EPA United States Environmental Protection Agency

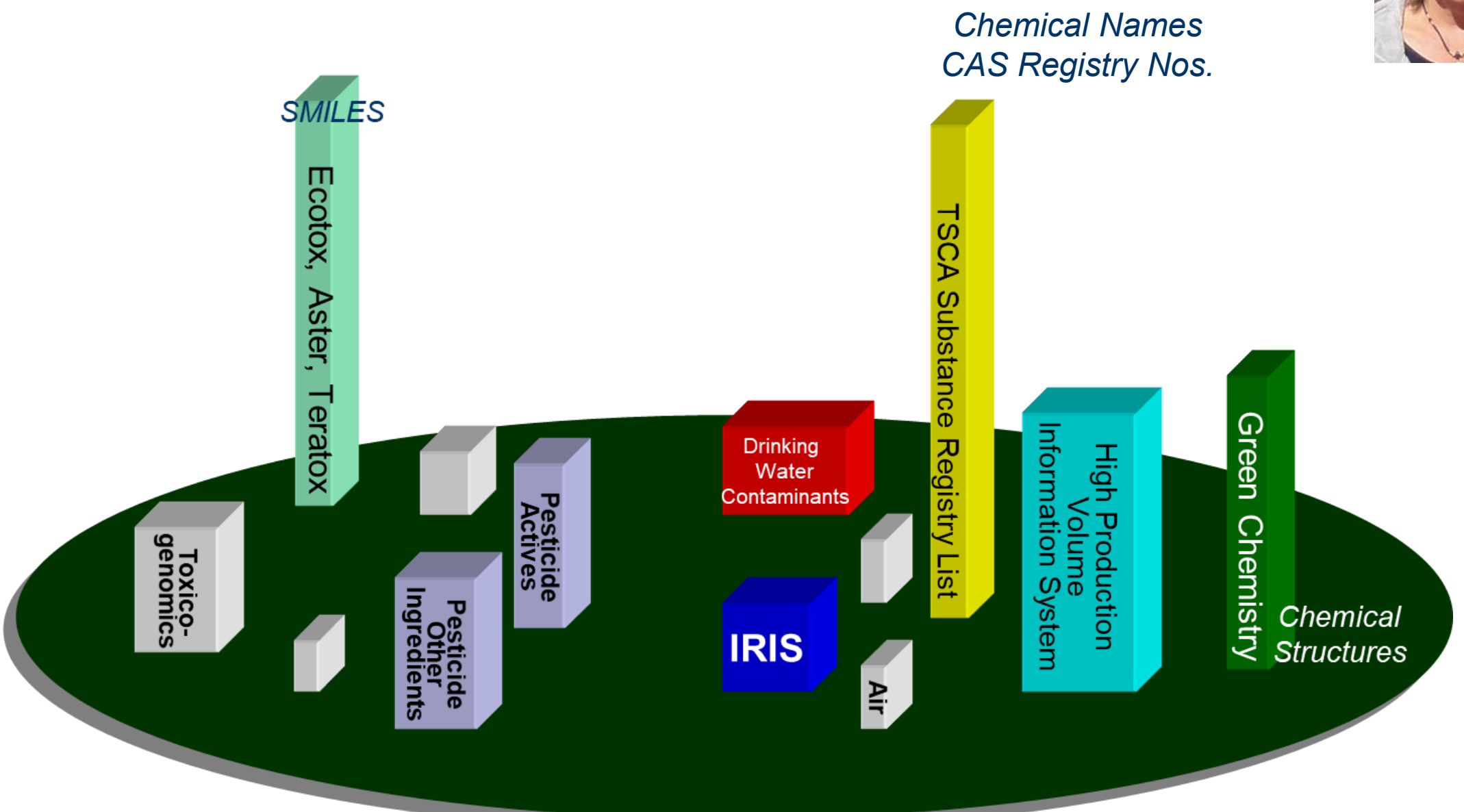
Crafting persistent identifiers and structure-based representations in DSSTox as surrogates for chemical names

Christopher Grulke[1]
Ann Richard[1]
Antony Williams[1]

1. National Center for Computational Toxicology, U.S. EPA
(soon to be the Center for Computational Toxicology and Exposure)

American Chemical Society Meeting, Fall 2019
26 August 2019, San Diego, FL

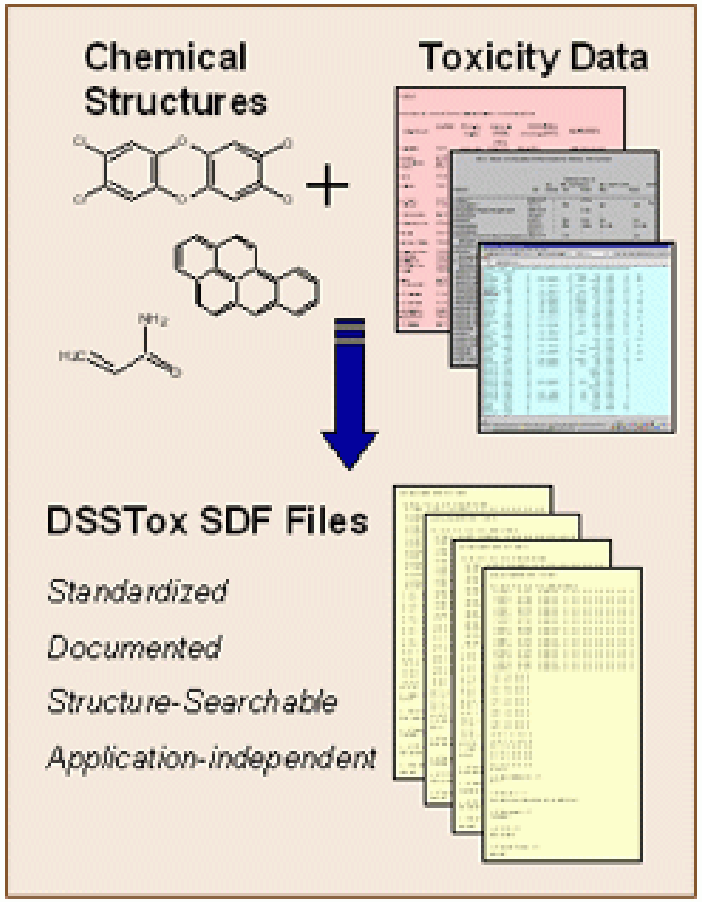
EPA's data islands … circa 2000



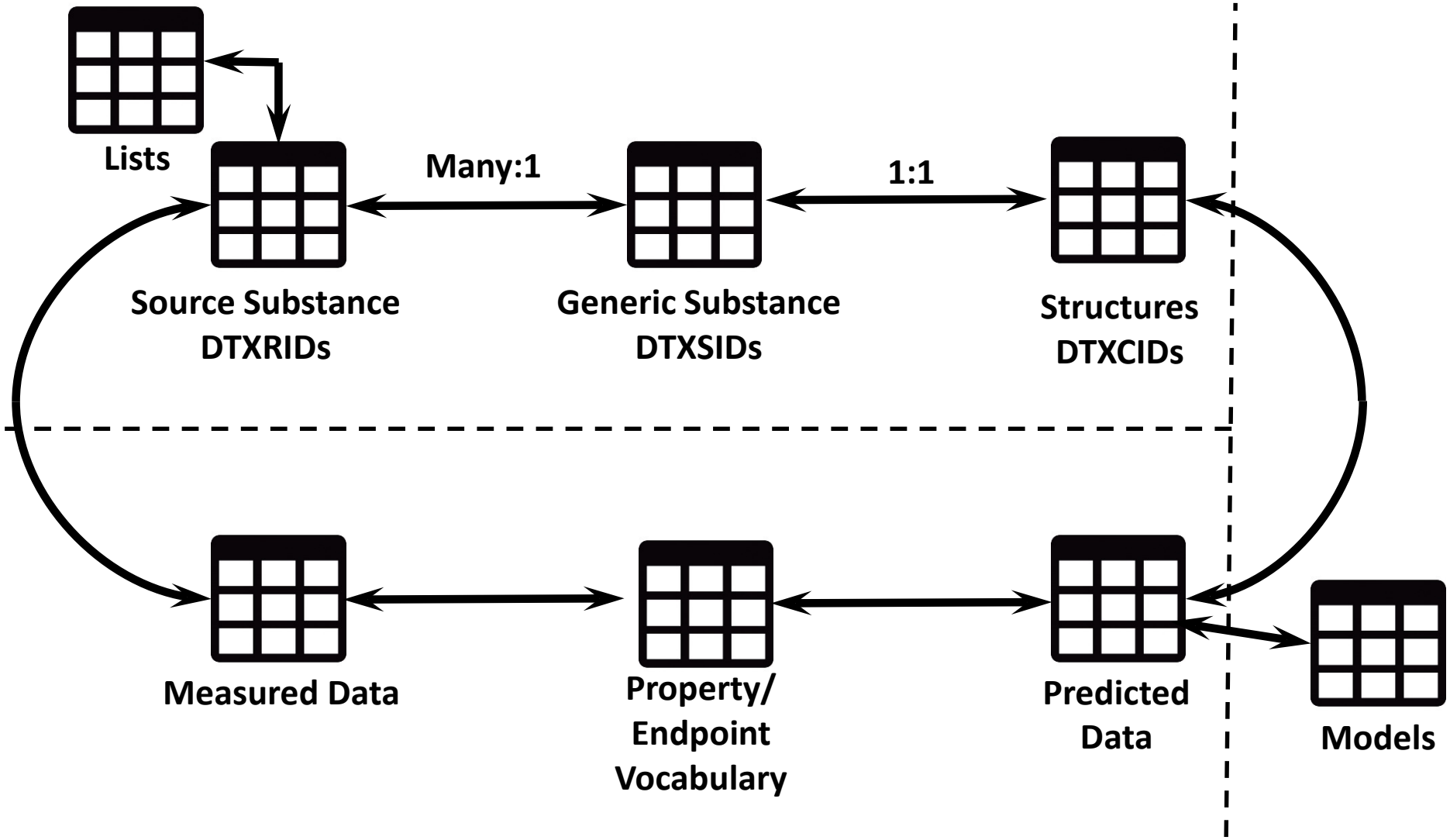
A Solution: DSSTox

Goal: Linking data to chemical structures enabling SAR

- First release of data files in 2004
- Focused on high impact sets of data
  - Carcinogenic Potency Database
  - Drinking water disinfection by-products
  - EPA's Integrated Risk Information System
  - FDA's Maximum Daily Dose dataset
  - EPA's Fat Head Minnow Toxicity dataset
  - ToxCast and Tox21 chemicals
- Currently contains: 876K records (32K manually curated)
- Check it out: https://comptox.epa.gov/dashboard

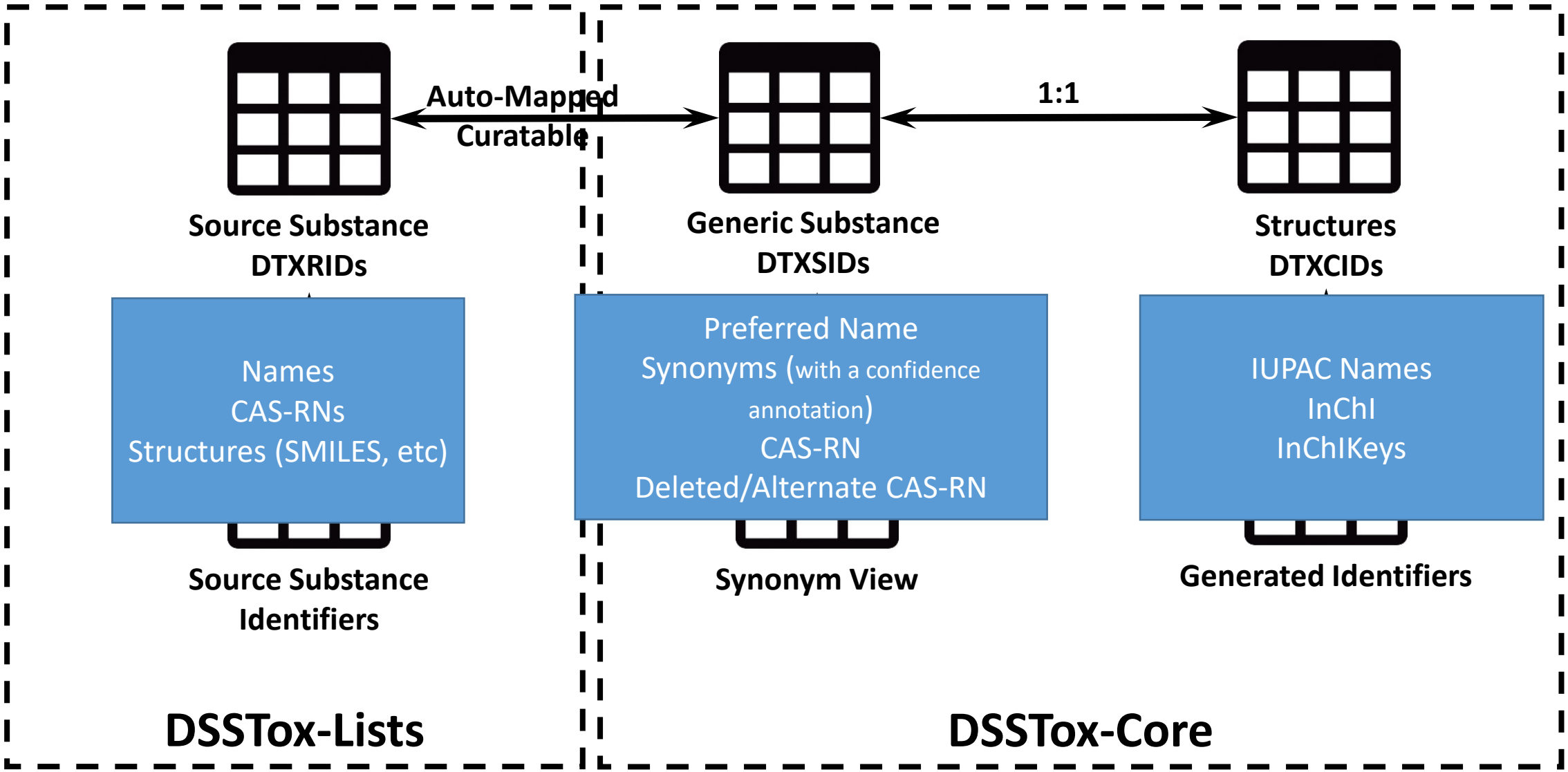


Data linkage in DSSTox



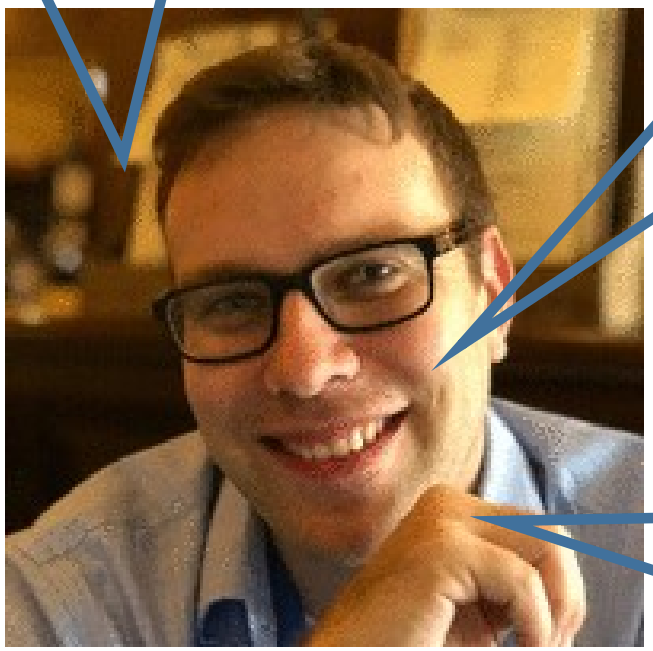
4

Chemical Identifiers in DSSTox

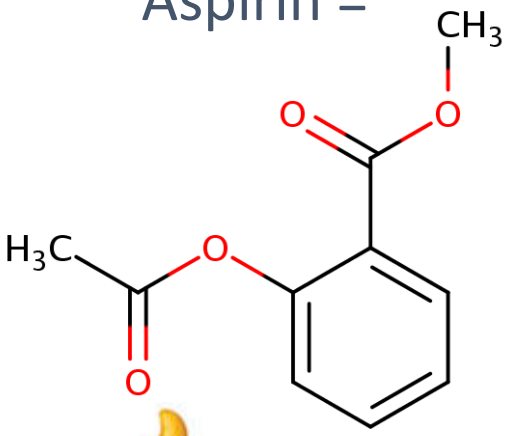


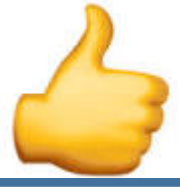
Presentation Preparation

Hmmmm… Chemical Nomenclature…



Aspirin =





Chemical Nomenclature = Names?

Presentation Preparation

Aims of chemical nomenclature

The primary function of chemical nomenclature is to ensure that a … chemical name leaves no ambiguity … each chemical name should refer to a single substance.

CAS numbers form an extreme example of names

Didn't even know what a name was…

Another system gaining popularity is the International Chemical Identifier (InChI) – which reflects a substance's structure and composition ~~making it more general than a CAS number.~~

… so most researchers simply use the informal names.

Crafting persistent identifiers and structure-based representations in DSSTox as ~~surrogates for~~ chemical names

Christopher Grulke[1]
Ann Richard[1]
Antony Williams[1]

1. National Center of Computational Toxicology, U.S. EPA
(soon to be the Center for Computational Toxicology and Exposure)

American Chemical Society Meeting, Fall 2019
26 August 2019, San Diego, FL

Chris' Understanding of Naming Terms

- Nomenclature = Structure 2 Name rules
- Lexicography = Using CompTox Chemicals Dashboard (https://comptox-prod.epa.gov/dashboard/) to look up names... or maybe some other chemical databases (probably not)

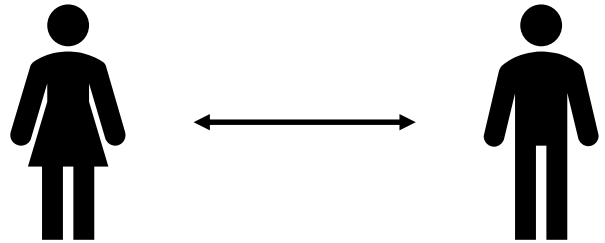
How did I figure this out???

I searched for it … so lexicography???

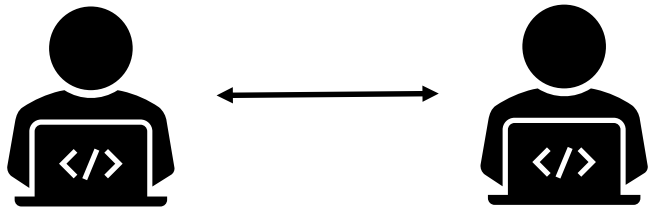
The Purpose(s) of Names: Communication

- Who Cares?
  - Chemists
  - Cheminformaticians
  - Toxicologists
  - Risk Assessors
  - Biochemists
- What do we want?
  - Chemically understandable
  - Uniqueness
  - Coverage
  - Open
  - Authoritative
  - Easy to Mint

- Mechanisms of Communication
  - Person to Person



  - Computer Mediated

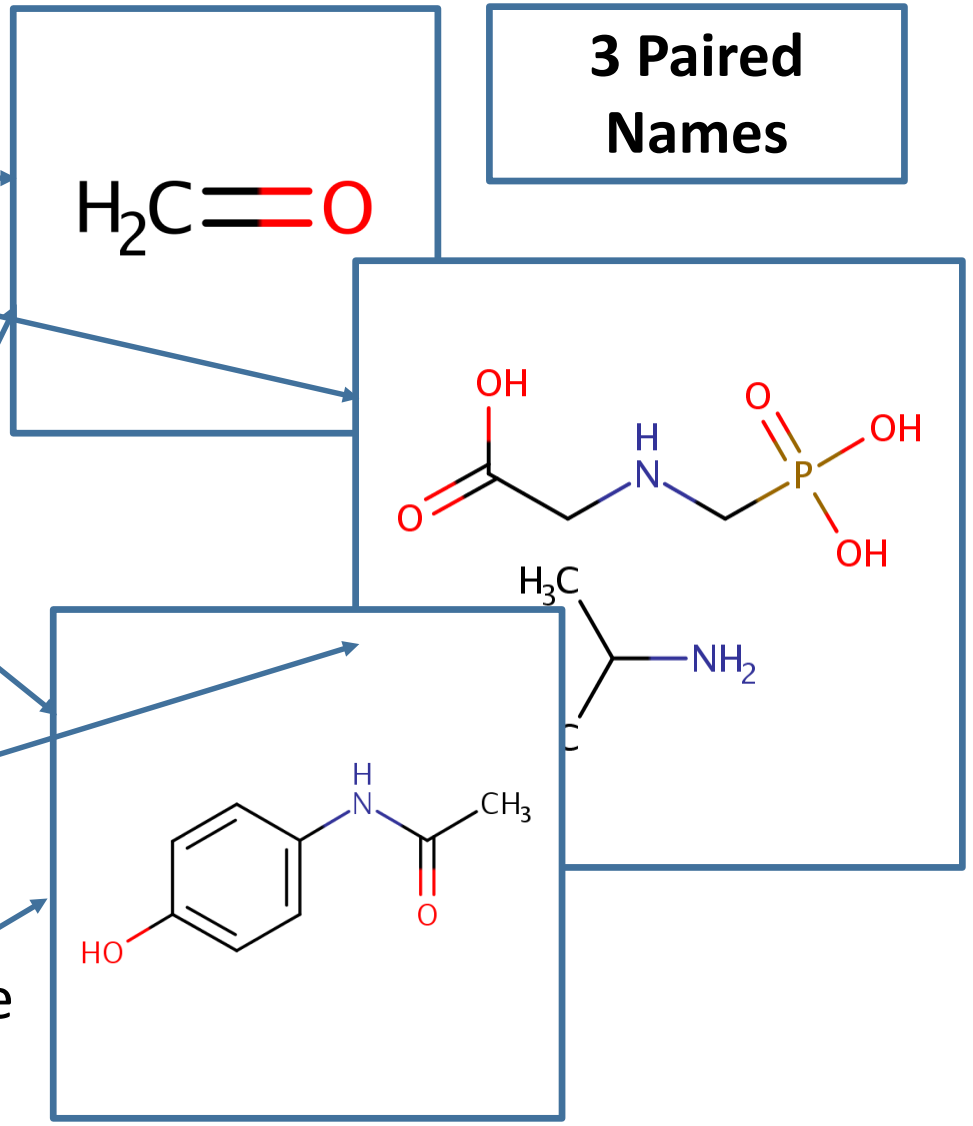


Person to Person Communication

https://sloanreview.mit.edu/article/when-communication-should-be-formal/

The Limitations of “names”

- Examples
  1. Formaldehyde
  2. RoundUp
  3. BPA
  4. Tylenol
  5. Stomoxi
  6. Atrazine
  7. Sorbitol
  8. Permethrin
  9. Glyphosate
  10. Formalin
  11. 4-hydroxyacetanilide



Used By
- Chemists
- ~~Cheminformaticians~~
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable
- Uniqueness
- Coverage
- Open
- Authoritative
- Easy to Mint

The Limitations of Systematic Names

- Examples (IUPAC taken from PubChem)
  1. propan-1-ol
  2. 4-nitrophenol
  3. *N*-(4-hydroxyphenyl)acetamide
  4. 2-(phosphonomethylamino)acetic acid;propan-2-amine
  5. 4,4,10,14-tetramethyl-17-(6-methylhept-5-en-2-yl)-1,2,3,5,6,7,11,12,13,15,16,17-dodecahydrocyclopenta[a]phenanthren-3-ol
  6. 4,4,14-Trimethyl-18-norcholesta-8,24-dien-3-ol
  7. 1,3-Bis[fluoro(dimethyl)silyl]-2,2,4,4-tetra(propan-2-yl)-1,3,2,4-diazadisiletidine
  8. (3-phenoxyphenyl)methyl 3-(2,2-dichloroethenyl)-2,2-dimethylcyclopropane-1-carboxylate

Used By
- Chemists
- ~~Cheminformaticians~~
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable*
- Uniqueness*
- Coverage
- Open
- Authoritative*
- Easy to Mint

The Limitations of CAS-RNs

- 50-00-0
- 38641-94-0
- 80-05-7
- 103-90-2
- 52645-53-1
- 1912-24-9
- 50-70-4
- 1071-83-6
- 71-23-8
- 100-02-7
- 175205-40-0
- 83312-37-2
- 48115-12-5

Used By
- Chemists
- ~~Cheminformaticians~~
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable
- Uniqueness*
- Coverage*
- Open
- Authoritative
- Easy to Mint*

Computer Mediated Communication

https://paolaespino.wordpress.com/2016/02/29/computer-mediated-communication-an-observation-of-gender-in-chat-rooms/

Limitation Reduction using a Computer

Common Names	Systematic Names	CAS-RNs
Used By	Used By	Used By
Chemists	Chemists	Chemists
Cheminformaticians	Cheminformaticians	Cheminformaticians
Toxicologists	Toxicologists	Toxicologists
Risk Assessors	Risk Assessors	Risk Assessors
Biochemists	Biochemists	Biochemists
Meets Criteria	Meets Criteria	Meets Criteria
Chemically understandable*	Chemically understandable*	Chemically understandable*
Uniqueness	Uniqueness	Uniqueness*
Coverage	Coverage**	Coverage*
Open	Open	Open
Authoritative	Authoritative	Authoritative
Easy to Mint	Easy to Mint	Easy to Mint

The Limitations of Structure Files

```
  1
  2    Mrv16111122917111132D
  3
  4     0  0  0     0  0            999 V3000
  5  M  V30 BEGIN CTAB
  6  M  V30 COUNTS 6 5 0 0 0
  7  M  V30 BEGIN ATOM
  8  M  V30 1 C 2.6618 -1.5325 0 0
  9  M  V30 2 O 2.6618 0 0 0
 10  M  V30 3 C 1.3289 -2.3107 0 0
 11  M  V30 4 N 1.3289 -3.8432 0 0
 12  M  V30 5 C 0 -1.5325 0 0
 13  M  V30 6 O 3.9909 -2.3107 0 0
 14  M  V30 END ATOM
 15  M  V30 BEGIN BOND
 16  M  V30 1 2 1 2
 17  M  V30 2 1 1 3
 18  M  V30 3 1 1 6
 19  M  V30 4 1 3 4
 20  M  V30 5 1 3 5
 21  M  V30 END BOND
 22  M  V30 END CTAB
 23  M  END
```

Used By
- Chemists
- Cheminformaticians
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable
- Uniqueness*
- Coverage**
- Open
- Authoritative
- Easy to Mint

Benzenamine, ethylenated, distn. residues

72207-55-7 | DTXSID8029022

The Limitations of InChIs

InChI String: InChI=1/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(H,5,6)
InChIKey: QNAYBMKLOCPYGJ-UHFFFAOYSA-N

Used By
- Chemists
- Cheminformaticians
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable
- Uniqueness*
- Coverage
- Open
- Authoritative
- Easy to Mint
- CANONICAL++++++

Common Names

Used By
- Chemists
- Cheminformaticians
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable*
- Uniqueness
- Coverage
- Open
- Authoritative
- Easy to Mint

Systematic Names

Used By
- Chemists
- Cheminformaticians
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable*
- Uniqueness
- Coverage**
- Open
- Authoritative
- Easy to Mint

CAS-RNs

Used By
- Chemists
- Cheminformaticians
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable*
- Uniqueness*
- Coverage*
- Open
- Authoritative
- Easy to Mint

Structures

Used By
- Chemists
- Cheminformaticians
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable
- Uniqueness*
- Coverage**
- Open
- Authoritative
- Easy to Mint

InChI

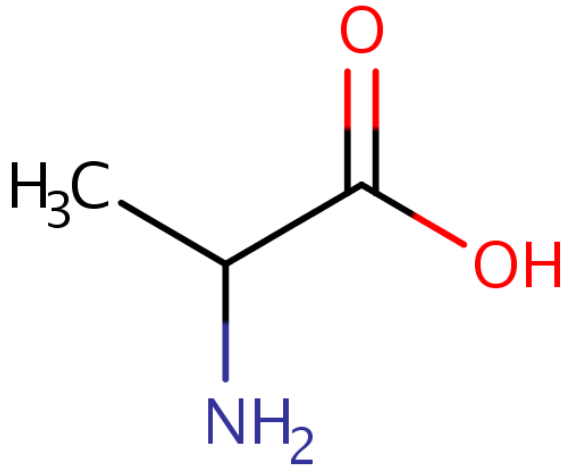
Used By
- Chemists
- Cheminformaticians
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable
- Uniqueness*
- Coverage**
- Open
- Authoritative
- Easy to Mint

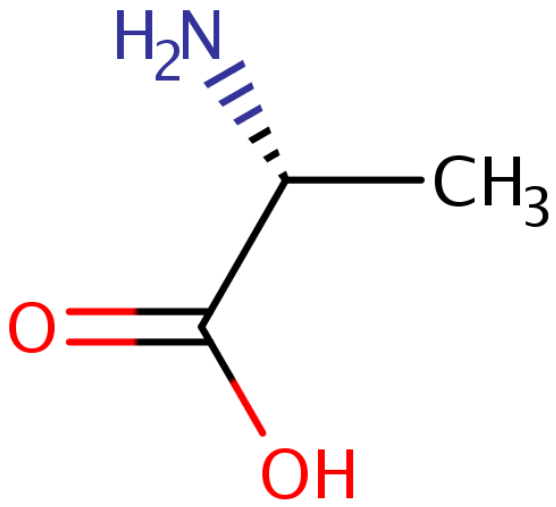
Why Not Use Them ALL???
The problems of conflicts

- An Example for the PhysProp Dataset: DTXRID202526400
- Name: ALANINE
- CAS-RN: 56-41-7

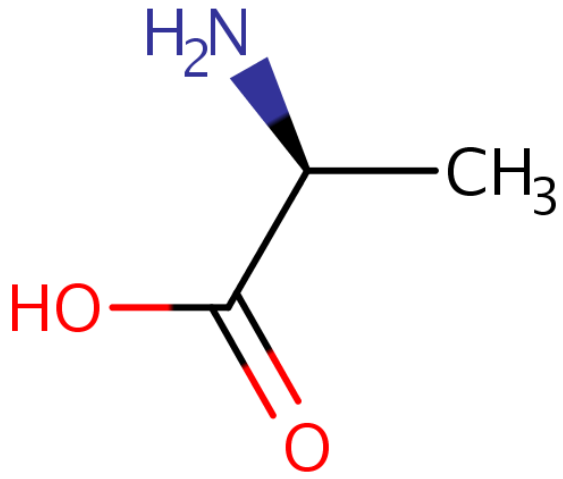
Structure (SMILES)

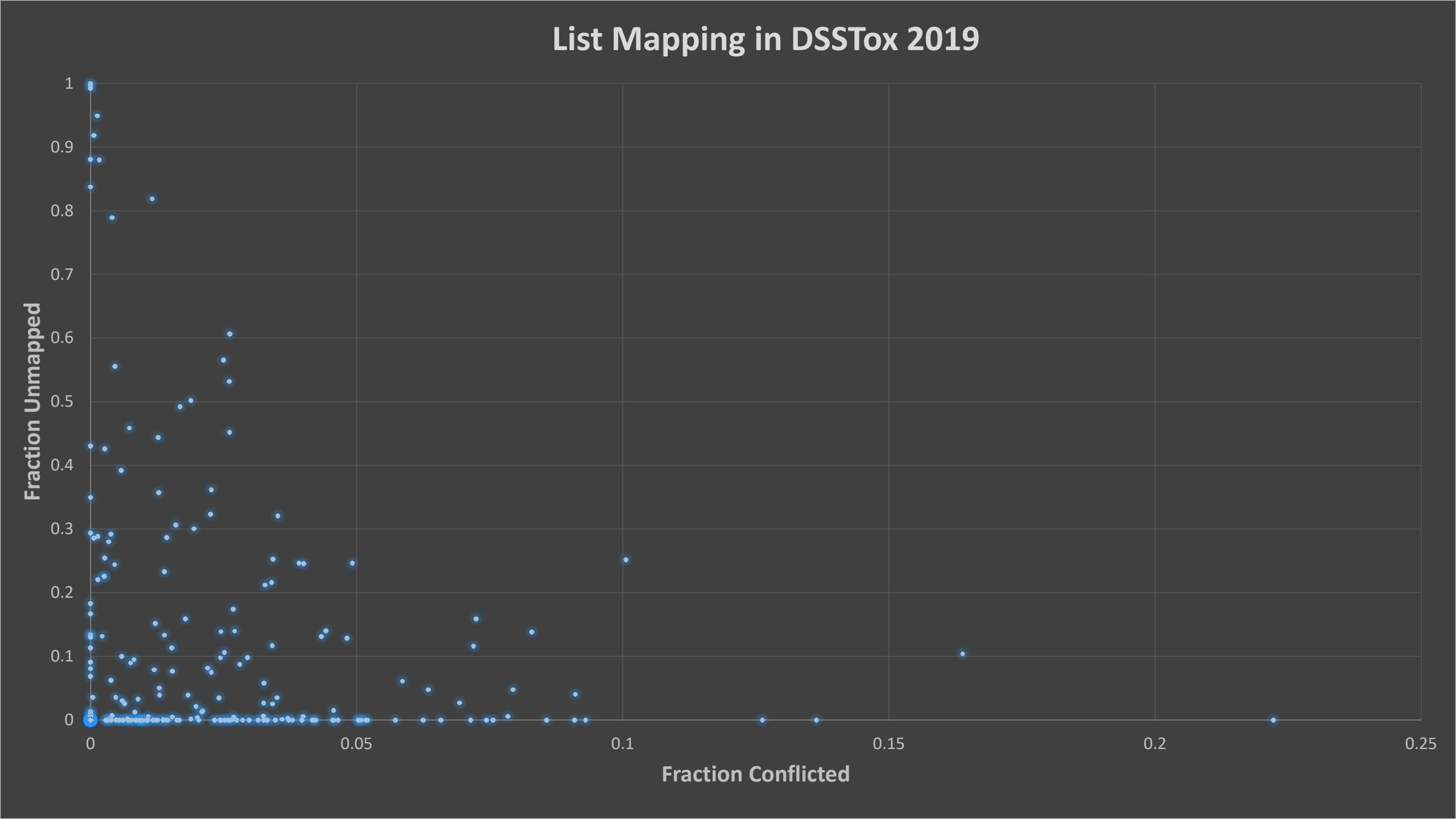


Structure (MolBlock)



Structure (Based on CAS-RN)




List Mapping in DSSTox 2019
Fraction Unmapped
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.05
0.1
0.15
0.2
0.25
Fraction Conflicted

Common Names

Used By
- Chemists
- Cheminformaticians
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable*
- Uniqueness
- Coverage
- Open
- Authoritative
- Easy to Mint

Systematic Names

Used By
- Chemists
- Cheminformaticians
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable*
- Uniqueness
- Coverage**
- Open
- Authoritative
- Easy to Mint

CAS-RNs

Used By
- Chemists
- Cheminformaticians
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable*
- Uniqueness*
- Coverage*
- Open
- Authoritative
- Easy to Mint

Structures

Used By
- Chemists
- Cheminformaticians
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable
- Uniqueness*
- Coverage**
- Open
- Authoritative
- Easy to Mint

InChI

Used By
- Chemists
- Cheminformaticians
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable
- Uniqueness*
- Coverage**
- Open
- Authoritative
- Easy to Mint

Leaning in to Lexicography>

- CAS-RNs are really names for the purpose of lexicography
- There is always a definitive meaning for the CAS-RN
- The lexicographic solution may be easier to implement that a comprehensive nomenclature
  - UNII
  - EC-Numbers
  - Etc.
- Registries require work to constantly curate substance content
- CAS needs funding to continue curation so access is restricted

EPA United States Environmental Protection Agency

Summary

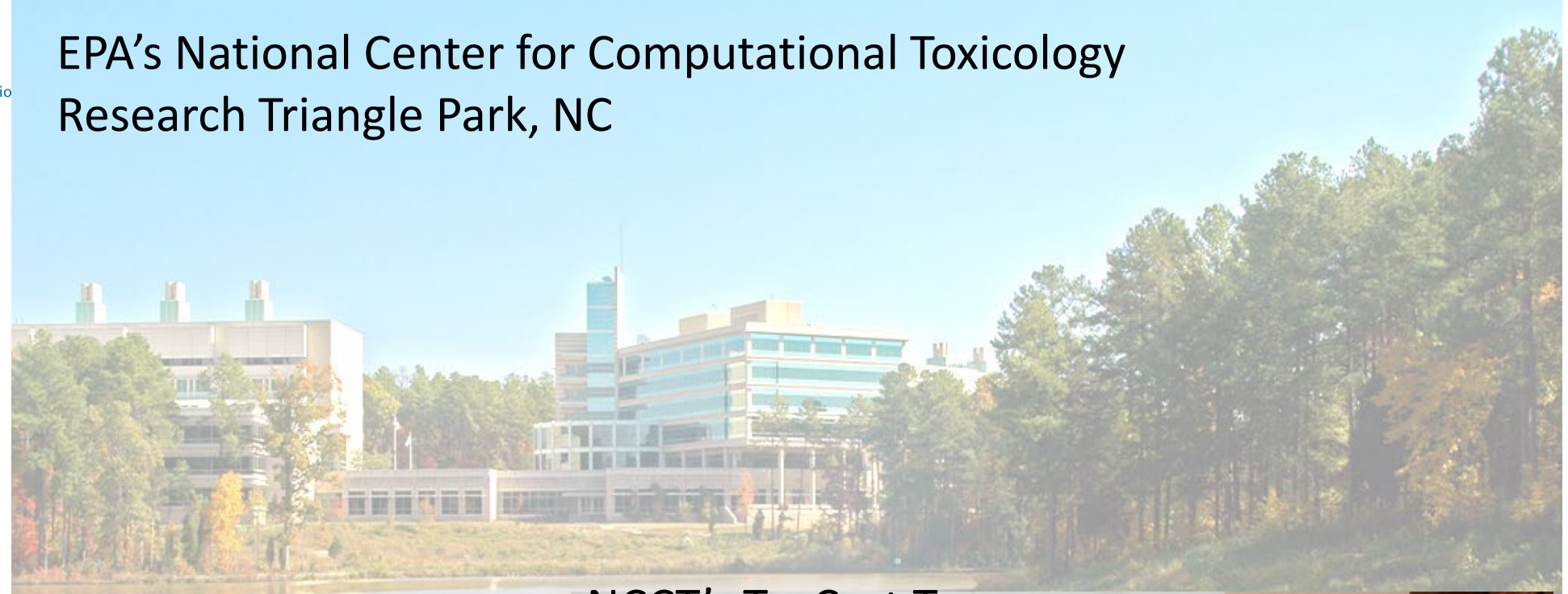
- Names in the public domain are a mess
- CAS provides an authoritative source of chemical information, but is restrictive in access
- The CAS-RN model of resolution requires work to constantly curate substance content
- Nomenclature and structure-based identifiers do not provide a solution for many chemicals of interest to EPA (Research Problem)
- Creating multiple substance registries leads to a lack of a definitive identifier (People Problem)
- Open definitive identifiers covering chemical substances requires funding for manual curation (Resource Problem)

Acknowledgements

Software Development
Jeff Edwards
Jeremy Dunne

DSSTox
Inthirany Thillainadarajah
Sakuntala Sivasupramaniam
Brian Meyer

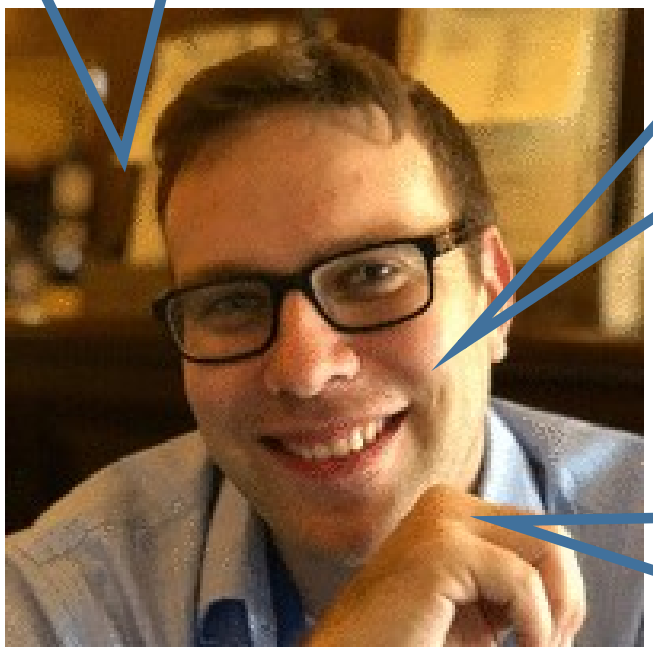
EPA's National Center for Computational Toxicology
Research Triangle Park, NC



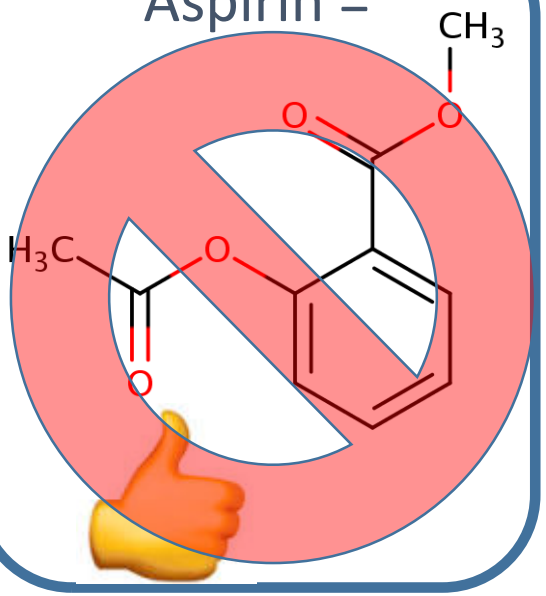
NCCT's ToxCast Team

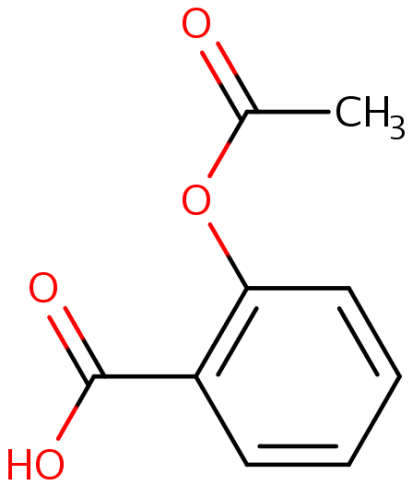
Error 1

Hmmmm… Chemical Nomenclature…



Aspirin =

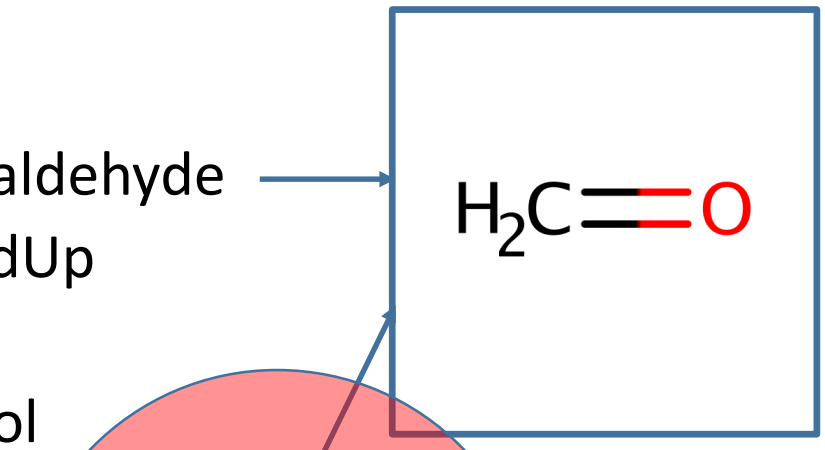




Chemical Nomenclature = Names?

Error 2

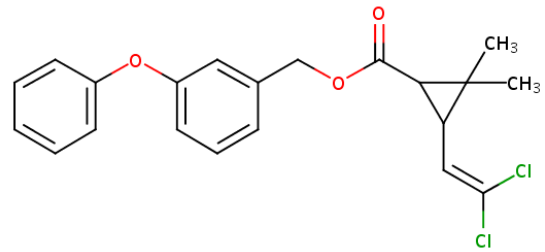
- Examples
  1. Formaldehyde
  2. RoundUp
  3. BPA
  4. Tylenol
  5. Stomoxi
  6. Atrazine
  7. Sorbitol
  8. Permethrin
  9. Glyphosate
  10. Formalin
  11. 4-hydroxyacetanilide



3 Paired Names

Used By
- Chemists
- ~~Cheminformaticians~~
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable
- Uniqueness
- Coverage
- Open



Error 3

- Examples (IUPAC taken from PubChem)
  1. propan-1-ol
  2. 4-nitrophenol
  3. *N*-(4-hydroxyphenyl)acetamide
  4. 2-(phosphonomethylamino)acetic acid;propan-2-amine
  5. 4,4,10,14-tetramethyl-17-(6-methylhept-5-en-2-yl)-1,2,3,5,6,7,11,12,13,15,16,17-dodecahydrocyclopenta[a]phenanthren-3-ol
  6. 4,4,14-Trimethyl-18-norcholesta-8,24-dien-3-ol
  7. 1,3-Bis[fluoro(dimethyl)silyl]-2,2,4,4-tetra(propan-2-yl)-1,3,2,4-diazadisiletidine
  8. (3-phenoxyphenyl)methyl 3-(2,2-dichloroethenyl)-2,2-dimethylcyclopropane-1-carboxylate

Used By
- Chemists
- ~~Cheminformaticians~~
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable*
- Uniqueness*
- Coverage
- Open
- Authoritative*
- Easy to Mint

Error 4

- 50-00-0
- 38641-94-0
- 80-05-7
- 103-90-2
- 52645-53-1
- 1912-24-9
- 50-70-4
- 1071-83-6
- 71-23-8
- 100-02-7
- 175205-40-0
- 83312-37-2
- 48115-12-5 **(Not a CAS-RN because it doesn't meet CheckSum)**

Used By
- Chemists
- ~~Cheminformaticians~~
- Toxicologists
- Risk Assessors
- Biochemists

Meets Criteria
- Chemically understandable
- Uniqueness*
- Coverage*
- Open
- Authoritative

EPA United States Environmental Protection Agency

Questions?