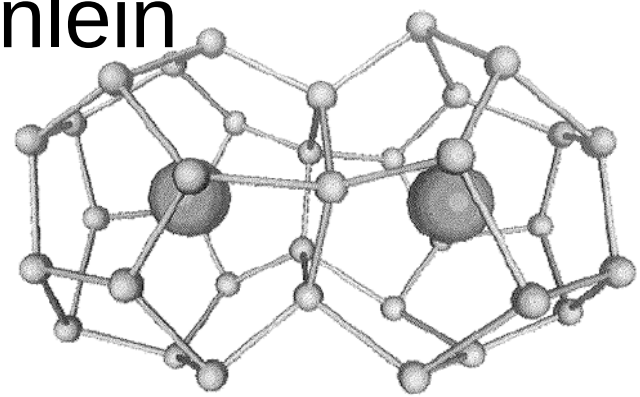
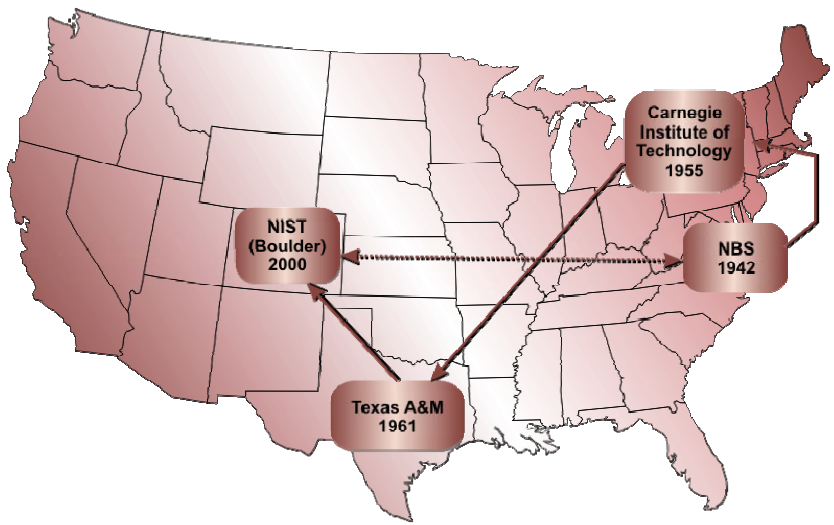
Gas Hydrate Database and Web Dissemination Channel

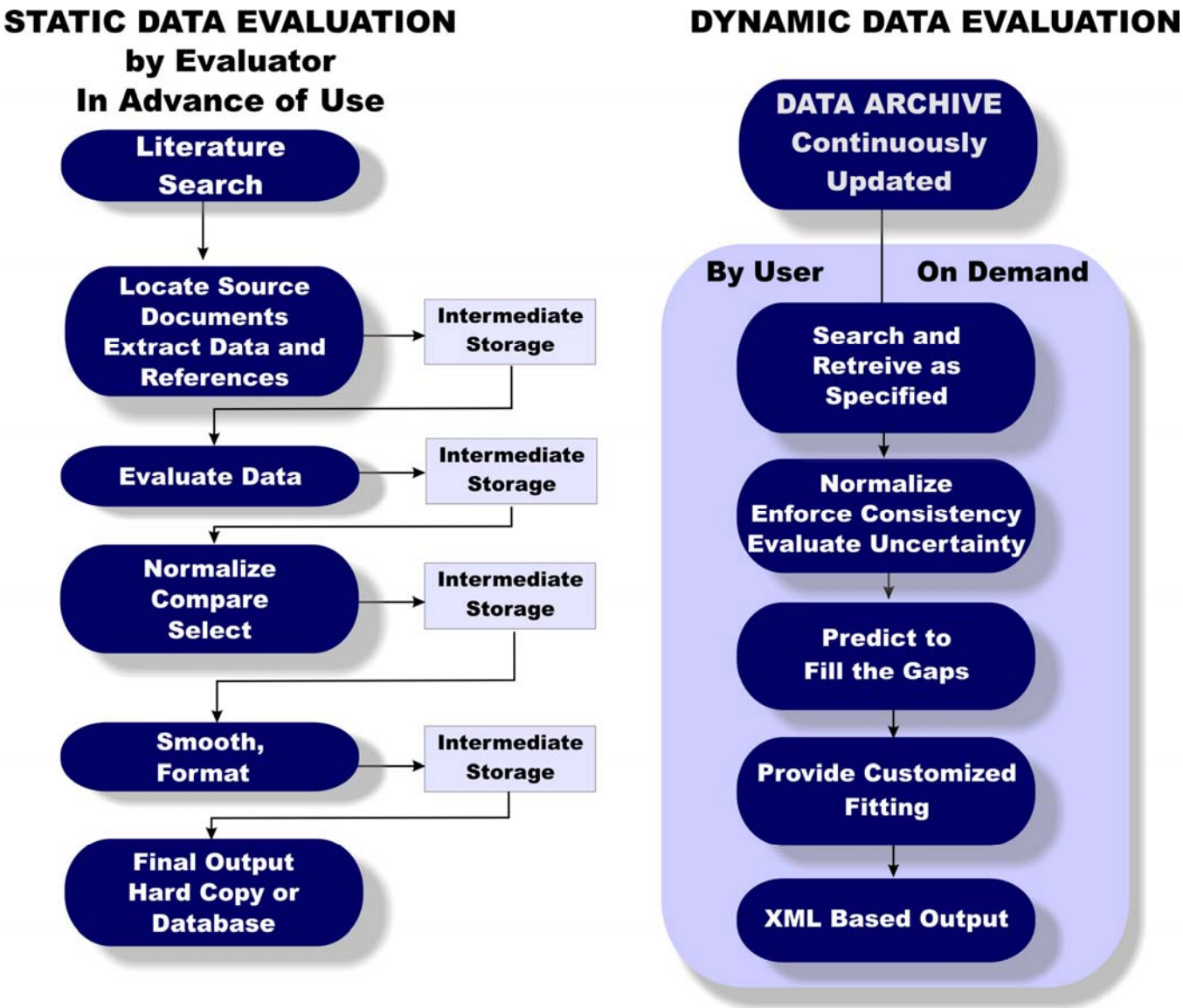
Ken Kroenlein



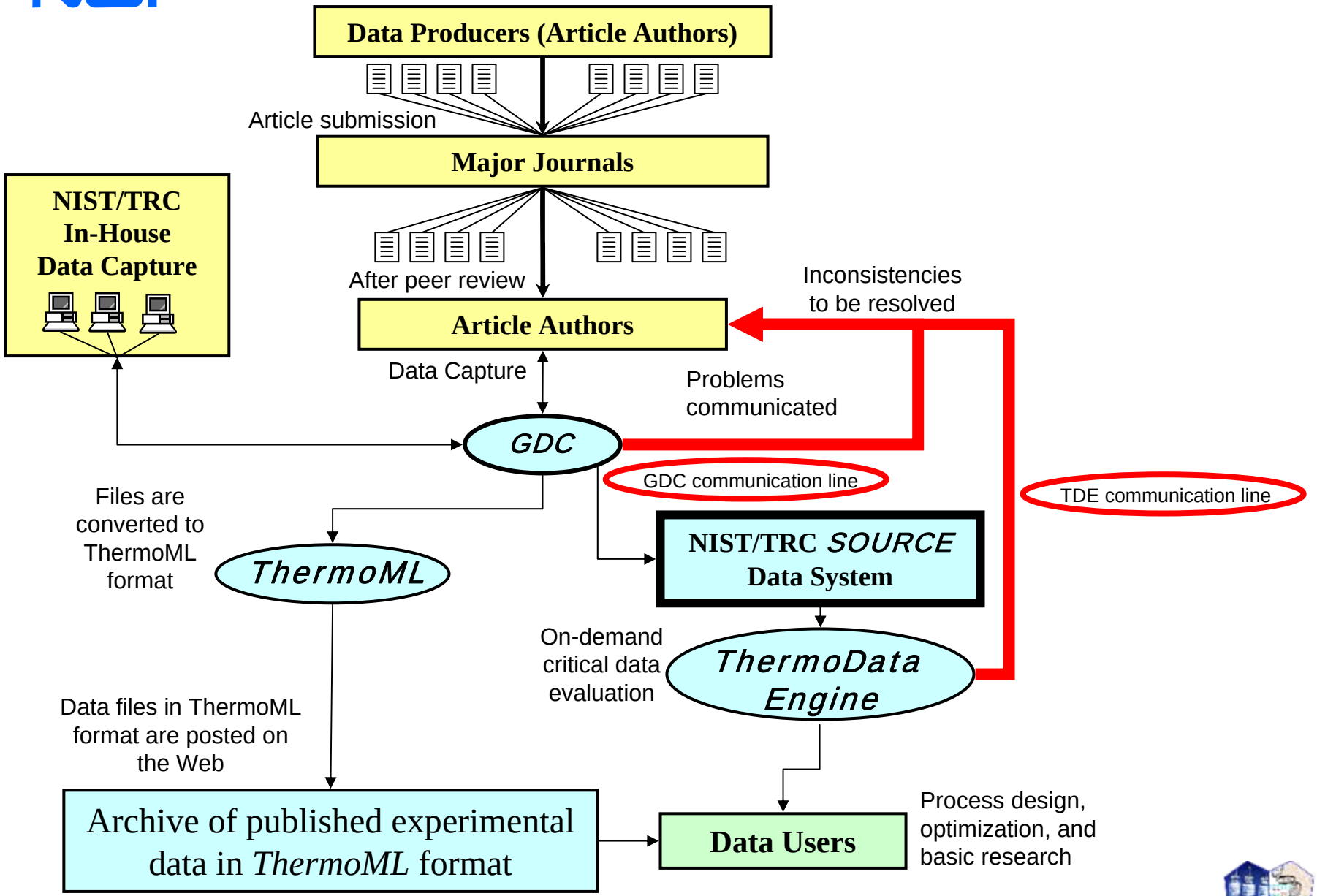
History of TRC

- Founded in 1942 to collect and disseminate thermodynamic and thermophysical property data on selected hydrocarbons and sulfur-containing derivatives
- Mission expanded in 1955 to include all organic compounds
- In 1986, the TRC SOURCE electronic database was designed and implemented
- In 2001, the TRC Data Entry Facility was established, with the goal of critically-evaluating 500,000 data points per year





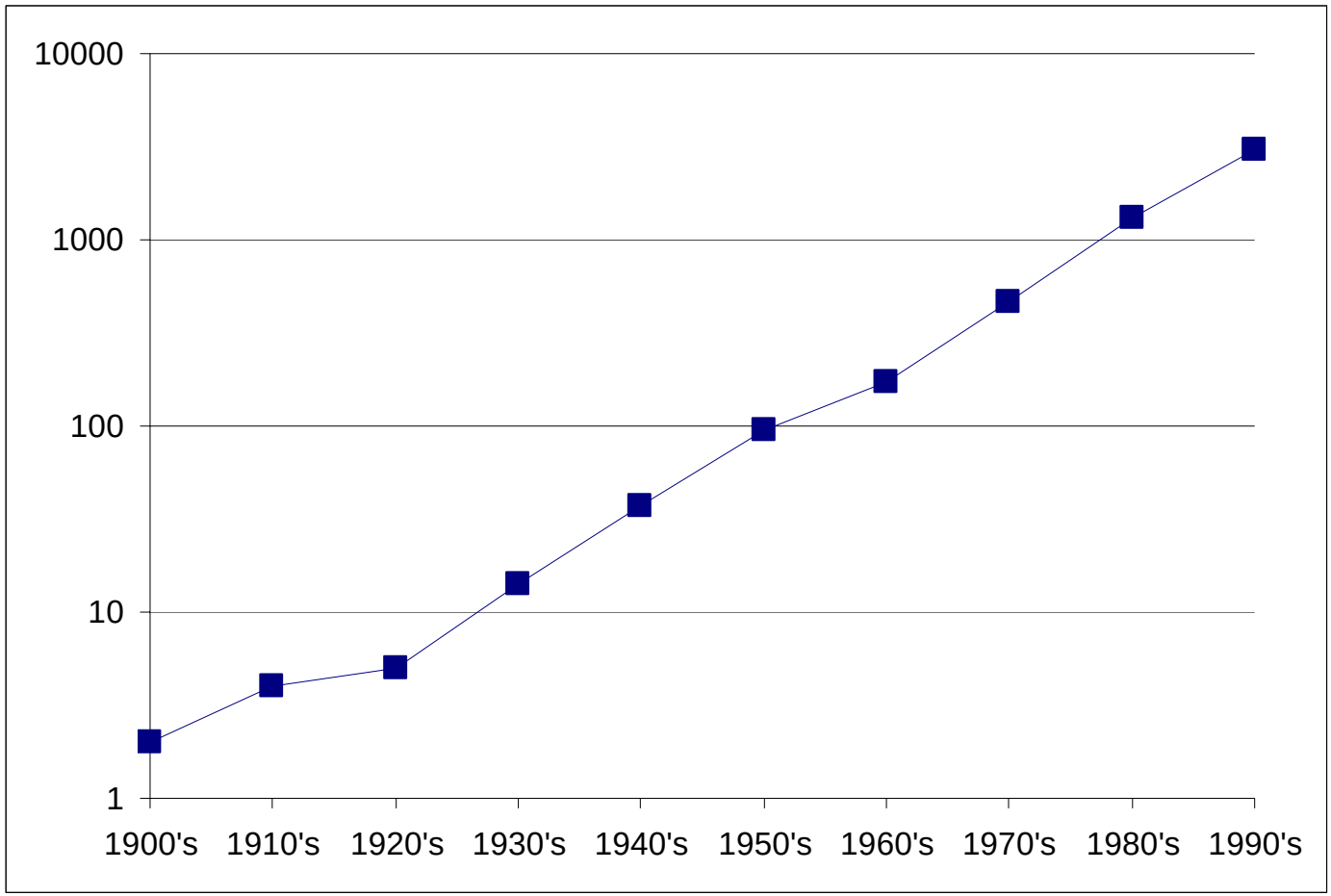
M. Frenkel (2005), *Pure Appl. Chem. 77,* p.1349


Data Producers (Article Authors)
Article submission
Major Journals
NIST/TRC
In-House
Data Capture
After peer review
Article Authors
Inconsistencies
to be resolved
Data Capture
Problems
communicated
GDC
GDC communication line
TDE communication line
Files are
converted to
ThermoML
format
ThermoML
NIST/TRC SOURCE
Data System
On-demand
critical data
evaluation
ThermoData
Engine
Data files in ThermoML
format are posted on
the Web
Archive of published experimental
data in ThermoML format
Data Users
Process design,
optimization, and
basic research

Motivation for Investigation

- Expanding TRC's mission to provide comprehensive, critically-evaluated sources of thermophysical property data for compounds and well-characterized mixtures
- Facilitate efforts aimed at developing technologies to recover naturally-occurring enclathrated natural gas deposits
- Faciliate efforts aimed at developing technologies utilizing clathrate hydrates in commercial technologies such as gas storage, cool storage or desalinization

Number of Publications



Sloan, 2004

Objectives

- Obtain a comprehensive collection of data from the open literature
- Store this data locally in an easily accessible and searchable format
- Develop technology to communicate this data reliably in an automatically-interpretable fashion
- Make this data available to the research community on open demand with free access and without any restrictions

Timeline

- Year 1
  - Establish Gas Hydrates Advisory Group
  - Establish the target database contents
- Year 2
  - Complete windows-based guided data capture (GDC) software windows for gas-hydrate data
  - Complete initial document collection (literature, reports, etc.)
  - Complete modifications for NIST SOURCE for gas-hydrate data accommodation
  - Process examples of all target gas-hydrate data types established in year 1 using the developed GDC software
  - Complete GHML with elements of ThermoML included on a modular basis
  - Complete building of the Web-Oracle infrastructure
- Year 3
  - Complete development of Web-Oracle interface
  - Complete Large scale data collection of gas-hydrate data
  - Development of data submission protocols
  - Extensive internal and external testing of web interface

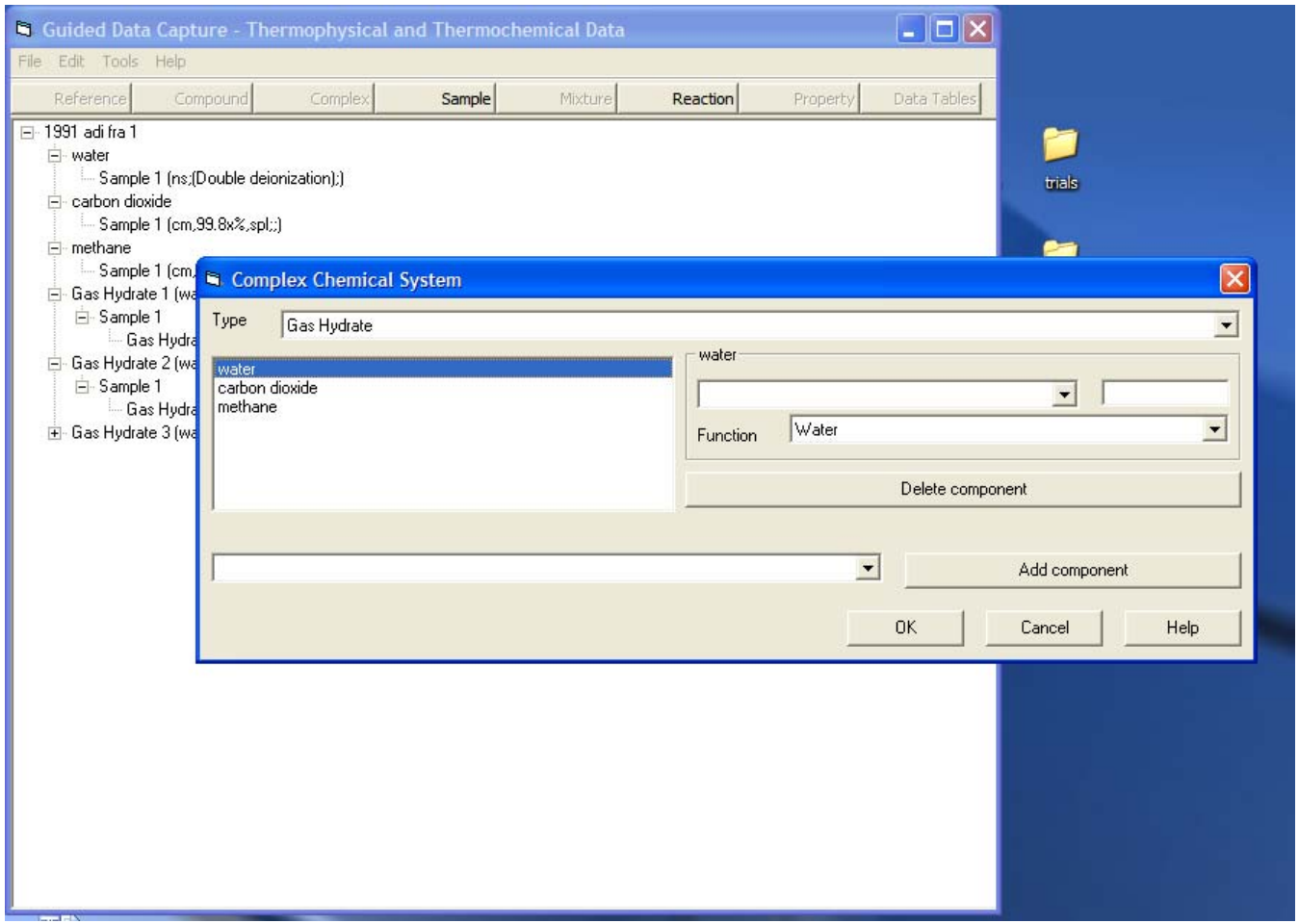
Elements of Infrastructure

- Facility to obtain and process literature documents
- Software to guide digitization of data capture efforts
- Well-constrained and well-documented intermediate storage and communication data format
- World Wide Web based technology for search and retrieval of data files with open access

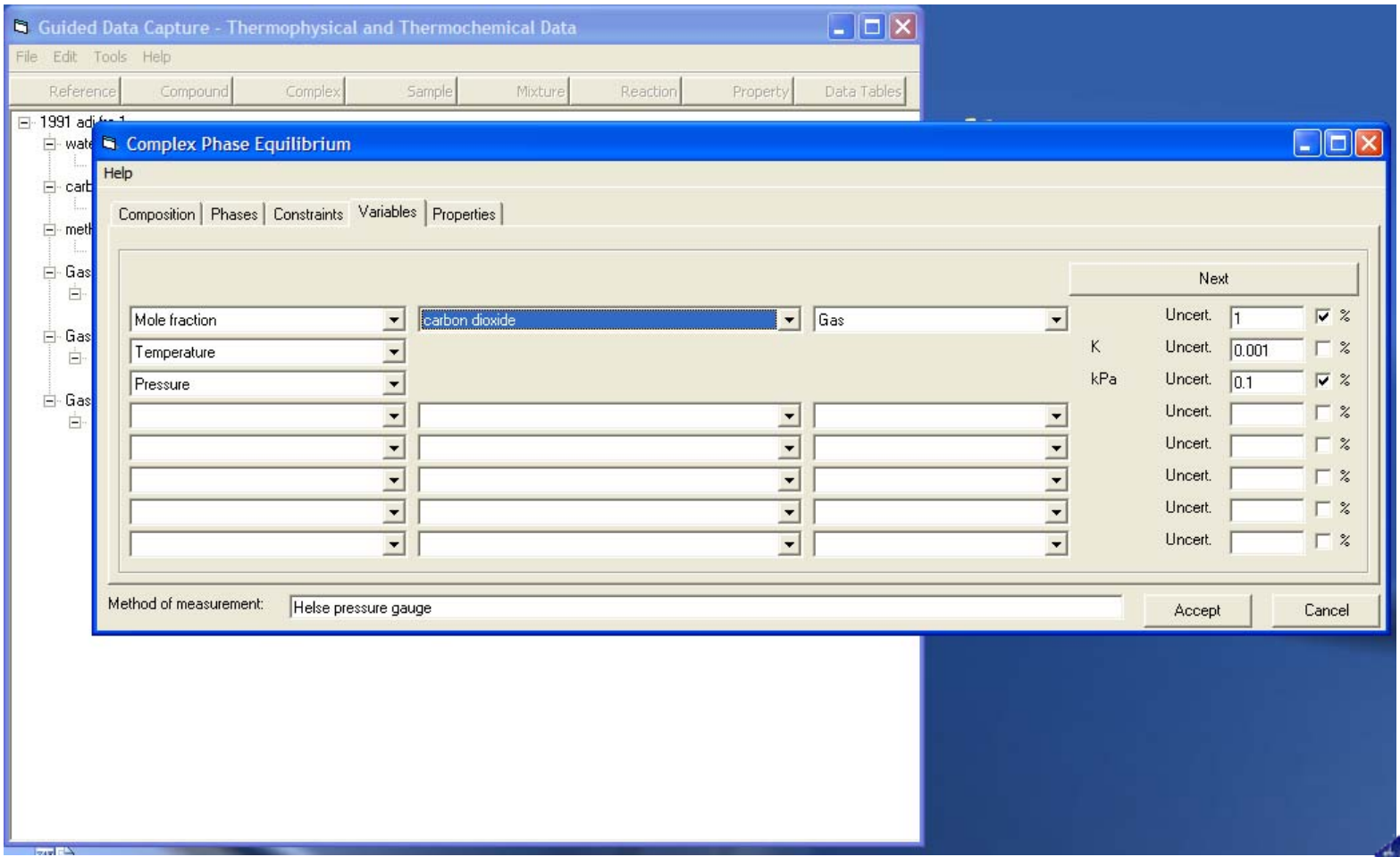
Status: Data Entry Capability

- Following the design approaches utilized in creation of the TRC Data Entry Facility, a separate Gas Hydrate Data Entry Facility has been created
- The facility is staffed by four student operators
- Each student has a personal work station with all necessary software and network connectivity
- Specific protocols have been developed, using Portable Document Format (pdfs) files as the primary system key and ensuring that each article passes through the system once
- Existing Guided Data Capture (GDC) software has had extensions added to facilitate data processing on complex phase equilibrium, bulk thermophysical properties and crystalline structural information

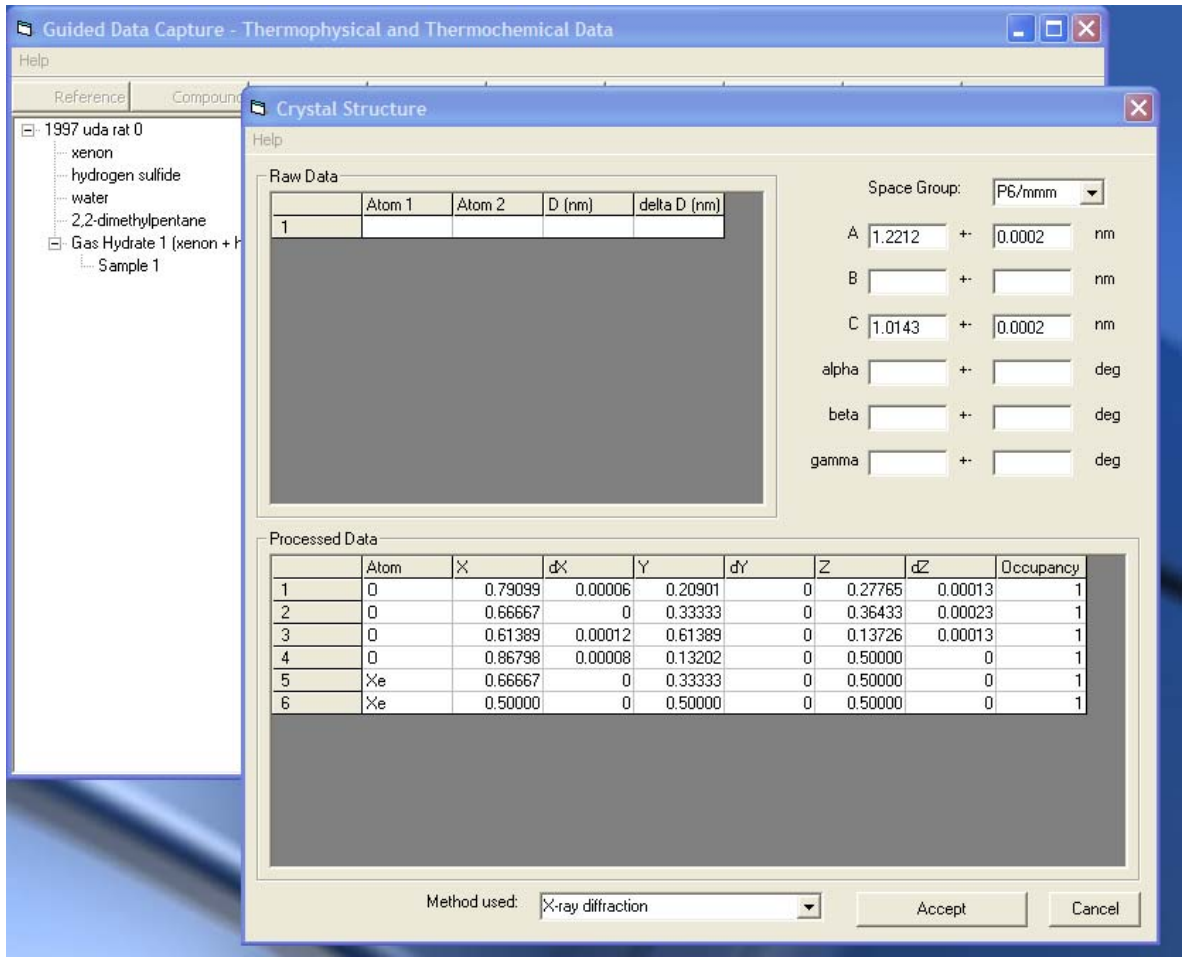
Gas Hydrate Definition



Phase Boundary Definition



Crystallographic Data



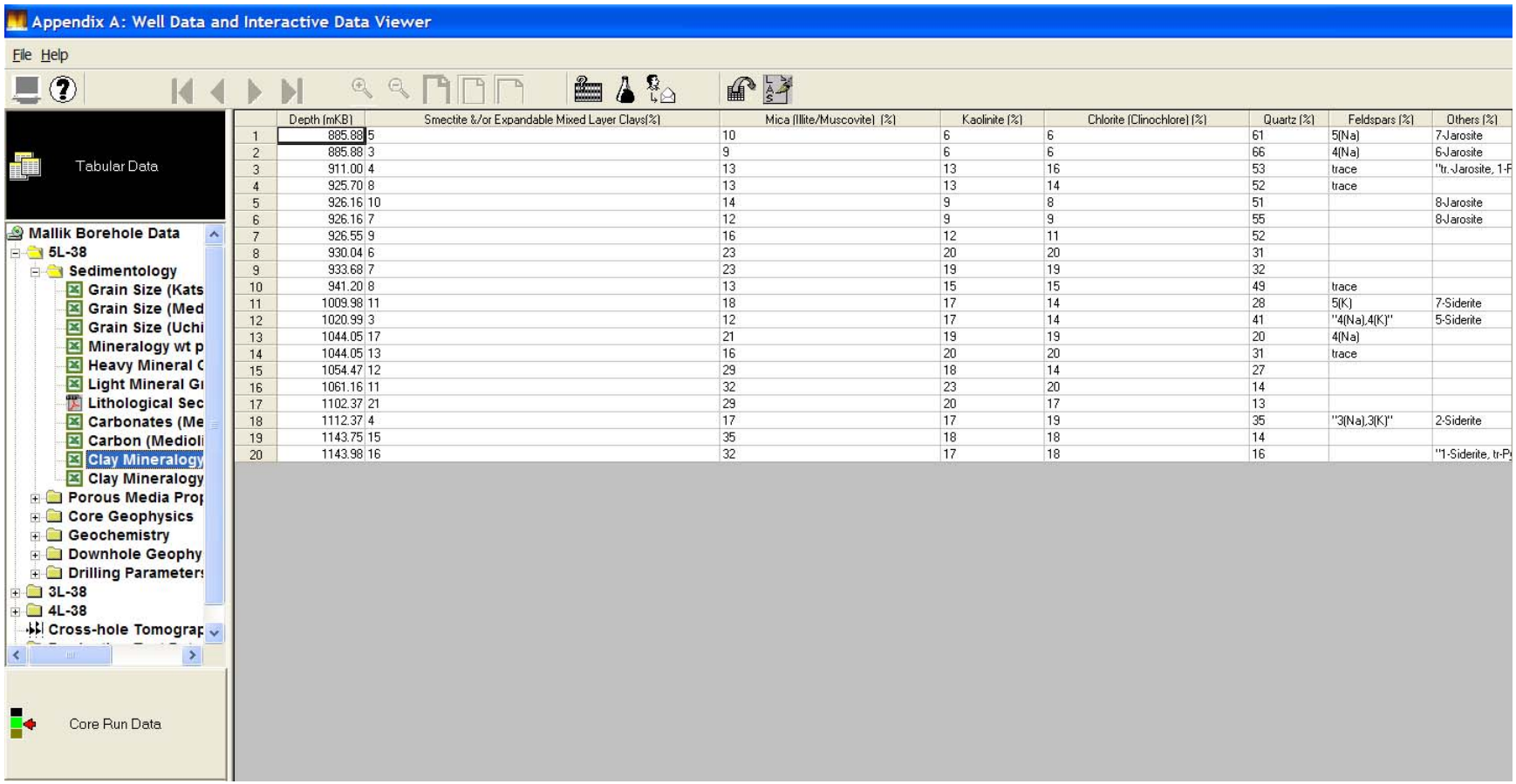
Status: Data Entry Progress

- A total document archive of more than 4,500 documents has been collected to date
- Data has been captured from more than 300 documents, including several dissertations and technical reports
  - This includes nearly all data sets cited in the well-respected "Clathrate Hydrates of Natural Gas, 3rd Ed." by Sloan and Koh, 2008
- More than 3,300 documents have been critically evaluated and found to be lacking in appropriate data
- Efforts have begun to use article citations to increase the comprehensiveness of our document archive

Status: GHML

- Due to the broad range of “data” associated with the gas hydrate field, as determined by consulted experts, it was necessary and appropriate to develop a new XML schema for data communication
- Care was taken to provide support for communicating thermophysical data in a fashion that is automatically interpretable while still being able to communicate datasets of highly variable type and content
- Development included multinational collaboration with GFZ (Germany) and CNIC (China)
- To avoid issues with copyright of the American Chemical Society, the universal primary identifier for chemical compounds was selected to be the IUPAC standard InChI.

Review of Mallik data

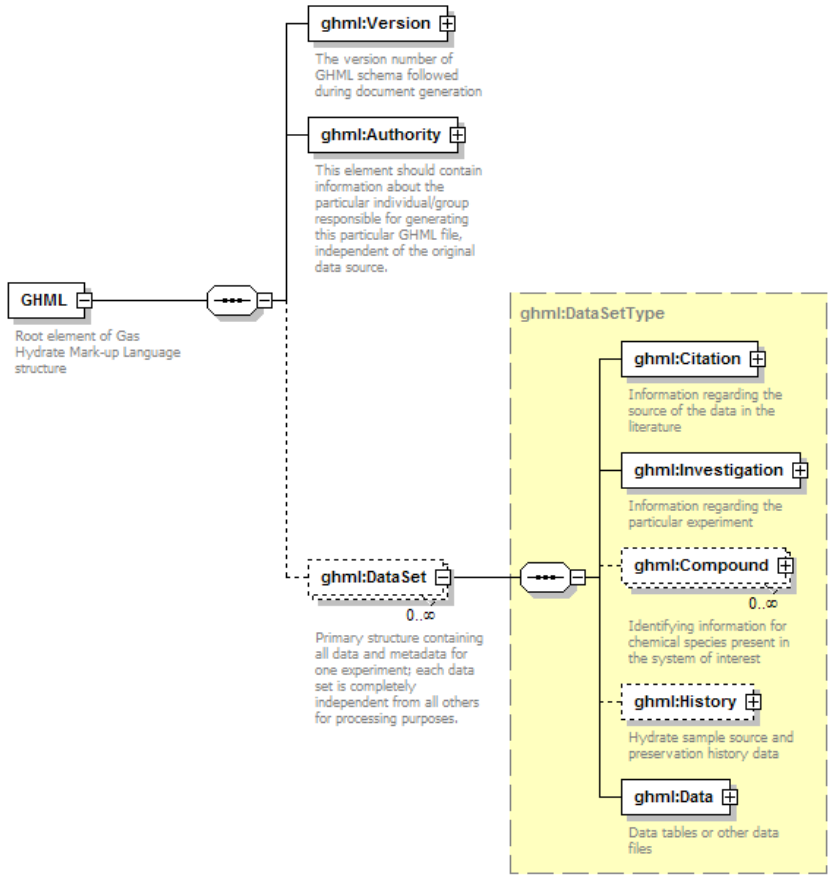


	Depth (mKB)	Smectite &/or Expandable Mixed Layer Clays(%)	Mica (Illite/Muscovite) (%)	Kaolinite (%)	Chlorite (Clinochlore) (%)	Quartz (%)	Feldspars (%)	Others (%)
1	885.88	5	10	6	6	61	5(Na)	7-Jarosite
2	885.88	3	9	6	6	66	4(Na)	6-Jarosite
3	911.00	4	13	13	16	53	trace	"tr-Jarosite, 1-F
4	925.70	8	13	13	14	52	trace	
5	926.16	10	14	9	8	51		8-Jarosite
6	926.16	7	12	9	9	55		8-Jarosite
7	926.55	9	16	12	11	52		
8	930.04	6	23	20	20	31		
9	933.68	7	23	19	19	32		
10	941.20	8	13	15	15	49	trace	
11	1009.98	11	18	17	14	28	5(K)	7-Siderite
12	1020.99	3	12	17	14	41	"4(Na),4(K)"	5-Siderite
13	1044.05	17	21	19	19	20	4(Na)	
14	1044.05	13	16	20	20	31	trace	
15	1054.47	12	29	18	14	27		
16	1061.16	11	32	23	20	14		
17	1102.37	21	29	20	17	13		
18	1112.37	4	17	17	19	35	"3(Na),3(K)"	2-Siderite
19	1143.75	15	35	18	18	14		
20	1143.98	16	32	17	18	16		"1-Siderite, tr-Py

Clay Mineralogy from Medioli et al, 2005

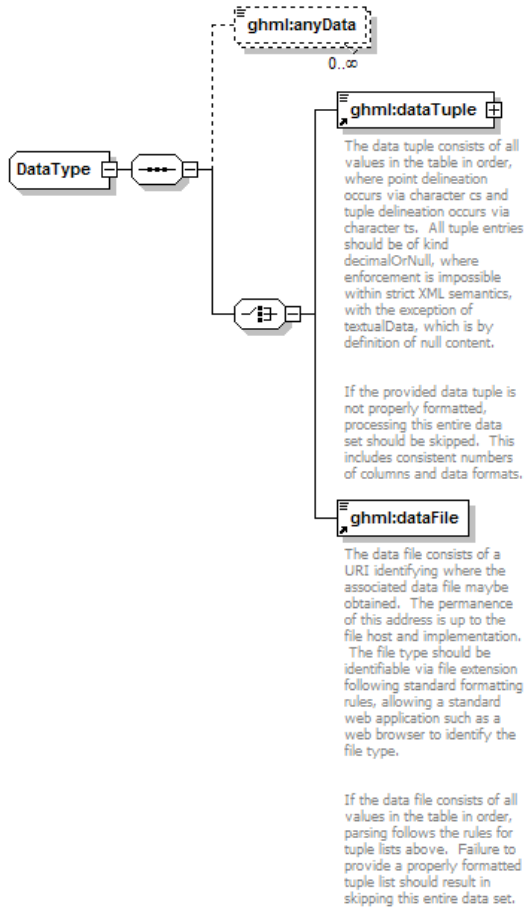
GHML Structure: Metadata

- Root elements represent traditional XML hierarchical storage for associated metadata



GHML Structure: Metadata

- Root elements represent traditional XML hierarchical storage for associated metadata
- Data elements themselves are communicated as a “data tuple”, or delimited ordered list

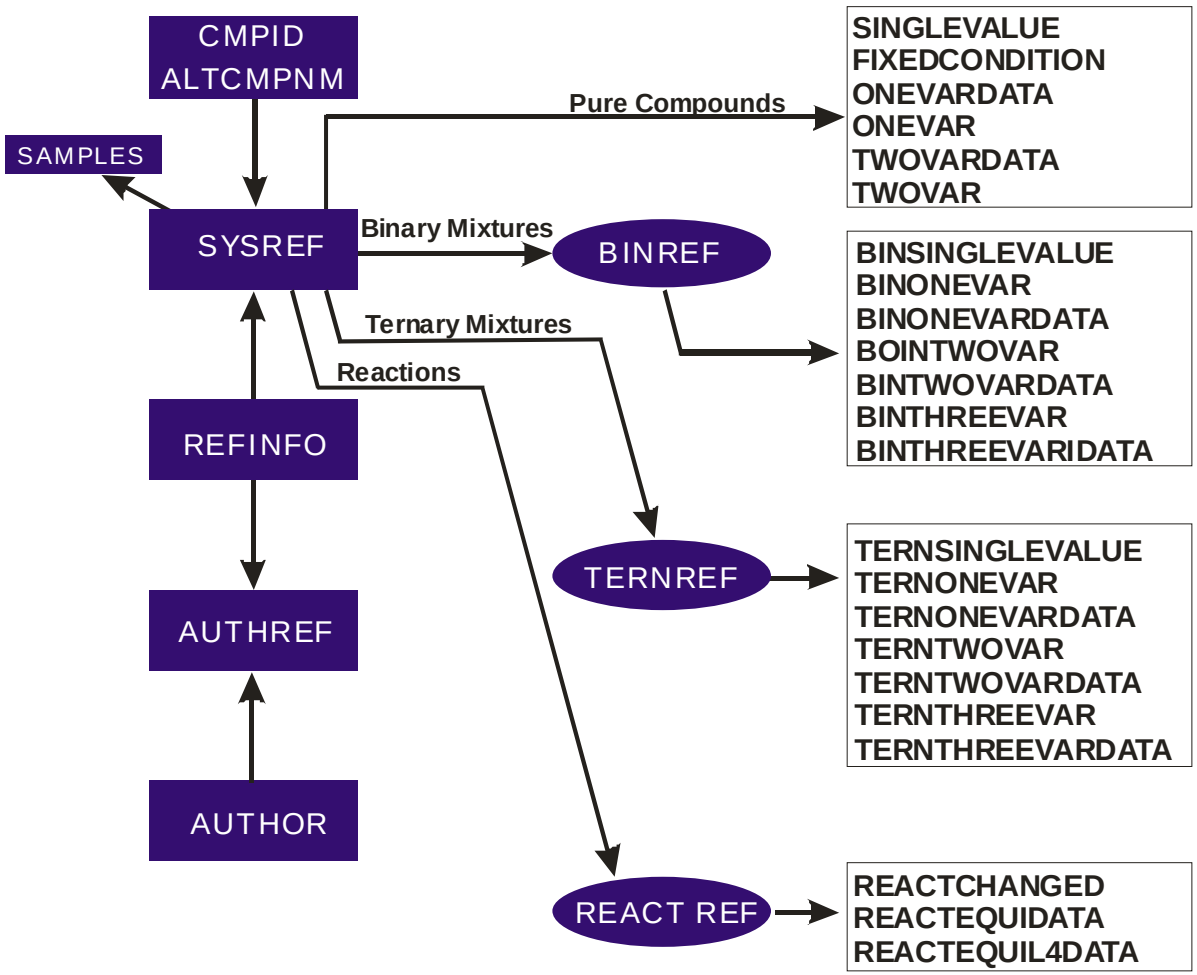


GHML Structure: Data

- We created enumerated lists of what are essentially a series of column headings for a given table
- We identified the following broad data categories:
  - Chemical
  - Crystallographic
  - Deep Biospheric
  - Mechanical
  - Petrological
  - Thermophysical
  - Spatial/Temporal
  - Uncertainty
  - Other/Text
- In this way, we have created a relational XML structure for well constrained communication of a very broad range of data types.

```
<element name="chemicalData">
  <annotation>
    <documentation>Broad Categories describing chemical properties of a sample.  See
  </annotation>
  <complexType>
    <simpleContent>
      <extension base="ghml:chemicalDataType">
        <attribute name="columnNumber" type="integer" use="required"/>
        <attribute name="descriptor" type="string" use="optional"/>
        <attribute name="sampleID" type="integer" use="optional"/>
        <attribute name="phaseID" type="ghml:phaseEnumeration" use="required"/>
      </extension>
    </simpleContent>
  </complexType>
</element>
<simpleType name="chemicalDataType">
  <restriction base="string">
    <enumeration value="Mole Fraction"/>
    <enumeration value="Mass Fraction"/>
    <enumeration value="Volume Fraction"/>
    <enumeration value="Mass Concentration (kg/m^3)"/>
    <enumeration value="Molar Concentration (mol/m^3)"/>
    <enumeration value="Salinity (ppt)"/>
    <enumeration value="Salinity Class"/>
  </restriction>
</simpleType>
```

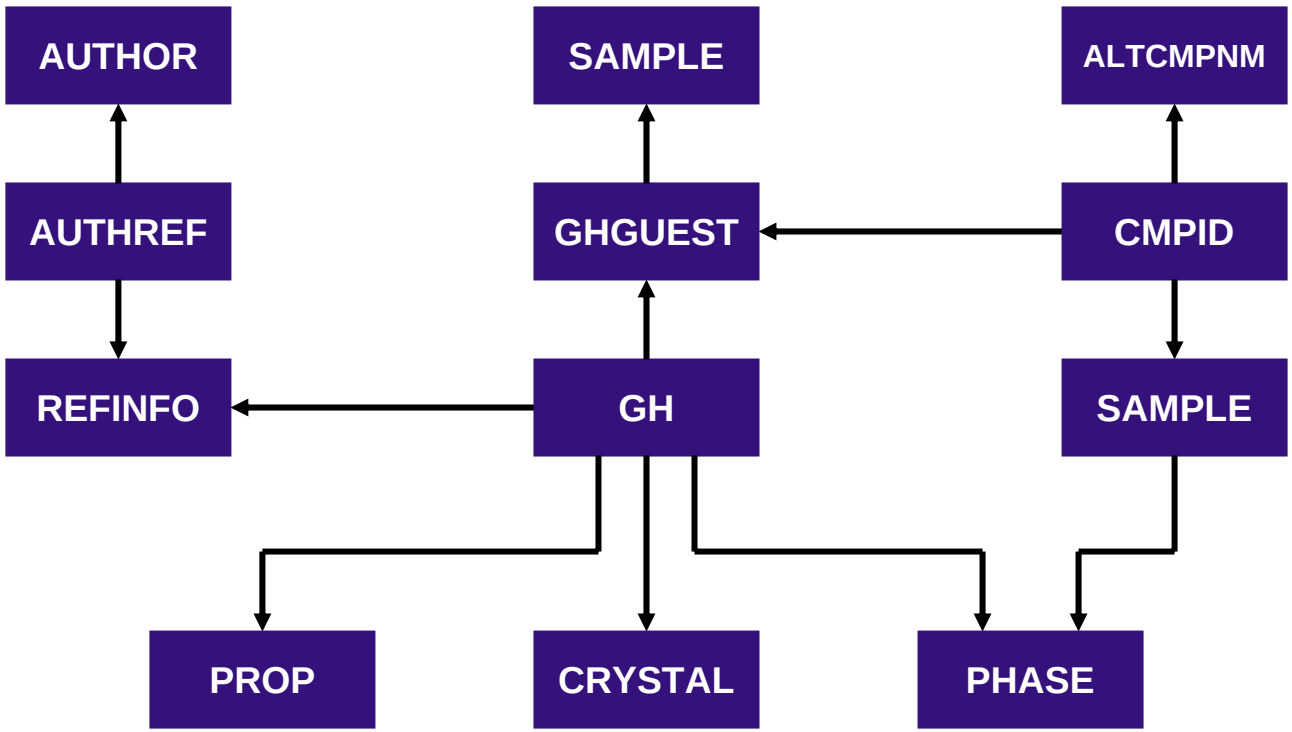

SOURCE Data Tables



- 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.*)

Frenkel *et al.*
Int. J. Thermophys.
2001, *22*, 215

Database Schematic



Summary of Progress

- The Gas Hydrates Data Entry Facility is operational
  - Modifications to Guided Data Capture (GDC) software for treatment of gas hydrate data sets have been successfully implemented
  - Undergraduate staff is trained and processing literature materials
- The desire to exchange both easily constrained, simple datasets and complex datasets with highly variable content has required creative design of the Gas Hydrate Markup Language (GHML)
  - Design of GHML ensures proper transmission of core sample and well log data sets in addition to thermophysical data sets
- Relevant materials were presented in a poster at the 6th International Conference on Gas Hydrates in Vancouver

Remaining Tasks

- Continue to obtain new documents and process documents for additional data sets
- Implement database design in practical system and populate it with captured data sets.
  - Develop software to import data from captured files into a database
- Develop appropriate web services for data collection and dissemination
  - This task is being performed in collaboration with the existing CODATA task force on Gas Hydrates