Data, data everywhere, nor any bit processable
Opportunities for amalgamating and opening up
hemical data and information relevant to hazar
recognition and safety planning

Jian Zhang, Paul Thiessen, Gang Fu, **Evan Bolton**, Gang Fu 1

Leah McEwen 2

Aug. 18, 2015

1 U.S. National Center for Biotechnology Information (NCBI)

2 Clark Library, Cornell University

U.S. National Library of Medicine

ease note …

am a chemist

am an informatician

work with data .lots. of it

help to create public interfaces to chemical (biology) information

work with health and safety professionals, authoritative sources of H&S nformation, (increasingly) those involved in H&S communities (including chemistry librarians

am not a chemical health and safety (H&S) professional

pecial thanks

Chemical Health and Safety collaborators

- Especially: Leah McEwen, Ye Li, Ralph Stuart

Software collaborators

- Especially: Daniel Lowe and Roger Sayle (NextMove Software .. LeadMine)

All Contributors and Collaborators

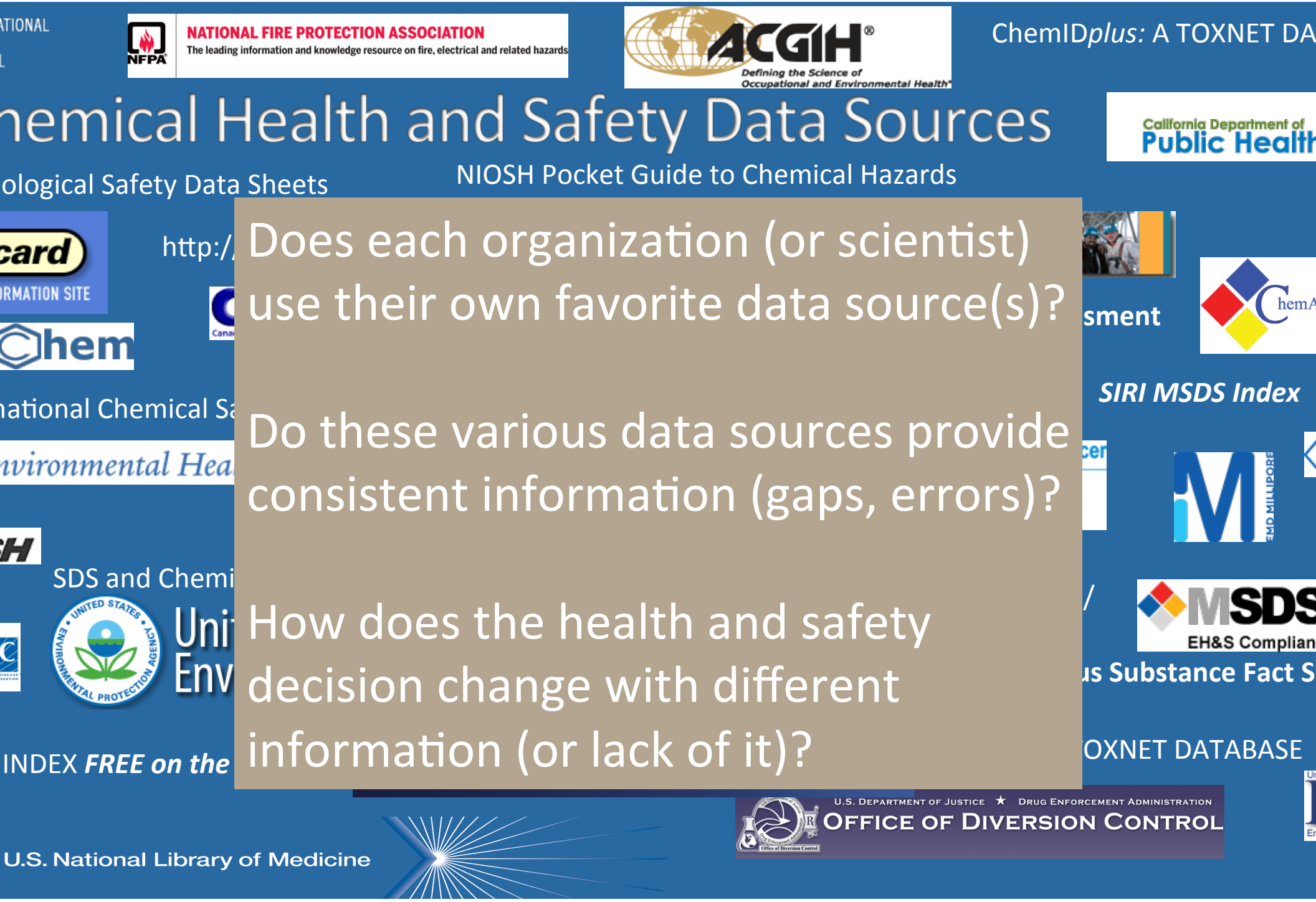
This research was supported [in part] by the Intramural Research Program o
the NIH, National Library of Medicine.

ubChem resource

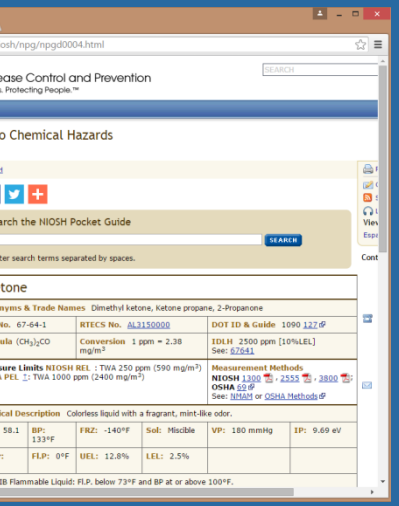
https://pubchem.ncbi.nlm.nih.g

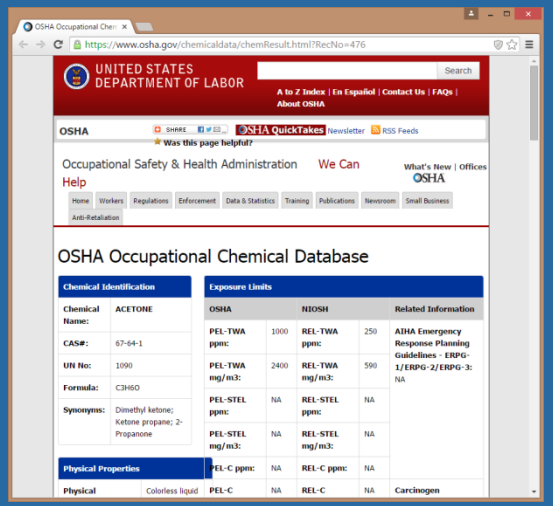
Primary aim is to provide information known about chemical substances

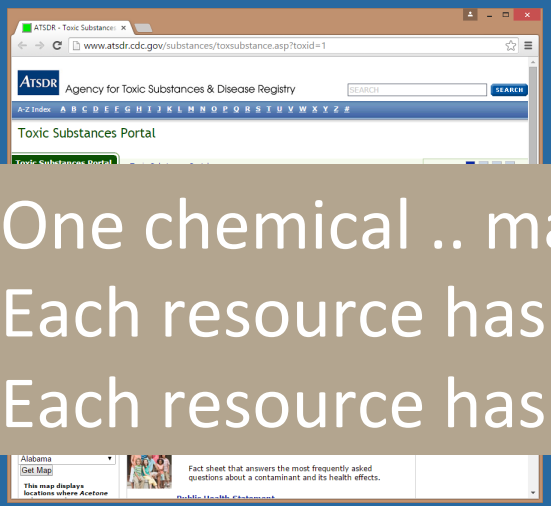
- Contains authoritative resources
- Contains manually curated resources
- Includes health and safety
- There are errors
  - Authoritative doesn't mean error free
  - Curated doesn't mean error free
- We DO NOT curate (known errors reported , filtered)
- We can leverage curated content

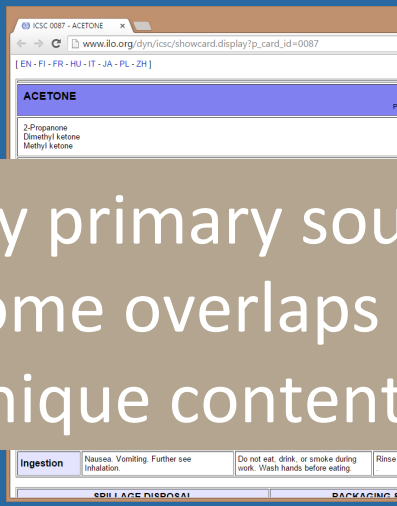

NATIONAL FIRE PROTECTION ASSOCIATION
The leading information and knowledge resource on fire, electrical and related hazards
ACGIH®
Defining the Science of Occupational and Environmental Health
ChemIDplus: A TOXNET DA
hemical Health and Safety Data Sources
California Department of Public Health
ological Safety Data Sheets
NIOSH Pocket Guide to Chemical Hazards
http:/
Does each organization (or scientist) use their own favorite data source(s)?
Do these various data sources provide consistent information (gaps, errors)?
How does the health and safety decision change with different information (or lack of it)?
sment
SIRI MSDS Index
national Chemical Sa
vironmental Hea
SDS and Chemi
MSDS
EH&S Complian
us Substance Fact S
INDEX FREE on the
OXNET DATABASE
U.S. Department of Justice ★ Drug Enforcement Administration
Office of Diversion Control
U.S. National Library of Medicine

nemical Health and Safety Information

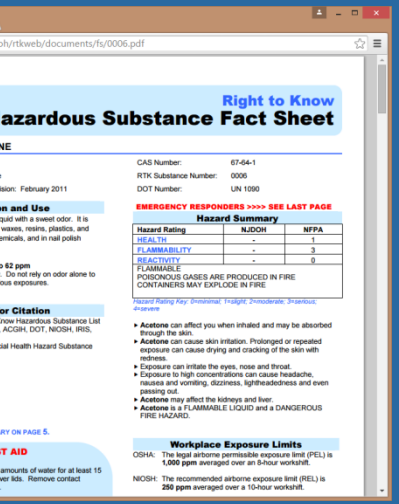




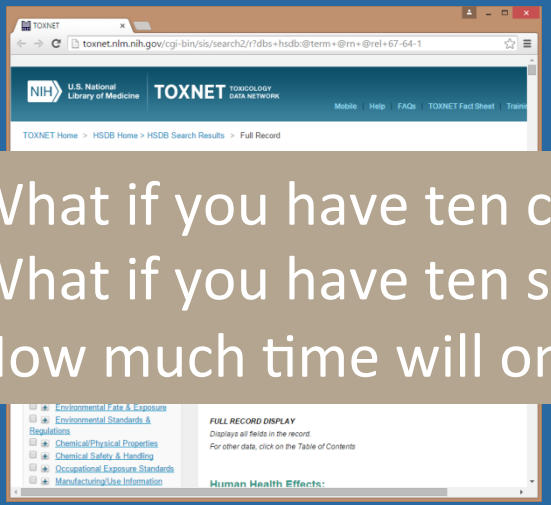


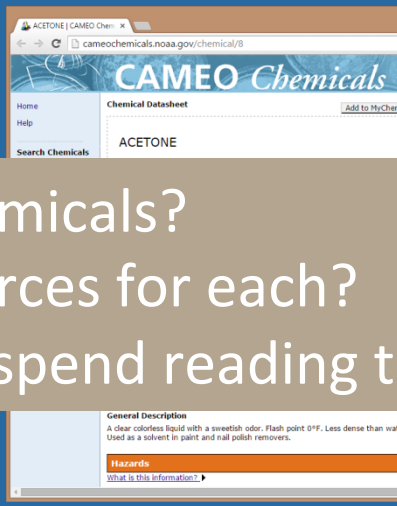


One chemical .. many primary sou
Each resource has some overlaps
Each resource has unique content



Acetone .. seven sources
CDC NIOSH, NJ HSFS, DOL
OSHA, WHO ICSC, NLM
HSDB, EPA ATSDR, EPA/
NOAA CAMEO





What if you have ten chemicals?
What if you have ten sources for each?
How much time will one spend reading t

U.S. National Library of Medicine

ubChem Laboratory Chemical Safety Sheet (LCSS)

https://pubchem.ncbi.nlm.nih.gov/l

Pull in primary chemical health and safety resources (as-is possible)

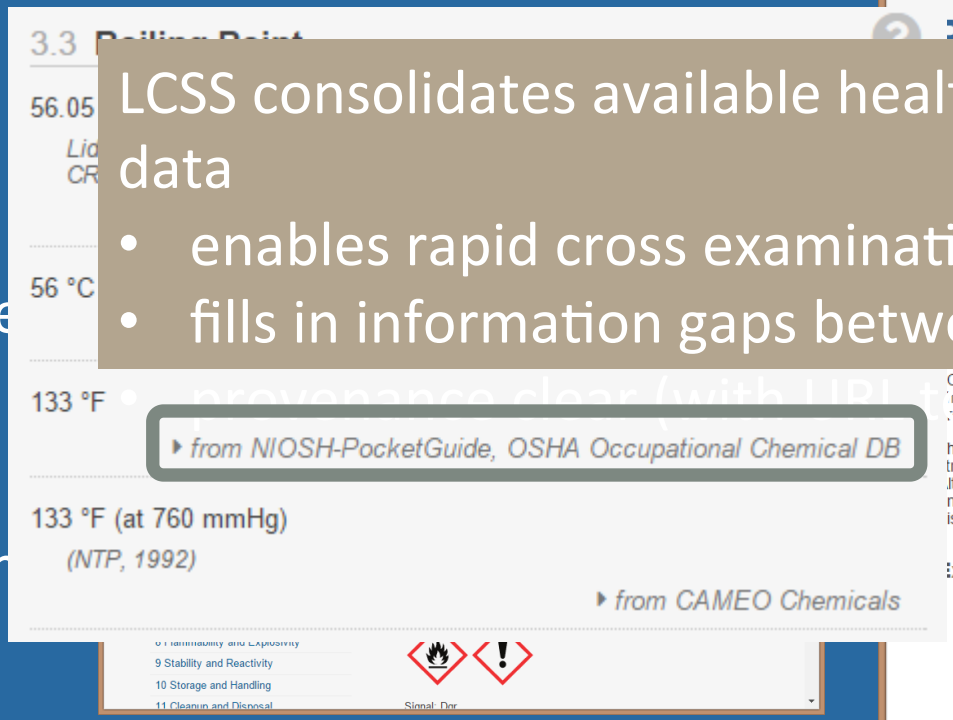
Organize it

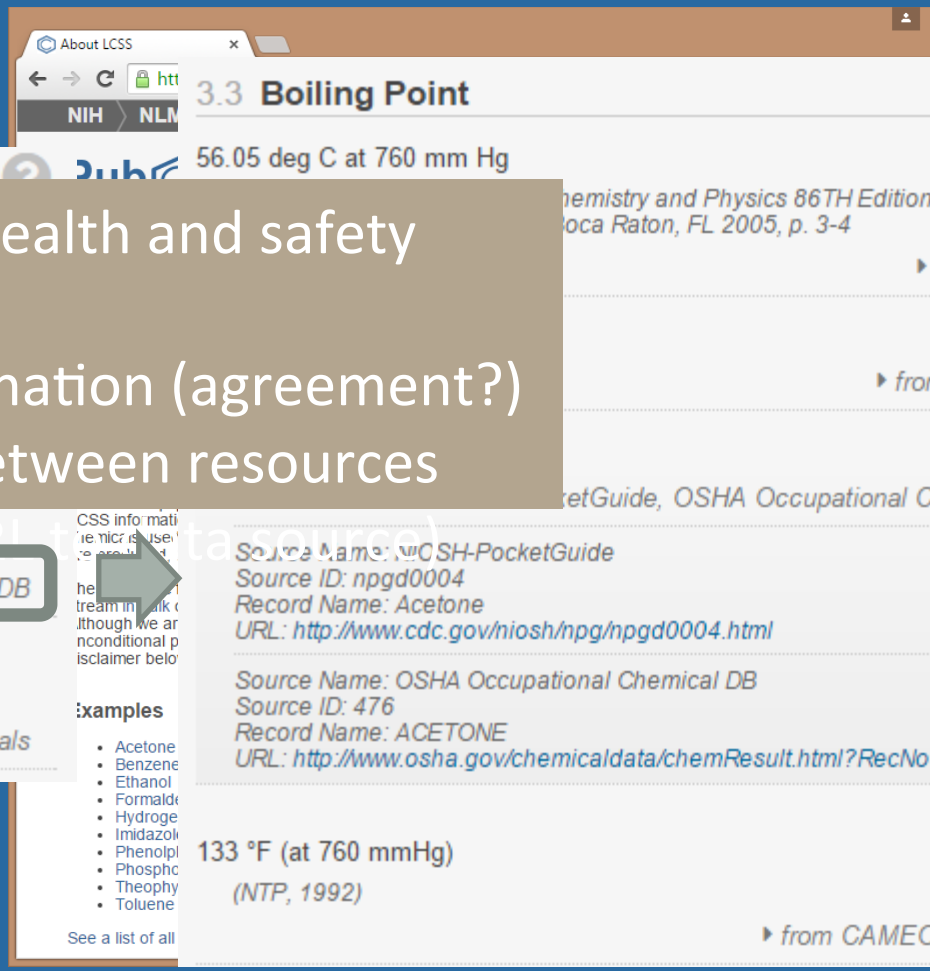
Make it available

- Interactive
- In bulk
- On-demar

LCSS consolidates available health and safety data

- enables rapid cross examination (agreement?)
- fills in information gaps between resources
- provenance clear (with URL to data source)





ealth and Safety Information summaries are for human

Yes, this is obvious .. and they are great! Succinct write-ups. What more can one ask for?!

Part of a CAMEO record for Acetone

Fire Hazard

Excerpt from GUIDE 127 [Flammable Liquids (Polar / Water-Miscible)]:

HIGHLY FLAMMABLE: Will be easily ignited by heat, sparks or flames. Vapors may form mixtures with air. Vapors may travel to source of ignition and flash back. Most vapors than air. They will spread along ground and collect in low or confined areas (sewers, ba tanks). Vapor explosion hazard indoors, outdoors or in sewers. Those substances desig (P) may polymerize explosively when heated or involved in a fire. Runoff to sewer may explosion hazard. Containers may explode when heated. Many liquids are lighter than 2012)

Health Hazard

INHALATION: vapor irritating to eyes and mucous membranes; acts as an anesthetic concentrations. INGESTION: low order of toxicity but very irritating to mucous membr prolonged excessive contact causes defatting of the skin, possibly leading to dermatitis 1999)

Reactivity Profile

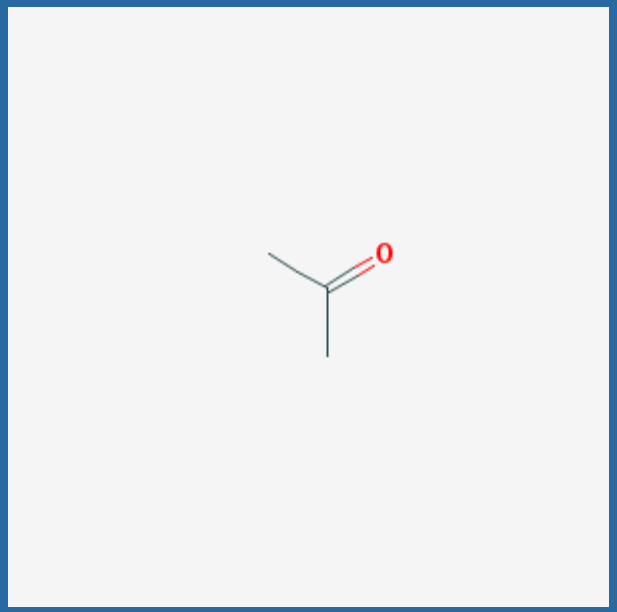
It was reported that a mixture of ACETONE and chloroform, in a residue bottle, explod addition of acetone to chloroform in the presence of base will result in a highly exothe it is thought that a base was in the bottle [MCA Case History 1661. 1970]. Also, Nitro sealed in a tube with a residue of acetone in the presence of platinum catalyst, gave a reaction [Chem. Eng. News 35(43):60. 1967]. The reaction of nitrosyl perchlorate and ignites and explodes. Explosions occur with mixtures of nitrosyl perchlorate and prima [Ann. Chem. 42:2031. 1909]. Reacts violently with nitric acid. Also causes exothermi when in contact with aldehydes.

uman understanding

As a chemist, you can understand and recognize that this picture is the chemical acetone

You can put a name or registry number next to it

s this not good enough?



Acetone

67-64-1

The picture has no association with the name or registry number

The computer ‘sees’ a binary image

omputer understanding

propan-2-one
58.07914 g/m

Give a computer a chemical
structure a (n
understand it

Computer ca
he structure

Computer ca
o the structu

Computer can generate other key
nformation from structure

Computer understanding can help provide human understanding

If the computer understands, we can leverage it for search, analysis, and more

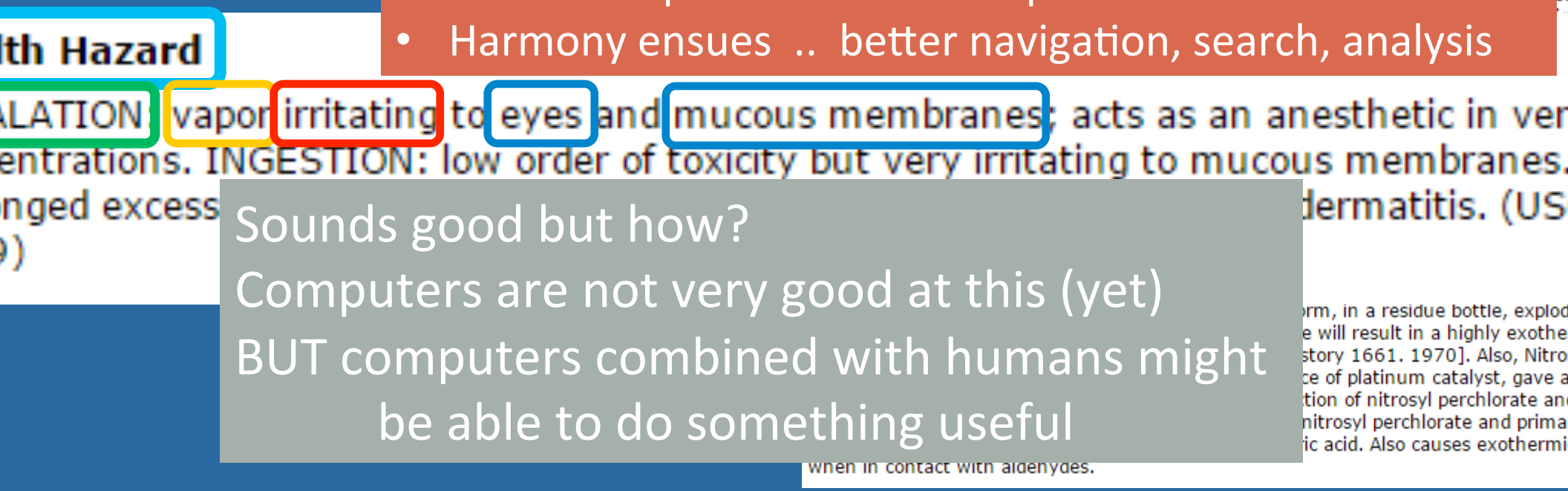
Ac
67

ealth and Safety Information summaries for computers

Can human readab
made computer un

Potential workflow involving Computers and Humans

- Humans need to provide the terms and organize them
- Computer can recognize these terms (NER)
- Identify relationships between terms (Human/NLP)
- Relationships fed back to computer
- Harmony ensues .. better navigation, search, analysis



Sounds good but how?
Computers are not very good at this (yet)
BUT computers combined with humans might be able to do something useful

Working with chemical information has CAVEATS!

Chemical information is a bit of a mess and can be rather nuanced

- Names, names, and more names (210M in PubChem)
  - Some standard names are not open and cannot be used/verified without $$$
- Name/structure associations vary by use case (many overlapping)
  - Acetic acid vs. Acetic acid tri-hydrate
  - Formaldehyde: (gas) vs. Formalin (liquid , 40% formaldehyde w/ water)
  - Sulfuric acid: SO_3 (gas) vs. H_2SO_4 (liquid)
  - Glucose: L/D, ring open/closed, alpha/beta/both vs. Glucose monohydrate
  - Large corpus in the 'wild' .. data source dependent nuances

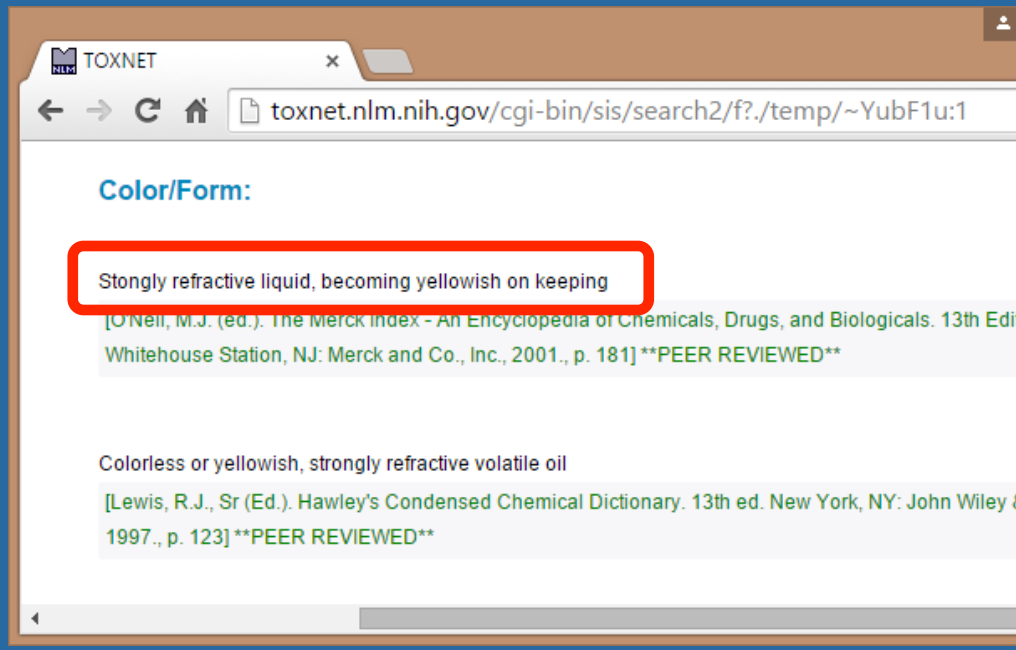
Verify with primary source(s) prior to information use

- i.e., is this the form of the chemical I care about?

Working with chemical information has CAVEATS!

Chemical annotation in PubChem is text locked (i.e., meant for humans)

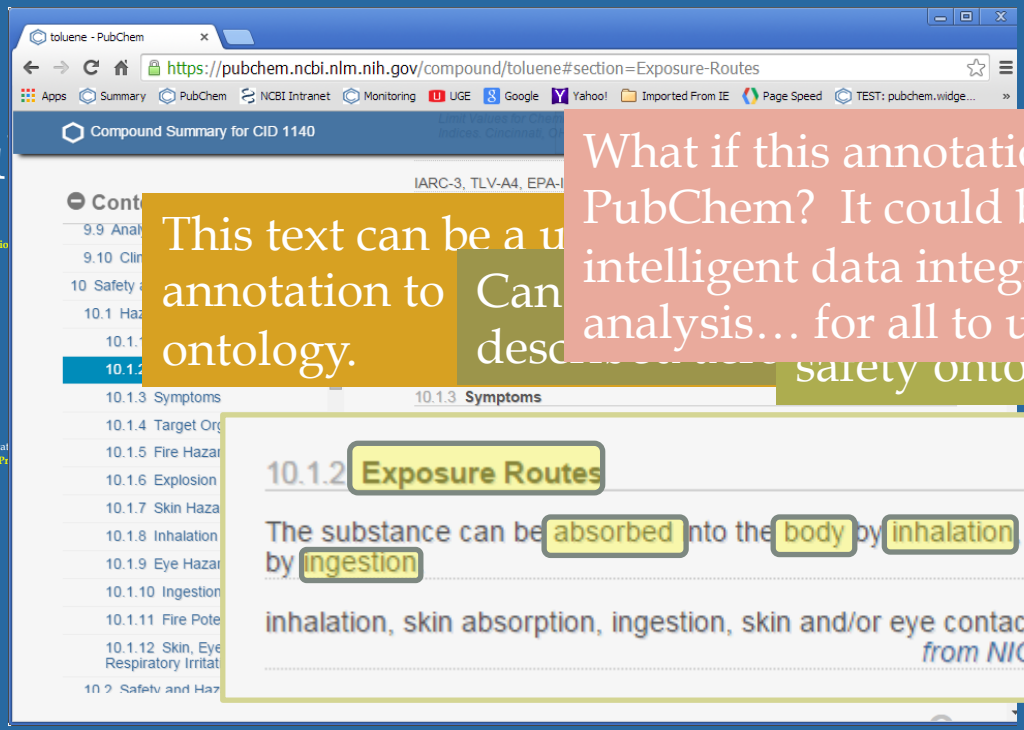
- Benzaldehyde
  - Color
    - Stongly refractive liquid, becoming yellowish on keeping
    - Colorless or yellowish, strongly refractive volatile oil
  - Boiling point
    - 179 deg C
    - 179 °C
    - 354 °F (at 760 mmHg)
  - Flash point
    - 145 deg F, 63 deg C (Closed cup)
    - 73.9 deg C (Open cup)
    - 63 °C



ow can PubChem help?

Well .. we have lots of data

Eureka!!!
Let's Make a connec
graph of knowledg



This text can be a u
annotation to
ontology.

What if this annotation is provided back to PubChem? It could be used to power more intelligent data integration, access and analysis… for all to use.

ostly
munity
ical

What is a knowledge graph?

Knowledge graphs could be envisaged as a network of all kind things which are relevant to a specific domain or to an organization.

https://blog.semantic-web.at/2014/07/15/from-taxonomies-over-ontologies-to-knowledge-graphs/

Why a knowledge graph?

(To be) able to make complex queries over the entirety of all kind of information. By breaking up the data silos there is a high probability that query results become more valid.

https://blog.semantic-web.at/2014/07/15/from-taxonomies-over-ontologies-to-knowledge-graphs/

rend towards community based annotation

"Chemical Safety Ontology" (CSO) participants request and examine a provided table of contents of chemical safety information in PubChem to identify areas of interest

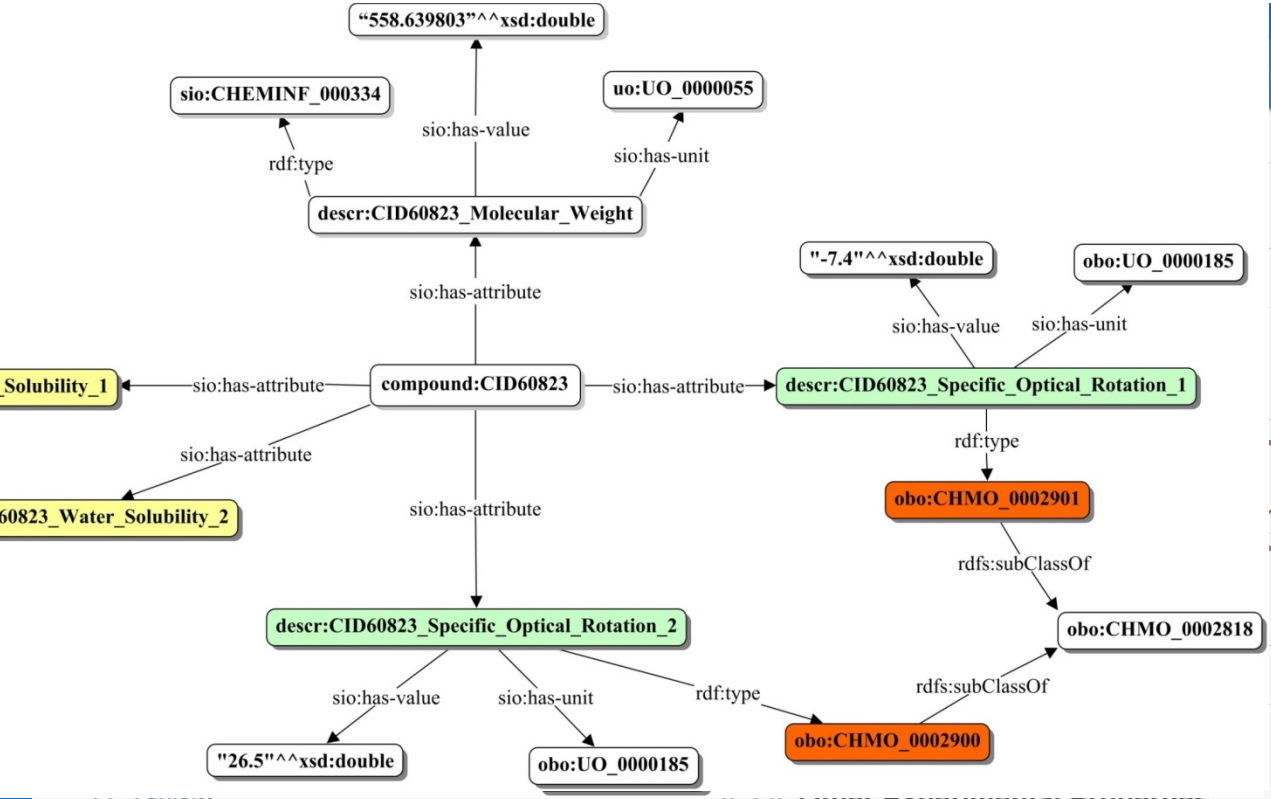
PubChem provides text for a given area of interest to CSO to train the ontology by adding appropriate concepts:

1. PubChem uses current set of ontology terms to auto-annotate the text
2. CSO manually reviews the auto-annotation for ontological gaps, makes fixes
3. Iterate

PubChem may be able to provide tools that help to automate this iterative process and could even serve as a locus for volunteers to semi-manually annotate chemical safety text

PubChem may be able to act as a repository for annotation

- Integration into the PubChemRDF project



Print Share Help

from DrugBank [19]

from DrugBank [19]

from DrugBank [19]

12 Literature

13 Patents

14 Biomolecular Interactions and Pathways

15 Biological Test Results

16 Classification

17 Information Sources

4.2.4 Other Experimental Properties

MW: 1155.36. Specific optical rotation at 25 deg C for D (sodium) line = -7.4 deg (c = 1 in DMSO) /Calcium salt/

O'Neil, M.J. (ed.). The Merck Index - An Encyclopedia of Chemicals, Drugs, and Biologicals. 13th Edition, Whitehouse Station, NJ: Merck and Co., Inc., 2001., p. 148

from HSDB [27]

pKa= 4.46. Solubility in water (30 deg C) = 20.4 ug/ml (pH 2.1); 1.23 mg/ml (pH 6.0) /Sodium salt/

O'Neil, M.J. (ed.). The Merck Index - An Encyclopedia of Chemicals, Drugs, and Biologicals. 13th Edition, Whitehouse Station, NJ: Merck and Co., Inc., 2001., p. 148

from HSDB [27]

MW: 540.62. MP: 159.2-160.7 deg C. Specific optical rotation at 25 deg C = +26.05 (c = 1 in chloroform) /Lactone/

O'Neil, M.J. (ed.). The Merck Index - An Encyclopedia of Chemicals, Drugs, and Biologicals. 13th Edition, Whitehouse Station, NJ: Merck and Co., Inc., 2001., p. 148

from HSDB [27]

hemical Reactivity

Many scenarios of acute interest

- Reaction planning
  - Anything to worry about? Proper protective clothing? Explosion hazards? Incompatibilities? Risks? Storage? Solvent incompatibility? Etc.

Chemical reactivity classification

- NOAA Chemical Reactivity Worksheet (CRW)
- Brethericks
- Knowledge base variants – CAMEO (CRW)

Wouldn't it be great to classify all of PubChem?

- Imagine tools where you could determine incompatibilities for any arbitrary chemical reaction and get an automated assessment

ummary

PubChem provides chemical health and safety information LCSS for chemica

- Clear provenance of information provided
- Assembling health and safety information for easy access
- Breadth and depth of available information varies greatly per chemical

Chemical information is (sadly) messy

- PLEASE double check the information that it deals with the same thing

Working with chemical health and safety community towards:

- Adding structure to textual data to make it computable
- Chemical reactivity classification

PubChem Crew …

Steve Bryant

Jie Chen

Tiejun Chen

Gang Fu

Renata Geer

Asta Gindulyte

Lianyi Han

Jane He

Siqian He

Sunghwan Kim

Ben Shoemaker

Paul Thiessen

Jiyao Wang

Yanli Wang

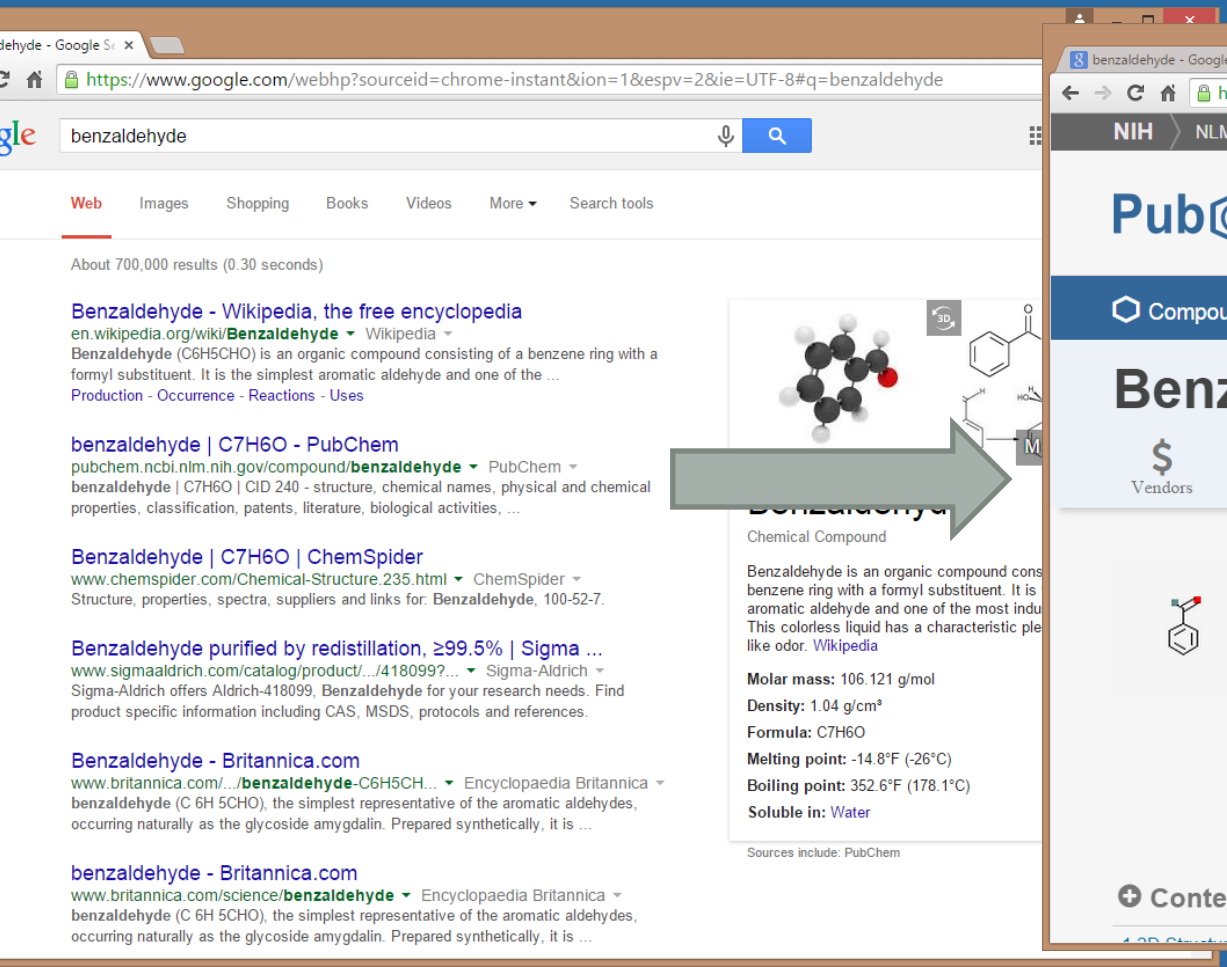
Bo Yu

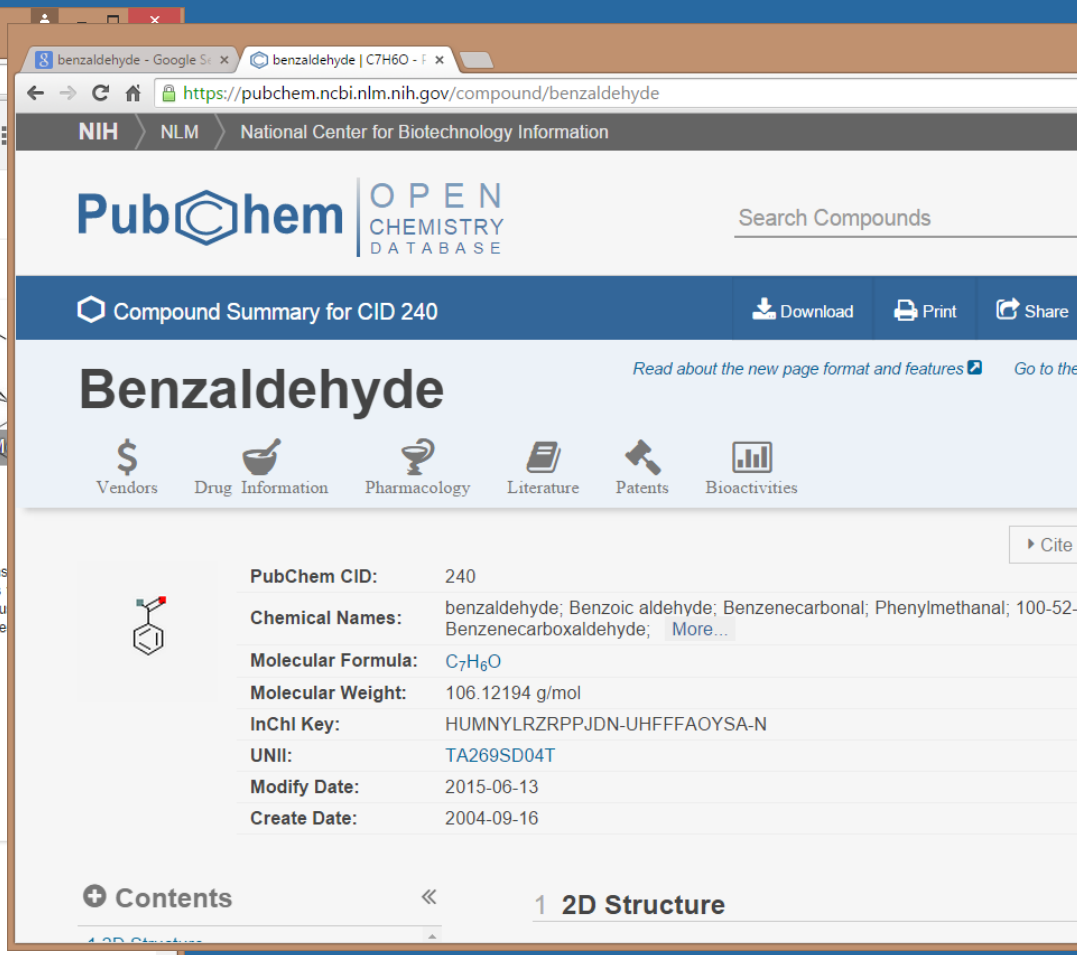
Jian Zhang

Special thanks to the NCBI Help Desk, especially Rana M

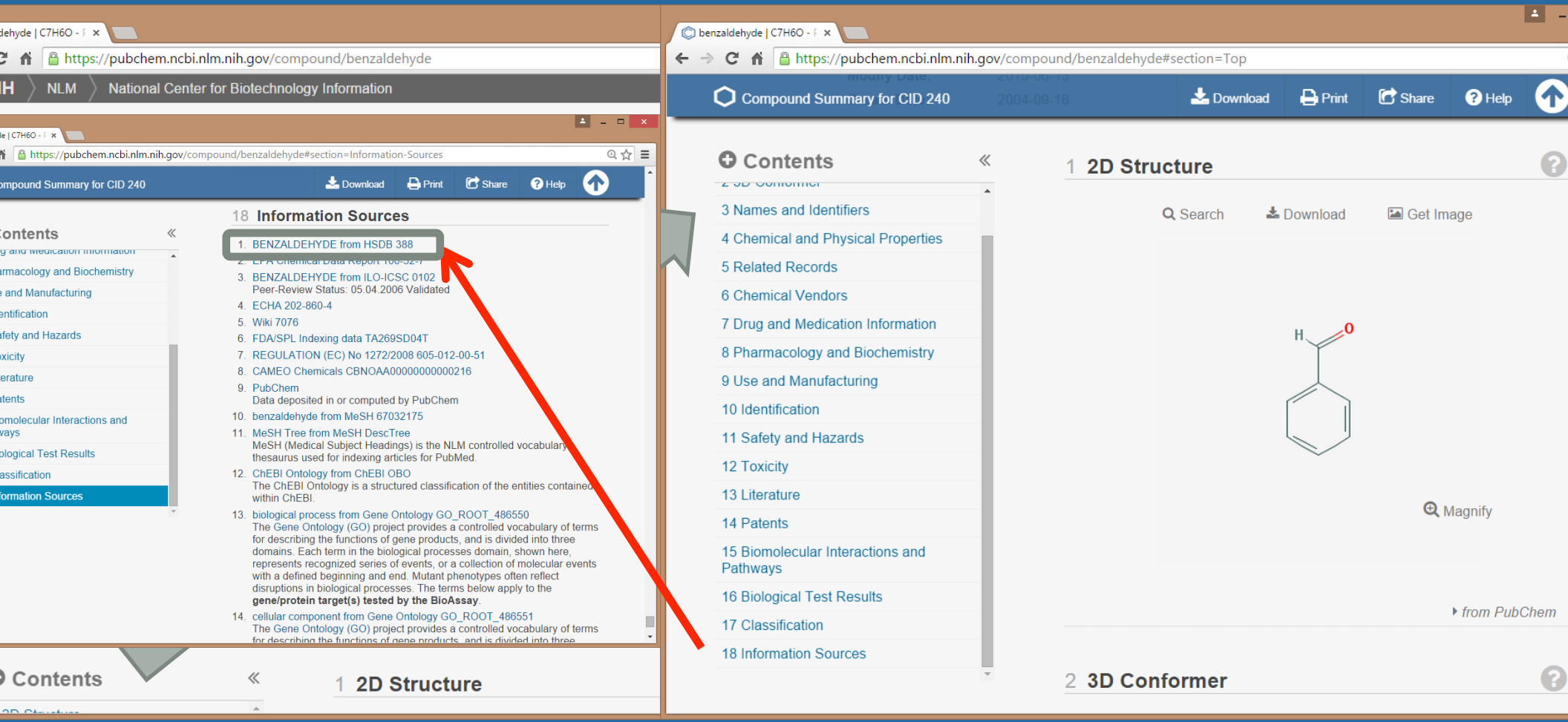
Have any questions?

nemical information is everywhere





enzaldehyde as a case study



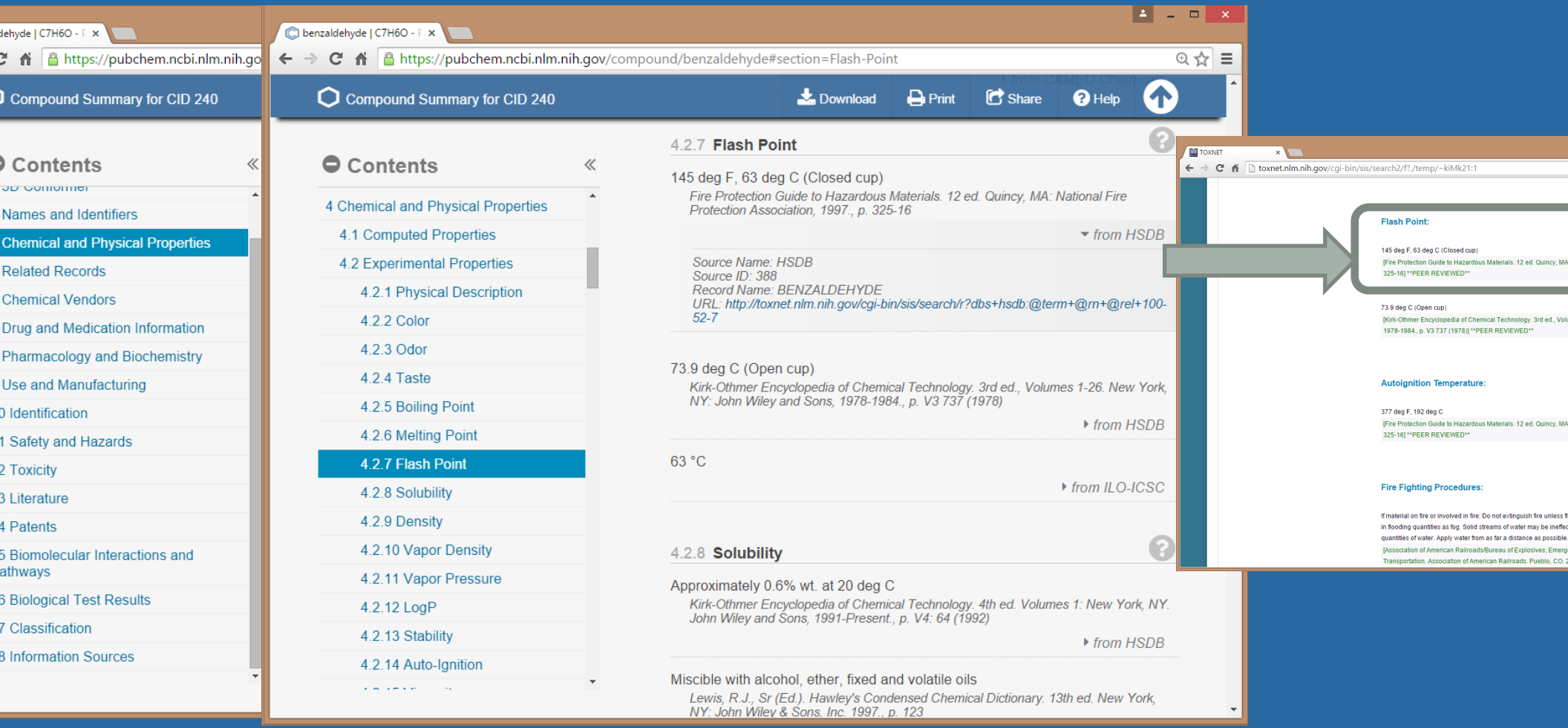
PubChem Compound TOC - Experimental Properties

ysical Description
lor
rm
dor
ste
iling Point
elting Point
ash Point
lubility
ensity
por Density
por Pressure
gP
gS
enrys Law Constant
mospheric OH Rate
onstant
ability

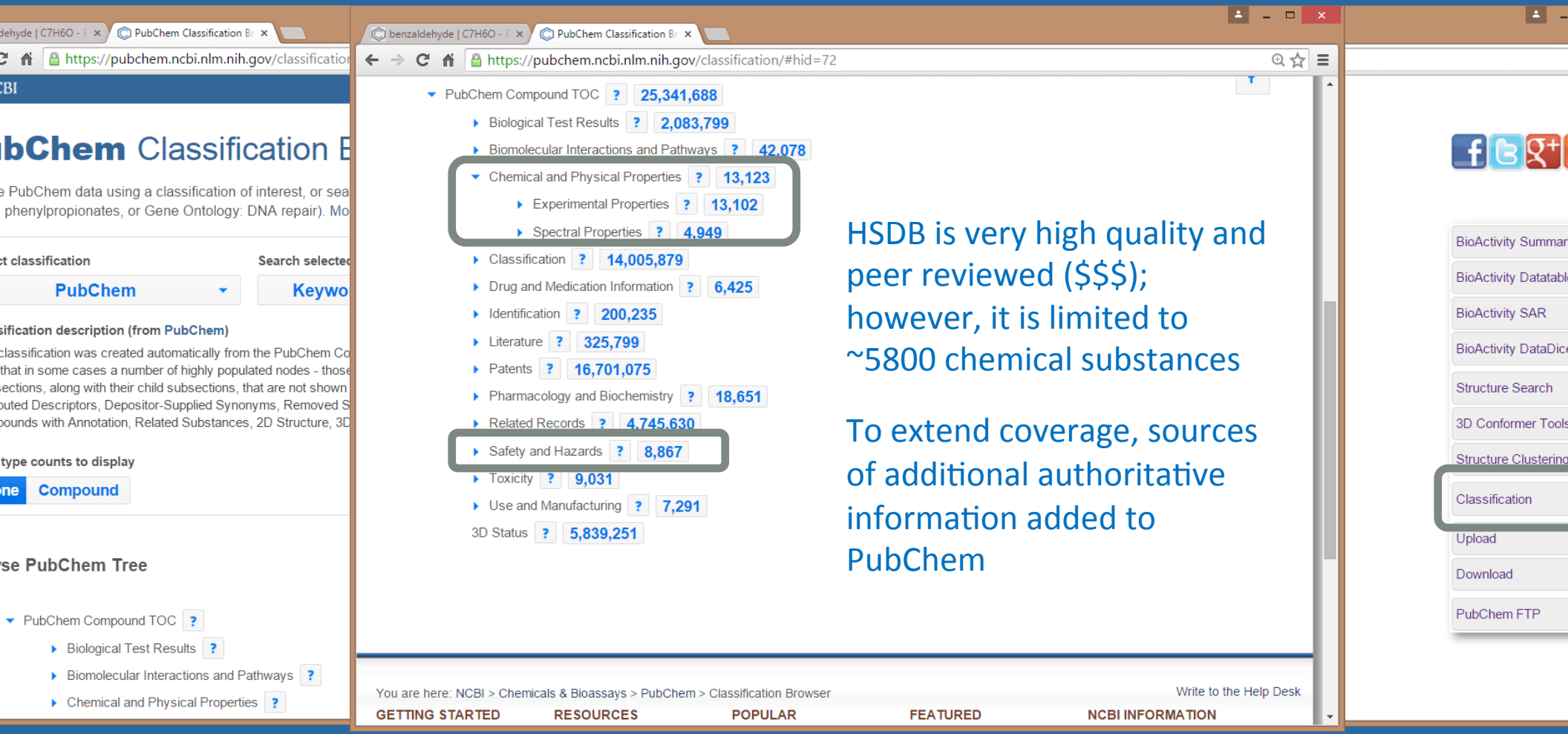
Optical Rotation
Auto-Ignition
Decomposition
Viscosity
Corrosivity
Volatility
Heat of Combustion
Heat of Vaporization
pH
Surface Tension
Evaporation Rate
Ionicity
Dispersion
Polymerization
Odor Threshold
Enthalpy of Formation
Entropy of Formation
Thermal Conductivity

Electrical Conductivity
Heat of Capacity
Bioavailability
Ghose Filter
Hydrophobicity
Isoelectric Point
Polarizability
Refractive Index
Caco2 Permeability
pKa
pKb
Dissociation Constants
Relative Evaporation Rate
Other Experimental Properties
…

enzaldehyde chemical and physical properties



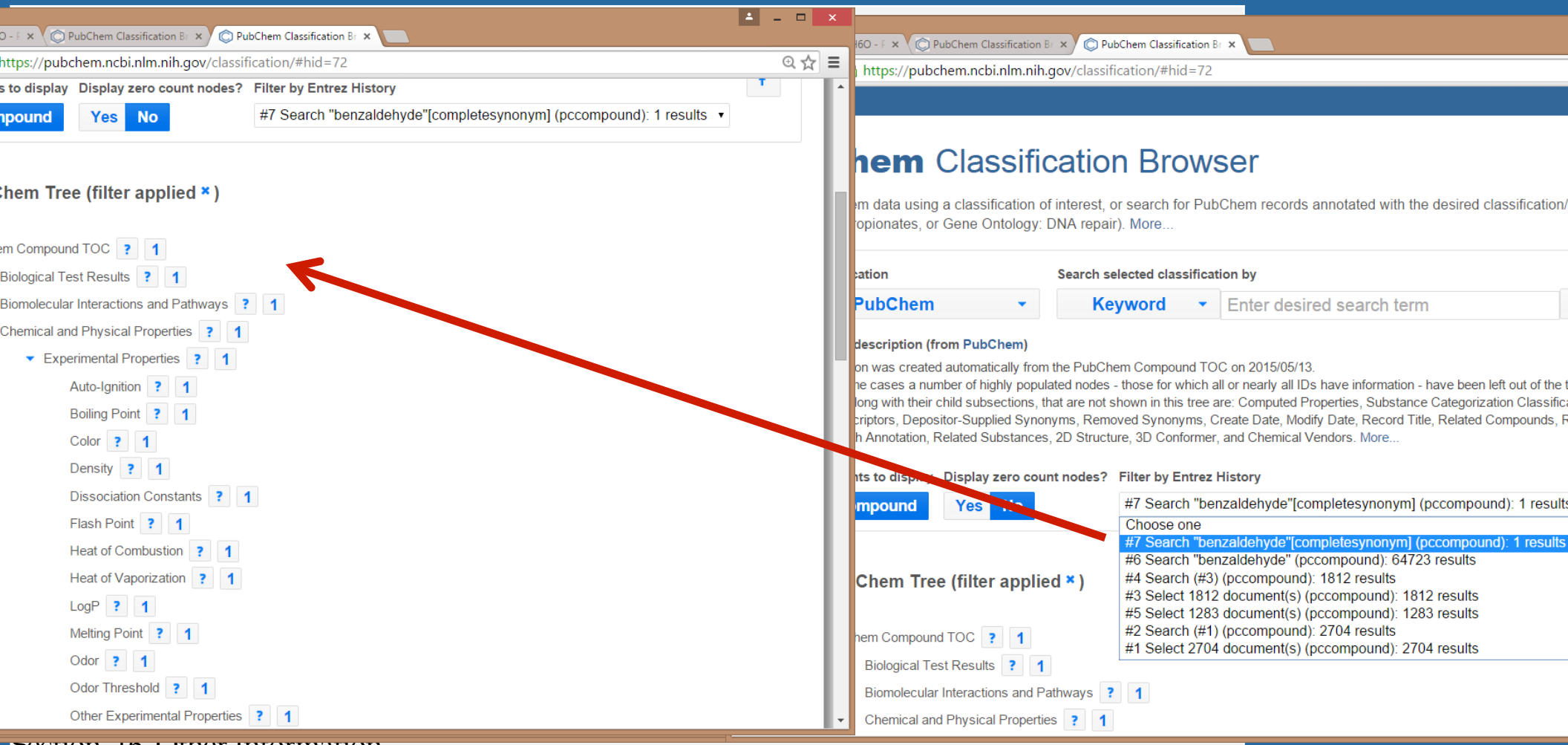
ubChem Compound Table of Contents



HSDB is very high quality and peer reviewed ($$$); however, it is limited to ~5800 chemical substances

To extend coverage, sources of additional authoritative information added to PubChem

OSHA SDS (Safety Data Sheet, formerly MSDS) standard



Section. 16. Other Information