

From chemical name to structure: finding a noodle in the haystack

Presented at
The CAS/IUPAC Conference on Chemical
Identifiers and XML for Chemistry,
July 1, 2002



"Haystacks - End of Summer"

1890-91

Claude Monet

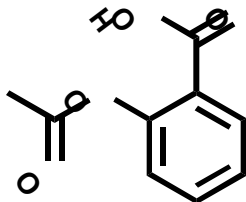
Jonathan Brecher
Director, Software Development
CambridgeSoft Corporation

CambridgeSoft

Names: The most common identifiers

“What is aspirin?”

- Chemist says:



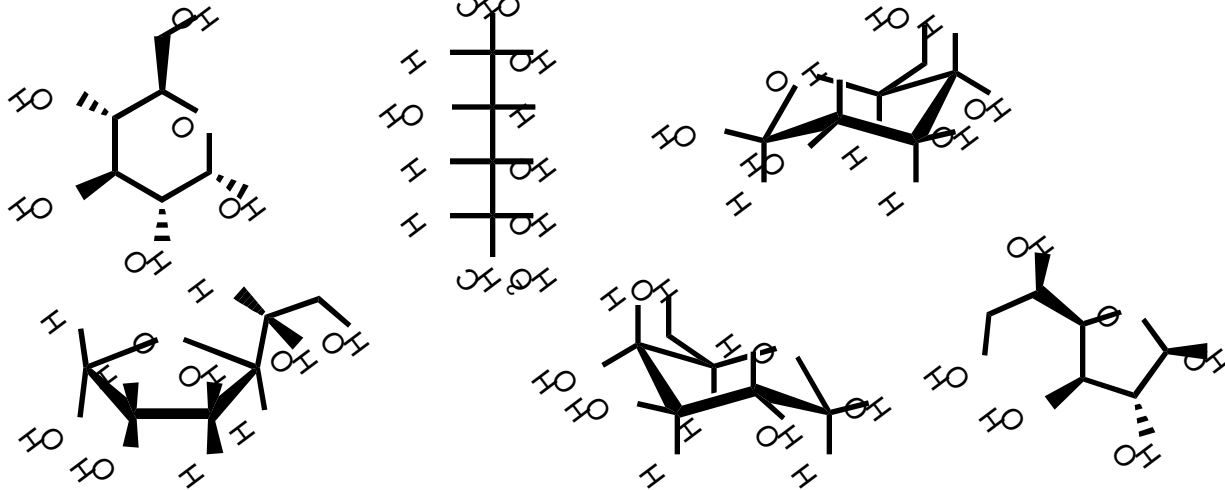
- Non-chemist says:

- Headache medicine
- Bayer makes it
- Good for the heart

Both are right!

Names are *substance* identifiers, not *structure* identifiers

“50 mg of glucose”



Don't read too much into a name

- **Two true statements:**

- Copper sulfate is a blue crystalline solid
- The CAS RN for copper sulfate is 7758-98-7

TRUE

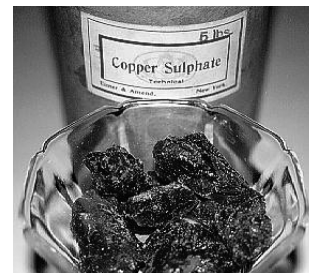
TRUE

- **...combine to make one false one:**

- CAS RN 7758-98-7 is a blue crystalline solid

FALSE!

- **Only the hydrated form is blue
(copper sulfate, pentahydrate
= CAS RN 7758-99-8)**



<http://www.chss.montclair.edu/~pererat/0000d.jpg>

So why do we care?

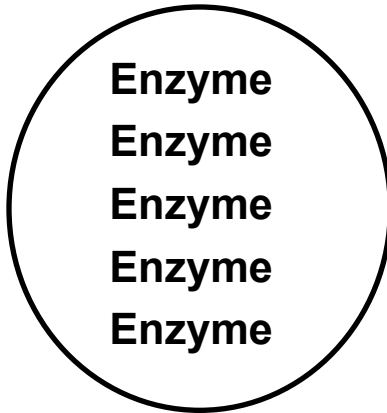
- **Sometimes names are the only identifier available**
 - Lots of existing data available in no other form
 - “Better” identifiers available only from trained chemists
- **People like names!**

The challenge:

When you have to interpret a name, how to best extract as much information as possible from it?

Names aren't always meaningful

<u>Catalog Num</u>	<u>Product Name</u>	<u>Price</u>
...		
84289	1-Eicosene	\$50/25g
...		
12345	Enzyme	\$100/g
23456	Enzyme	\$5/mg
34567	Enzyme	\$32/g
45678	Enzyme	\$70/g
56789	Enzyme	\$100/mg

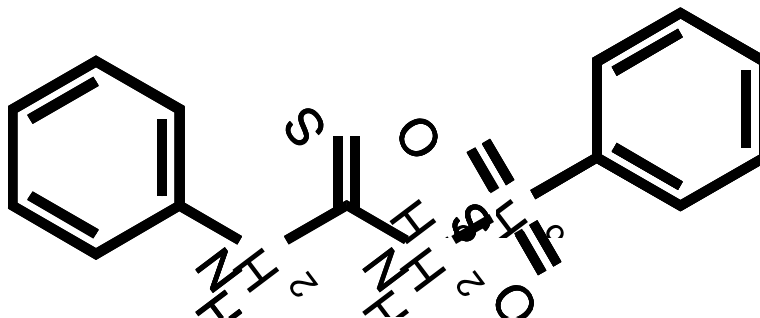


Names aren't always sensical

<u>Catalog Num</u>	<u>Product Name</u>	<u>Price</u>
...		
84289	5-Aminosalicylic acid	\$100/100g
84289	α -Aminosuberic acid	\$150/50mg
...		
12345	5-Amino-2-sulfonic acid	\$15/25g

Names aren't always reasonable

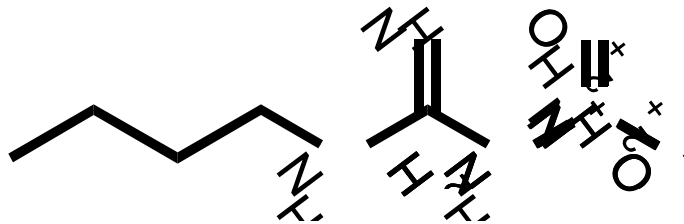
[(anilinothioyl)amino](dioxo)phenyl-lambda6-sulfane



N-phenyl-*N'*-(phenylsulfonyl)thiourea

Names aren't always reasonable, 2

2-[(butylamino)(imino)methyl]-1-oxohydrazinium-1-olate



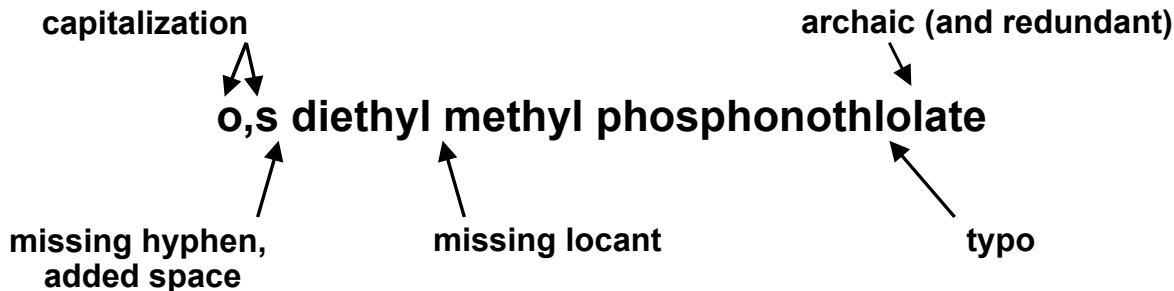
N-butyl-*N'*-nitroguanidine

“Why can't I find any information?”

“I've been trying to look for the structure (and CAS number) of o,s diethyl methyl phosphonothiolate

I've looked on the free online databases as well as the CAS substance index, but I could'nt find anything.”

-- posted to the CCL mailing list May 21, 2002



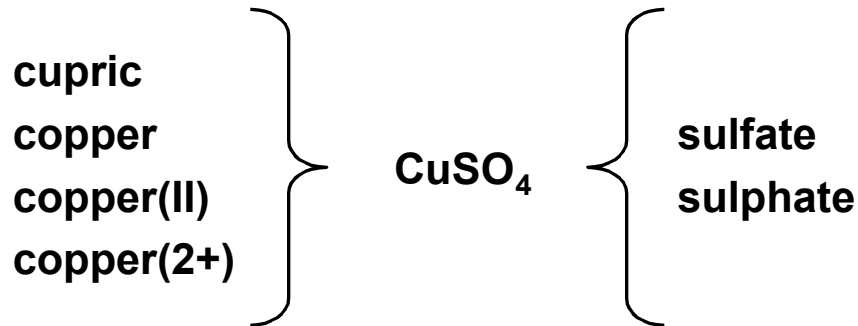
Best approach for dealing with chemical names

- **Assume the name was intended to be accurate**
 - Recognize standard nomenclature systems (IUPAC, CAS)
 - Be flexible about minor typographical variations (spaces, capitals)
 - Don't second-guess! (benzyne and benzine aren't necessarily typos for benzene)

- **If the name can't be interpreted as provided, then be a little more flexible**
 - But only if the original name couldn't be interpreted!
 - (Assumes a very accurate interpreter for names)

- **Don't expect perfection**

Many naming conventions



Other complications

- **Capitalization**
 - copper sulfate, Copper sulfate, Copper Sulfate, COPPER SULFATE
- **Spacing**
 - copper sulfate, coppersulfate
- **Punctuation**
 - copper sulfate, “copper, sulfate”
- **Parentheses**
 - copper(II) sulfate, copper[II] sulfate
- **“Commentary”**
 - Copper sulfate, 99%

Copper sulfate searches on ChemFinder.com

Normalized names Structures from typos



Exact Structures from names

70 copper sulfate
 12 copper sulfate
 11 cupric sulfate
 8 Cooper Sulfate
 8 copper sulphate
 7 copper (ii) sulfate
 7 copper 2 sulfate
 7 copper 2 sulphate
 7 coppersulfate
 6 Kupfersulfat
 5 Cooper Sulphate
 5 copper(2) sulfate
 4 *copper sulfate*
 4 cofer sulfate
 4 copper sulfate
 4 copper sulfate
 4 copper sulfate*
 4 coppersulfate
 4 cuprum sulfate
 3 Copper (2) sulfate
 3 copper(2)sulphate
 3 coppersulphate
 3 cuperic sulfate
 3 cupric sulfate crystals
 3 kopersulfat
 2 *coppersulfate*
 2 cooper (ii) sulfate
 2 copper sulfat
 2 copper sulfate
 2 copper sulphate crystals
 2 copper sulphate
 2 copper(11) sulphate
 2 Copper(2)Sulfate
 2 copper(ii) sulfate

2 Copper2Sulfate
 2 coppersulfate
 2 coppersulfat
 2 Cupric (II) sulfate
 2 cupric sulphate
 2 curpic sulfate
 2 Kupfer(II)-Sulfat
 2 sulfate copper
 2 sulfate cupric
 1 *copper(ii)sulphate*
 1 *coppersulfate*
 1 blue copper sulfate
 1 cupric sulphate
 1 cobber*sulfat
 1 cobbersulfat
 1 cobbersulfate
 1 Cooper (II) sulfate anhydrous
 1 cooper+sulfate
 1 copper sulfate
 1 copper(II) sulfate
 1 copper sulfate
 1 Copper (11) Sulfate
 1 copper (11) sulphate
 1 copper (2) sulphate
 1 copper (2)sulfate
 1 Copper (II) sulphate
 1 copper (ii) sulphate
 1 copper (ii)sulphate
 1 copper 11 sulfate
 1 copper and sulfate
 1 copper II sulfate
 1 COPPER SELFATE
 1 Copper Silphate

1 copper sulphate
 1 copper slate
 1 copper sifate
 1 copper sulfate
 1 copper sulfate
 1 copper sulfate crystal
 1 copper sulfate crystals
 1 copper sulfate II
 1 Copper Sulfate Powder
 1 copper sulfate structure
 1 copper sulfate(4)
 1 copper sulfate
 1 copper sulfated
 1 copper sulfates
 1 copper sulfatr
 1 copper sulfayte
 1 copper sulfate
 1 copper sulfate
 1 copper sulfste
 1 copper sulfte
 1 copper sulfate
 1 copper sulfate
 1 copper sulphare
 1 Copper sulphate 0.1M
 1 copper sulphate II
 1 Copper sulphate physical data
 1 copper sulphate supplier
 1 copper sulphate(VI)
 1 copper sulphate*
 1 copper sulphate
 1 copper(2) sulphate
 1 copper(20)sulphate
 1 copper(ii) sulphate
 1 copper(II) sulphate(VI)

1 copper(ii)sulfate
 1 copper(II)sulfate
 1 copper+ sulphuric acid
 1 copper+sulfate
 1 copper+sulfuric acid
 1 copper+sulphate
 1 copper-2-sulphate
 1 copper-sulphate crystals
 1 copper2 sulfate
 1 copper2sulphate
 1 coppre sulfate
 1 COPPSULPHATE
 1 coprum sulfuricum
 1 COPSULPHATE
 1 cu2+sulfate
 1 cubber(II)sulfate
 1 cubbersulfate
 1 cubric sulfate
 1 cuperic sulphate
 1 cupfersulphate
 1 cupic sulfate
 1 cupic sulfate
 1 cuppe sulphate
 1 copper (II) sulfate
 1 copper sulfure
 1 copper sulphate
 1 cupperic sulfate
 1 coppersulphat
 1 coppersulphate
 1 Cupric (copper)Sulfate
 1 Cupric sulfate
 1 cupric sulfate
 1 cupric sulfathe

1 cupric sulfhate
 1 cupric(I)sulfate
 1 cupric+sulfate
 1 cupris sulfate
 1 cuprisulfat
 1 cuprit sulfate
 1 cuprite sulfate
 1 cuprium sulphate
 1 cupro sulfate
 1 curite sulfate
 1 curpic sulphate
 1 curpic sulfate
 1 Dicopper sulfate
 1 distilling Cuperous sulfate
 1 EPA Registered Copper Sulfate
 1 granular copper sulphate
 1 hazards of copper sulphate
 1 kobber(II)-sulfat
 1 koppersulfat
 1 Kupfer(II)sulfat
 1 kupric sulfate
 1 sulfate cooperate
 1 sulfate cupic
 1 sulfato cobre
 1 sulfato cuprico
 1 sulfato de cobre
 1 sulfato de cobre 2
 1 sulfuric cooperate

Copper sulfate, hydrated

40 hydrated copper sulfate	2 cupric sulfate pentahydrate	1 Copper Sulfate octahydrate	1 copper sulphate pentahydrate	1 cupric sulfate pntahydrate
19 copper sulfate hydrate	2 hydrated copper (II) sulfate	1 copper sulfate p*	1 copper sulphate pentra hydrate	1 cupric sulfate quadrahydrate
16 hydrated cupric sulfate	2 hydrous cupric sulfate	1 copper sulfate penta hydrate	1 copper sulphate pentrahydrate	1 cupric sulfate-5h2O
9 hydrated copper sulphate	1 "copper sulfate hydrate"	1 copper sulfate penta-hydrate	1 copper sulphate pentrahydrate	1 cupric sulfatte pentahydrate
8 copper 2 sulfate pentahydrate	1 aqueous copper sulfate	1 Copper Sulfate penta-water	1 copper sulphate septahydrate	1 cupric sulphate pentahydride
8 copper sulfate pentahydrate	1 aqueous copper(II)sulfate	1 Copper Sulfate Pentahidrate	1 copper sulphate solution	1 cuproc sulfate pentahydrate
8 Copper(II) sulfate hydrate	1 cooper sulfate pentahyd	1 copper sulfate pentahydattte	1 copper(2)sulfate pentahydrate	1 cuprous sulfate pentahydrate
5 copper sulfate solution	1 coper (2) sulfate pentahydrate	1 copper sulfate pentahydiat	1 copper(I) sulfate pentahydrate	1 hydrated copper sulphate
5 hydrous copper sulfate	1 copper sulfate penahydrate	1 copper sulfate pentahydrae	1 copper(II) sulfate dihydrate	1 hydrate copper sulfate
4 cupric sulfate hydrate	1 copper (2) sulfate hydrate	1 copper sulfate pentahydrous	1 copper(II) sulphate hydrate	1 hydrate copper sulphide
3 copper (II) Sulfate hydrate	1 copper (II) sulfate dihydrate	1 copper sulfate pentahydrate	1 Copper(II)Sulfate Decahydrate	1 hydrate copper(II) sulfate
3 copper ii sulfate pentahydrate	1 copper (II) sulphate (hydrate)	1 copper sulfate pentanhydride	1 copper(II)sulfate Monohydrate	1 hydrated Copper II Sulfate
3 copper sulfate dihydrate	1 copper 1 sulfate pentahydrate	1 Copper Sulfate Pentahidrate	1 copper2 sulphate pentahydrate	1 hydrated copper(II) Sulphate
3 copper sulfate heptahydrate	1 copper I sulphate pentahydrate	1 Copper Sulfate Pentahidrate§	1 copperas heptahydrate	1 hydrated cupric sulphate
3 copper(2) sulfate pentahydrate	1 copper II sulfate penahydrate	1 copper sulfate pentahydrate	1 copperII sulfate penahydrate	1 hydrated cuprous sulfate
3 cupric sulfate penta hydrate	1 copper II sulfate petahydrate	1 Copper Sulfate Tetra Hydrate	1 copperIIsulfatepentahydrate	1 hydratedcoppersulphate
3 hydrated copper(II) sulfate	1 copper II sulphate hydrate	1 Copper Sulfate(aq)	1 coppersulfate heptahydrate	1 hydrdated copper sulfate
2 copper sulfate aqueous	1 copper pentasulfate	1 copper sulfate(hydrous)	1 coppersulfate+5H2O	1 hydrolyzed copper sulfate
2 copper sulfate pentahydride	1 copper sufate pentahydrate	1 copper sulphare monohydrate	1 coppersulfatepentahydrat	1 Hydrous copper sulphate
2 copper sulfate penthydrate	1 Copper Sulfate + 5 Water	1 copper sulphate hdrate	1 coppersulfteperwater	1 Kupfersulfat-Pentahydrat
2 Copper Sulfate Tetrahydrate	1 copper sulfate 6hydrate	1 copper sulphate heptahydrate	1 cuppic sulfate monohydrate	1 kupfersulfatepentahydrat
2 copper sulfate trihydrate	1 Copper Sulfate decahydrate	1 copper sulphate penta hydrated	1 cupric pentasulfate	1 pentahidrate copper sulfate
2 copper sulfate water	1 Copper Sulfate Di-hydrate	1 copper sulphate penta hydrous	1 cupric sulfate dihydrate	1 pentahidrated copper sulfate
2 copper sulphate hydrate	1 copper sulfate five hydride	1 copper sulphate penta-hydrated	1 cupric sulfate hydrated	1 pentahidrated sulfate copper
2 copper sulphate penta hydrate	1 copper sulfate hydrated	1 Copper sulphate pentahidrat	1 cupric sulfate monohydrate	1 Penthydrous Copper Sulfate
2 copper2 sulfate pentahydrate	1 copper sulfate hydrous	1 Copper Sulphate Pentahidrate	1 cupric sulfate pentahdrate	
2 cupric sulfate heptahydrate	1 Copper Sulfate nonahydrate	1 copper sulphate pentahydrate	1 cupric sulfate pentahydrated	

Other copper-and-sulfur searches

copper (I) sulfate

20 Copper (I) sulfate
 19 Cuprous Sulfate
 13 copper(I) sulfate
 6 copper I sulfate
 5 Copper (1) sulfate
 3 copper(1) sulfate
 3 copper(I) sulphate
 3 copper(I)sulfate
 2 cyprous sulfate
 1 copper (1) sulfide
 1 copper (I) and sulfate
 1 copper (I) sulphate
 1 copper 1 sulphate
 1 copper I sulphate
 1 COPPER SULFATEI
 1 copper(1) sulphate
 1 copper(I)\ sulphate
 1 cuperous sulfate
 1 Cuprous Sulfate
 1 cupro(I)sulfate

copper bisulfate

3 Copper disulfate
 2 COPPER sulfate basic
 2 copper(I) hydrogen sulfite
 2 copper(II) hydrogensulfate
 2 CUPRIC SULFATE BASIC
 2 cuprous bisulfate
 1 copper bisulfate pentahydrate
 1 copper bisulphate
 1 COPPER HYDRAZIDE SULFATE
 1 copper hydrogen sulfite
 1 Copper hydrogensulfite
 1 copper sulfate acidic
 1 copper(I) hydrogen sulfate
 1 COPPERsulfatebasic
 1 cupric bisulfate
 1 hydrogen copper sulphate
 1 tribasic copper sulfate
 1 Tribasic Copper Sulphate

copper sulfide

2 copper(III) sulfide
 2 coppo sulfide
 2 cuprous sulfide
 1 "cupric sulfide"
 1 cooper (II) sulfide
 1 cooper 2 sulfide
 1 cooper disulfide
 1 cooper II sulfide
 1 cooper sulfide
 1 cooper(II)sulfide
 1 cooper(II)sulfide
 1 cooperic sulfide
 1 copersulfide
 1 copper (I) sulfider
 1 copper (ii) hydrogensulfide
 1 copper di-sulfide
 1 copper disulfide
 1 copper hydrogensulfide
 1 copper hydrogensulphide
 1 copper hydrosulfide
 1 copper sulfide
 1 Copper Sulfde
 1 copper sulfide
 1 Copper Sulfide Ore
 1 copper sulfied
 1 copper tetrasulfide
 1 copper(2)sulfide
 1 copper2 sulfide
 1 copperic sulfide
 1 coppersulfied
 1 copperus sulfide
 1 cooperic sulfide
 1 Kupfersulfid

copper sulfite

34 copper sulfite
 7 cupric sulfite
 5 copper (I) sulfite
 4 COPPER(I) SULFITE
 4 cuprous sulfite
 3 Copper II sulfite
 2 Copper(II) sulfite
 2 Coppersulfit
 1 Copper (I) sulfite
 1 copper (1) sulfite
 1 copper (II) sulfite
 1 copper 1 sulfite
 1 copper I sulfite
 1 Copper II sulfite
 1 Copper II sulfite pentahydrate
 1 Copper II sulfite pentahydrate
 1 copper sulphite
 1 Copper(I)Sulfite
 1 Copper(II)sulfite
 1 copper* sulfite
 1 copper1 sulfite
 1 coppersulfite
 1 curpric sulfite
 1 sulfite cuprous

other

4 copper persulfate
 3 copper III sulfate
 3 copper sulfer
 2 copper sulfur
 1 copper (III) sulfate
 1 copper + sulfur
 1 copper 3 sulfate
 1 COPPER PYROSULFATE
 1 copper sulfate anhydride
 1 copper sulfon
 1 copper sulfonate
 1 Copper(III) Sulfate
 1 coppersulfur
 1 cupric persulfate
 1 cuprous persulfate
 1 cuprous sulfonate

Second-guessing can be embarrassing

CAS/IUPAC Conference on Chemical
Identifiers and XML for Chemistry
1 July 2002, Columbus, Ohio, USA

The conference will cover the following topics:

- From chemical name to structure: finding a needle in the haystack



Chemistry International, 2002, Vol 24, No. 2 23

Case study: Google

- If possible, answer the question that was asked



Category: [Business](#) > [Industries](#) > [Wholesale](#) > [Materials](#) > [Chemicals](#)

EXTOXNET PIP - COPPER SULFATE

... **Copper sulfate**. Trade and Other Names: **Copper sulfate** is also called Agritox, Basicap, BSC **Copper** Fungicide, CP Basic **Sulfate** and Tri-Basic **Copper Sulfate**. ...

ace.ace.orst.edu/info/extoxnet/pips/coppersu.htm - 14k - [Cached](#) - [Similar pages](#)

Copper Sulfate

Case study: Google

- If not possible to answer the direct question, use some intelligence to try to figure it out



Google [Advanced Search](#) [Preferences](#) [Lan](#)

copper sulfate

Google Search

[Web](#) [Images](#) [Groups](#) [Directory](#)

Searched the web for **copper sulfate**. Results 1 - 10 of about 108,000.

Your original search: **coppir sulfate** returned **zero results**.
The alternate spelling: **copper sulfate** returned the results below.

Category: [Business](#) > [Industries](#) > [Wholesale](#) > [Textile Materials](#) > [Res](#)

[EXTOXNET PIP - COPPER SULFATE](#)
... **Copper sulfate**. Trade and Other Names: **Copper sulfate** is also called Anritox Basican

Case study: Google

- Ideally, look for “unusual” terms and try to offer alternatives – but always answer the question asked, regardless
 - May be impossible for chemical names in the general case!

Advanced Search Preferences Lan

Google™ coppr sulfate

Google Search




Web Images Groups Directory

Searched the web for **coppr sulfate**. Results 1 - 10 of about 26. Search

Did you mean: **copper sulfate**

[PDF] [Proceedings of the 3rd AEC Air Cleaning Conference, WASH-170](#)
File Format: PDF/Adobe Acrobat - [View as HTML](#)
... As shown in Table 1, tests 1 and 3 with atmospheric dust and tests 5, 6, 8 and 9 with **coppr sulfate** indicate average efficiency increasos of 15 and 11 portent ...
[tis.eh.doe.gov/hepa/Nureg_3rd/262.pdf](#) - [Similar pages](#)

Names are often incorrect

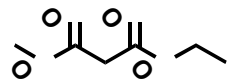
Address  http://www.google.com/jobs/britney.html  Go  Links >>

The data below shows some of the misspellings detected by our spelling correction system for the query [britney spears], and the count of how many different users spelled her name that way. Each of these variations was entered by at least two different unique users within a three month period, and was corrected to [britney spears] by our spelling correction system (data for the correctly spelled query is shown for comparison).

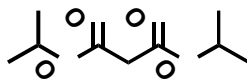
488941 britney spears	29 britent spears	9 brinttany spears	5 brney spears
40134 brittany spears	29 brittnany spears	9 britanay spears	5 broitney spears
36315 brittney spears	29 britttany spears	9 britinany spears	5 brotny spears
24342 britany spears	29 btiney spears	9 britn spears	5 bruteny spears
7331 britny spears	26 birttney spears	9 britnew spears	5 btiyney spears
6633 briteny spears	26 breitney spears	9 britneyn spears	5 btrittney spears
2696 britteny spears	26 brinity spears	9 britrney spears	5 gritney spears
1807 briney spears	26 britenay spears	9 brtiny spears	5 spritney spears
1635 brittny spears	26 britneyt spears	9 brtittney spears	4 bittny spears
1479 brintey spears	26 brittan spears	9 brtny spears	4 bnritney spears
1479 britanny spears	26 brittne spears	9 brytny spears	4 brandy spears
1338 britny spears	26 btittany spears	9 rbitney spears	4 brbritney spears
1211 britnet spears	24 beitney spears	8 birtiny spears	4 breatiny spears
1096 britiney spears	24 birteny spears	8 bithney spears	4 breetney spears
991 britaney spears	24 brightney spears	8 brattany spears	4 bretiney spears
991 britnay spears	24 brintiny spears	8 breitny spears	4 brfitney spears
811 brittney spears	24 britanty spears	8 breteny spears	4 briattany spears
811 brtiney spears	24 britenny spears	8 brightny spears	4 brieteny spears
664 birtney spears	24 britini spears	8 brintay spears	4 briety spears
664 brintney spears	24 britnwy spears	8 brinttey spears	4 briitny spears
664 briteney spears	24 brittni spears	8 briotney spears	4 briittany spears
601 bitney spears	24 brittnie spears	8 britanys spears	4 brinie spears

Chemical names are more difficult to interpret than regular English

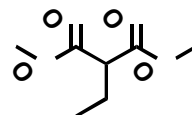
- Lots of very similar “words” used in similar contexts
 - -ane / -ene / yne
 - -ol / al / yl
 - methyl / ethyl / menthyl
- Even spaces change the meaning of a name



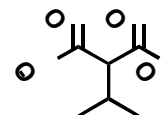
methyl ethyl malonate



methylethyl malonate



methyl ethylmalonate



methylethylmalonate

- Google can use some techniques (Soundex, etc.) that don't work with chemical names

Existing systems

- **CambridgeSoft**
 - ChemDraw Ultra, and batch version
- **ACD/Labs**
 - ACD/Name, and batch version
- **[MDL]**
 - In development, announced in 2001
- **[Chemical Abstracts]**
 - Mentioned in several journal articles, but not publicly available

Benchmarks for Name Structure interpretation

- **>> 90% of organic nomenclature rules**
 - Depends on how you count the rules!
- **Can generate structures for 70-90% of most real-life lists of chemical names**
 - Remainder are generally not systematic and/or have no structure that could possibly be generated
- **> 99% accurate for the structures that are generated**
 - Remainder are generally ambiguous names to start with
- **Can process > 10,000 names/minute**
 - Pentium III, 933 MHz

Future growth: Algorithms

- **Not much!**
- **Could implement a few more rules**
 - All remaining ones are obscure
 - Would have little practical effect
- **Support for some classes currently disabled**
 - Not a limitation of the name interpretation, but difficult to generate appealing structures: metallocenes, fullerenes, etc

Future growth: Intelligence

- Typo recognition
 - Often difficult to guess what was intended
 - Can be very slow
- Context recognition

[New Query](#)[Edit Query](#)[Simple Query](#)[Browse Reactions](#)[Preferences](#)[Submit Comment](#)[Printer-Friendly Format](#)

Record 10 of > 100 hits

<<

<

Rec#

>

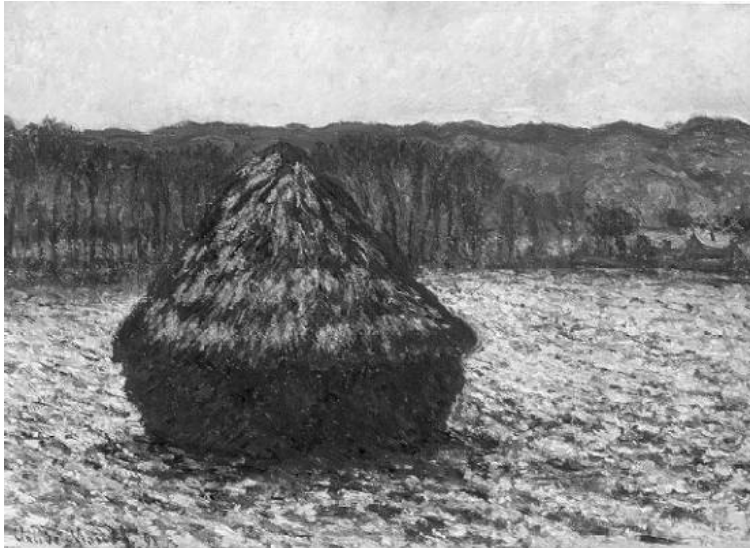
>>

[Get More Hits](#)[Return to List](#)

1. Procedure

A. *trans*-2-(2-Propenyl)cyclopentanol. A 500-mL, three-necked, round-bottomed flask equipped with a magnetic stirring bar, reflux condenser with a stopcock, and a 250-mL addition funnel is charged with 18.3 g (750 mmol) of magnesium turnings (Note 1). The system is evacuated and placed under argon, then 100 mL of ethyl ether (Note 2) is added to the system via cannula. The system is placed in an ice-water bath, and 2 mL of allyl bromide (Note 3) is added via syringe to the magnesium suspension to initiate Grignard formation. The addition funnel is charged with 45.5 g (375 mmol) of allyl bromide and 30 mL of ethyl ether. Another 100 mL of ethyl ether is added to the reaction flask. Stirring is begun, and the allyl bromide-ethyl ether mixture is added dropwise to the cooled reaction flask over a period of about 2 hr. After the addition is complete, the dark-gray

- Integration with other types external resources



“Haystack” 1890-91 Claude Monet