ontochem
IT SOLUTIONS

# Name-2-Chemistry (N2C)

November 16, 2016

Recognizes, classifies and converts chemical terms into normalized chemical terms and if possible into chemical structures with a CX-SMARTS or SMILES, InChI, and InChI Key. N2C comes as an executable jar file that could be run on any Java enabled platform and integrated into any other Java program. We also provide custom installations as an enterprise Web-Service.

Recognized chemical types are:

| | |
|---|---|
| COMPOUND | Defined chemical compounds that have a SMILES, InChI and InChI Key. If the compound is known in the included database, a preferred name is also returned. |
| CLASS | Chemical classes are converted into normalized classes, with SMILES if possible. In case of a class term which also may describe a compound, the SMILES, InChI and InChI Key of that compound is also returned (e.g. “imidazoles” will return N1C=NC=C1), otherwise a Markush or CX-SMARTS |
| GROUP | Chemical groups are converted into normalized groups, with SMILES or CX-SMARTS if possible. |
| FRAGMENT | Chemical fragments into normalized fragments, with the recognized text string and the SMILES or SMILES + * if possible. |
| POLYMER | Polymeric compounds as defined in the polymer taxonomy e.g. heparin, polyethylene without a defined SMILES string, possibly a CX-SMARTS |
| SUBSTANCE | Chemical compound without a defined SMILES string, possibly a CX-SMARTS |
| INORGANIC_MATERIAL | Inorganic material, with the SMILES, InChI and InChI Key if possible, possibly a CX-SMARTS |

This N2C implementation integrates various modules and annotators such as: regular name-2-structure, inorganic materials, polymer recognition, chemical class, group and fragment recognition, chemical half formula recognition, chemical dictionary based lookup, coordinated entity recognition.

- Name-2-structure engines: Our N2C integrates various rule-based systems for the recognition of systematic and IUPAC-like chemical names and half trivial names, e.g. “1,2-diiodopyridine”.

  The standard N2C package contains Opsin developed by Daniel Lowe (NextMove Software) - this module has been modified and extended to allow for specific chemical terms to be recognized (such as e.g. 2,3-disubstituted pyridine being a CLASS with a SMILES *C1=C(*)N=CC=C1 ).

Optionally, ChemAxon's or any other vendors name-2-structure software can be added as well, in this case the module is activated if Opsin does not return a structure.

- Chemical dictionary based lookup: OntoChem's dictionary of chemical names is built from a selectable database of chemical names (with up to 450 million chemical terms) which is accessed using OntoChem's proprietary fast dictionary technology. The dictionary contains a large variety of chemical expressions including brand names and proprietary terms collected from 16 publicly available sources such as PubChem, MeSH, DrugBank, ChEMBL, among others.
- Coordinated entity recognition: Expressions like "Vitamin B6 and B12" are recognized as a coordinated entities comprising the compounds "vitamin B6" and "vitamin B12".
- An optional chemical half formula and inorganic materials recognizing module developed at OntoChem. This module can be optionally switched on and requires a ChemAxon license (e.g. "Na2SO4" is recognized as sodium sulfate or "Pd77.5Cu6Si16.5" as an alloy.

Example: ascorbate

| | |
|---|---|
| State | SUCCESS |
| chemType | CLASS |
| recognizedTerm | ascorbate |
| smiles | O=C1C(O)=C([O-])[C@H](O1)[C@@H](O)CO |
| preferredName | L-ascorbates |
| inchi | InChI=1S/C6H8O6/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-10H,1H2/p-1/t2,5+/m0/s1 |
| inchiKey | CIWBSHSKHKDKBQ-JLAZNSOCSA-M |

Example: paracetamol

SUCCESS
COMPOUND
Paracetamol
CC(=O)Nc1ccc(O)cc1
p-acetaminophenol
InChI=1S/C8H9NO2/c1-6(10)9-7-2-4-8(11)5-3-7/h2-5,11H,1H3,(H,9,10)
RZVAJINKPMORJF-UHFFFAOYSA-N

Example: butanoyloxy

SUCCESS
GROUP
Butanoyloxy
CCCC(O*)=O
butanoyloxy
null
null

Example: dimethicone

SUCCESS
POLYMER
dimethicone
[#6][Si]([#6])(-[*])[#8]-[*] |$;;;star_e;;star_e$,Sg:any:1,2,0,4::ht|
polydimethylsiloxane polymer
null
null

In case of chemType FRAGMENT, a list of objects is returned:

Example: ethyl acetate:heptane

SUCCESS

FRAGMENT
Fragment1:
ethyl acetate
CCOC(C)=O
acetic acid ethyl ester
InChI=1S/C4H8O2/c1-3-6-4(2)5/h3H2,1-2H3
XEKOWRVHYACXOJ-UHFFFAOYSA-N
Fragment2:
heptane
CCCCCCC
heptane-1,2,3,-d7
InChI=1S/C7H16/c1-3-5-7-6-4-2/h3-7H2,1-2H3/i1D3,3D2,5D2
IMNFDUFMRHMDMM-HJHJEWFGSA-N

Example: Vitamin B6 and B12

SUCCESS
FRAGMENT
Fragment1:
Vitamin B6
COMPOUND
Cc1ncc(COP(O)(O)=O)c(C=O)c1O
pyridoxal 5'-phosphate
InChI=1S/C8H10NO6P/c1-5-8(11)7(3-10)6(2-9-5)4-15-16(12,13)14/h2-3,11H,4H2,1H3,(H2,12,13,14)
NGVDGCNFYWLIFO-UHFFFAOYSA-N
Fragment2:
B12
COMPOUND
N#C[Co]N([C@@H]1[C@H](CC(N)=O)[C@@]2(C)CCC(=O)NC[C@@H](C)OP(O)(=O)O[C@H]3[C@H]([C@H](O[C@@H]3CO)N3C4=CC(C)=C(C)C=C4N=C3)O)\C2=C(C)/C([C@H](C\2(C)C)CCC(N)=O)=N/C/2=C\C([C@H]([C@@]/2(CC(N)=O)C)CCC(N)=O)=N\C\2=C(C)/C2=N[C@]1(C)[C@@](C)(CC(N)=O)[C@@H]2CCC(N)=O
vitamin B12
InChI=1S/C62H90N13O14P.CN.Co/c1-29-20-39-40(21-30(29)2)75(28-70-39)57-52(84)53(41(27-76)87-57)89-90(85,86)88-31(3)26-69-49(83)18-19-59(8)37(22-46(66)80)56-62(11)61(10,25-48(68)82)36(14-17-45(65)79)51(74-62)33(5)55-60(9,24-47(67)81)34(12-15-43(63)77)38(71-55)23-42-58(6,7)35(13-16-44(64)78)50(72-42)32(4)54(59)73-56;1-2;/h20-21,23,28,31,34-37,41,52-53,56-57,76,84H,12-19,22,24-27H2,1-11H3,(H15,63,64,65,66,67,68,69,71,72,73,74,77,78,79,80,81,82,83,85,86);;/q;;+1/p-1/t31-,34-,35-,36-,37+,41-,52-,53-,56-,57+,59-,60+,61+,62+;;/m1../s1
SEKGMJVHSBBHRD-WZHZPDAFSA-M

## Add-Ons:

## Chemistry-Parent Classifications

Assigns chemical attributes (ancestor classes) to the chemical terms from above by using optionally or together

- OntoChem classifications
- ChEBI classifications
- MeSH classifications

**Example**:  ***is_a***

Benzenes; 5-7-membered cyclic compounds; 6-membered carbocycles; aromatic hydrocarbons; carbocycles; cyclic compounds; heteroarenes; hydrocarbons; monocyclic carbocycles; organic compounds; lead like molecules; lipinski molecules; lipophilic molecules; small molecules;

CHEBI:33655; CHEBI:33658; CHEBI:22712; CHEBI:33598; CHEBI:33595; CHEBI:33833; CHEBI:24632; CHEBI:33661;

MeSH:D006838 (hydrocarbons); MeSH:D001554 (benzene);

Requires an existing ChemAxon JChem Base license.