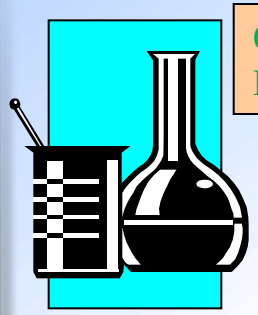
(IUPAC Recommendations 2006)

Prepared for publication by
MICHAEL FRENKEL[1,‡], ROBERT D. CHIRICO[1], VLADIMIR DIKY[1], QIAN DONG[1], KENNETH N. MARSH[2], JOHN H. DYMOND[3], WILLIAM A. WAKEHAM[4], STEPHEN E. STEIN[5], ERICH KÖNIGSBERGER[6], AND ANTHONY R. H. GOODWIN[7]

[1]Physical and Chemical Properties Division, National Institute of Standards and Technology, Boulder, CO 80305-3328, USA; [2]Department of Chemical and Process Engineering, University of Canterbury, Private Bag 4800, Christchurch, New Zealand; [3]Chemistry Department, University of Glasgow, Glasgow G12 8QQ, UK; [4]School of Engineering Sciences, University of Southampton, Highfield, Southampton SO17 1BJ, UK; [5]Physical and Chemical Properties Division, National Institute of Standards and Technology, Gaithersburg, MD 20899-8380, USA; [6]School of Engineering Science, Murdoch University, Murdoch, WA 6150, Australia; [7]Schlumberger Technology Corporation, 125 Industrial Blvd., Sugar Land, TX 77478, USA

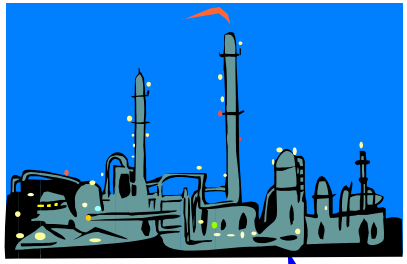
Thermophysical Properties DIVISION

Global Data Communications



Committee on Printed & Electronic Publication (CPEP)

International Association of Chemical Thermodynamics (IACT)



IUPAC

Experimental Measurements

NIST Guided Data Capture (GDC)

products

Users

- J. Chem. Eng. Data *(ACS)*
- J. Chem. Thermodyn. *(Elsevier)*
- Fluid Phase Equilib.
- Int. J. Thermophys. *(Springer)*
- Thermochimica Acta

ThermoML

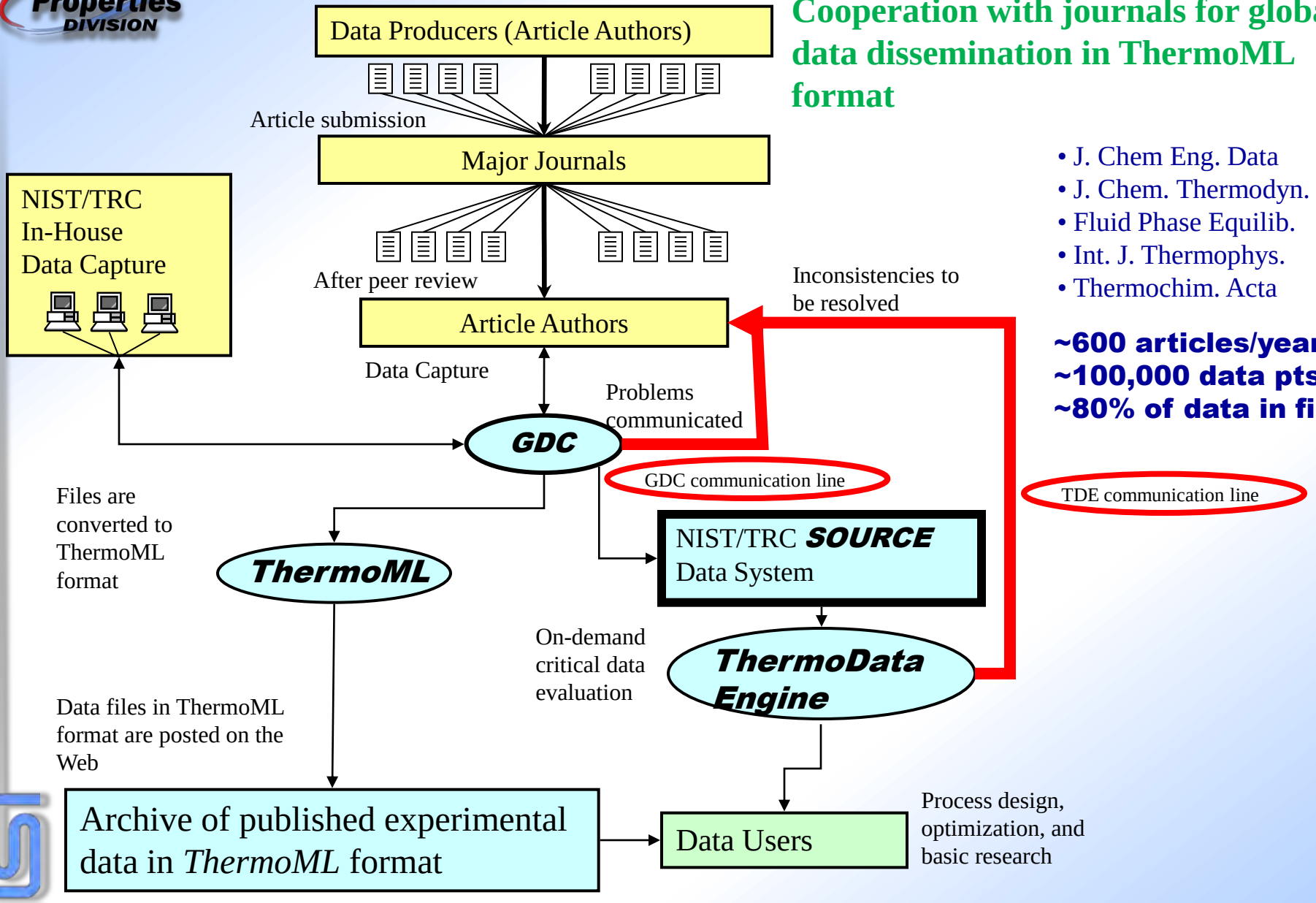
"Reader" Software

NIST ThermoML Archive

WEB

- AspenTech *(USA)*
- SimSci *(USA)*
- Virtual Materials Group *(Canada)*
- Nat'l Engineering Lab *(UK)*
- Fiz Chemie *(Germany)*
- DIPPR *(USA*

Cooperation with journals for global data dissemination in ThermoML format



- J. Chem Eng. Data
- J. Chem. Thermodyn.
- Fluid Phase Equilib.
- Int. J. Thermophys.
- Thermochim. Acta

~600 articles/year
~100,000 data pts/yr
~80% of data in field

From 2008 TRC Study

Approximately 10-15% of all published articles include typographical errors in the tables of experimental data

Problems:

- Number transposition
- Unit errors
- Component switch for mixtures
- Wrong compound
- Erroneous column duplication in tables
- Hidden errors (calibration errors)
- + many others

Note: We do not quibble with authors concerning poor uncertainty estimates and terminology, but the poor estimates are **not** transferred to SOURCE.

Thermophysical Properties DIVISION

New mandatory submission process via NIST/TRC for major journals in the field

Fluid Phase Equilibria 276 (2009) 165–166

Contents lists available at ScienceDirect

Fluid Phase Equilibria

journal homepage: www.elsevier.com/locate/fluid

Editorial

Joint Statement of Editors of Journals Publishing Thermophysical Property Data Process for Article Submission for *The Journal of Chemical Thermodynamics*, *Fluid Phase Equilibria*, *International Journal of Thermophysics*, *Thermochimica Acta*, and *Journal of Chemical Engineering Data*

2 *J. Chem. Eng. Data* **2009**, *54*, 2–3

Joint Statement of Editors of Journals Publishing Thermophysical Property Data

Process for Article Submission for *The Journal of Chemical Thermodynamics*, *Fluid Phase Equilibria*, *International Journal of Thermophysics*, *Thermochimica Acta*, and *Journal of Chemical and Engineering Data*

A requirement for submission of a manuscript describing properties is a literature search and comparison of the results with previously reported literature values. Often, reviewers cannot make informed decisions regarding the manuscript because the authors have made only a minimal literature review and comparisons. It is then an unacceptable burden to require reviewers to research previously published literature data to ensure a proper comparison has been made and hence determine the ultimate worth of the manuscript. To accommodate this, a new arrangement has been made with the Thermodynamics Research Center (TRC) of the National Institute of Standards and Technology (NIST). Specifically, thermophysical property data for systems reported in a newly submitted manuscript will be compared against the NIST TRC databases. TRC will provide a report to the Editors who at their discretion will forward it to the reviewers and/or the authors. This new procedure is mandatory and will operate by collaborative agreement with Journals in this field including *The Journal of Chemical Thermodynamics*, *Fluid Phase Equilibria*, *International Journal of Thermophysics*, *Thermochimica Acta*, and *Journal of Chemical and Engineering Data*. The Editors will adhere strictly to this policy and there will be no exceptions. This new procedure becomes effective January 2009

A requirement for submission of a manuscript describing properties is a literature search and comparison of the results with previously reported literature values. Often, reviewers cannot make informed decisions regarding the manuscript because the authors have made only a minimal literature review and comparisons. It is then an unacceptable burden to require reviewers to research previously published literature data to ensure a proper comparison has been made and hence determine the ultimate worth of the manuscript. To accommodate this, a new arrangement has been made with the Thermodynamics Research Center (TRC) of the National Institute of Standards and Technology (NIST). Specifically, thermophysical property data for systems reported in a newly submitted manuscript will be compared against the NIST TRC databases. TRC will provide a report to the Editors who at their discretion will forward it to the reviewers and/or the authors. This new procedure is mandatory and will operate by collaborative agreement with all Journals in this field including *The Journal of Chemical Thermodynamics*, *Fluid Phase Equilibria*, *International Journal of Thermophysics*, *Thermochimica Acta*, and *Journal of Chemical Engineering Data*. The Editors will adhere strictly to this policy and there will be no exceptions. This new procedure becomes effective January 2009

Fluid Phase Equilib. 2009, *276,165–166*
Int. J. Thermophys. 2009, *30, 371–373*
J. Chem. Eng. Data 2009, *54, 2-3*
J. Chem. Thermodyn. 2009, *41, 575–576*
Thermochim. Acta 2008, *484, vii–viii*

"We, the Editors, are convinced that this additional data review will substantially benefit the scientific and engineering communities because of the increase in quality and usefulness of the reported experimental data."

ThermoData Engine (TDE) software

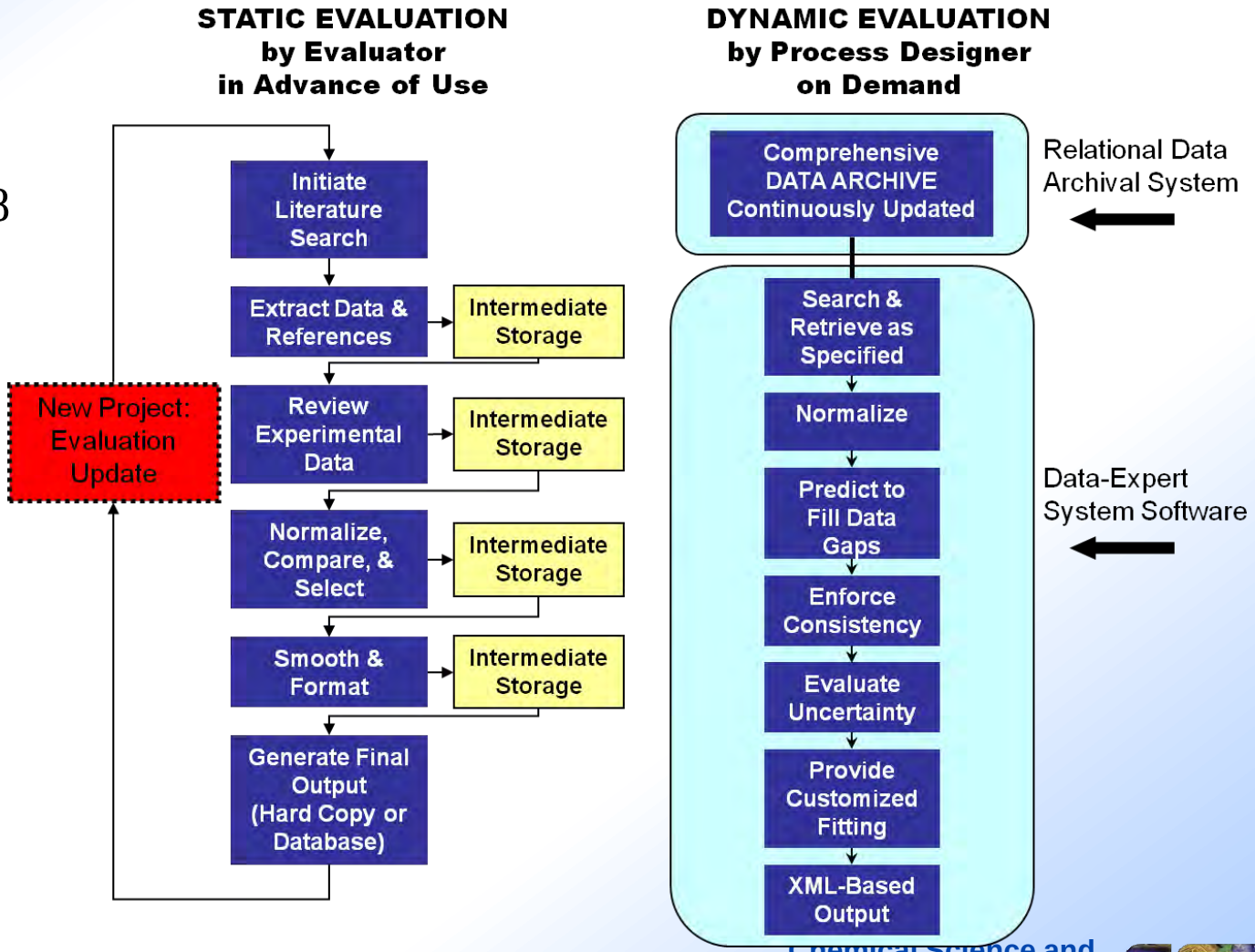
Used for > 200 years in all critical data evaluation worldwide

Developed at NIST: Implemented for the first time in 2004

Concept development 1988 to 1997

Concept Implementation 1997 to present

Framework completion target: 2012



Application & Advantages of Global Data System (in Thermodynamics)

- Automated generation of consistent recommended values
  - **(*on-demand*: *results in minutes vs. months or years for traditional static methods*)**
- Can be applied to hypothetical compounds
  - **Requests for compound data can be input as drawn structures**
- Full set of properties for pure compounds are always generated (predictions w/ +/-)
- Estimated uncertainties for all recommended data
- Can be used to develop new and validate old models
- Reveals published experimental errors
- Provides a comprehensive data source for process simulation *via* the ThermoML-formatted output

Principal Questions

- How to Navigate to Find Relevant Information in the Entire Body of Knowledge Pertaining Thermophysical Property Data (Needle in a Haystack) ?
- How to Develop Experimental Plans Assuring “Significant Return on the Investment”?
- Is it Possible to Validate New Experimental Data?
- How to Deliver Efficiently Information from “Data Producers” to “Data Users”?
- Is it Possible to Discover Engineering Models that have Fewer Parameters than Currently Used Forms, but which still Work Well over a Broad Range of Fluids?

A Vision for NIST Measurement Services

NIST is seen by its staff and stakeholders as providing value to its customers with the right measurement services at the right time.

We succeed by:

- ***Identifying customers and their needs** in an open and consistent process.*
- ***Maintaining a world-class research program** in measurement science to enable our ability to respond to respond to dynamic wide-ranging customer needs in a timely manner.*
- ***Leveraging our resources through strategic partnerships***
  - Utilize strategic and more effective partnering with other NMIs
  - Leverage private sector and other agency capabilities and resources
- ***Consistently assessing** the direction, value and impact of our work.*
- ***Having staff be rewarded** for their contributions to measurement service delivery as well as research.*

NIST's Interests

- NIST's databases are a national resource representing an investment of several hundred million $ since the 1950s.
- Realizing their value into the future will require appropriate policies and solutions for:
  - Archiving
  - Preservation
  - Maintenance
  - Protocols for Database Quality
  - Cyberinfrastructure (metadata standards, …)
  - Dissemination

Conclusions

- Further partnerships among agencies should be considered
- Data structure should reflect underlying scientific principles of field
- Critical data evaluation necessary for integrity of data systems
- Appropriate metrics for database quality should be established
- Policies and data management plans must be appropriate to the community of practice

Questions??