

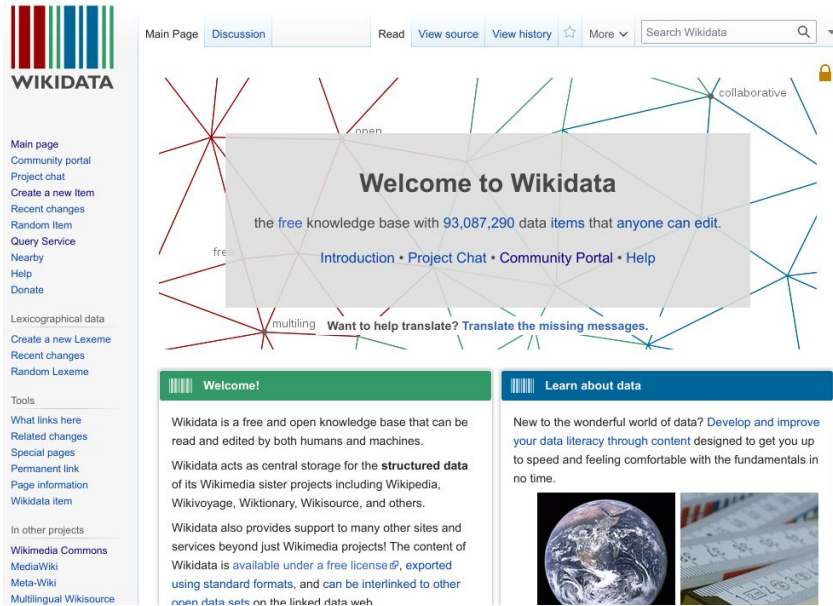
InChI and InChIKey in Wikidata and Scholia

Egon Willighagen
NIH Virtual Workshop on InChI
March 22-24, 2021

@egonwillighagen
0000-0001-7542-0286



Wikidata and Scholia



Wikidata is a free and open knowledge base that can be read and edited by both humans and machines. Wikidata acts as central storage for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others. Wikidata also provides support to many other sites and services beyond just Wikimedia projects! The content of Wikidata is available under a free license, exported using standard formats, and can be interlinked to other open data sets on the linked data web.

wikidata.org

SCHOLIA Author Work Organization Location Event Project Award Topic Tools Help

Scholia is a service that creates visual scholarly profiles for topic, people, organizations, species, chemicals, etc using bibliographic and other information in Wikidata. [More info...](#)

Scholia relies on Wikidata, and Wikidata contains only a limited albeit growing subset of the corpus of scholarly literature, its authors and citations. Read more about the limitations in the [FAQ](#).

Search

Search for a scientist, topic, publication, organization, award, event, etc.

Examples

Profiles

Denny Vrandečić
View the researcher profile for the Semantic Web researcher Denny Vrandečić. It shows his papers, co-authors, etc.

Technical University of Denmark
View the profile for an organization: People

Combinations

Scholia can show multiple items together.

Technical University of Denmark and University College London
Compare two or more organizations. Here a comparison between two universities with collaborating researchers

Redirects

If you know the external identifier of a concept, then Scholia can make a lookup based on it:

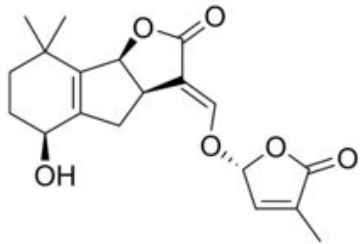
twitter/utafriith
Look up by Twitter username @utafriith. This will identify the London-based researcher Uta Frith and redirect to her Scholia page

scholia.toolforge.org

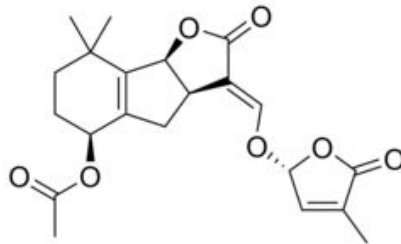
Strigolactones (in Wikipedia and Wikidata?)

Chemical structures [\[edit \]](#)

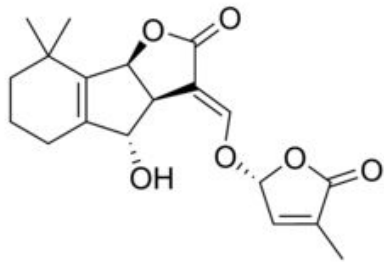
Some examples of strigolactones include:



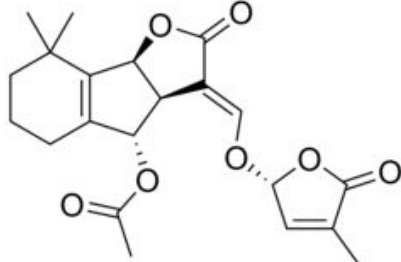
(+)-Strigol



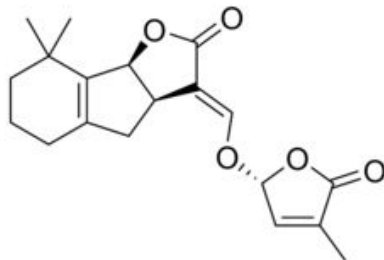
(+)-Strigyl acetate



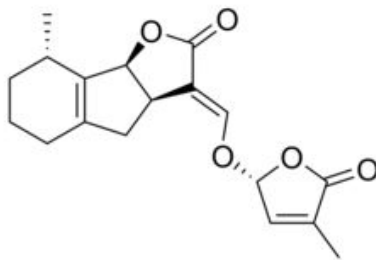
(+)-Orobanchol



(+)-Orobanchyl acetate



(+)-5-Deoxystrigol



Sorgolactone

strigolactones ([Q2157332](#))

Strigolactones are a group of chemical compounds produced by a plant's roots. Due to their mechanism of action plant hormones or phytohormones. So far, strigolactones have been identified to be responsible for three different promote the germination of parasitic organisms that grow in the host plant's roots, such as *Striga lutea* and other [English Wikipedia](#)

Class Hierarchy

✕ ✎ ✏ ✎ ✎

WikiProject Chemistry



Main page
Community portal
Project chat
Create a new Item
Recent changes
Random Item
Query Service
Nearby
Help
Donate

Lexicographical data
Create a new Lexeme
Recent changes
Random Lexeme

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Wikidata item

In other projects
Wikimedia Commons

Project page

Discussion

Read

Edit

View history



More ▾

Search Wikidata



Wikidata:WikiProject Chemistry

Translate this page

Other languages:

Deutsch • English • Nederlands • català • dansk • español • français • polski • português do Brasil • svenska • čeština • русский • українська • հայերէս • اردو • العربية • 中文

Home

Guidelines

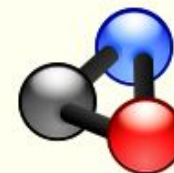
Properties

References

Tools



Welcome to WikiProject Chemistry



Contents [hide]

- Goals
- Participants
 - Bots
- How to contribute
- See also

Goals [edit]

- Define properties for items related to chemistry and the rules of use for these properties (qualifiers, datatypes, ...)
- Define references policy and especially ranking for references in order to ensure a high quality for chemical related data

WikiProject Chemistry

English
Español
Français
日本語
Polski
Svenska
Türkçe

🌐 22 more

 Edit links

Custom tools

Participants [edit]

The participants listed below can be notified using the following template in discussions:

[\[+ Add yourself to the list\]](#)

```
{{Ping project|Chemistry}}
```

- Saehrimnir
- Leyo
- Snipre
- Jasper Deng
- Dcirovic
- Walkerma
- Egon Willighagen
- Denise Slenter
- Daniel Mietchen
- Kopiersperre
- Emily Temple-Wood
- Pablo Busatto (*Almondaga*)
- Antony Williams (EPA)
- TomT0m
- Wostr
- Devon Fyson
- User:DePiep
- User:DavRosen
- Benjaminabel
- 99of9
- Kubaello
- Fractaler
- Sebotic
- Netha
- Hugo
- Samuel Clark
- Tris T7
- Leiem
- Christianhauck
- SCIdude
- Binter
- Photocyte
- Robert Giessmann
- Cord Wiljes
- Jonathan Bisson
- GrndStt
- Ameisenigel
- Charles Tapley Hoyt
- ChemHobby
- Peter Murray-Rust
- Erfurth

Bots [edit]

- SamoaBot - task 6 - set a property "atomic number" based on Wikipedia) -  On hold

WikiProject Chemistry



[Project page](#)

[Discussion](#)

[Read](#)

[Edit](#)

[Add topic](#)

[View history](#)



[More](#) ▾



Wikidata talk:WikiProject Chemistry



Old discussions are archived in [Archive 2013](#), [Archive 2014](#), [Archive 2015](#), [Archive 2016](#), [Archive 2017](#), [Archive 2018](#), [Archive 2019](#).

Contents [\[hide\]](#)

- [A lot of duplicate data](#)
 - [Tautomer/zwitterion](#)
 - [Non-standard InChI](#)
- [GZWDer added all \(most?\) of the US EPA CompTox dashboard](#)
- [New property proposals](#)
- [Difference between CAS numbers](#)
- [Introduction round](#)
- [Q5173335](#)
- [604 duplicate InChIKeys](#)
- [Difference between CAS numbers \(bis\)](#)
 - [CAS 28519-04-2 vs. CAS 7134-06-7](#)
 - [CAS 40102-60-1 vs. CAS 1439-07-2](#)
 - [CAS 64047-16-1 vs. CAS 6588-17-6](#)
 - [CAS 13455-34-0 vs. CAS 60459-08-7](#)
 - [CAS 103-26-4 vs. CAS 1754-62-7](#)
 - [CAS 1701-77-5 vs. CAS 7021-09-2](#)
 - [CAS 36393-56-3 vs. CAS 37577-07-4](#)
- [CAS and unspecified stereochemistry](#)

[Main page](#)

[Community portal](#)

[Project chat](#)

[Create a new Item](#)

[Recent changes](#)

[Random Item](#)

[Query Service](#)

[Nearby](#)

[Help](#)

[Donate](#)

[Lexicographical data](#)

[Create a new Lexeme](#)

[Recent changes](#)

[Random Lexeme](#)

[Tools](#)

[What links here](#)

[Related changes](#)

[Special pages](#)

[Permanent link](#)

[Page information](#)

[Custom tools](#)

Modelling the Chemistry in Wikidata

acetic acid (Q47512)

chemical compound [edit](#)

ethanoic acid | methanecarboxylic acid | CH₃-COOH | Acetic acid, glacial | HOAc | Vinegar | Essigsäure | Glacial acetic acid | Ethanoate | acide acétique | Ethylic acid | Ethoic acid | Methanecarboxylic acid | Aceticum acidum | Ethanoic acid | Acetic acid | Ethanoat | E260 | CH₃COOH

► **Recoin:** Most relevant properties which are absent

▼ **In more languages**

Language	Label	Description	Also known as
English	acetic acid	chemical compound	ethanoic acid methanecarboxylic acid CH ₃ -COOH Acetic acid, glacial HOAc Vinegar Essigsäure Glacial acetic acid Ethanoate acide acétique Ethylic acid Ethoic acid Methanecarboxylic acid Aceticum acidum Ethanoic acid Acetic acid Ethanoat E260 CH ₃ COOH
German	Essigsäure	Ethansäure, einprotonige Carbonsäure	Haushaltessig E260 Methancarbonsäure Acidum aceticum



Wikipedia (85 entries) [edit](#) [move](#)

af	Asynsuur	edit
ar	حمض الخليك	edit
ast	Ácidu acético	edit
azb	استیک اسید	edit
az	Sirke turşusu	edit
bcl	Asidong asetiko	edit
be	Воцатная кіслата	edit
bg	Оцетна киселина	edit

Typing: chemical compound and more

Statements

instance of

by F705i and James Hare (NIOSH)
and ProteinBoxBot and Egon
Willighagen and Chire and Antoni
Salvà and Thomas11 and Infovarius



carboxylic acid

edit

▼ 0 references

+ add reference



Class II combustible liquid

edit

▶ 1 reference



medication

edit

▼ 1 reference

stated in	DrugBank
DrugBank ID	03166
language of work or name	English
title	Acetic acid (English)
publication date	17 November 2015

+ add reference



metabolite

edit

Chemical structure

chemical formula

C2H4O2

  edit

by ProteinBoxBot and Ivan A.

Krestinin and The chemists and

Infovarius and Sebotic and

SoCalChemBot and 87.68.189.9

▶ 1 reference

+ add value

and Wostr

canonical SMILES

CC(=O)O

  edit

by ProteinBoxBot and Sebotic and

SoCalChemBot

▶ 1 reference

+ add value



**CDK
DEPICT**

InChI

InChI=1S/C2H4O2/c1-2(3)4/h1H3,(H,3,4)

  edit

by Happy5214 and KrBot and

ProteinBoxBot and Sebotic and

SoCalChemBot and Scidubot

▼ 1 reference

stated in	PubChem	
PubChem CID	176	
language of work or name	English	
title	acetic acid (English)	
retrieved	19 October 2016	

+ add reference

+ add value

InChIKey

QTBSBXVTEAMEQO-UHFFFAOYSA-N

  edit

by Happy5214 and KrBot and

ProteinBoxBot and Sebotic and

SoCalChemBot

▶ 2 references

+ add value



Physicochemical properties

boiling point

by Emily Temple-Wood (NIOSH) and James Hare (NIOSH) and Egon Willighagen and Wostr

 244 ± 1 degree Fahrenheit   edit

pressure 760 ± 1 torr

▼ 1 reference

reference URL <http://www.cdc.gov/niosh/npg/npgd0002.html>



+ add reference

 117.9 ± 0.1 degree Celsius    edit

▼ 1 reference

stated in Basic laboratory and industrial chemicals: a CRC quick reference handbook

+ add reference

 117.9 ± 0.2 degree Celsius   edit

pressure $101,325\pm 1$ pascal

▼ 1 reference

stated in CRC Handbook of Chemistry and Physics (95th edition)

page(s) 3-4

+ add reference

+ add value



```

1 SELECT ?typeLabel ?count WITH {
2   SELECT ?type (COUNT(DISTINCT ?chemical) AS ?count) WHERE {
3     ?chemical wdt:P31 ?type ;
4               wdt:P235 [] .
5   } GROUP BY ?type
6 } AS %TYPES {
7   INCLUDE %TYPES
8   SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
9 } ORDER BY DESC(?count)

```



2443 results in 30536 ms

</> Code

Download

Link

Search



typeLabel	count
chemical compound	1135341
group of stereoisomers	111523
chemical entity	14665
medication	2356
organic anion	826
carcinogen	487

Chemical Types

typeLabel	count
chemical compound	1135341
group of stereoisomers	111523
chemical entity	14665
medication	2356
organic anion	826
carcinogen	487
diacylglycerophosphocholine	485
wax monoester	409
pair of enantiomers	409
intermetallic	376
lipid	342
flavonoid	342
essential medicine	302
heterocyclic compound	275
unsaturated fatty acids	263
fatty acyl-CoA	254
developmental toxicant	253
carboxylic acid	246
diacylglycerophosphoinositols	239
insecticide	212
herbicide	205

typeLabel	count
chemical compound	917519
group of stereoisomers	9647
medication	2656
intermetallic	677
carcinogen	495
chemical entity	459
wax monoester	393
pair of enantiomers	357
essential medicine	327
monoclonal antibody	325
flavonoid	292
heterocyclic compound	274
developmental toxicant	262
carboxylic acid	238
insecticide	217
family of isomeric compounds	207
herbicide	205
mixture	192
biopharmaceutical	173
unsaturated fatty acids	164
fungicide	162

Chemical type guidance

Compounds without fully defined isomerism or isotopic composition [\[edit \]](#)

In the following rules *structural need* is defined as having at least one external identifier to a reliable database (InChI, InChIKey or SMILES are not regarded as such) or at least one valid sitelink to a page on a Wikimedia project (cf. point 1 of [Wikidata:Notability](#)).

See [discussion about this topic](#) (2018) in [WikiProject Chemistry](#).

Inclusion criteria [\[edit \]](#)

1. Every chemical compound with fully defined isomerism (*cis–trans* isomerism; *ortho*, *meta*, *para* isomerism; enantiomerism; etc.) or isotopic composition can be described in a separate item (hereafter **item A**).
2. Chemical compound with fully (**item B**) or partially (**item C**) undefined isomerism or isotopic composition can be described in a separate item if it fulfils some structural need.
3. Item about a [racemic mixture](#) (Q467717) (or more generally: about mixture of isomers) can be described in a separate item (**item D**) if it fulfils some structural need.
4. Item about a compound being an [atropisomer](#) (Q757764) can be described in a separate item if it can be isolated or if it fulfils some structural need.

- [\(S\)-2-pentanol](#) (Q20680358) describes a compound with fully defined stereochemistry (has one stereogenic centre and it is defined). Thus, it can be described in a separate item.
- [2-Bromobenzaldehyde](#) (Q33859440) describes a compound with defined positions of two substituents of the benzene ring (*ortho* position), so it can be described in a separate item.
- [DL-methamphetamine](#) (Q44909815) describes a compound with fully undefined stereochemistry (has one stereogenic centre and it is undefined). However, it has some external identifiers in reliable databases, like [ChemSpider ID](#) (P661) or [DSSTox substance ID](#) (P3117), so it can be described in a separate item.
- [\(2S\)-homocystine](#) (Q27161892) describes a compound with partially undefined stereochemistry (has two stereogenic centres and only one is defined). It has some external identifiers like [ChEBI ID](#) (P683) or [PubChem CID](#) (P662), so it can be described in a separate item.
- [Bromobenzaldehyde](#) (Q33859433) describes a compound with undefined position of two substituents of the benzene ring (i.e. it is a group of three isomers: *ortho*, *meta* and *para*). It has an article in Wikipedia, so it can be described in a separate item.
- [\(±\)-nicotine](#) (Q56697247) describes a racemic mixture of [\(–\)-nicotine](#) (Q28086552) and [\(+\)-nicotine](#) (Q27119762) and it is different from [nicotine](#) (Q12144) that describes a compound with undefined stereochemistry. It can be described in a separate item, because there is an external identifier ([ChEBI ID](#) (P683)).
- *no example yet*

Visualize Wikidata Schema

racemic mixture

Language en

Info about schema entity

E47 - racemic mixture

mixture of chemicals with the same structure but different stereochemistry

<https://www.wikidata.org/wiki/EntitySchema:E47>

Methodology article | Open Access | Published: 22 January 2021

A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses

Andra Waagmeester, Egon L. Willighagen, Andrew I. Su, Martina Kutmon, Jose Emilio Labra Gayo, Daniel Fernández-Alvarez, Quentin Groom, Peter J. Schaap, Lisa M. Verhagen & Jasper J. Koehorst

BMC Biology 19, Article number: 12 (2021) | Cite this article

1073 Accesses | 48 Altmetric | Metrics

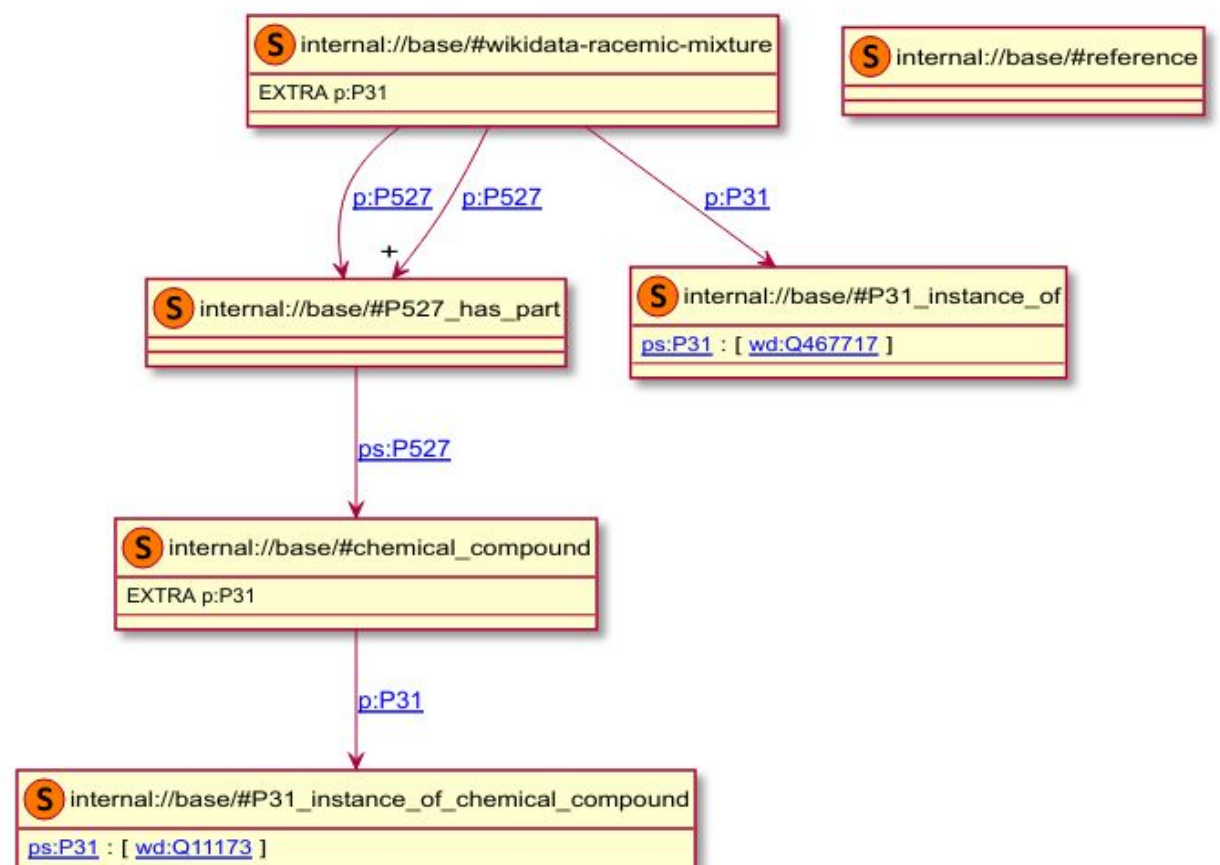
Abstract

Background

Pandemics, even more than other medical problems, require swift integration of knowledge. When caused by a new virus, understanding the underlying biology may help finding solutions. In a setting where there are a large number of loosely related projects and initiatives, we need common ground, also known as a “commons.” Wikidata, a public knowledge graph aligned with Wikipedia, is such a commons and uses unique identifiers to link knowledge in other knowledge bases. However, Wikidata may not always have the right schema for the urgent questions. In this paper, we address this problem by showing how a data schema required for the integration can be modeled with entity schemas represented by Shape Expressions.

Results

As a telling example, we describe the process of aligning resources on the genomes and proteomes of the SARS-CoV-2 virus and related viruses as well as how Shape Expressions can be defined for Wikidata to model the knowledge, helping others studying the SARS-CoV-2 pandemic. How this model can be used to make data between various resources interoperable is demonstrated by integrating data from NCBI National Center for Biotechnology



ShEx validation: E46 → chemical element

WikiShape Entity ▾ Schema ▾ Property ▾ Query ▾ Help ▾

Validate Wikidata entities

New result

Id ↑↓	Node ↑↓	Shape ↑↓	Status ↑↓	Details
0	wd:Q623	<#wikidata-element>	conformant	► Details
1	wds:q623-6FA2E9FD-D3B8-4CCB-A6CA-949B88B383FB	<#P246_chemical_symbol>	conformant	► Details
2	wds:Q623-B81E578D-49CE-45B9-A924-C2BF9EC802DB	<#P31_instance_of>	conformant	► Details
3	wds:Q623-eee42e14-46e0-c18c-76e3-af9b87475c7d	<#P1086_atomic_number>	conformant	► Details

► Details

Permalink

Q623 (carbon) ×

Language en

Wikidata schema

ShEx

chemical element

Language en

Shape <#wikidata-element>

Validate wikidata entities

Adding chemical compounds to Wikidata

Workhorse: Bioclipse scripts + the CDK

10.1186/1471-2105-8-59,
10.1186/1471-2105-10-397

Journal of Cheminformatics

[Home](#) [About](#) [Articles](#) [Submission Guidelines](#) [About The Editors](#) [Calls For Papers](#)

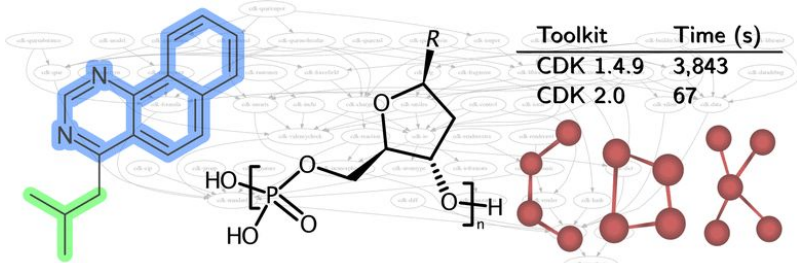
Software | [Open Access](#) | Published: 06 June 2017

The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching

Egon L. Willighagen , John W. Mayfield, Jonathan Alvarsson, Arvid Berg, Lars Carlsson, Nina Jeliakova, Stefan Kuhn, Tomáš Pluska, Miquel Rojas-Chertó, Ola Spjuth, Gillean Torrance, Chris T. Evelo, Rajarshi Guha & Christoph Steinbeck

Journal of Cheminformatics 9, Article number: 33 (2017) | [Download Citation](#) ↓

7825 Accesses | 50 Citations | 55 Altmetric | [Metrics](#) >>



Bacting: Bioclipse on the command line

```
@Grab(group='io.github.egonw.bacting', module='managers-cdk', version='0.0.9')

workspaceRoot = "."
def cdk = new net.bioclipse.managers.CDKManager(workspaceRoot);


println cdk.fromSMILES("COC")
```

- Wikicite/findConcepts.groovy
- Wikidata/createWDItemsFromSMILES.groovy
- LipidMaps/classifyLipids.groovy
- ExtIdentifiers/comptox.groovy
- MeltingPoints/createQuickStatements.groovy
- ...

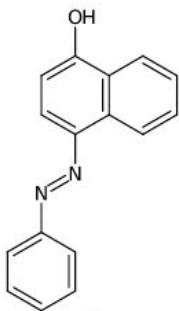
Generate depictions of molecules and reactions from [SMILES](#) or [SDF](#).

```

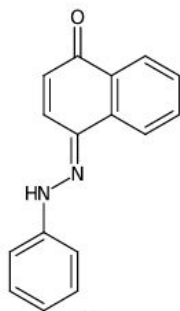
c1(c2cccc1)C(=O)=CC=C2\N=N\c1cccc1
c1(c2cccc1)C(=O)C=C\C2=N/Nc1cccc1
c1(c2cccc1)C(=O)=CC=C2N=Nc1cccc1
c1(c2cccc1)C(=O)C=CC2=NNc1cccc1
    
```



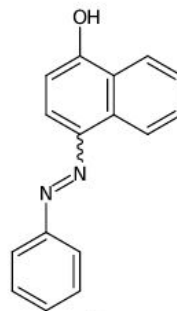
...



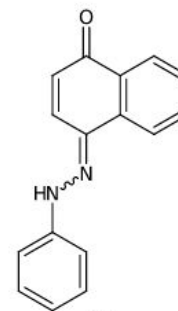
#1



#2



#3



#4

Built with the [Chemistry Development Kit](#). Depict v1.6-SNAPSHOT, CDK v2.4-SNAPSHOT.

Compare against Wikidata (with InChIKey)

```
egonw@debian:~/var/Projects/hub/ons-wikidata/Wikidata$ groovy createWDitemsFromSMILES.groovy
```

```
=====  
C16H12N2O is not yet in Wikidata
```

```
Full stereochemistry is defined  
=====  
=====  
=====  
C16H12N2O is not yet in Wikidata
```

```
Full stereochemistry is defined  
=====  
=====  
=====  
C16H12N2O is not yet in Wikidata
```

```
Compound has missing stereo on # of centers: 2  
=====  
=====  
=====  
C16H12N2O is not yet in Wikidata
```

```
Compound has missing stereo on # of centers: 1  
=====  
=====  
=====  
C16H12N2O is not yet in Wikidata
```

```
Compound has missing stereo on # of centers: 1  
=====  
=====  
=====
```

```
egonw@debian:~/var/Projects/hub/ons-wikidata/Wikidata$ more output.quickstatements
```

```
CREATE
```

```
LAST P31 Q11173
```

```
LAST Den "chemical compound"
```

```
LAST P2017 "c1(c2cccc1)C(=O)=CC=C2\N=N\c1cccc1"
```

```
LAST P274 "C16H12N2O"
```

```
LAST P234 "InChI=1S/C16H12N2O/c19-16-11-10-15(13-8-4-5-9-14(13)16)18-17-12-6-2-
```

```
LAST P235 "CQYDCXNJLAOBIF-ISLYRVAYSA-N"
```

```
CREATE
```

```
LAST P31 Q11173
```

```
LAST Den "chemical compound"
```

Use QuickStatements to add to Wikidata

QuickStatements

English



New batch

Last batches

Chat

Git

Help

Batch on Wikidata by Egon Willighagen [Batches]

Status:

1 init

CREATE Item

en:chemical compound

instance of [P31]:chemical compound [Q11173]

isomeric SMILES [P2017]:"c1(c2cccc1)C(=O)=CC=C2\N\c1cccc1"

chemical formula [P274]:"C16H12N2O"

InChI [P234]:"InChI=1S/C16H12N2O/c19-16-11-10-15(13-8-4-5-9-14(13)16)18-17-12-6-2-1-3-7-12/h1-11,19H/b18-17+"

InChIKey [P235]:"CQYDCXNJLAOBIF-ISLYRVAYSA-N"

2 init

CREATE Item

en:chemical compound

instance of [P31]:chemical compound [Q11173]

isomeric SMILES [P2017]:"c1(c2cccc1)C(=O)C=C\C2=N\Nc1cccc1"

chemical formula [P274]:"C16H12N2O"

InChI [P234]:"InChI=1S/C16H12N2O/c19-16-11-10-15(13-8-4-5-9-14(13)16)18-17-12-6-2-1-3-7-12/h1-11,17H/b18-15+"

InChIKey [P235]:"NZZPXZGSADNPOR-OBGWFSINSA-N"

3 init

CREATE Item

en:chemical compound

instance of [P31]:chemical compound [Q11173]

canonical SMILES [P233]:"c1(c2cccc1)C(=O)=CC=C2=Nc1cccc1"

chemical formula [P274]:"C16H12N2O"

InChI [P234]:"InChI=1S/C16H12N2O/c19-16-11-10-15(13-8-4-5-9-14(13)16)18-17-12-6-2-1-3-7-12/h1-11,19H"

InChIKey [P235]:"CQYDCXNJLAOBIF-UHFFFAOYSA-N"

PubChem CID [P662]:"77214"

4 init

CREATE Item

en:chemical compound

instance of [P31]:chemical compound [Q11173]

canonical SMILES [P233]:"c1(c2cccc1)C(=O)C=CC2=NNc1cccc1"

chemical formula [P274]:"C16H12N2O"

InChI [P234]:"InChI=1S/C16H12N2O/c19-16-11-10-15(13-8-4-5-9-14(13)16)18-17-12-6-2-1-3-7-12/h1-11,17H"

InChIKey [P235]:"NZZPXZGSADNPOR-UHFFFAOYSA-N"

First

Page

1

Last

Run

Run in background

Use QuickStatements to add to Wikidata

QuickStatements

English



New batch

Last batches

Chat

Git

Help

Batch on Wikidata by Egon Willighagen [Batches] [Discuss/revert batch](#)

Status: DONE

1	done Q106156511 [Q106156511]	CREATE Item	<p>en:chemical compound instance of [P31]:chemical compound [Q11173] isomeric SMILES [P2017]:"<chem>c1(c2cccc1)C(=O)=CC=C2\N=c1cccc1</chem>" chemical formula [P274]:"<chem>C16H12N2O</chem>" InChI [P234]:"<chem>InChI=1S/C16H12N2O/c19-16-11-10-15(13-8-4-5-9-14(13)16)18-17-12-6-2-1-3-7-12/h1-11,19H/b18-17+</chem>" InChIKey [P235]:"<chem>CQYDCXNJLAOBIF-ISLYRVAYSA-N</chem>"</p>
2	done Q106156512 [Q106156512]	CREATE Item	<p>en:chemical compound instance of [P31]:chemical compound [Q11173] isomeric SMILES [P2017]:"<chem>c1(c2cccc1)C(=O)C=C\C2=N\Nc1cccc1</chem>" chemical formula [P274]:"<chem>C16H12N2O</chem>" InChI [P234]:"<chem>InChI=1S/C16H12N2O/c19-16-11-10-15(13-8-4-5-9-14(13)16)18-17-12-6-2-1-3-7-12/h1-11,17H/b18-15+</chem>" InChIKey [P235]:"<chem>NZZPXZGSADNPOR-OBGWFSINSA-N</chem>"</p>
3	done Q106156514 [Q106156514]	CREATE Item	<p>en:chemical compound instance of [P31]:chemical compound [Q11173] canonical SMILES [P233]:"<chem>c1(c2cccc1)C(=O)=CC=C2=Nc1cccc1</chem>" chemical formula [P274]:"<chem>C16H12N2O</chem>" InChI [P234]:"<chem>InChI=1S/C16H12N2O/c19-16-11-10-15(13-8-4-5-9-14(13)16)18-17-12-6-2-1-3-7-12/h1-11,19H</chem>" InChIKey [P235]:"<chem>CQYDCXNJLAOBIF-UHFFFAOYSA-N</chem>" PubChem CID [P662]:"77214"</p>
4	done Q106156515 [Q106156515]	CREATE Item	<p>en:chemical compound instance of [P31]:chemical compound [Q11173] canonical SMILES [P233]:"<chem>c1(c2cccc1)C(=O)C=CC2=NNc1cccc1</chem>" chemical formula [P274]:"<chem>C16H12N2O</chem>" InChI [P234]:"<chem>InChI=1S/C16H12N2O/c19-16-11-10-15(13-8-4-5-9-14(13)16)18-17-12-6-2-1-3-7-12/h1-11,17H</chem>" InChIKey [P235]:"<chem>NZZPXZGSADNPOR-UHFFFAOYSA-N</chem>"</p>

First

Page

1

Last

Wikidata Quickstatements v2

qid,P921,#

Q26801490,Q70828631,Activities and Effects of Ergot Alkaloids on ...

Q28082319,Q70828631,Diversification of ergot alkaloids in natural and ...

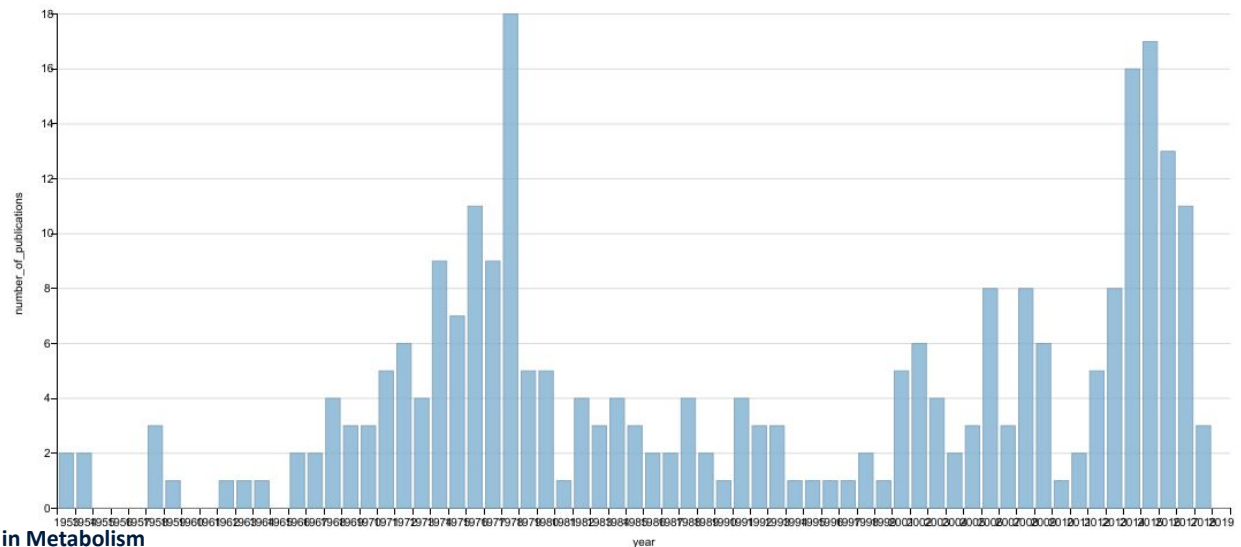
Q28214648,Q70828631,Biotechnology and genetics of ergot alkaloids

Q28276288,Q70828631,Ergot alkaloids--biology and molecular biology

Q28287164,Q70828631,Occurrence of peptide and clavine ergot alkaloids ...


...

Publications per year



Dr. Magnus Manske
Sanger Institute

Jenkins for Wikidata quality control

 **Jenkins** Willighagen, Egon (BIGCAT) | [log out](#)

[Jenkins](#) > [Wikidata Checks for Metabolomics](#) [ENABLE AUTO REFRESH](#)

[Back to Dashboard](#)
[Status](#)
[Changes](#)
[Workspace](#)
[Build Now](#)
[Delete Project](#)
[Configure](#)
[GitHub Hook Log](#)
[GitHub](#)
[Rename](#)

Project Wikidata Checks for Metabolomics

[add description](#)
[Disable Project](#)

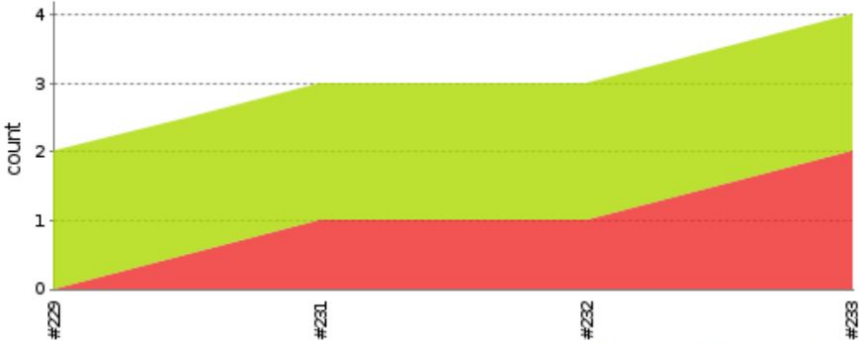
[Workspace](#)
[Recent Changes](#)
[Latest Test Result \(2 failures / +1\)](#)

Upstream Projects

[bacting](#)

Permalinks

Test Result Trend



Build	Failures	Passes	Total
#229	2	0	2
#231	1	2	3
#232	1	2	3
#233	2	2	4

[\(just show failures\)](#) [enlarge](#)

[Build History](#) [trend](#)

ChemCuration example: InChIKeys

Jenkins > Wikidata Checks for Metabolomics > #233 > Test Results > (root) > InChITests > InChIKeyMismatch

[ENABLE AUTO REFRESH](#)

 Edit Build Information

 History

 Git Build Data

 No Tags

 Test Result

 Previous Build

Error Message

The InChIKey computed from the isomeric SMILES and InChIKey in Wikidata does not match

Stacktrace

<http://www.wikidata.org/entity/Q421291> with isomeric SMILES '[Fe+2].O[C@H]([C@H](O)C([O-])=O)[C@H](O)[C@H](O)CO.[O-]C(=O)[C@H](O)[C@H](O)[C@H](O)[C@H](O)CO' has a calculated InChIKey VRIVJOXICYMTAG-IYEMJOQQA-L that does not match the given QDUZQ0IJXPPTLY-GMBKLUGCSA-N

<http://www.wikidata.org/entity/Q7777226> with isomeric SMILES 'Oc1cc(cc(0)c10)C(=O)Oc5c(0)cc([C@H]20c3cc(0)cc(0)c3C[C@H]20)c6\C=C(/C=C(/OC(=O)c4cc(0)c(0)c(0)c4)C(=O)c56)[C@H]70c8cc(0)cc(0)c8C[C@H]70' has a calculated InChIKey FJYGFTHLNNVPHY-BBXLVSEPSA-N that does not match the given TUJOKWPTOVJHLY-JBJHRQGLSA-N

<http://www.wikidata.org/entity/Q15427926> with isomeric SMILES 'CC1(C)C([C@H](OC(C)=O)C[C@@]2(C)[C@](C([C@H](OC(C)=O)CC2)=C)([H])[C@H]3OC(C)=O)=C(C)[C@H](OC([C@H](C)[C@H](C)O)=O)C[C@]13[H])' has a calculated InChIKey ULHOQE0TAJVICR-SJJKDWJASA-N that does not match the given FMPIEMVVEJGMCY-IRWPHOLZSA-N

<http://www.wikidata.org/entity/Q568> with isomeric SMILES '[Li]' has a calculated InChIKey WHXSMMKQMYFTQS-UHFFFAOYSA-N that does not match the given SIAPCJWMELPYOE-UHFFFAOYSA-N

<http://www.wikidata.org/entity/Q5278705> with isomeric SMILES 'C[C@@]130[C@]1(/C=C/C(C)=C/C=C/C(C)=C/C=C/C(C)/C=C/C(C)/[C@H]=C=C2C(C)(C)C[C@H](OC(C)=O)C[C@]2(C)O)C(C)(C)C[C@H](O)C3' has a calculated InChIKey PVNVIBOWBAPFOE-RWNIHPGNSA-N that does not match the given GJFBHWJTMIDLNX-UWCSZFODSA-N

Scholia



Research Ideas and Outcomes 5: e35820
doi: 10.3897/rio.5.e35820



Grant Proposal

Robustifying Scholia: paving the way for knowledge discovery and research assessment through Wikidata

Lane Rasberry[‡], Egon L. Willighagen[§], Finn Årup Nielsen[|], Daniel Mietchen[‡]

[‡] Data Science Institute, University of Virginia, Charlottesville, United States of America

[§] Dept of Bioinformatics - BIGCaT, NUTRIM, Maastricht University, Maastricht, Netherlands

[|] Technical University of Denmark, Kongens Lyngby, Denmark

Corresponding author: Daniel Mietchen (daniel.mietchen@virginia.edu)

Reviewable v1

Received: 29 Apr 2019 | Published: 02 May 2019

Citation: Rasberry L, Willighagen E, Nielsen F, Mietchen D (2019) Robustifying Scholia: paving the way for knowledge discovery and research assessment through Wikidata. Research Ideas and Outcomes 5: e35820. <https://doi.org/10.3897/rio.5.e35820>

Abstract

Knowledge workers like researchers, students, journalists, research evaluators or funders need tools to explore what is known, how it was discovered, who made which contributions, and where the scholarly record has gaps. Existing tools and services of this kind are not available as Linked Open Data, but Wikidata is. It has the technology, active contributor

Wikidata / Scholia



Redirecting

If you know the identifier then Scholia can make a lookup based on the identifier:

cas/50-00-0

Lookup CAS 50-00-0. This will identify formaldehyde and redirect to its Scholia page.

inchikey/QTBSBXVTEAMEQO-UHFFFAOYSA-N

Redirect also works for InChIKeys, here for acetic acid.

Show entries

Search:

Mol	InChIKey	CAS	ChemSpider	PubChem CID
acetic acid	QTBSBXVTEAMEQO-UHFFFAOYSA-N	64-19-7	171	176
deuterated acetic acid	QTBSBXVTEAMEQO-GUEYOVJQSA-N	1186-52-3	2006083	2723903
acetic acid c-14	QTBSBXVTEAMEQO-HQMMCQRPSA-N	2845-03-6	144444	164769
acetic acid c-13	QTBSBXVTEAMEQO-VQEHIDDOSA-N	1563-79-7	8329490	10153982
acetic acid c-11	QTBSBXVTEAMEQO-JVVVGQRLSA-N	78887-71-5	396653	450349
acetate ion	QTBSBXVTEAMEQO-UHFFFAOYSA-M	71-50-1	170	175

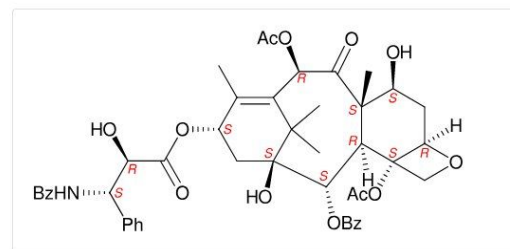
[Edit on query.Wikidata.org](#)

Showing 1 to 6 of 6 entries

Previous Next

paclitaxel (Q423762)

Paclitaxel (PTX), sold under the brand name Taxol among others, is a chemotherapy medication used to treat a number of types of cancer. This includes ovarian cancer, breast cancer, lung cancer, Kaposi sarcoma, cervical cancer, and pancreatic cancer. It is given by injection into a vein. ... (from the [English Wikipedia](#))



Identifiers

Show entries

Search:

IDpred Id

ATC code L01CD01

2019: 10.3897/rio.5.e35820
2017: 10.6084/m9.figshare.6356027.v1

Redirecting

If you know the identifier then Scholia can make a lookup based on the identifier:

[cas/50-00-0](#)

Lookup CAS 50-00-0. This will identify formaldehyde and redirect to its Scholia page.

[inchikey/QTBSBXVTEAMEQO-UHFFFAOYSA-N](#)

Redirect also works for InChIKeys, here for acetic acid.

Show entries

Search:

Mol	InChIKey	CAS	ChemSpider	PubChem CID
acetic acid	QTBSBXVTEAMEQO-UHFFFAOYSA-N	64-19-7	171	176
deuterated acetic acid	QTBSBXVTEAMEQO-GUEYOVJQSA-N	1186-52-3	2006083	2723903
acetic acid c-14	QTBSBXVTEAMEQO-HQMMCQRPSA-N	2845-03-6	144444	164769
acetic acid c-13	QTBSBXVTEAMEQO-VQEHIDDOSA-N	1563-79-7	8329490	10153982
acetic acid c-11	QTBSBXVTEAMEQO-JVVVGQRLSA-N	78887-71-5	396653	450349
acetate ion	QTBSBXVTEAMEQO-UHFFFAOYSA-M	71-50-1	170	175

[Edit on query.Wikidata.org](#)

Showing 1 to 6 of 6 entries

Previous

1

Next

Wikidata / Scholia

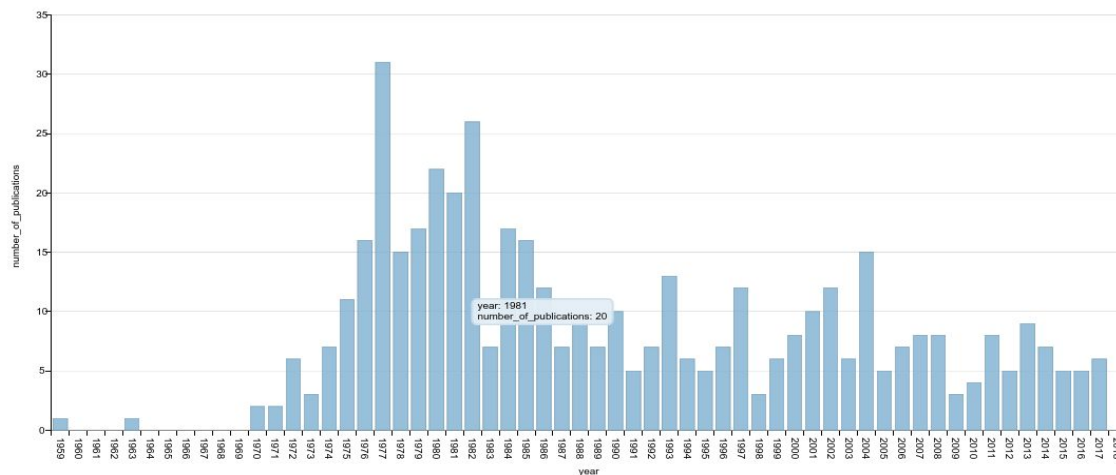
Physchem Properties

Show entries

Search:

PropEntity	Value	Units	Qualifiers	Source	Doi
acid dissociation constant	4.74	1		Small Scale Determination of the pKa Values for Organic Acids	10.1021/ED071PA6
mass	60.021129	atomic mass unit		PubChem	
acid dissociation constant	4.756	1	temperature: 25	CRC Handbook of Chemistry and Physics (95th edition)	
boiling point	117.9	degrees Celsius	pressure: 101325	CRC Handbook of Chemistry and Physics (95th edition)	
density	1.0446	gram per cubic centimetre	temperature: 25	CRC Handbook of Chemistry and Physics (95th edition)	

Publications per year



Recently published works on the chemical

Show entries

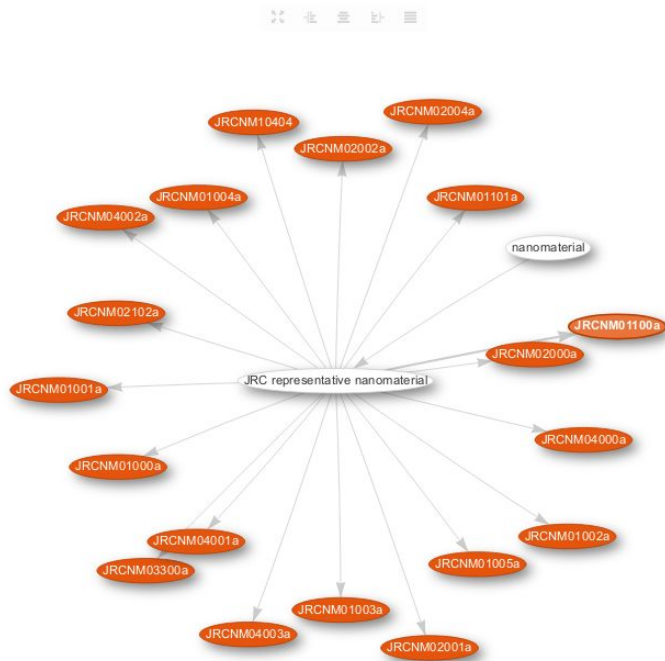
Date	Work	Type	Topics
2017-08-09	In vitro human skin permeation of benzene in gasoline: effects of concentration, multiple dosing and skin preparation	scholarly article	oil and gas extraction // benzene
2017-04-27	Nicotine, aerosol particles, carbonyls and volatile organic compounds in tobacco- and menthol-flavored e-cigarettes	scholarly article	toluene // benzene

Scholia: JRC representative industrial nanomaterials

topic chemical

JRC representative nanomaterial (Q47461491)

Class Hierarchy



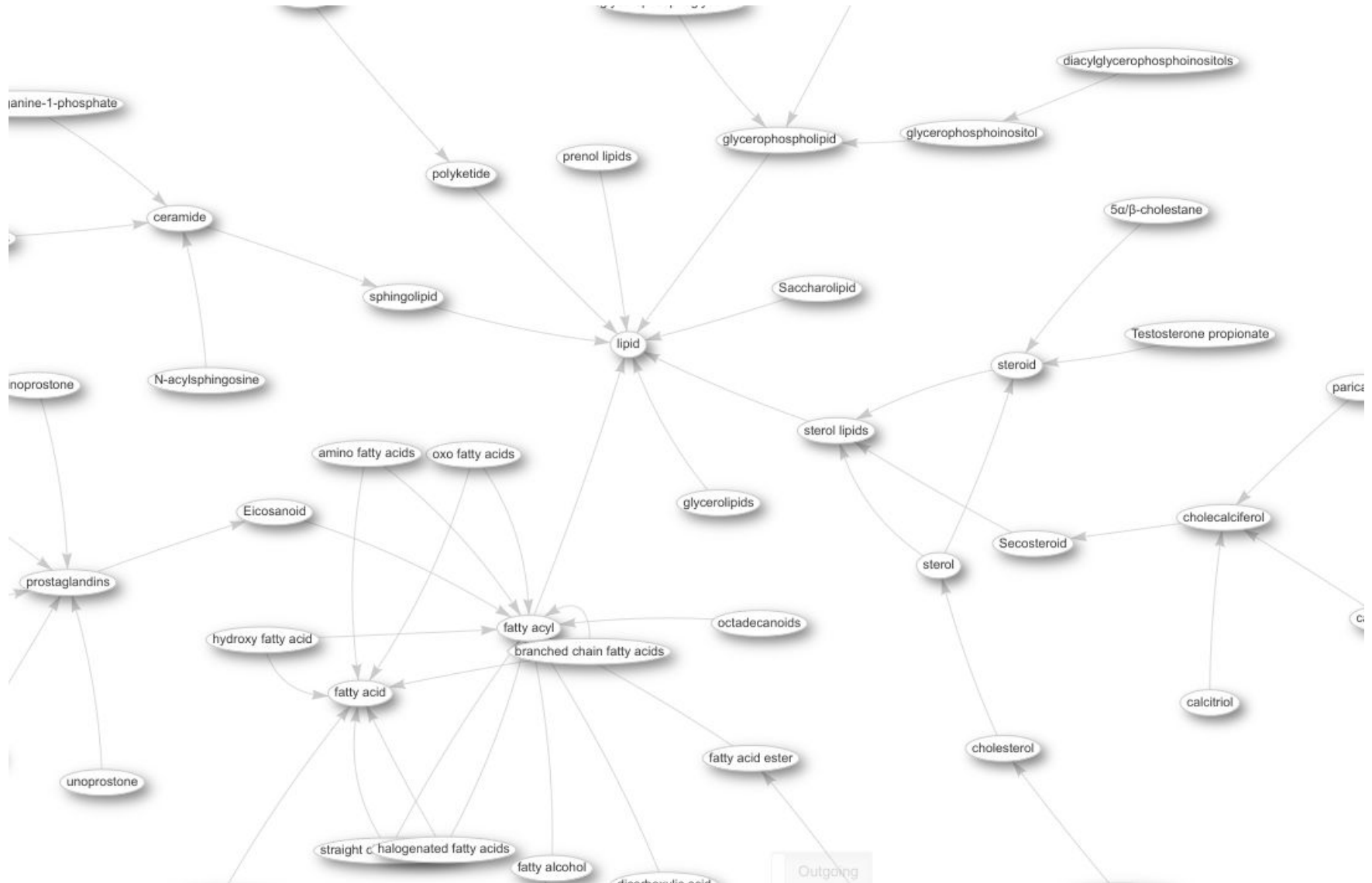
Recently published works on the chemical

Show entries

Search:

Date	Work	Type	Topics
2017-09-28	Fish cell lines as a tool for the ecotoxicity assessment and ranking of engineered nanomaterials.	scholarly article	JRCNM02000a // JRCNM04000a // JRCNM01101a // JRCNM01100a // JRCNM02102a // nanomaterial // toxicology
2017-06-01	Graphistrength® C100 MultiWalled Carbon Nanotubes (MWCNT): thirteen-week inhalation toxicity study in rats with 13- and 52-week recovery periods combined with comet and micronucleus assays	scholarly article	JRCNM04002a // Brown Rat // toxicology
2017-05-19	Elucidating the Role of Dissolution in CeO2 Nanoparticle Plant Uptake by Smart Radiolabeling.	scholarly article	JRCNM02102a // general chemistry // catalysis // nanoparticle
2017-04-05	Multi-walled carbon nanotube-physicochemical properties predict the systemic acute phase response following pulmonary exposure in mice.	scholarly article	JRCNM04003a // JRCNM04001a // JRCNM04000a // carbon nanotube
2017-01-03	Negligible cytotoxicity induced by different titanium dioxide nanoparticles in fish cell lines.	scholarly article	JRCNM01005a // JRCNM01004a // JRCNM01003a
2016-11-01	The JRC Nanomaterials Repository: A unique facility providing representative test materials for nanoEHS research	scholarly article	JRC representative nanomaterial // Directorate-General for Joint Research Centre // nanomaterial // toxicology
2015-11-12	Towards the standardization of nanoecotoxicity testing: Natural organic matter 'camouflages' the adverse effects of TiO2 and CeO2 nanoparticles on green microalgae.	scholarly article	JRCNM02102a // JRCNM01003a

The LIPID MAPS hierarchy (in Wikidata)



class	classLabel	Imid	count
Q63433687	fatty acyl	LMFA	0
Q63434442	straight chain fatty acids	LMFA0101	37
Q24901874	branched chain fatty acids	LMFA0102	79
Q61737535	unsaturated fatty acid	LMFA0103	279
Q40211102	hydroxy fatty acid	LMFA0105	184
Q63435564	oxo fatty acids	LMFA0106	56
Q63436532	halogenated fatty acids	LMFA0109	24
Q63434663	amino fatty acids	LMFA0110	39
Q422050	dicarboxylic acid	LMFA0117	78
Q61716319	octadecanoids	LMFA02	82
Q407680	Eicosanoid	LMFA03	83
Q209717	prostaglandins	LMFA0301	89
Q4198767	isoprostane	LMFA0311	5
Q378871	fatty alcohol	LMFA05	156

In which species is this lipid found?

lipid	lipidLabel	lmid	species	speciesLabel	source	sourceLabel	doi
Q26840883	(-)-methyl jasmonate	LMFA02020010	Q23501	Solanum lycopersicum	Q33228063	Induced defences in plants reduce herbivory by increasing cannibalism	10.1038/S41559-017-0231-6
Q27158341	quercetin 5,7,3',4'-tetramethyl ether	LMPK12112771	Q22701	Sambucus nigra	Q39812430	Elderberry flavonoids bind to and prevent H1N1 infection in vitro.	10.1016/J.PHYTOCHEM.2009.06.003
Q55620521	(R)-1,7-Dioxaspiro[5.5]undecane	LMPK09000012	Q2207329	olive fruit fly	Q55645881	Sex-specific activity of (R)-(-) and (S)-(+)-1,7-dioxaspiro[5.5]undecane, the major pheromone of Dacus oleae	10.1007/BF01012372
Q55620476	(S)-1,7-Dioxaspiro[5.5]undecane	LMPK09000013	Q2207329	olive fruit fly	Q55645881	Sex-specific activity of (R)-(-) and (S)-(+)-1,7-dioxaspiro[5.5]undecane, the major pheromone of Dacus oleae	10.1007/BF01012372
Q27135687	geranylacetone	LMFA11000696	Q16528	Nelumbo nucifera	Q902623	ChEBI	
Q27135687	geranylacetone	LMFA11000696	Q16528	Nelumbo nucifera	Q43240571	Comparative analysis of essential oil components and antioxidant activity of extracts of Nelumbo nucifera from various	10.1021/JF902643E



Wikidata and Scholia as a hub linking chemical knowledge

Egon Willighagen^A, Denise Slenter^A, Daniel Mietchen^B, Chris Evelo^{A,C}, Finn Nielsen^D

^A Department of Bioinformatics - BIGCaT, Maastricht University, The Netherlands, ^BData Science Institute, University of Virginia, Charlottesville, Virginia, USA, ^C Maastricht Centre for Systems Biology - MaCSBio, Maastricht University, The Netherlands, ^D Cognitive Systems, DTU Compute, Technical University of Denmark, Denmark

Introduction

Making chemical databases more FAIR (findable, accessible, interoperable, and reusable) benefits computational chemistry and cheminformatics. We here discuss Wikidata, a young sister project of Wikipedia, with one key difference: it is a machine readable database, making it far more useful for interoperability of molecular databases in systems biology [1,2]. Thanks to the WikiProject Chemistry community on Wikidata, there is a growing amount of information about chemical compounds.



Methods

Scholia is a Python/Flask-based server system that creates webpages using a template approach [5]. It defines templates for concepts around knowledge exchange, such as publications, journals, publishers, but also topics. It uses SPARQL queries against the Wikidata Query Service (WDQS,

Results

We here introduce our contributions to the WikiProject Chemistry to support FAIR-ification of open chemical knowledge. For example, we proposed new Wikidata properties to annotate compounds with external database identifiers for the EPA CompTox Dashboard [3], the SPLASH [4], and MetaboLights. We also introduced a Scholia extension [5], visualizing data about chemicals and chemical classes:

<https://tools.wmflabs.org/scholia/>

Provenance: "stated in"

Related compounds

Lookup by identifier

Redirecting

<https://tools.wmflabs.org/scholia/>

Identifiers

QID	Identifier	Count
Q423762	PubChem	182000
Q423762	ChEMBL	182000
Q423762	PubChem CID	182000
Q423762	ChemSpider ID	182000
Q423762	KEGG	182000
Q423762	PubChem Substance ID	182000
Q423762	PubChem CAS Registry Number	182000
Q423762	PubChem EINECS	182000
Q423762	PubChem InChI	182000
Q423762	PubChem InChI Key	182000
Q423762	PubChem IUPAC	182000
Q423762	PubChem Molecular Weight	182000
Q423762	PubChem SMILES	182000
Q423762	PubChem XRef	182000

Literature-backed (PhysChem) Facts



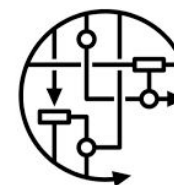
Linking Databases

Identifiers

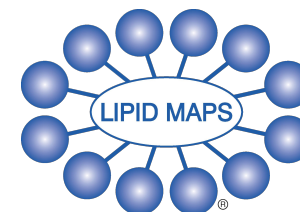
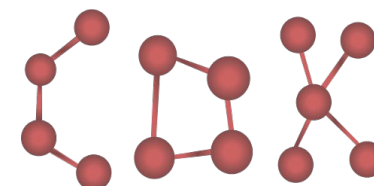
Identifier	Count
PubChem	182000
ChEMBL	182000
PubChem CID	182000
ChemSpider ID	182000
KEGG	182000
PubChem Substance ID	182000
PubChem CAS Registry Number	182000
PubChem EINECS	182000
PubChem InChI	182000
PubChem InChI Key	182000
PubChem IUPAC	182000
PubChem Molecular Weight	182000
PubChem SMILES	182000
PubChem XRef	182000

Acknowledgements

- the many people of WikiProject Chemistry
- Denise Slenter (PhD candidate, metabolites in WikiPathways)
- the Blue Obelisk community
 - Chemistry Development Kit
 - Bioclipse
 - InChI wrapping in Java
- Scholia team
- Roger Sayle (for sharing his slides)



WIKIPATHWAYS
Pathways for the People



List of publications

Show entries

Search:

Date	Work	Type	Pages	Venue	Authors
2021-03-01	Open Natural Products Research: Curation and Dissemination of Biological Occurrences of Chemical Structures through Wikidata	scholarly article	36	bioRxiv	Adriano Rutz , Jonathan Bisson , Jiří Vondrášek , Pierre-Marie Allard , Jean-Luc Wolfender , Guido F Pauli , Ralf Stephan , Maria Sorokina , James G Graham , Egon Willighagen , Daniel Mietchen , Christoph Steinbeck , Roderic D. M. Page , Jakub Galgonek
2021-01-22	A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses	scholarly article	14	BMC Biology	Martina Summer-Kutmon , Lisa M. Verhagen , Daniel Fernández-Álvarez , Jasper Koehorst , Quentin Groom , Peter J. Schaap , Andra Waagmeester , José Emilio Labra Gayo , Andrew I. Su , Egon Willighagen
2020-04-07	A protocol for adding knowledge to Wikidata, a case report	preprint		bioRxiv	Martina Summer-Kutmon , Lisa M. Verhagen , Daniel Fernández-Álvarez , Jasper Koehorst , Peter J. Schaap , Andra Waagmeester , José Emilio Labra Gayo , Andrew I. Su , Egon Willighagen
2020-03-17	Wikidata as a knowledge graph for the life sciences	scholarly article	15	eLife	Sebastian Burgstaller-Muehlbacher , Elvira Mittraka , Lynn Schriml , Kristina Hanspers , Henning Hermjakob , Katherine Thornton , Núria Queralt Rosinach , Gregory Stupp , Anders Riutta , Chunlei Wu , Alexander R. Pico , Toby Hudson , Ginger Tsueng , Andra Waagmeester , Kevin Hybiske , Sarah M Keating , Thomas Shafee , Sabah Ul-Hasan , Michael Mayers , Roger Tu , Ralf Stephan , Timothy Elliott Putman , Andrew I. Su , Benjamin M. Good , Egon Willighagen , Malachi Griffith , Daniel Mietchen , Magnus Manske , Obi Griffith , Denise Slenter

InChI and InChIKey in Wikidata and Scholia

Egon Willighagen
NIH Virtual Workshop on InChI
March 22-24, 2021

@egonwillighagen
0000-0001-7542-0286

