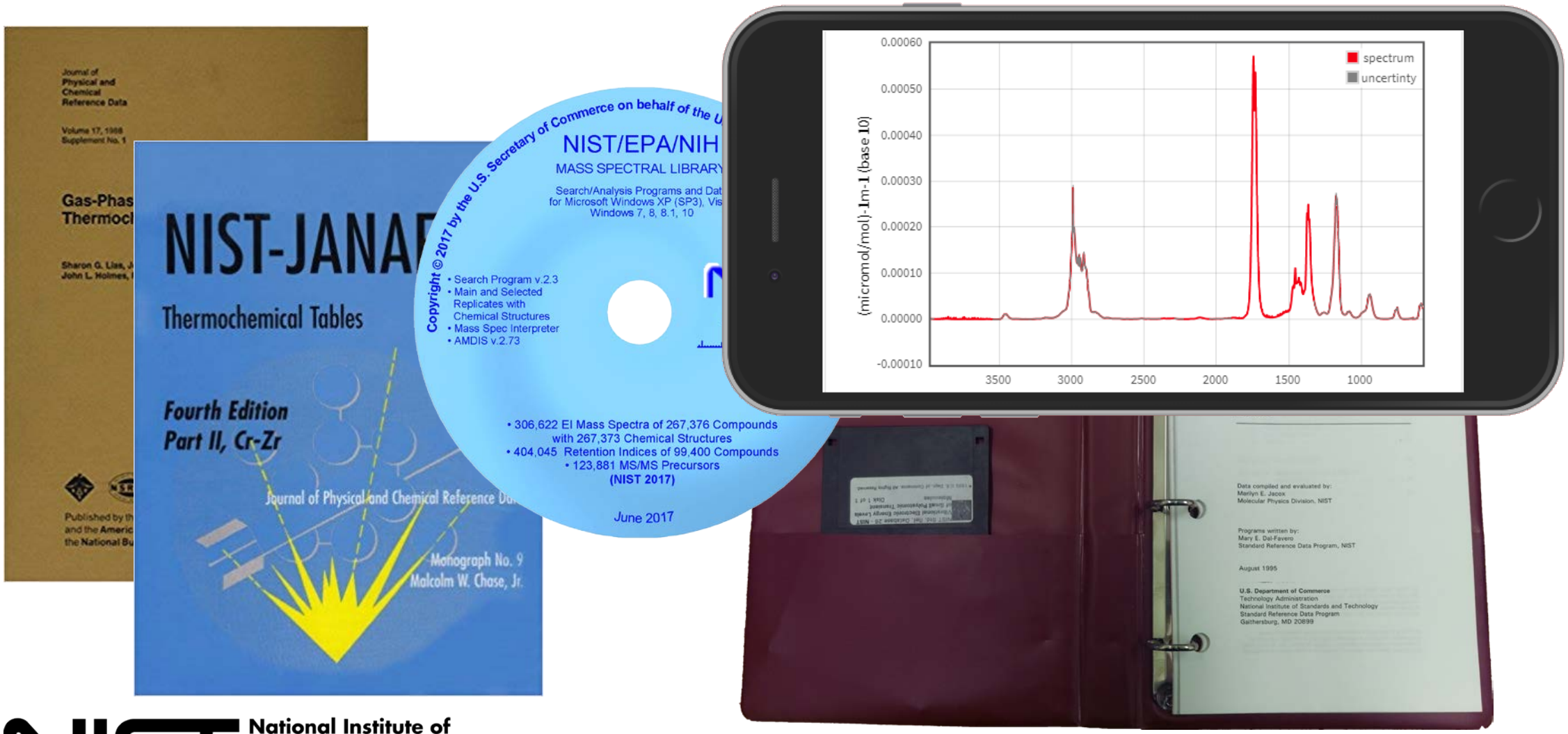
InChI and Data Management

Peter J. Linstrom

NIST, Office of Data and Informatics

NIST – History of ~~publishing~~ managing data



National Institute of Standards and Technology
U.S. Department of Commerce

InChI and data management

- Data applies to *something*.
  - Is there any value in data which corresponds to an undefined object?
  - Types of errors
    - Numerical
    - Units of measure
    - **Wrong object! ← Bad, really bad!**
- When *something* is an atom or molecule.
  - Good news: people have been working on describing these systems for some time.
  - Bad news: there is a reason people have been working on this problem for some time

InChI is a tool!

- Well known useful properties
  - Canonical representation of molecule
  - Modular structure enables discovery of related species
  - Search engine friendly hash
- Limitations
  - Labor: drawing or finding structures
  - Can’t deal with poorly defined systems
  - Chemists are human

- A craftsman chooses and uses tools carefully!

NIST Chemistry WebBook

- A collection of data compilations
  - Many developed by small groups of scientists
  - Historical data – not designed to integrate with other databases
- Challenges
  - Combining data from databases with vastly different designs
  - Helping users find data (SEO)

NIST National Institute of Standards and Technology U.S. Department of Commerce — NIST Chemistry WebBook, SRD 69

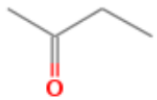
Search ▾ | NIST Data ▾ | About ▾

2-Butanone

- **Formula:** C_4H_8O
- **Molecular weight:** 72.1057
- **IUPAC Standard InChI:**
  - InChI=1S/C4H8O/c1-3-4(2)5/h3H2,1-2H3
  - Download the identifier in a file.

InChI TRUST CERTIFIED 2011

- **IUPAC Standard InChIKey:** ZWEHNKRNPOVVGH-UHFFFAOYSA-N
- **CAS Registry Number:** 78-93-3
- **Chemical structure:**



This structure is also available as a 2d Mol file or as a computed 3d SD file
The 3d structure may be viewed using Java or Javascript.

- **Other names:** Butan-2-one; Butanone; Ethyl methyl ketone; Ketone, methyl ethyl; Methyl ethyl ketone; MEK; C2H5COCH3; Acetone, methyl-; Aethylmethylketon; 3-Butanone; Butanone 2; Ethyl methyl cetone; Ethylmethylketon; Ketone, ethyl methyl; Meetco; Methyl acetone; Metiletilchetone; Metyloetyloketon; Rcra waste number U159; UN 1193; 2-Oxobutane; 2-Butanal; 2-butanone (MEK; methyl ethyl ketone); 2-butanone (MEK)
- **Permanent link** for this species. Use this link for bookmarking this species for future reference.

InChI and the NIST Chemistry WebBook

- Database development
  - Draw structures, find matches
- Identify species across data collections
  - Are these species the same?
- Create invariant URLs
  - Inbound links
- Identify isotopologues
- Help users find data
  - Internet search engines

Database development

Skilled chemist

- Abstracts data (including identifiers: name, CAS #, etc.)
- Draws or obtains structures

Chemical informatics person

- Identifies duplicate structures ← InChI!
- Identifies "bad" structures ← InChI makes this easier (disconnections, undefined stereo)
- May identify values inconsistent with the structure

Skilled chemist

- Acts on analyses: updates data or finds error in analysis

Identify species across data collections

Merging data collections based on molecular identifiers can be difficult.

If we are lucky the collections have structures

- Most likely with different drawing conventions
- InChI can still find matching molecules!
- Unless the conventions are really different
  - Acid drawn for conjugate base
  - Coordination compounds
  - Chrages on nitro groups, ring tautomers, etc.

InChI is not immune to Murphy's Law

Nothing is…

- Bad structures can come from numerous causes
- Unexpected hydrogens may present in the structure but not shown on the screen
- Where possible, confirm match with other identifiers
- **Knowledge of chemistry and likely failure modes is still essential**

Invariant URLs – Use InChI to create link

Pros

- Stable
- Can provide some response for valid InChIs that don't resolve

Cons

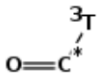
- URLs look ugly
  - Characters in InChI need to be URL-encoded
  - URLs are long (~2000 character limit in Internet Explorer)
- Why not use InChIKey?

Invariant URLs – InChI does not resolve

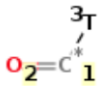
Information from the InChI

There are no matching entries in the database for this IUPAC International Chemical Identifier. The following information was obtained from the identifier.

- **Formula:** CTO
- **Molecular weight:** 31.0261
- **IUPAC Standard InChI:**
  - `InChI=1S/CHO/c1-2/h1H/i1T`
- **IUPAC Standard InChIKey:** `CFHIDWOYWUOIHU-CNRUNOGKSA-N`
- Connectivity:

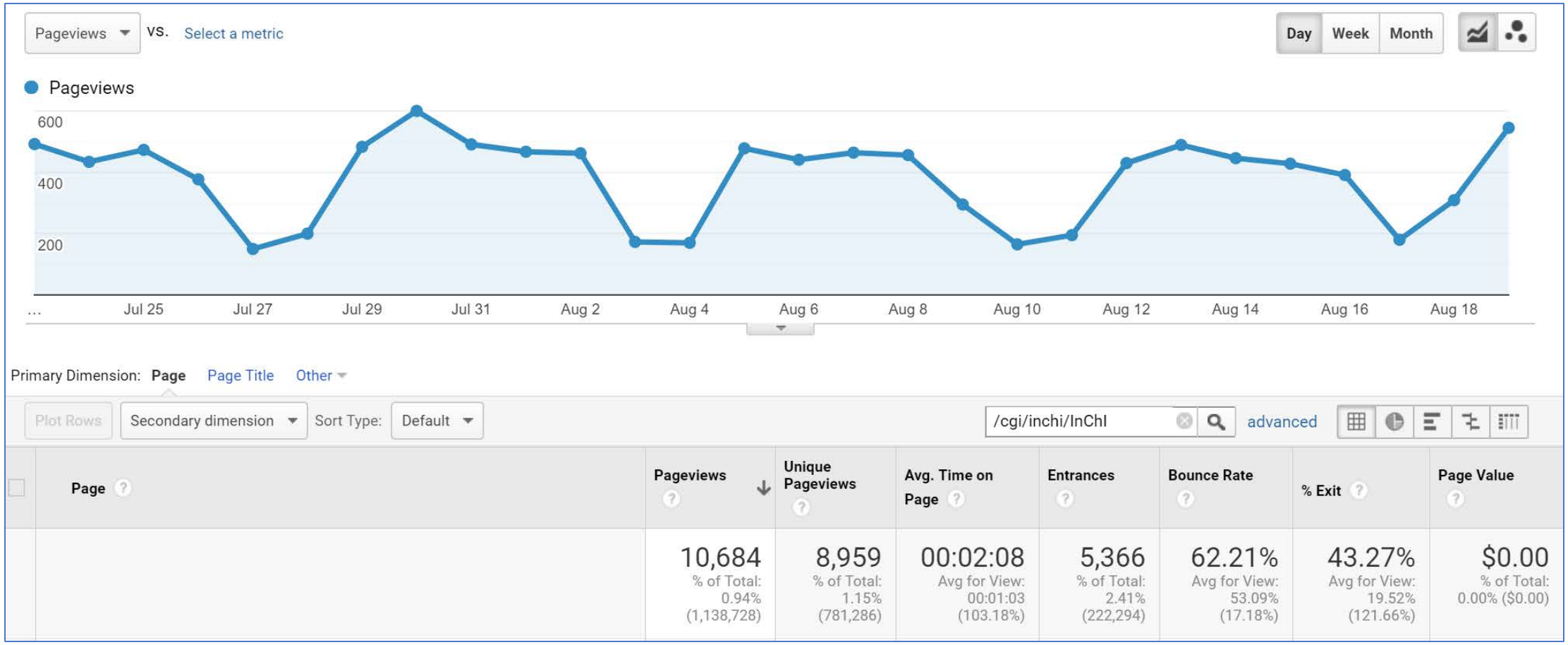


- 2-d Mol File from the identifier
- Canonical atom numbers:



  Note: stereochemistry is currently not indicated in the items above.
- **Permanent link** for this search. Use this link for bookmarking this species for future reference.
- **Isotopologues:**
  - Methyl-d1 radical, oxo-
  - Formyl radical

Invariant URLs – Less than 1% of page views



Page	Pageviews	Unique Pageviews	Avg. Time on Page	Entrances	Bounce Rate	% Exit	Page Value
	10,684 % of Total: 0.94% (1,138,728)	8,959 % of Total: 1.15% (781,286)	00:02:08 Avg for View: 00:01:03 (103.18%)	5,366 % of Total: 2.41% (222,294)	62.21% Avg for View: 53.09% (17.18%)	43.27% Avg for View: 19.52% (121.66%)	$0.00 % of Total: 0.00% ($0.00)

Screenshot from Google Analytics

Isotopologues – Parsing InChIs

Formyl radical

- **Formula:** CHO
- **Molecular weight:** 29.0180
- **IUPAC Standard InChI:**
  - `InChI=1S/CHO/c1-2/h1H`
  - Download the identifier in a file.

InChITRUST CERTIFIED 2011

- **IUPAC Standard InChIKey:** `CFHIDWOYWUOIHU-UHFFFAOYSA-N`
- **CAS Registry Number:** 2597-44-6
- **Chemical structure:**

This structure is also available as a 2d Mol file or as a computed 3d SD file

The 3d structure may be viewed using Java or Javascript.

- **Isotopologues:**
  - Methyl-d1 radical, oxo-

Search engines

- Marketing hyperbole:

  InChIKey is an important part of a search engine optimization strategy for web pages dealing with molecules or atoms!

- But:
  - Using “InChIKey + relevant search text” can be a useful search strategy
  - “relevant search text” is part of a SEO strategy!
- Search engines don’t know some synonyms

Search engine example

CFHIDWOYWUOIHU-UHFFFAOYSA-N microwave spectra

Web Images Videos News Maps Settings

All Regions Safe Search: Moderate Any Time

PDF Microwave (Rotational) Spectroscopy

sci.tanta.edu.eg/files/Microwave spectroscopy BSc-Lect-2.pdf

From **microwave** spectroscopy, bond lengths can be determined with a correspondingly high precision, as illustrated in this example. From the rotational **microwave** spectrum 35of 1H Cl, we find that B = 10.59342 cm-1. Given that 35the masses of 1H and Cl are 1.0078250 and 34.9688527 amu, respectively, determine the bond

Formyl radical - webbook.nist.gov

https://webbook.nist.gov/cgi/cbook.cgi?ID=C2597446

IUPAC Standard InChIKey: **CFHIDWOYWUOIHU-UHFFFAOYSA-N**; CAS Registry Number: 2597-44-6; Chemical structure: This structure is also available as a 2d Mol file or as a computed 3d SD file The 3d structure may be viewed using Java or Javascript. Isotopologues: Methyl-d1 radical, oxo-Permanent link for this species. Use this link for bookmarking this species for future reference.

Screenshot from Duck Duck Go

Integrating InChI into work flows

- Chemists should never have to see InChI strings.

- Drawing programs:
  - Control them programmatically
  - Support chemistry beyond mol file v2000
  - Make it easy to show hydrogens?

- In the future this may all be done in web browsers?

A real problem – Ag^{+47}, InChI=1S/Ag/q+47

Molecular Formula "Ag" > substances (183) > **181588-83-0** > get references (3)

SUBSTANCE DETAIL | **Get References**

Return | Previous

97. CAS Registry Number 181588-83-0

~6

Ag

Silver, ion (Ag^{47+})

Other Names

Silver(47+)

Ag^{47+}

Screenshot from CAS SciFinder

Our most valuable resource

Some of our contributors (NIST Chemistry WebBook): H.Y. Afeefy, J.F. Liebman, S.E. Stein, Glushko Thermocenter (Moscow), E.S. Domalski, E.D. Hearing, S.G. Lias, H.M. Rosenstock, K. Draxl, B.W. Steiner, J.T. Herron, J.L. Holmes, R.D. Levin, J.F. Liebman, S.A. Kafafi, M. Meot-Ner (Mautner), E.P. Hunter, M.E. Jacox, T. Shimanouchi, K.P. Huber, G. Herzberg J. A. Martinho Simões, J.S. Chickos, W.E. Acree, Jr., Students of University of Missouri – St .Louis (Chem 202 – Introduction to the Literature of Chemistry), R.L. Brown, R. Sander, E.W. Lemmon, M.O. McLinden, D.G. Friend, V. Talrose, E.B. Stern, A.A. Goncharova, N.A. Messineva, N.V. Trusova, M.V. Efimkina, A.N. Yermakov, A.A. Usov, A.A. Leskin, NIST Thermodynamics Research Center (C. Muzny director), NIST Mass Spectrometry Data Center (W.E. Wallace director), P.M. Chu, F.R. Guenther, G.C. Rhoderick, W.J. Lafferty, Coblentz Society, Inc., T.L. Myers, R.G. Tonkyn, A.M. Oeck, T.O. Danby, J.S. Loring, M.S. Taubman, S.W. Sharpe, J.C. Birnbaum, T.J. Johnson, Y-F Su, M.R.K. Kelly-Gorham, E.J. Heilweil, M. Campbell, K.K. Irikura, T.C. Allison, K.C. Hafner, S.C. Ness, D.X. Du, J.W. Qiu, A.H. Yang, H.M. Park, J.K. Skerritt, M.S. Shevchuk, M.Y. Liou, N.B. Ravi, S.F. Dermer, E.N. Ho, E.W. Jin, S.N. Aggarwal, S.N. Pan, D.S. Graham, S.E. Wollman, Y. Niyonzima, A.M. Newman, A.M. Zhang, A.M. Martin, N. Patel, A. Tran, J. Tseytlin, N. Kau, P.J. Christian, D.H. Frizzell, J.J. Reed

The last slide

Links

- NIST Chemistry WebBook: https://webbook.nist.gov/chemistry/
- NIST Standard Reference Data (SRD): https://www.nist.gov/srd

Disclaimer

Any mention of commercial products is for information only; it does not imply recommendation or endorsement by NIST.