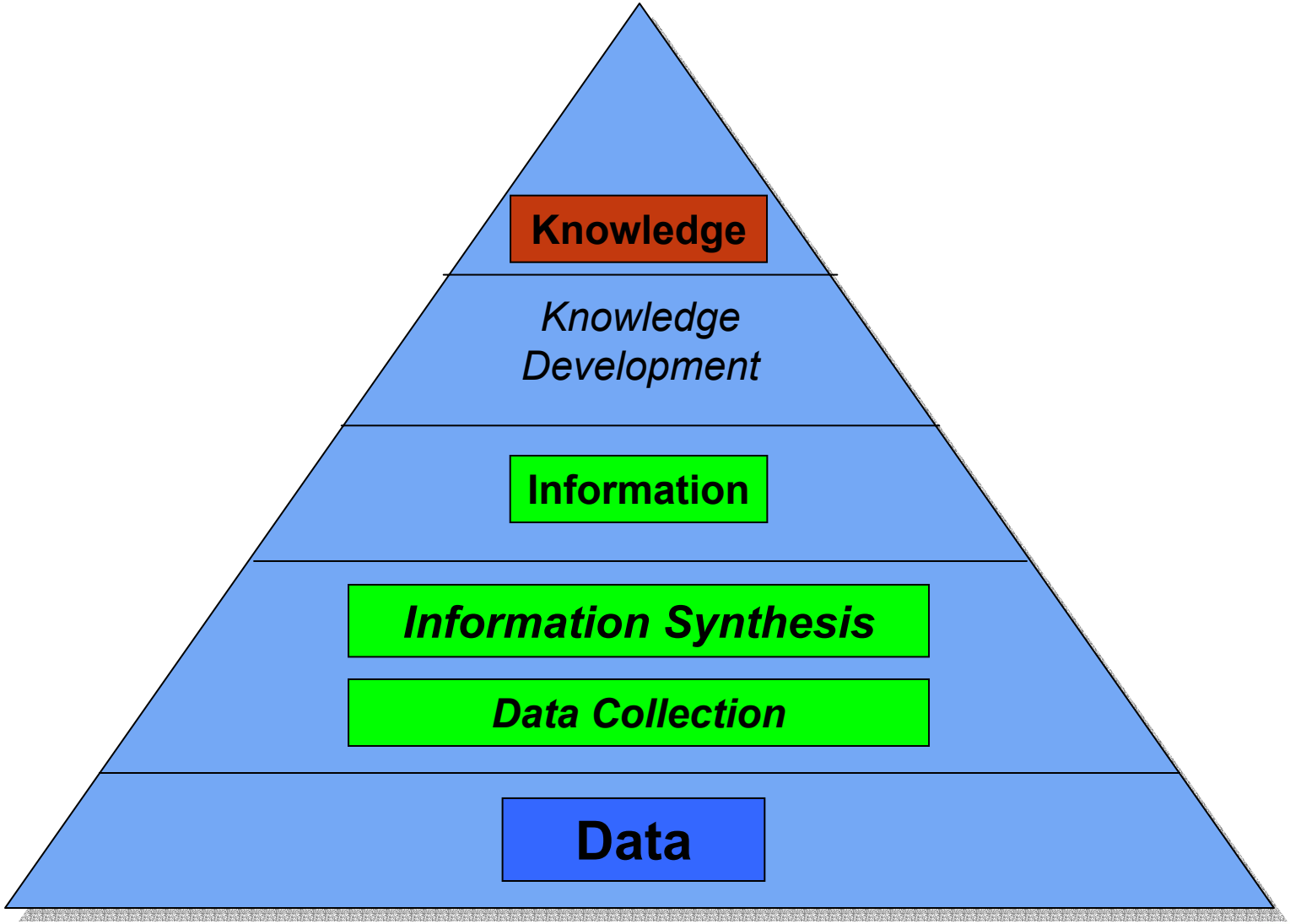
Intelligent Data Infrastructure

Greg Rosasco
Chief, Physical and Chemical Properties Division
Chemical Science and Technology Laboratory

Information & Knowledge Management



Vision

- **NIST will be the world's foremost and best resource for physical, chemical, biochemical, and materials property information**
  - **Intelligent Information Infrastructure**
    - **Data Collection**
    - **Information Synthesis**
    - **Information Dissemination**

Major Industrial Drivers

Industries of the Future - Roadmaps

- Aluminum Technology Roadmap

  "enabling technologies: comprehensive process models, integrate product design and processing"

- Technology Vision 2020: The U.S. Chemical Industry

  "Throughout the chemical industry, the ways in which data are turned into information and used, managed, transmitted, and stored will be critical to its ability to compete.

  Improved and enhanced information systems are at the heart of our vision..."

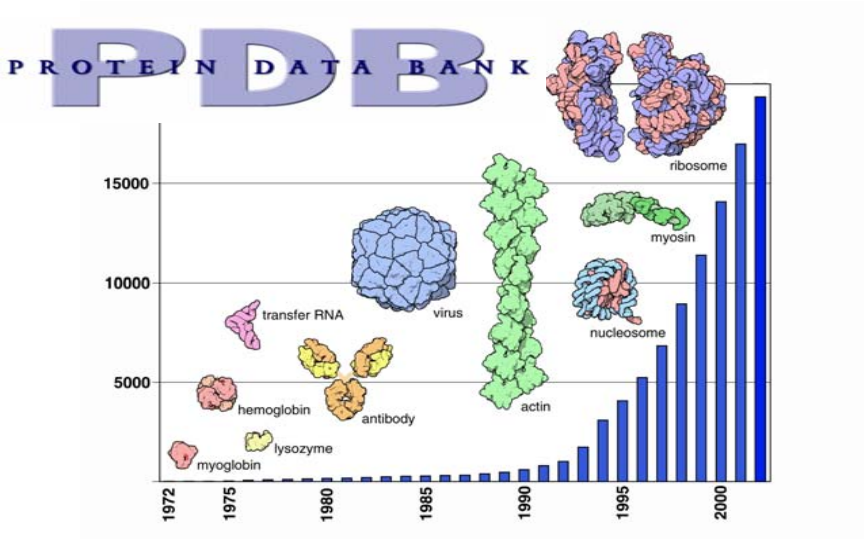
Meeting the challenge

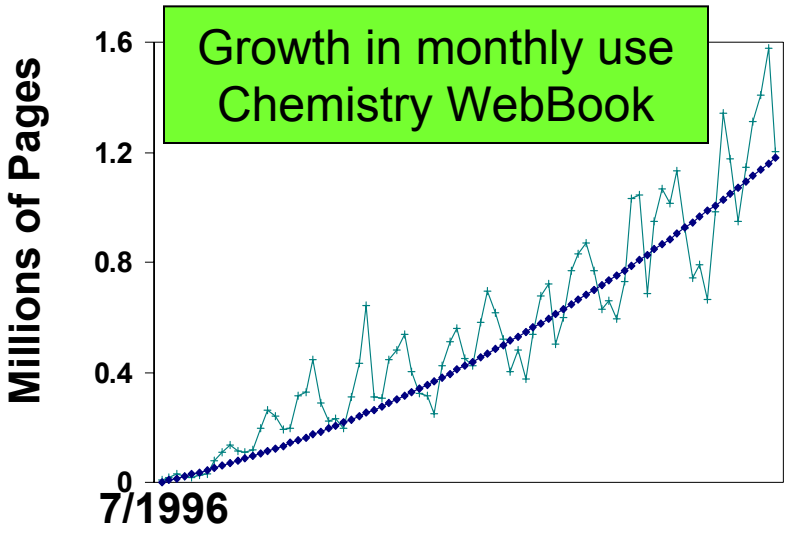
- *"The desire to improve manufacturing processes and the need to design new materials will be major driving forces in the chemical industry over the next two decades."*
- *"Advances in modeling and simulation...could have a significant impact on reducing the cost and time involved in designing chemical processes and new materials or catalysts."*

Roadmap for Computational Chemistry: Technology Vision 2020: The U.S. Chemical Industry

Impact on Data Infrastructure

- We see an explosive growth in the need for **reliable** physical, chemical, biochemical, and materials property **data** to support process and product optimization and discovery.
  - combinatorial methods
    - NIST Combinatorial Methods Center (MSEL)
  - high throughput experimentation
  - doubling of the volume of published data in the last 10 yrs





Reliable Data → **Information**

- Reliable Data
  - suitably documented
  - established uncertainty
  - first step in the creation of information
  - core of NIST mission

"...critically evaluated data on well-characterized substances..."

Standard Reference Data Act

"...NIST may be the ONLY organization that could effectively carry out [Virtual Measurements and Dynamic Data Evaluation] because of its scope, experience in the disciplines required, its independence, and lack of short-term financial return requirement (which has killed private attempts...)." CEO, Kaufman Associates

"...The best reason for NIST to do this and not another group or agency is its reputation for quality and high standards..." Director of R&D and Chief Scientist, INEEL

"I feel that NIST is uniquely positioned and qualified to provide high-quality data products, tools, and support that is needed by U.S. industry, and I am convinced that NIST needs a continued mandate and continued support." Lab. Head and Principal Engineer, Eastman Chemical Company

Situation Analysis

- Effective utilization of the dramatically increasing volume of scientific data requires **advances in data evaluation, virtual measurements** (computational estimation/prediction)**, data management, and data mining**

"The need for thermophysical data in industry has reached a level that overwhelms traditional resources...It is not only impossible to generate all the necessary data given the financial concerns, it often is not possible to know for certain what data are necessary sufficiently far in advance to permit conventional measurement."

Dean, Chemical Engineering, Texas A&M University

Thrust of NIST Efforts

- Robust, secure, autonomous intelligent systems adapted to information needs of the customer
  - **Data collection**
    - volume of data, dispersed resources world-wide
  - **Information Synthesis**
    - rapid pace of industrial innovation
    - presently unknown systems and conditions
  - **Information Dissemination**
    - disaggregated, disparate customers
    - direct interface to applications
    - information current and available on-demand

Most examples taken from efforts in CSTL; similar efforts throughout the Institute

Data Collection

- **Data exchange standards**
  - **facilitate collection**
    - XML Standards -developed in collaboration with industry stakeholders
    - structure identification standards and software
  - **support evaluation**
    - adequate specification of substance, conditions, uncertainties
- **Interactive, self-checking systems for data collection**
  - Guided Data Capture developed by the TRC Group
    - improved data quality
    - collection rate of 300,000+ points per year

  Goal: eliminate backlog in data entry from major sources of published thermophysical data by 2006.
  (more than 80% of all such data)

Data Collection

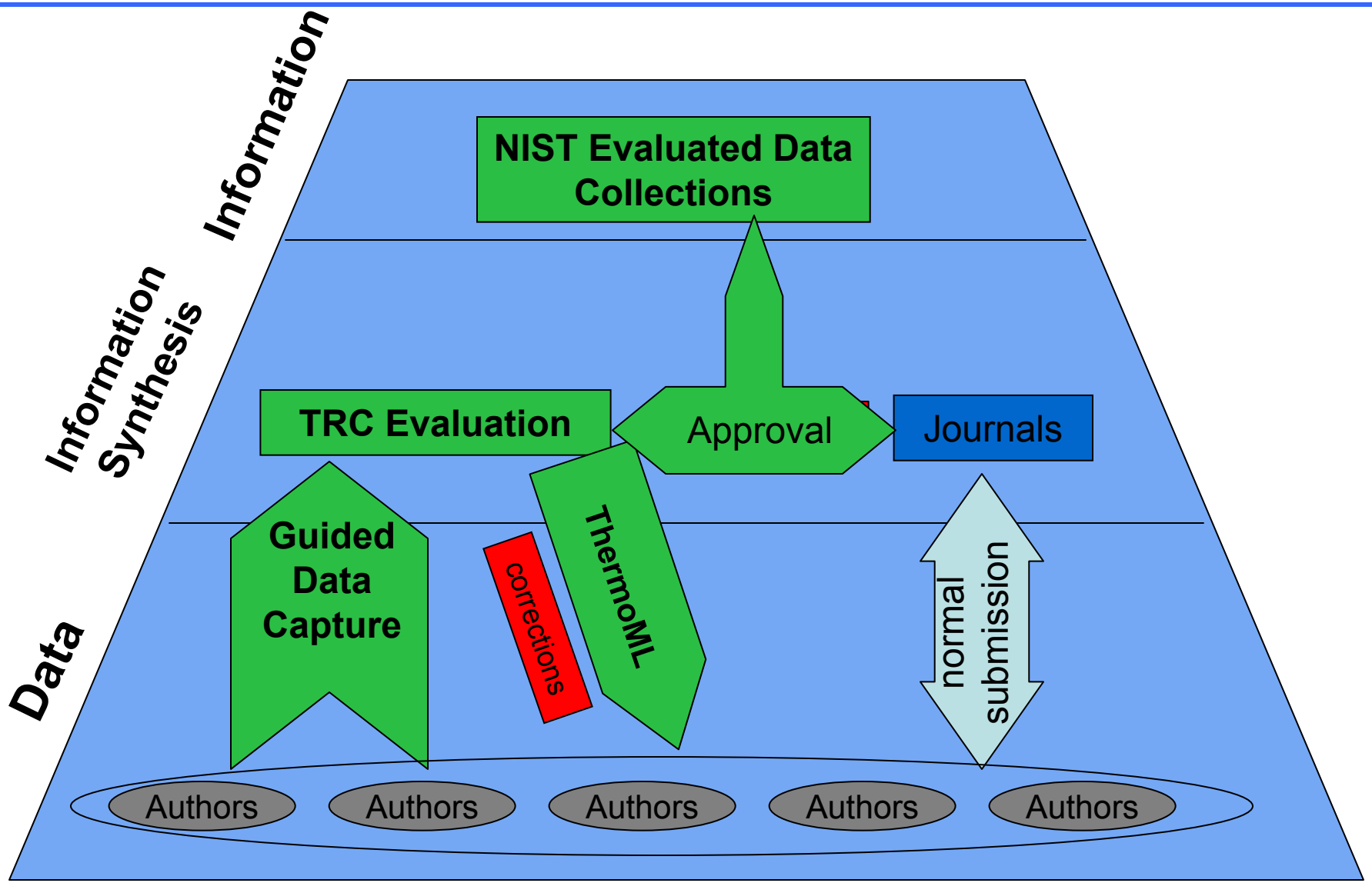
The Protein Data Bank (Rutgers, UC San Diego, NIST)

- structural archive for atomic coordinates of biological macromolecules and assemblies
  - informatics for Structural Genomics (large scale, high-throughput)
- **collect all data prior to publication for projects supported by NSF, NIH, and other government agencies (req'd for funding!)**
- interactive self-checking data submission interfaces
  - user adapted data transfer standards
- quality control protocols at all levels
- data deposition rate in synch with demands
- continuous improvements in query, reporting, and access

PDB PROTEIN DATA BANK

NIST CSTL

Thermophysical Data Collection



Data Collection

- Strategic partnerships with external data resources
  - **J. Chemical and Engineering Data**
    - all data approved for publication
      - improved data quality via GDC system
      - available on TRC Group website
      - incorporated into TRC-SOURCE Database

Goal: Expand journal coverage to acquire automatically 80% of newly published thermophysical data by 2006.

Data Collection

- Strategic partnerships with external data resources
  - Industry willing to share data
    - TRC Consortium
    - donation of previously proprietary data
  - Alliances with other established suppliers of data
    - Fiz Chemie - Berlin; Chinese Academy of Sciences; Russian National Institute for Standardization

Partnerships driven by recognized quality of NIST Data Resources

Information Synthesis

Thrust of NIST Efforts

Data collection

volume of data, dispersed resources world-wide

Information Synthesis

rapid pace of industrial innovation

presently unknown systems and conditions

Information Dissemination

- Rese… …nded data …
  - …
  - …stimation
  - …matched to the pace of innovation

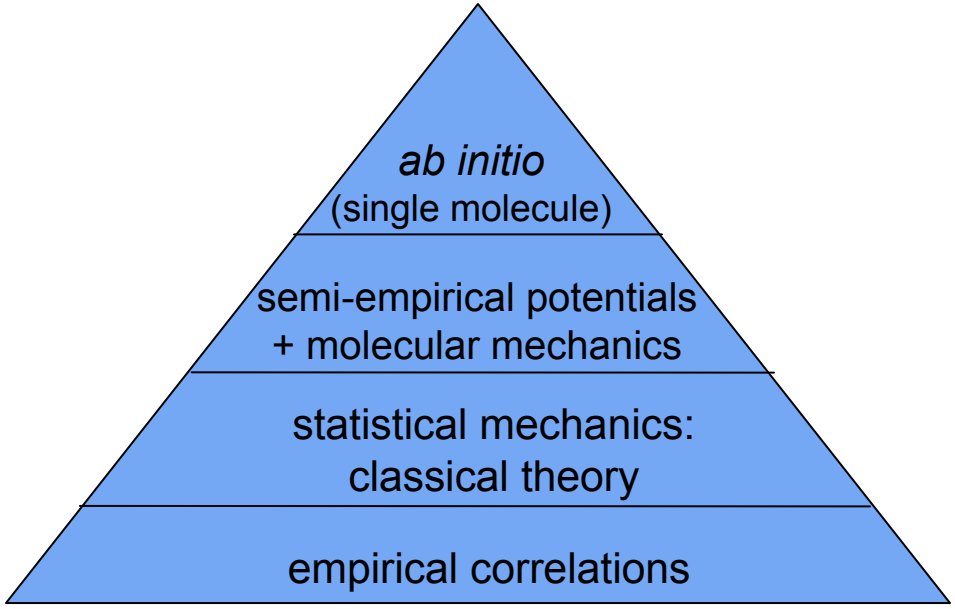
"This tool would provide a repository for data and information, query tools, and delivery of new data based on known. The new data infrastructure as outlined...is an outstanding start at addressing...very real problems." *BPAmoco Naperville Complex*

"...software for processing of parameter uncertainty and error propagation will create a quantum change in the way chemical and process engineering are practiced and it will change the perception of value of ... data in the eyes of commercial enterprises, large and small." *CTO, Virtual Materials Group*

- Added benefit: *Provide tools to customers that will facilitate their assessment of data quality*
- software versions of MSEL's NIST Recommended Practice Guides

Information Synthesis

- Researching Virtual Measurement (Expert) Systems
  - filling the gaps using computational tools
    - new chemical or system
    - conditions preclude direct measurement
  - "fit-for-purpose" predictions
    - matched to the level of uncertainty required by the user
    - based on knowledge of uncertainty limits of predictive methods



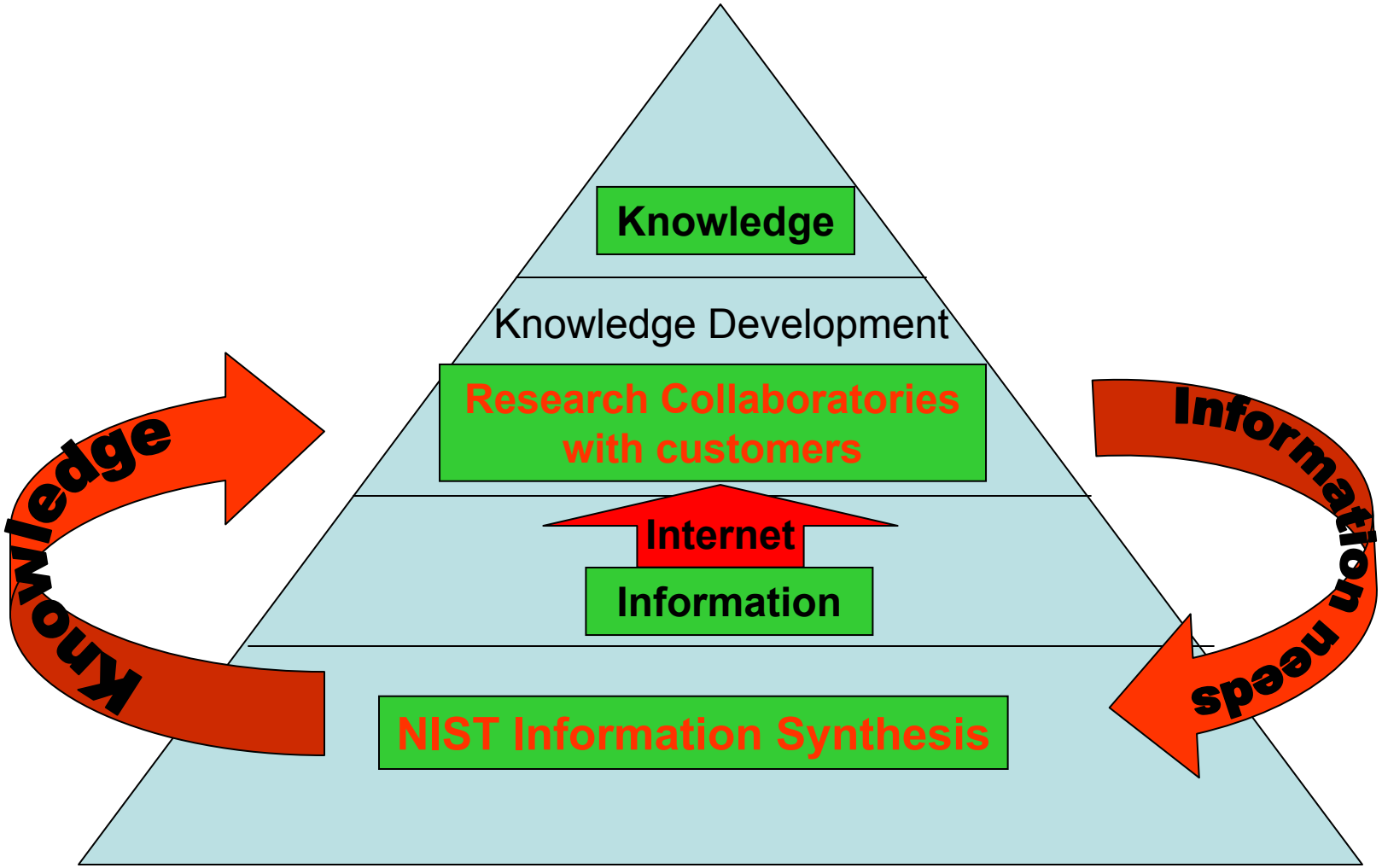
"I am very enthusiastic about the Virtual Measurements... I think there will be increasingly dramatic opportunities for sophisticated computations to generate estimations/predictions of physical properties that are worthy of comparison with the finest experimental results."

Research Partnerships Leader, Dow Chemical Company, Corporate R&D

Information Dissemination

- **rapid migration to web-based dissemination**
  - over 70 websites at NIST
  - real time updates
  - adapted to user platform and application needs
  - utilizing information exchange standards
  - feedback for assessing information needs of the customer
- **establishing research collaboratories with specific user communities**
  - rapid feedback on efficacy and adequacy of information
  - establish priorities for development efforts
    - Research Collaboratory on Structural BioInformatics
    - NIST Combinatorial Methods Center
    - Research Collaboratory on Multi-scale Molecular Science

Research Collaboratories



+ Real-time, on-line peer review

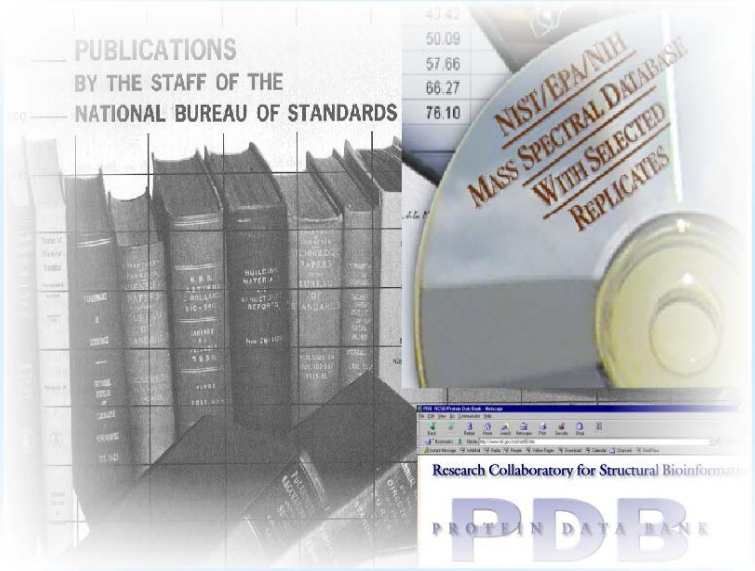
Enabling Our Vision

- Provide **Trustworthy Information Infrastructure** for the **Information Age**
  - three broad, high-level strategic opportunities
    - Infrastructural Technologies for Intelligent Interconnected Systems
      - Trustworthy Computing
    - Interoperability Technologies for Collaboration and Sharing
    - Virtual Measurements and Dynamic Data Evaluation

Information and Knowledge Management-Strategic Focus Area, Team Report, April 2002

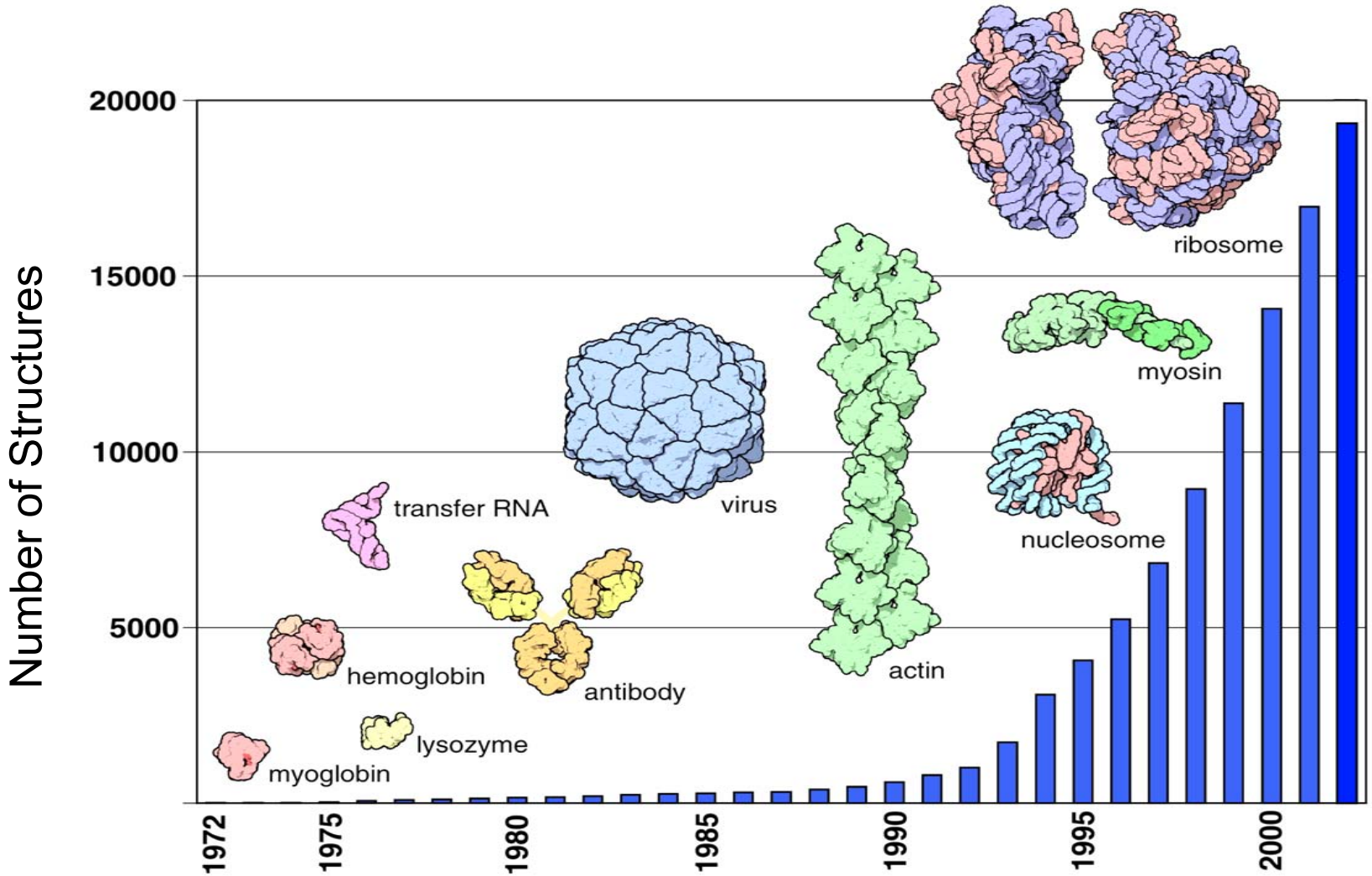
END

Trends in provision of Reference Data

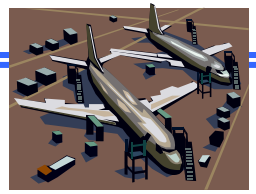


- Traditional approaches
  - extensive time of experts
  - long lead-times
  - static information
    - not based on most current data
    - cannot anticipate current needs

- **Dynamic Data Evaluation**
  - expert systems reduce load
  - based on comprehensive and current data sets
  - able to address user needs on demand
  - real time peer-review of NIST data products
  - real time assessment of user needs and gaps


Number of Structures
20000
15000
10000
5000
1972
1975
1980
1985
1990
1995
2000
ribosome
myosin
nucleosome
virus
transfer RNA
actin
hemoglobin
antibody
lysozyme
myoglobin

Enterprise Integration for the Chemicals Industry



Automotive Industry

Aerospace Industry

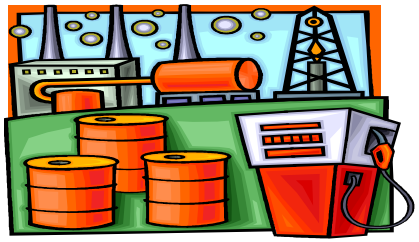
Construction Industry

Chemical Industries

PDXI, CIDX

Construction Industry

Properties Data are essential specifications in many transactions



Raw Materials Suppliers

Examples of XML Standards

- ThermoML: thermophysical property data
  - AIChE-DIPPR
- MatML: materials property data
  - industry, academia steering committee
- SpectroML: spectroscopic data
  - commercial spectral analysis software companies
  - commercial instrument mfgs.
  - American Society for Testing and Materials
- AniML: analytical laboratory data
  - integrates attributes from wide-range of XML standards
- MML: Microanalysis Markup Language
  - analytical instrument makers and vendors

Combinatorial and High-Throughput Techniques

NIST Combinatorial Methods Center (NCMC) launched in January 2002

- Lower Barriers to Widespread Adoption of Combinatorial Methods in Materials Research
- 14 members to date

COVER STORY

November 11, 2002

Volume 80, Number 45
CENEAR 80 45 pp. 58-60
ISSN 0009-2347

TAPPING INTO NIST'S COMBI EXPERTISE

Chemical companies see value in participating in the fledgling Combinatorial Methods Center

RON DAGANI, C&EN WASHINGTON

Speed and efficiency. That's what's driving everything in industry today. It's why pharmaceutical makers adopted the combinatorial or high-throughput approach, which allows them to synthesize large numbers of compounds and screen them for useful medicinal properties--all in record time. And it's why chemical and materials companies increasingly are embracing the same strategy.

High-throughput experimentation, a term often used synonymously with the combinatorial approach, enables researchers to do their work more quickly, while also broadening the range of chemical substances they can examine, explains chemist John S. Sadowski, director of corporate research services at Air Products & Chemicals. "That should enable you to get products to market more quickly and help increase your probability of success."

GRADIENT GURUS Karim (left) and Amis with the flow-coater used to make thin-film polymer libraries. PHOTO BY RON DAGANI

Rhodia · The Gillette Company · Bayer · Rohm and Haas · Akzo Nobel · Dow · Air Products · 3M · BASF

Air Force Research Laboratory · Honeywell · ICI · ExxonMobil · Procter&Gamble

- Parallel Experiments
- Automated Specimen Array Fabrication
- Automated Analysis
- Iterative Approach

Faster, Cheaper, Better Product Discovery

70% of the worlds 30 largest chemical companies have substantial investment in combinatorial programs" - Peter Cohen, CTO, Symyx Corp. June, 2002

Practical Metrology

NIST Recommended Practice Guides

MSEL

- Feedback from customers

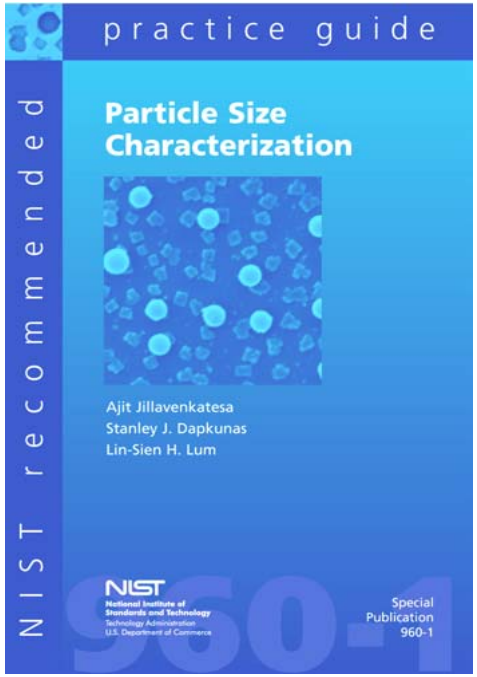
  "We will distribute to all our customers"

  – Reynolds Metal Company

  – Malvern Instruments

  "Concise, clear, and well-rounded"

  – Sympatec



- 6800 hard copies distributed
  - -equipment makers and users
  - -accreditors such as NVLAP
  - -auditors for traceability
  - -trainers/educators

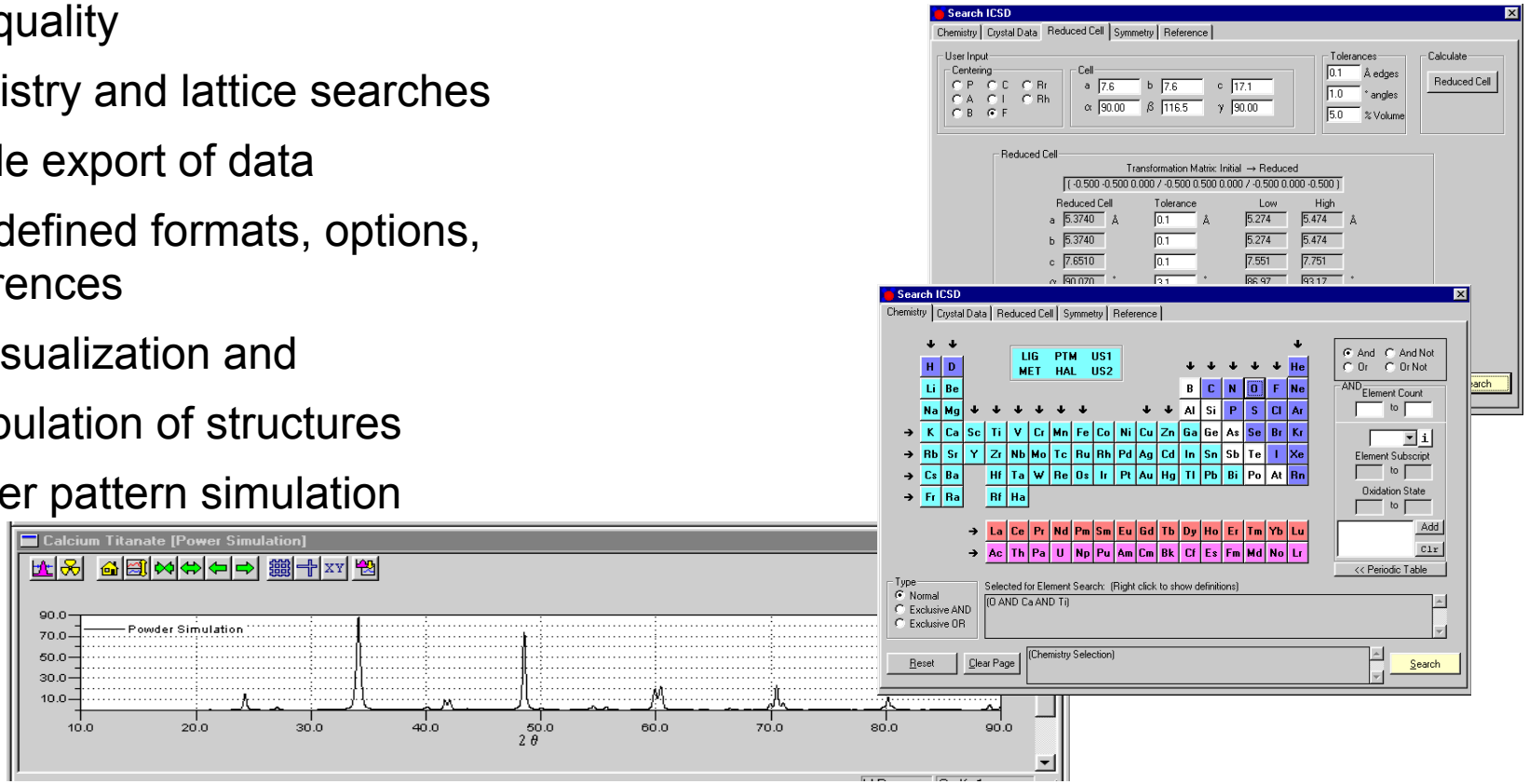
NIST

CSTL

Inorganic Crystal Structure Database

MSEL

- Windows-based PC product presented to the worldwide crystallographic community
  - –data quality
  - –chemistry and lattice searches
  - –flexible export of data
  - –user-defined formats, options, preferences
  - –3-d visualization and
  - –manipulation of structures
  - –powder pattern simulation



The critically evaluated data and modern data structures and interfaces for the ICSD represent a first step towards interoperability with other data sources and scientific software tools.

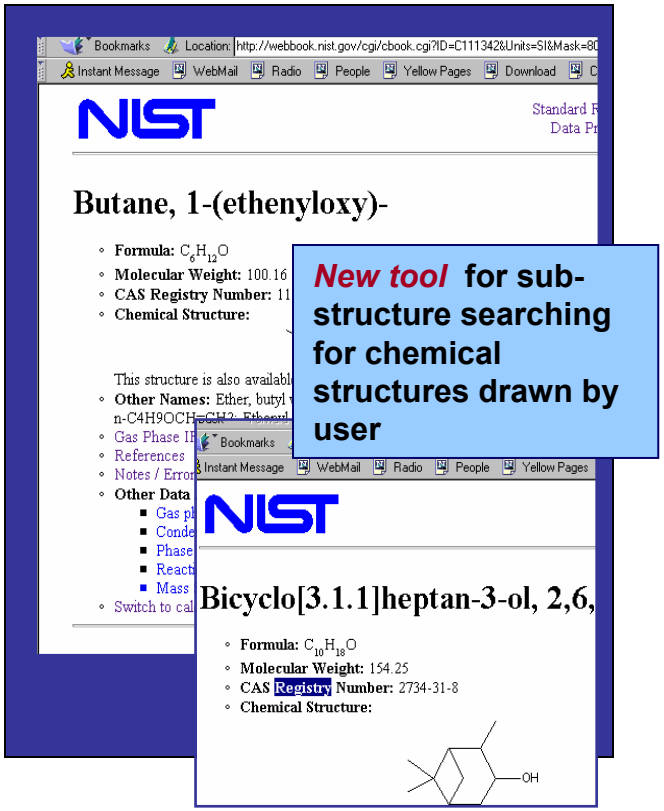
Web-Based Dissemination of Reference Data

The NIST Chemistry WebBook

http://webbook.nist.gov

- Data for ~40,000 species
  - New enthalpy of fusion database
  - New fluid property models and capabilities
- User Profile: *80,000 hits per month*

Chemical and Physical Data Resources on the Internet



WebBook voted the *"Best Chemical Site on the Web"*
sponsored by: ChemIndustry.com, Inc; John Wiley and Sons, Inc; and the Royal Society of Chem.

Challenges in the Creation of Knowledge

Knowledge Systems	Data Mining	Modeling and Presentation	Intelligent Systems
-automated knowledge/data classification and cataloging -seamless integration of geographically distributed metadata sets -advanced system learning/ automated reasoning systems -virtual measurements from models and data -data evaluation and representation -formal logic-based knowledge models and ontologies -integrated data measurement uncertainties	-automated knowledge/data integration, classification, and cataloging – from other sources -intelligent integration of multiple heterogeneous data sources -automated spatial data and image -autonomous information-gathering agents -indexing and pattern recognition -data/info gap recognition and identification	-automated model certification: validation and verification -automatic data culling, filtering, conversion, and synthesis -automated spatial and image analysis -fusing of theoretical and empirical models -new methods for modeling and presentation -new visual interpretations of non-spatial data -new human- computer interfaces -symbolic analysis	-automated system knowledge capture, classification, and cataloging -automated system certification: validation and verification -integrated artificial intelligence and learning systems -automatic intelligent systems/sensor integration and data management -condition-based/ anticipatory system prognostics and maintenance