Expressing Measurements and Chemical Systems for Physical Property Data

Peter Linstrom

National Institute of Standards and Technology

Gaithersburg, MD, USA

Outline

- Nature of physical property data
- Historical record or interpretation
- Limitations of automated systems
- Problem areas
- Summary

Physical Property Data

- A numeric tuple which applies to a physical system
- Describing how the numeric value was obtained from the system is difficult
  - Identification of techniques, equipment, ancillary data used in calculations and calibrations.

Physical Property Data

- Describing the system is difficult
  - Identification of sample: chemical species and concentrations

- Recording the numeric tuple is easy

Historical Record or Interpretation

Two goals (non-exclusive) goals:

1.) Historical record
- What was measured, computed, or estimated?
- How was this done?

2.) Basic knowledge
- What do we know about this property?
- What is the probable range of the numeric value?

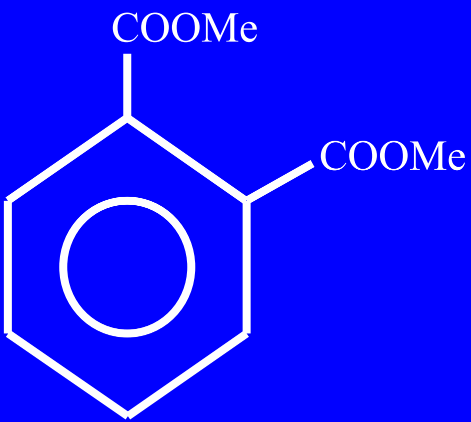
Historical Record or Interpretation

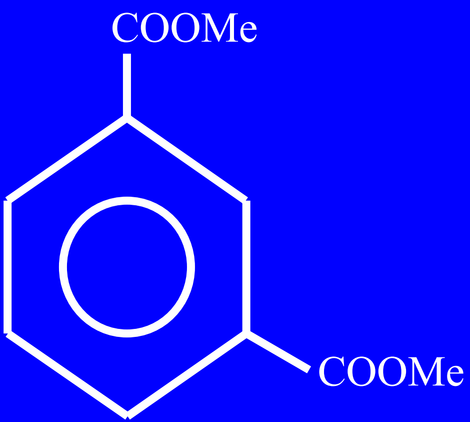
- Historical record
  - Does not change
  - Applies to specific physical systems and measurements

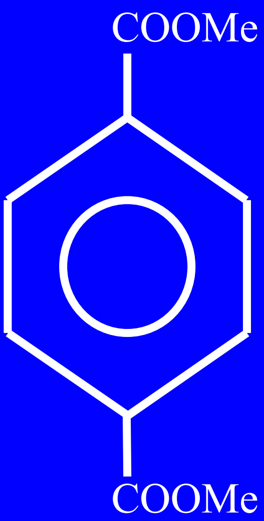
- Basic knowledge
  - Built on analysis of the historical record
  - Applies to an "idealized" physical system
  - Improves through scientific processes

Example

- 1998 Roux et al, *Fraday Trans.*
- Stability of dimethyl benzenedicarboxolates







Example

$$2\ C_6H_5CO_2Me \rightarrow C_6H_4(CO_2Me) + C_6H_6$$

Endothermicity of gas phase reaction:

ortho	52.3 kJ/mol
meta	29.2 kJ/mol
para	30.4 kJ/mol

Quite different from dinitro and dicyano benzenes

Example

- 2002 Roux et al, *Phys. Chem. Chem. Phys.*
- $\Delta_f H^o$ methyl benzoate gas (kJ/mol):

1971 Hall et al	-299.8
1980 Guthrie et al	-269.3 ± 5.1
1994 Pedley	-287.9 ± 2.4
1998 Maksimuk et al	-277.74 ± 1.2
2002 Roux et al	-276.1 ± 4.0

Example

$$2\ C_6H_5CO_2Me \rightarrow C_6H_4(CO_2Me) + C_6H_6$$

Revised endothermicity of gas phase reaction:

ortho	28.7 kJ/mol
meta	5.6 kJ/mol
para	6.8 kJ/mol

Similar to dinitro and dicyano benzenes

Automated Systems

- Automated systems require data with well defined semantics
- Portions of physical property data are recorded in natural language (literature)

- Need procedures to map information to a form appropriate for automated systems

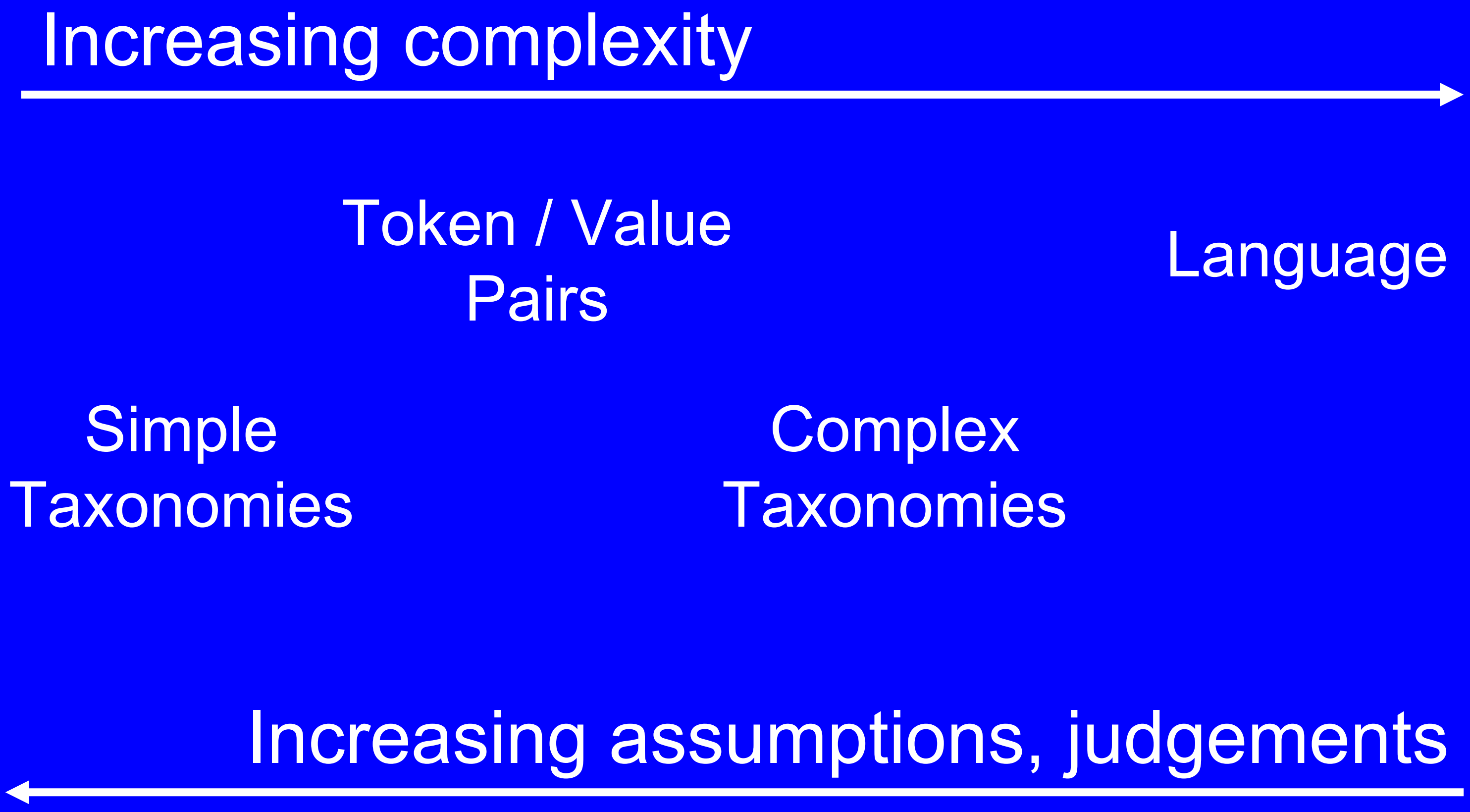
Automated Systems

- Mapping of information to computer friendly semantics may involve
  - Loss of information
  - A judgement on the part of the archivist (introduction of information not explicitly contained in the original source)
  - Blurring of the line between historical record and interpretation

Automated Systems

- Some options for expressing information
  - Develop taxonomy of codes
  - Token value pairs
  - Incorporate into database design
  - Text comments (loss of data processing capability)
  - Ignore the information

Automated Systems



Increasing complexity

Token / Value Pairs

Language

Simple Taxonomies

Complex Taxonomies

Increasing assumptions, judgements

Automated Systems

- Proper design of systems for expressing data requires significant domain knowledge
  - Definition of appropriate taxonomies, codes, etc.
  - Knowledge of what will be important to future investigators
  - Knowledge of what can be safely ignored

Some Problem Areas

- Chemical identification
- Taxonomies for methods
- Describing domain-specific meta-data

Chemical Identification

- Identification of pure species can be difficult
- Identification of mixtures is a superset of the problem for single species
- Chemical nomenclature is too complex for most data systems to handle

Chemical Identification

- Registry of species
  - Simplifies identification to an integer number
  - Maintained by third parties
  - Species may not be in registry
  - Identification may not be precise (isomers)
  - Deprecated entries
  - Users consult secondary sources – errors propagate

Chemical Identification

- Chemical structure
  - No third party
  - Less ambiguity, but more complex semantics
  - Expensive to draw or look up
  - Costs decreasing with modern technology

Chemical Identification

- Purity / uncertainty of composition
  - May not be known
  - Purification / synthesis technique may be provided
  - Often omitted from database

Taxonomies for Methods

- Classification of the manner in which a value was obtained
- Instrument type, model form natural divisions
  - Appropriate resolution determined by archivist
- How does one handle unique methods?
  - Science is not static – taxonomies will grow

Taxonomies for Methods

- Lias, et al, Ionization Potential Database
  - Compiled over many years
  - Taxonomy for basic measurement types
  - Additional codes added to supplement supplement taxonomy for new methods which cross existing hierarchical boundaries (e.g. electron impact and laser spectroscopy)

Domain Specific Meta-Data

- Meta-data recognized by archivist (domain specialist) as significant
- Need method to encode in computer friendly format
  - Taxonomies
  - Token value pairs

Domain Specific Meta-Data

- Affefy, Liebman, and Stein – Neutral Thermochemistry Archive
  - Meta-data options expanded as archive grew
  - Correction to current CODATA heats of formation: done, not-done, or not-possible
  - Data disagrees with previously published data: acknowledged by author(s), or not acknowledged

Summary

- Two pairs of trade-offs
  - Historical record vs. interpretation
  - Semantic complexity vs. loss of information

- Important for archivists and researchers to be aware of the compromises that are made