

# Wikidata and Scholia as a hub linking chemical knowledge

Egon Willighagen<sup>A</sup>, Denise Slenter<sup>A</sup>, Daniel Mietchen<sup>B</sup>, Chris Evelo<sup>A,C</sup>, Finn Nielsen<sup>D</sup>

<sup>A</sup> Department of Bioinformatics - BiGCaT, Maastricht University, The Netherlands,

<sup>B</sup> Data Science Institute, University of Virginia, Charlottesville, Virginia, USA,

<sup>C</sup> Maastricht Centre for Systems Biology - MaCSBio, Maastricht University, The Netherlands,

<sup>D</sup> Cognitive Systems, DTU Compute, Technical University of Denmark, Denmark

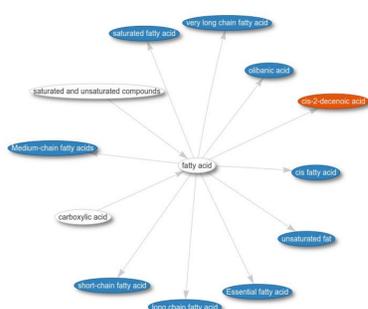
## Introduction

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.

## Results

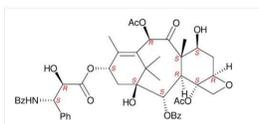
We here introduce our contributions to the WikiProject Chemistry to support FAIR-ification of open chemical knowledge. For example, we proposed new Wikidata properties to annotate compounds with external database identifiers for the EPA CompTox Dashboard [3], the SPLASH [4], and MetaboLights. We also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

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



## 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

IDpred	Id
ATC code	L01CD01
CAS Registry Number	33069-42-4

## Provenance: "stated in"

DSSTOX substance identifier	DTXSID
DTXSID30678817	DTXSID30678817

## Related compounds

SMILES	INChIKey	CAS	ChemSpider	PubChem CID
CC(=O)O <td>CC(=O)O <td>64-19-7 <td>171 <td>176 </td></td></td></td>	CC(=O)O <td>64-19-7 <td>171 <td>176 </td></td></td>	64-19-7 <td>171 <td>176 </td></td>	171 <td>176 </td>	176
CC(=O)OC <td>CC(=O)OC <td>118-50-3 <td>2000003 <td>2723803 </td></td></td></td>	CC(=O)OC <td>118-50-3 <td>2000003 <td>2723803 </td></td></td>	118-50-3 <td>2000003 <td>2723803 </td></td>	2000003 <td>2723803 </td>	2723803
CC(=O)OC(=O)C <td>CC(=O)OC(=O)C <td>2845-02-6 <td>14444 <td>14479 </td></td></td></td>	CC(=O)OC(=O)C <td>2845-02-6 <td>14444 <td>14479 </td></td></td>	2845-02-6 <td>14444 <td>14479 </td></td>	14444 <td>14479 </td>	14479
CC(=O)OC(=O)OC <td>CC(=O)OC(=O)OC <td>1861-76-7 <td>820460 <td>919392 </td></td></td></td>	CC(=O)OC(=O)OC <td>1861-76-7 <td>820460 <td>919392 </td></td></td>	1861-76-7 <td>820460 <td>919392 </td></td>	820460 <td>919392 </td>	919392
CC(=O)OC(=O)OC(=O)C <td>CC(=O)OC(=O)OC(=O)C <td>2881-71-6 <td>28450 <td>40289 </td></td></td></td>	CC(=O)OC(=O)OC(=O)C <td>2881-71-6 <td>28450 <td>40289 </td></td></td>	2881-71-6 <td>28450 <td>40289 </td></td>	28450 <td>40289 </td>	40289
CC(=O)OC(=O)OC(=O)OC <td>CC(=O)OC(=O)OC(=O)OC <td>71-90-1 <td>170 <td>175 </td></td></td></td>	CC(=O)OC(=O)OC(=O)OC <td>71-90-1 <td>170 <td>175 </td></td></td>	71-90-1 <td>170 <td>175 </td></td>	170 <td>175 </td>	175

## Lookup by identifier

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:Q7BSBXVTEAMEQO-UHFFFAOYSA-N](#)  
Redirect also works for InChIKeys, here for acetic acid.

## Identifiers

IDpred	IDpredLabel	count
Q:wd:P235	InChIKey	152393
Q:wd:P233	canonical SMILES	152233
Q:wd:P234	InChI	149944
Q:wd:P662	PubChem CID	145798
Q:wd:P661	ChemSpider ID	125510
Q:wd:P2017	isomeric SMILES	84844
Q:wd:P683	CHEBI ID	84011
Q:wd:P231	CAS Registry Number	72475
Q:wd:P652	UNII	59293
Q:wd:P592	ChEMBL ID	49622
Q:wd:P3117	DSSTOX substance identifier	36373
Q:wd:P232	EC ID	20335
Q:wd:P1579	Beilstein Registry Number	19083
Q:wd:P665	KEGG ID	15065
Q:wd:P2566	ECHA InfoCard ID	12362
Q:wd:P715	Drugbank ID	7786
Q:wd:P595	Guide to Pharmacology Ligand ID	5950
Q:wd:P2057	HMDB ID	5705
Q:wd:P2064	KNAPSack ID	4272

Identifier mappings are made available via BridgeDb.

## 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, query.wikidata.org) and visualizes the data in various forms. Furthermore, we used a combination of Bioclipse (bioclipse.net) and QuickStatements to add missing chemical compounds for biological pathways from WikiPathways [6]. Where needed, new Wikidata properties were proposed.

## Literature-backed (PhysChem) Facts

**Physchem Properties**

Property	Value	Units	Qualifiers
acid dissociation constant	4.74	1	
mass	90.021128	atomic mass unit	
acid dissociation constant	4.756	1	temperature 25
boiling point	117.9	degrees Celsius	pressure: 101325
density	1.0446	gram per cubic centimetre	temperature 14

**Recently published works on the chemical**

Date	Work	Type	Topics
2017-09-09	In vitro human skin permeation of benzoin in gelatin: effect of concentration, moisture energy and skin preparation	scholarly article	oil and gas extraction // benzoin
2017-04-27	Negative enthalpy partitioning: carbonic and volatile organic	scholarly article	toluene // benzene

## Linking Databases

### Identifiers

IDpred	Id
ATC code	G01AD02
ATC code	S02AA10
Beilstein Registry Number	506007
CAS Registry Number	64-19-7
CHEBI ID	15366
CHEMBL ID	CHEMBL539
ChemSpider ID	171
Cosing number	31572
DSSTOX substance identifier	DTXSID5024394
Drugbank ID	03166

## References

- [1] Enabling Open Science: Wikidata for Research (Wiki4R), Research Ideas and Outcomes, 1, **2015**, doi:[10.3897/RIO.1.E7573](https://doi.org/10.3897/RIO.1.E7573) [2] WikiGenomes: an open Web application for community consumption and curation of gene annotation data in Wikidata, Database, 2017:1, **2017**, doi:[10.1101/102046](https://doi.org/10.1101/102046) [3] The CompTox Chemistry Dashboard: a community data resource for environmental chemistry, Journal of Cheminformatics, 9(1), **2017**, doi:[10.1186/S13321-017-0247-6](https://doi.org/10.1186/S13321-017-0247-6) [4] SPLASH, a hashed identifier for mass spectra, Nature Biotechnology, 34(11), **2016**, doi:[10.1038/NBT.3689](https://doi.org/10.1038/NBT.3689) [5] Scholia, Scientometrics and Wikidata, The Semantic Web: ESWC 2017 Satellite Events, **2017**, doi:[10.1007/978-3-319-70407-4\\_36](https://doi.org/10.1007/978-3-319-70407-4_36) [6] WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research", Nucleic Acids Research, 46(D1), **2018**, doi:[10.1093/NAR/GKX1064](https://doi.org/10.1093/NAR/GKX1064)

**Funding** Scholia has received funding from the Alfred P. Sloan Foundation under grant number G-2019-11458.