Need and benefits for structure standardization to facilitate integration and connectivity between government databases

Valery Tkachenko[2], Christopher Grulke[1], Antony Williams[1]

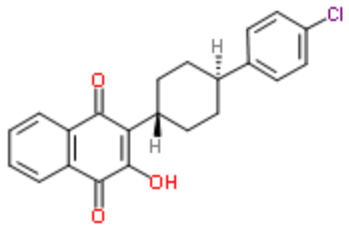
The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. EPA

[1] National Center for Computational Toxicology, EPA, NC, United States;
[2] SCIENCE DATA SOFTWARE, Rockville, MD, United States

Fall ACS 2017, Washington, DC

ChemSpider – an example of chemical database

Search term: **atovaquone** (Found by approved synonym)



2D 3D Save Edit Zoom

- 2 of 2 defined stereocentres

Atovaquone

ChemSpider ID: **10482034**

Molecular Formula: $C_{22}H_{19}ClO_3$

Average mass: 366.837494 Da

Monoisotopic mass: 366.102264 Da

▼ Systematic name

2-[trans-4-(4-Chlorophenyl)cyclohexyl]-3-hy

▶ SMILES and InChIs

▶ Cite this record

Wikibox

Embed

Deprecate

Watch this record

Manage data slice

▼ Names and Identifiers

Names and Synonyms | Database ID(s)

Validated by Experts, Validated by Users, Non-Validated, ~~Removed by Users~~, Redirected by Users, **Redirect Approved by Experts**

(-)-Cholesterol

(3b)-cholest-5-en-3-ol

(3S,8S,9S,10R,13R,14S,17R)-10,13-Dimethyl-17-[(2R)-6-methyl-2-heptanyl]-2,3,4,7,8,9,10,11,12,13,14,15,16,17-tetradecahydro-1H-cyclopenta[a]phenanthren-3-ol

(3S,8S,9S,10R,13R,14S,17R)-10,13-Diméthyl-17-[(2R)-6-méthyl-2-heptanyl]-2,3,4,7,8,9,10,11,12,13,14,15,16,17-tétradécahydro-1H-cyclopenta[a]phénanthrén-3-ol *[French]*

(3β)-cholest-5-en-3-ol *[ACD/IUPAC Name]*

(3β)-Cholest-5-en-3-ol *[German] [ACD/IUPAC Name]*

(3β)-Cholest-5-én-3-ol *[French] [ACD/IUPAC Name]*

3b-Hydroxy-5-cholestene

3β-Hydroxycholest-5-ene

5:6-Cholesten-3b-ol

More...

▼ Properties

Experimental data | Predicted - ACD/Labs | Predicted - EPISuite | Predicted - ChemAxon

Data supplied by datasources and users.

- **Experimental Physico-Chemical Properties**

  Experimental Melting Point:
  149 °C **Tokyo Chemical Industry Ltd** C0318
  147-150 °C **Alfa Aesar**
  148-150 °C **Oxford University Chemical Safety Data** http://msds.chem.ox.ac.uk/CH/cholesterol.html
  147-150 °C **Alfa Aesar** A11470

  Experimental Boiling Point:
  360 °C **Alfa Aesar**
  360 °C **Oxford University Chemical Safety Data** http://msds.chem.ox.ac.uk/CH/cholesterol.html
  360 °C **Alfa Aesar** A11470

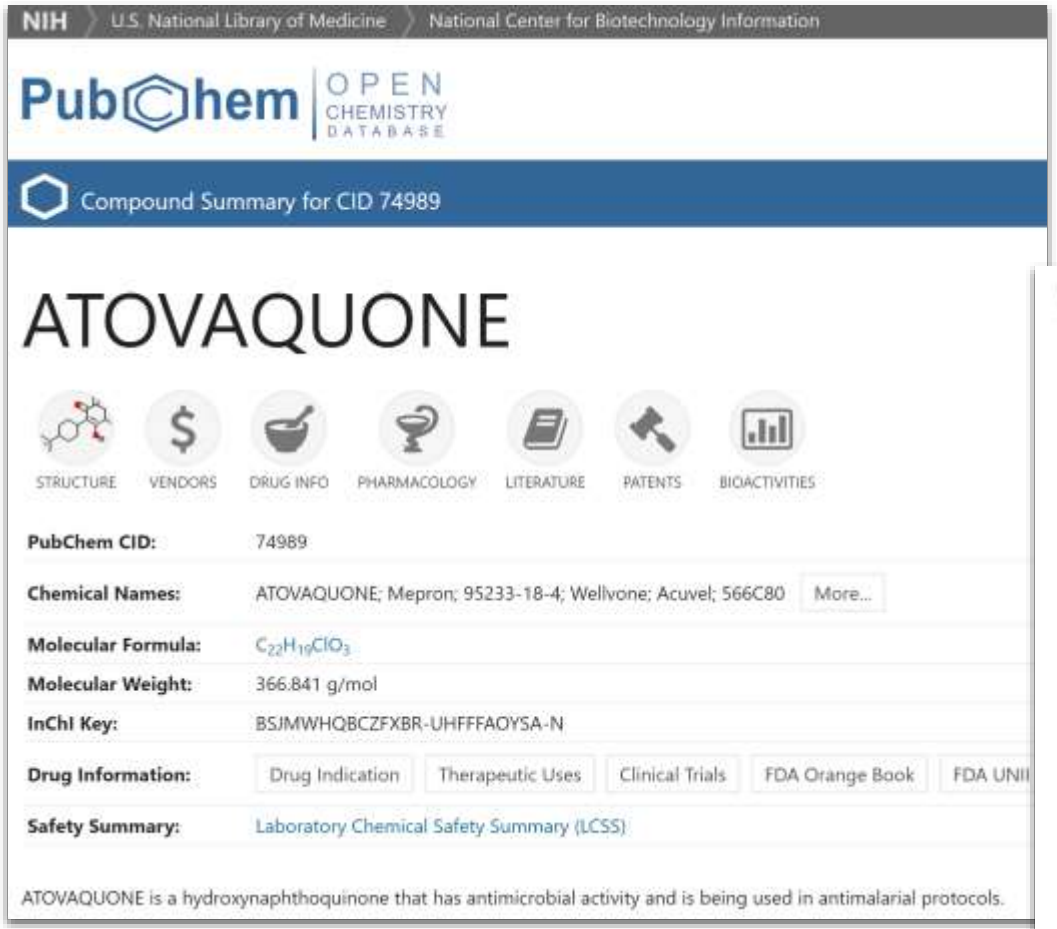
  Experimental Optical Rotation:
  -36 **Alfa Aesar** A11470

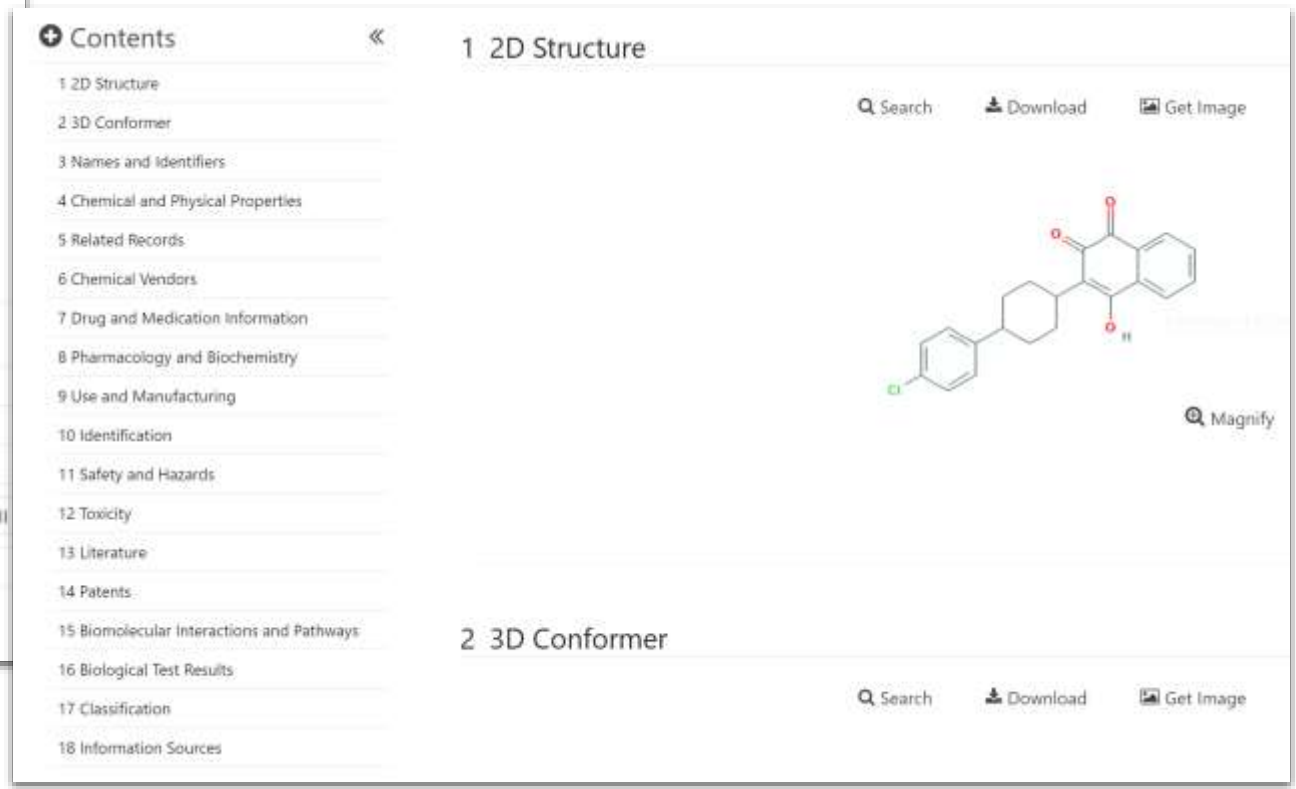
  Experimental Gravity:
  1.067 g/mL **Alfa Aesar** A11470

- **Predicted Physico-Chemical Properties**

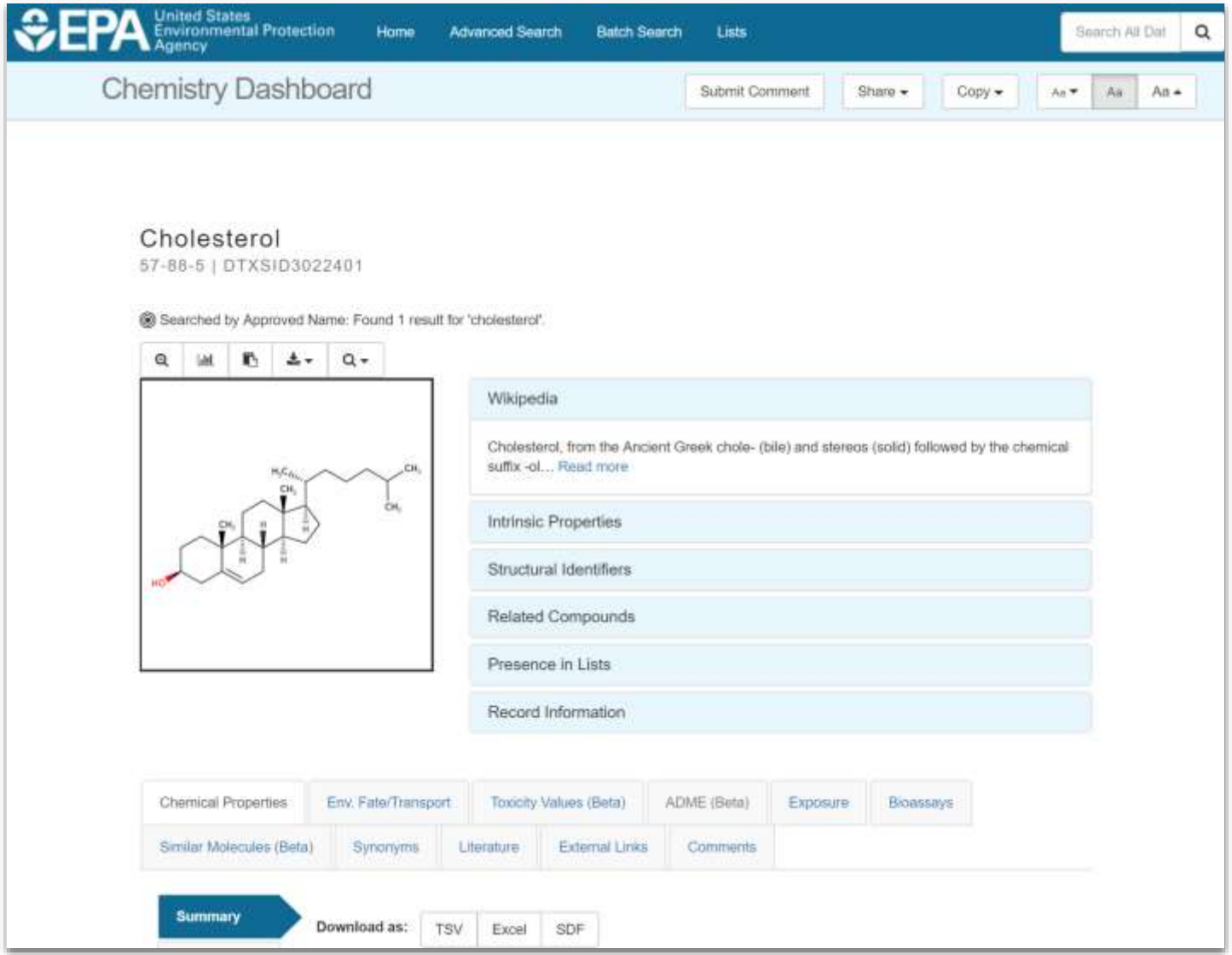
  Predicted Melting Point:
  149 °C **Tokyo Chemical Industry Ltd**
  149 °C **Tokyo Chemical Industry Ltd** C0318

PubChem


NIH
U.S. National Library of Medicine
National Center for Biotechnology Information
PubChem
OPEN CHEMISTRY DATABASE
Compound Summary for CID 74989
ATOVAQUONE
STRUCTURE
VENDORS
DRUG INFO
PHARMACOLOGY
LITERATURE
PATENTS
BIOACTIVITIES
PubChem CID: 74989
Chemical Names: ATOVAQUONE; Mepron; 95233-18-4; Wellvone; Acuvel; 566C80 More...
Molecular Formula: $C_{22}H_{19}ClO_3$
Molecular Weight: 366.841 g/mol
InChI Key: BSJMWHQBCZFXBR-UHFFFAOYSA-N
Drug Information: Drug Indication | Therapeutic Uses | Clinical Trials | FDA Orange Book | FDA UNII
Safety Summary: Laboratory Chemical Safety Summary (LCSS)
ATOVAQUONE is a hydroxynaphthoquinone that has antimicrobial activity and is being used in antimalarial protocols.


Contents
1 2D Structure
2 3D Conformer
3 Names and Identifiers
4 Chemical and Physical Properties
5 Related Records
6 Chemical Vendors
7 Drug and Medication Information
8 Pharmacology and Biochemistry
9 Use and Manufacturing
10 Identification
11 Safety and Hazards
12 Toxicity
13 Literature
14 Patents
15 Biomolecular Interactions and Pathways
16 Biological Test Results
17 Classification
18 Information Sources
1 2D Structure
Search
Download
Get Image
Magnify
2 3D Conformer
Search
Download
Get Image

NCCT Chemistry Dashboard



Data quality issues

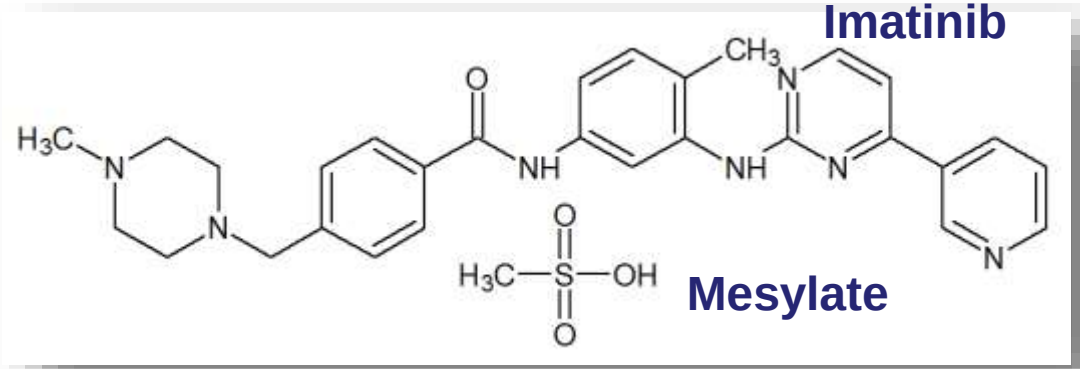
Robochemistry

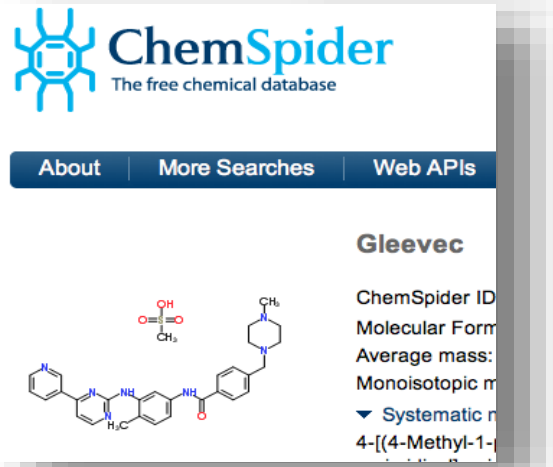
Proliferation of errors in public and private databases

Automated quality control system

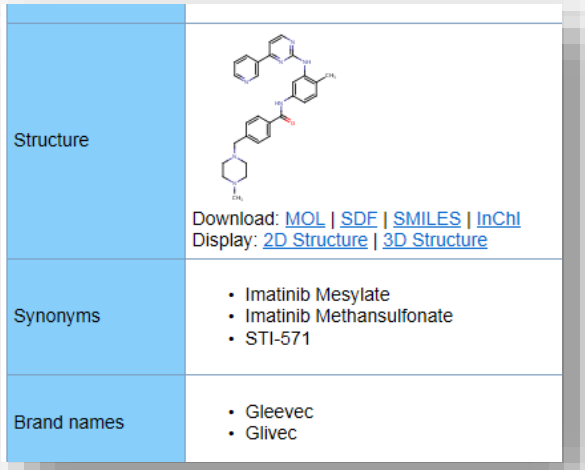
Ambiguities

What Is Gleevec?

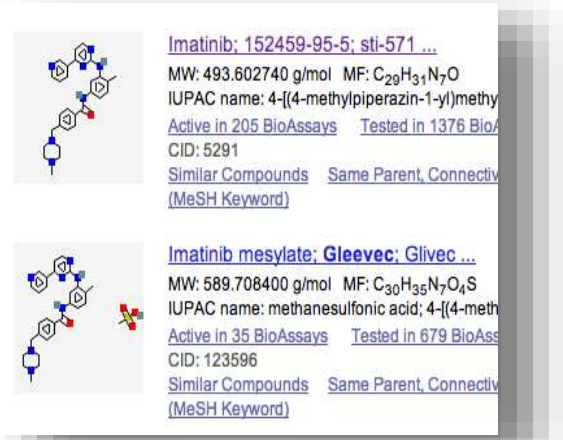




ChemSpider



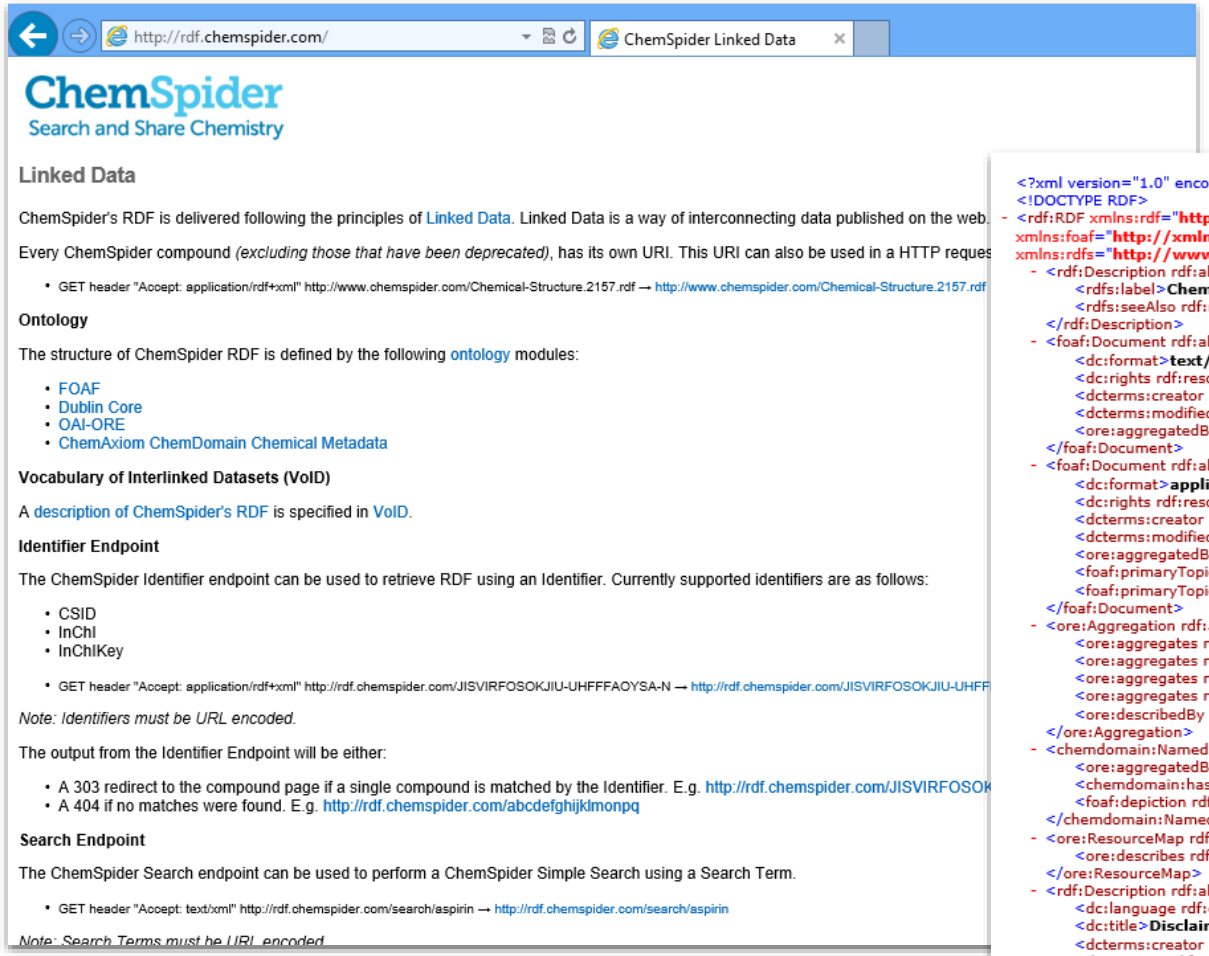
Drugbank



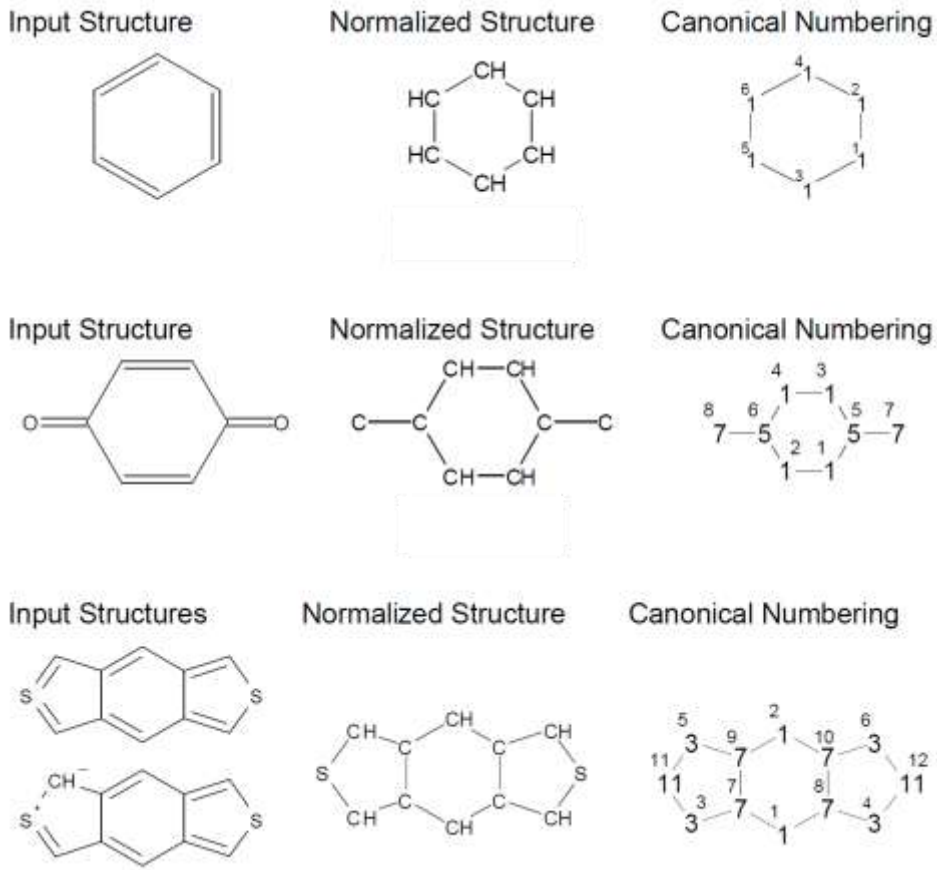
PubChem

We live in a hyperconnected World

What is “sharing in a proper way”?



InChI (http://www.inchi-trust.org/)



{InChI version}

1. Main Layer (M):

/{formula}
/c{connections}
/h{H_atoms}

2. Charge Layer

/q{charge}
/p{protons}

3. Stereo Layer

/b{stereo:dbond}
/t{stereo:sp3}
/m{stereo:sp3:inverted}
/s{stereo:type (1=abs, 2=rel, 3=rac)}

4. Isotopic Layer (MI):

/i{isotopic:atoms}*
/h{isotopic:exchangeable_H}
/b{isotopic:stereo:dbond}
/t{isotopic:stereo:sp3}
/m{isotopic:stereo:sp3:inverted}
/s{isotopic:stereo:type (1=abs, 2=rel, 3=rac)}

5. Fixed H Layer (F):

/f{fixed_H:formula}*
/h{fixed_H:H_fixed}
/q{fixed_H:charge}
/b{fixed_H:stereo:dbond}
/t{fixed_H:stereo:sp3}
/m{fixed_H:stereo:sp3:inverted}
/s{fixed_H:stereo:type (1=abs, 2=rel, 3=rac)}

(6.) Fixed/Isotopic Combination (FI)

/i{fixed_H:isotopic:atoms}*
/b{fixed_H:isotopic:stereo:dbond}
/t{fixed_H:isotopic:stereo:sp3}
/m{fixed_H:isotopic:stereo:sp3:inverted}
/s{fixed_H:isotopic:stereo:type (1=abs, 2=rel, 3=rac)}
/o{transposition}

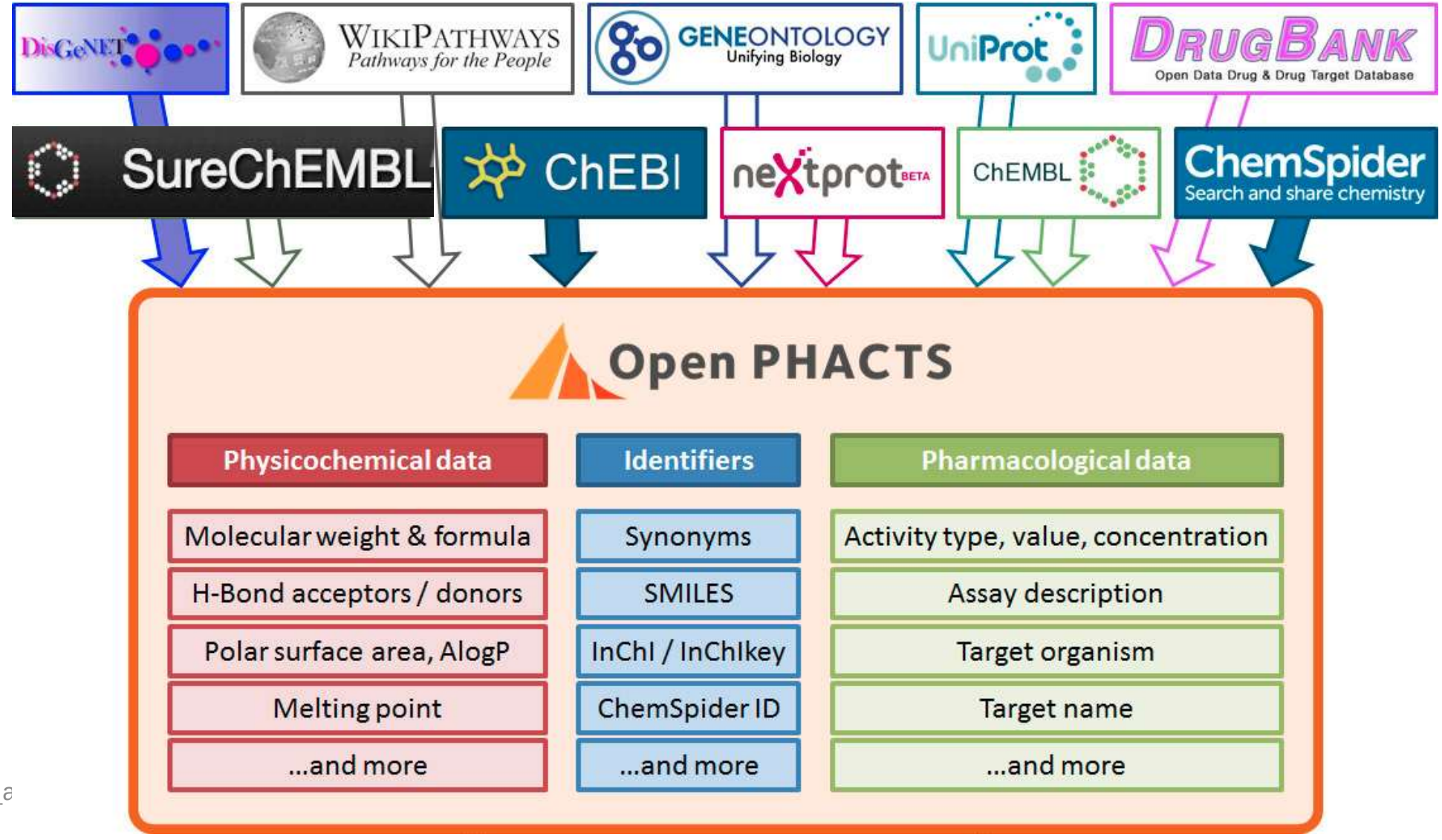
How is this a semantic web problem? Why can’t people just be clear?

People may be working with faulty data.

Salts, say, may make little difference to the effects of an active ingredient.

People may assume a one-to-one mapping between a gene and the gene product (protein, ncRNA) that it codes for.

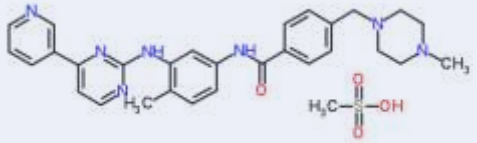
Knowledge is federated



What is lenses?

- Equivalence rules
- The BridgeDB vocabulary adds metadata that provides a justification for treating two URIs alike, thus allowing the researcher to determine whether their circumstances fit.
- owl:sameAs ≤ skos:exactMatch ≤ skos:closeMatch ≤ rdfs:seeAlso
- The ChEBI and CHEMINF ontologies provide a rich set of relations (many of which developed for this project) to relate one molecule to another.

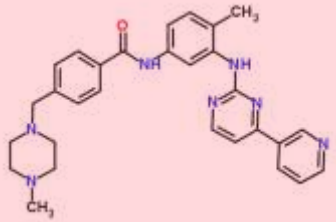
ChemSpider
Search and share chemistry



Link: skos:closeMatch
Reason: non-salt form

Link: skos:exactMatch
Reason: drug name

DrugBank
Open Data Drug & Drug Target Database



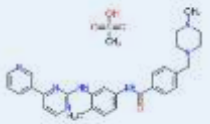
Strict

Relaxed

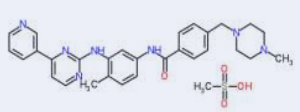
Analysing

Browsing

ChEMBL

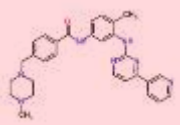


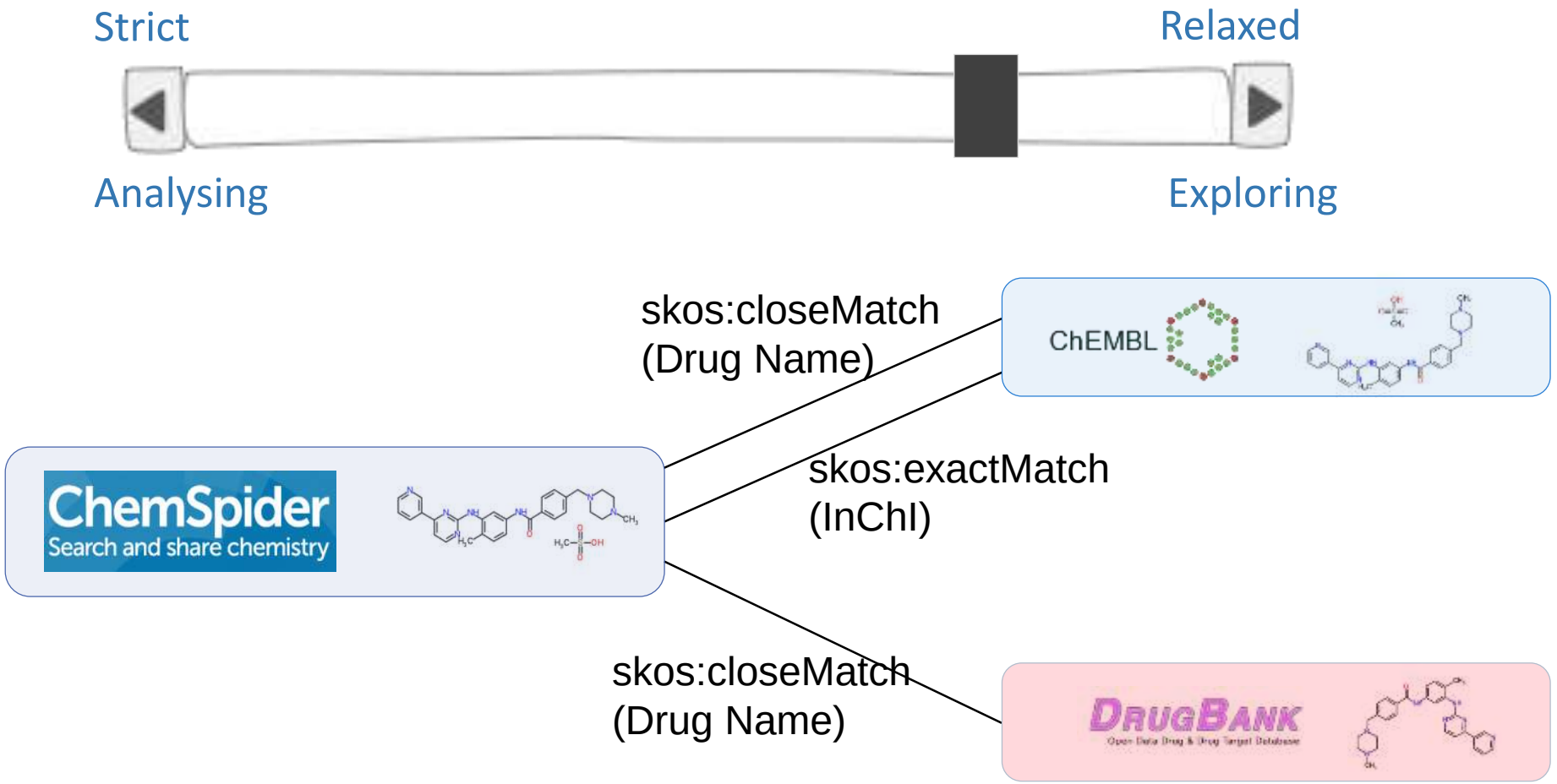
ChemSpider
Search and share chemistry



skos:exactMatch
(InChI)

DrugBank




Strict
Relaxed
Analysing
Exploring
skos:closeMatch
(Drug Name)
ChEMBL
skos:exactMatch
(InChI)
ChemSpider
Search and share chemistry
skos:closeMatch
(Drug Name)
DrugBank

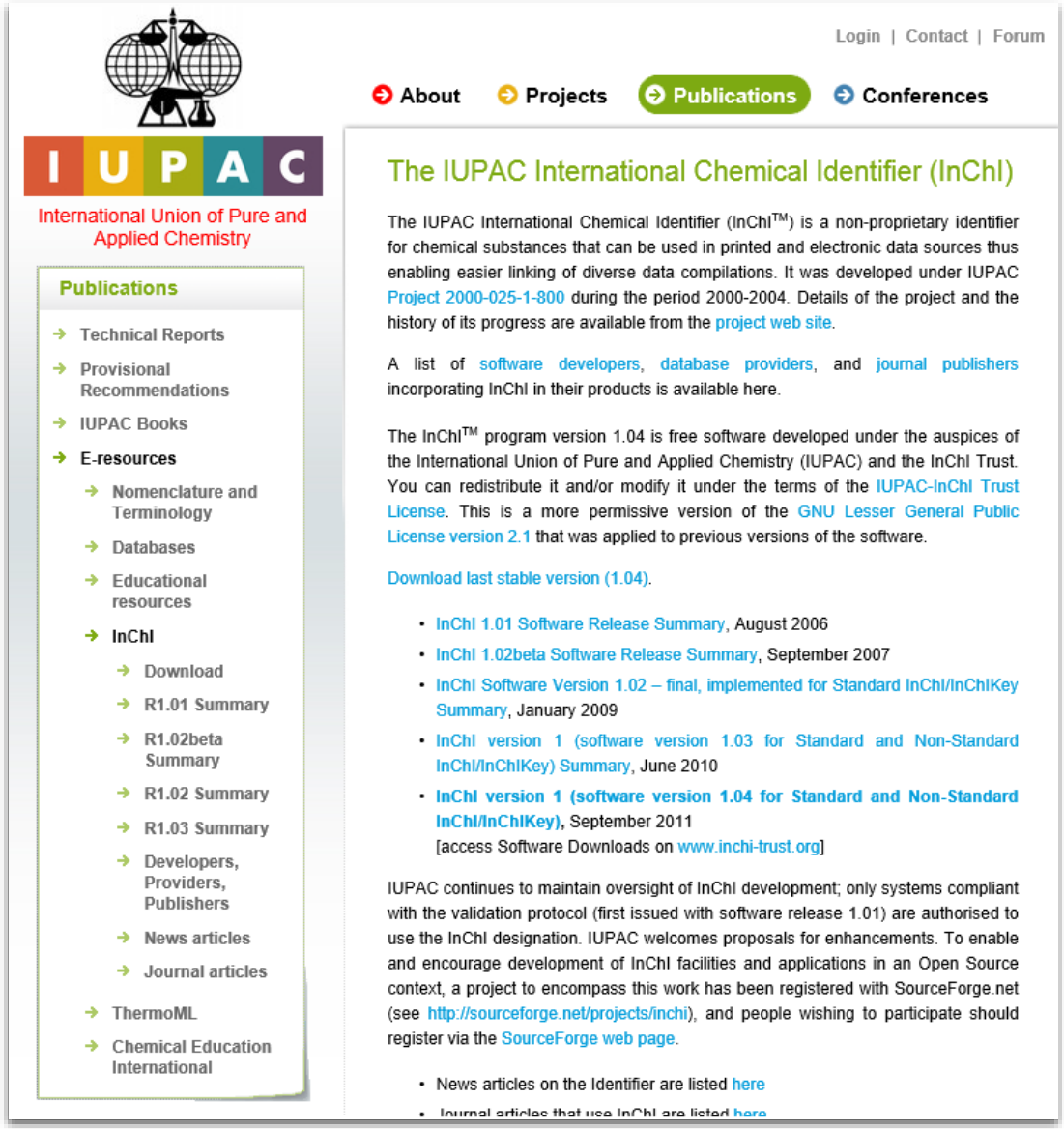
What does the Open PHACTS Chemistry Registration System do?

Takes in structures from ChEMBL, ChEBI, DrugBank, PDB, Thomson Reuters.

Normalizes structures according to rules based on FDA guidelines.

Generates counterpart molecules: without charge, fragments

Standards



IUPAC

International Union of Pure and Applied Chemistry

Login | Contact | Forum

About | Projects | Publications | Conferences

Publications

- Technical Reports
- Provisional Recommendations
- IUPAC Books
- E-resources
  - Nomenclature and Terminology
  - Databases
  - Educational resources
  - InChI
    - Download
    - R1.01 Summary
    - R1.02beta Summary
    - R1.02 Summary
    - R1.03 Summary
    - Developers, Providers, Publishers
    - News articles
    - Journal articles
  - ThermoML
  - Chemical Education International

The IUPAC International Chemical Identifier (InChI)

The IUPAC International Chemical Identifier (InChI™) is a non-proprietary identifier for chemical substances that can be used in printed and electronic data sources thus enabling easier linking of diverse data compilations. It was developed under IUPAC Project 2000-025-1-800 during the period 2000-2004. Details of the project and the history of its progress are available from the project web site.

A list of software developers, database providers, and journal publishers incorporating InChI in their products is available here.

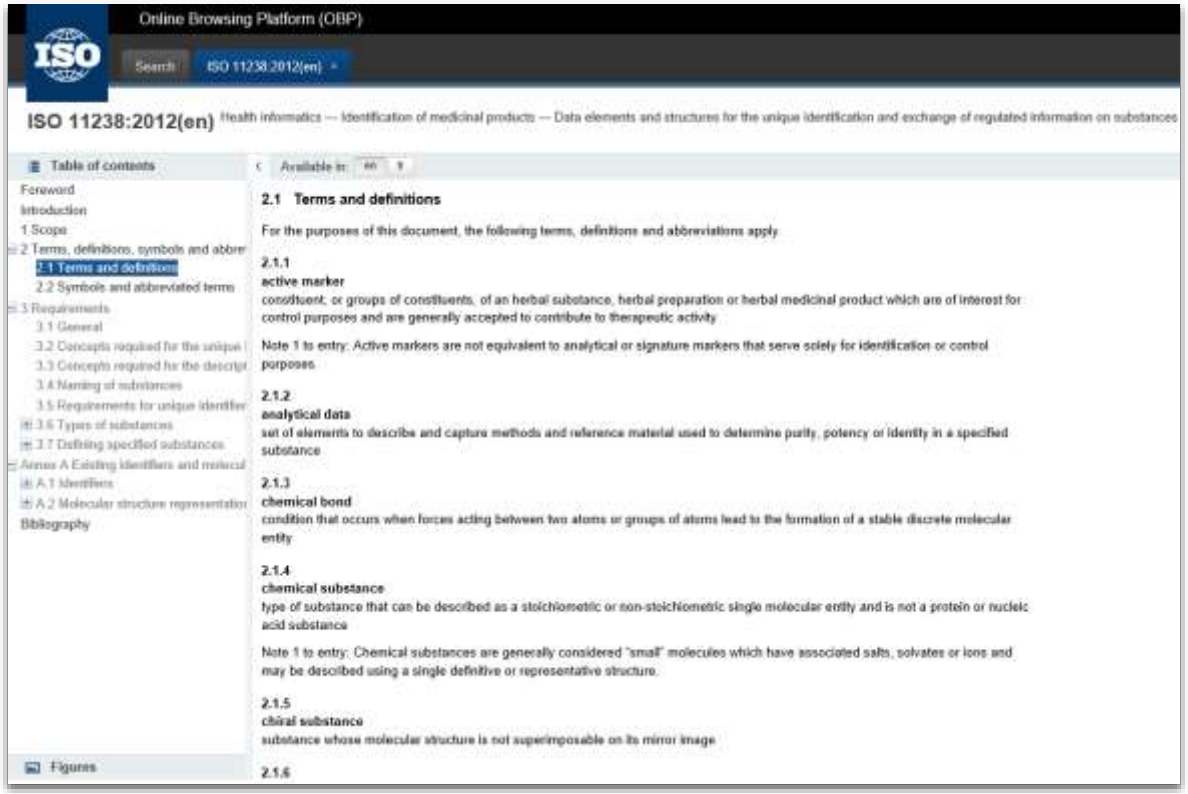
The InChI™ program version 1.04 is free software developed under the auspices of the International Union of Pure and Applied Chemistry (IUPAC) and the InChI Trust. You can redistribute it and/or modify it under the terms of the IUPAC-InChI Trust License. This is a more permissive version of the GNU Lesser General Public License version 2.1 that was applied to previous versions of the software.

Download last stable version (1.04).

- InChI 1.01 Software Release Summary, August 2006
- InChI 1.02beta Software Release Summary, September 2007
- InChI Software Version 1.02 – final, implemented for Standard InChI/InChIKey Summary, January 2009
- InChI version 1 (software version 1.03 for Standard and Non-Standard InChI/InChIKey) Summary, June 2010
- **InChI version 1 (software version 1.04 for Standard and Non-Standard InChI/InChIKey),** September 2011
  [access Software Downloads on www.inchi-trust.org]

IUPAC continues to maintain oversight of InChI development; only systems compliant with the validation protocol (first issued with software release 1.01) are authorised to use the InChI designation. IUPAC welcomes proposals for enhancements. To enable and encourage development of InChI facilities and applications in an Open Source context, a project to encompass this work has been registered with SourceForge.net (see http://sourceforge.net/projects/inchi), and people wishing to participate should register via the SourceForge web page.

- News articles on the Identifier are listed here
- Journal articles that use InChI are listed here



Online Browsing Platform (OBP)

ISO 11238:2012(en) Health informatics — Identification of medicinal products — Data elements and structures for the unique identification and exchange of regulated information on substances

Table of contents

Foreword
Introduction
1 Scope
2 Terms, definitions, symbols and abbre
2.1 Terms and definitions
2.2 Symbols and abbreviated terms
3 Requirements
3.1 General
3.2 Concepts required for the unique
3.3 Concepts required for the descript
3.4 Naming of substances
3.5 Requirements for unique identifier
3.6 Types of substances
3.7 Defining specified substances
Annex A Existing identifiers and molecul
A.1 Identifiers
A.2 Molecular structure representation
Bibliography

Figures

2.1 Terms and definitions

For the purposes of this document, the following terms, definitions and abbreviations apply

2.1.1
active marker
constituent, or groups of constituents, of an herbal substance, herbal preparation or herbal medicinal product which are of interest for control purposes and are generally accepted to contribute to therapeutic activity

Note 1 to entry: Active markers are not equivalent to analytical or signature markers that serve solely for identification or control purposes

2.1.2
analytical data
set of elements to describe and capture methods and reference material used to determine purity, potency or identity in a specified substance

2.1.3
chemical bond
condition that occurs when forces acting between two atoms or groups of atoms lead to the formation of a stable discrete molecular entity

2.1.4
chemical substance
type of substance that can be described as a stoichiometric or non-stoichiometric single molecular entity and is not a protein or nucleic acid substance

Note 1 to entry: Chemical substances are generally considered "small" molecules which have associated salts, solvates or ions and may be described using a single definitive or representative structure.

2.1.5
chiral substance
substance whose molecular structure is not superimposable on its mirror image

2.1.6

[Very incomplete] list of common problems

- Violations of chemical and common sense
- Violations of valence bond theory
- Unsupported format and chemical model features
- Information loss during conversion
- Tautomers
- Stereochemical issues
- Mixtures
- Other classes of chemicals (materials, formulations, biologicals, structurally diverse, etc)
- Equivalence/mapping issues
- Identifiers/names issues
- Etc, etc, etc…

…problems (continued)

- Multiple [historical, proprietary, shortcoming] formats
  - ChemDraw, ChemSketch, AccelrysDraw
  - MOL, SDF
  - SMILES
  - Identifiers
  - Names and Synonyms
- Multiple toolkits/models
  - Open Source (alphabetical)
    - CDK
    - RDKit
    - Indigo
    - OpenBabel
    - Etc…
  - Commercial (alphabetical)
    - CACTVS
    - ChemAxon
    - OpenEye
    - Etc…
- Hystorical software
- No [machine-readable] standards
- ~~No authorities~~ No coordinated efforts!!!

How to link and integrate various resources

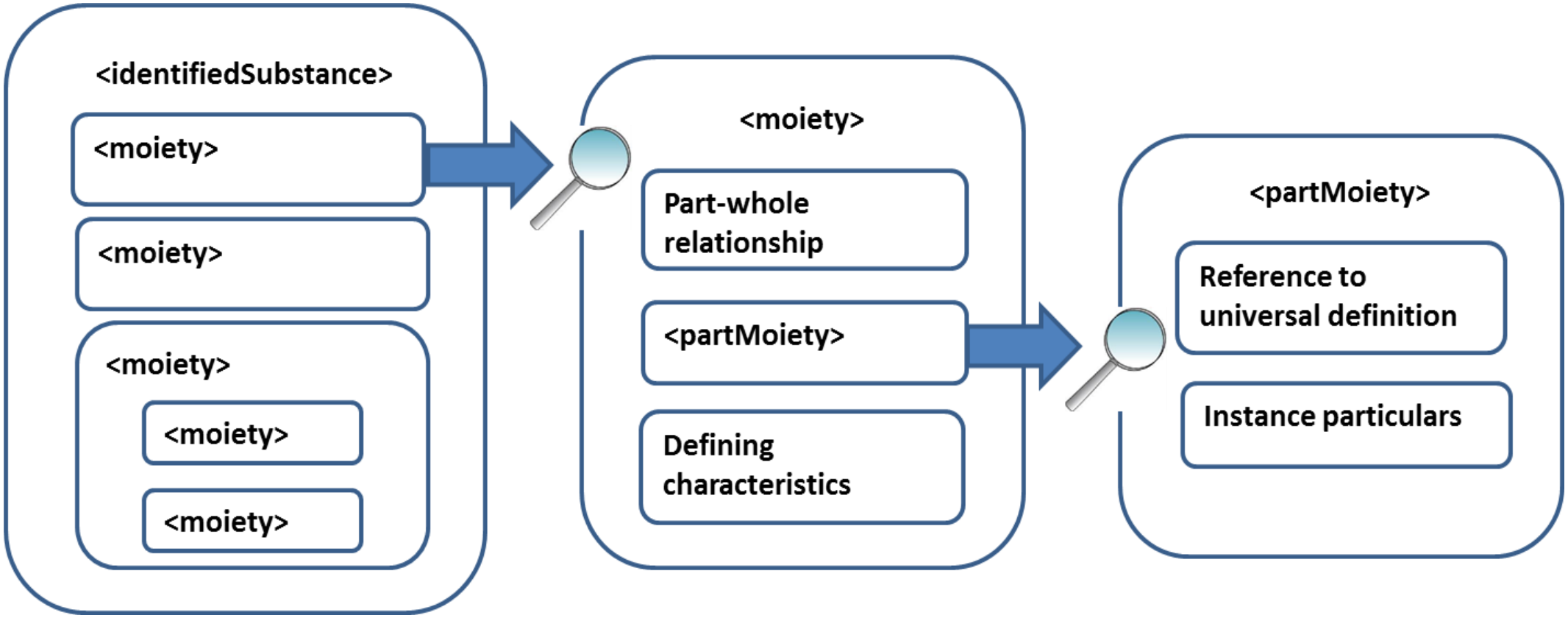
- Available in a variety of databases
- Expressed in a variety of formats
- Some data types are too complex to be exchanged by standard formats. Specific examples are
  - Complex mixtures with defined, or ill-defined concentrations
  - Biological substances
  - Polymers

Structured Product Labeling (SPL)

Health Level Seven (HL7) Structured Product Labeling (SPL)

- an ANSI-accredited data exchange standard
- adopted in 2004 by FDA for the exchange of health and regulatory product and facility data

SPL model


<identifiedSubstance>
<moiety>
<moiety>
<moiety>
<moiety>
<moiety>
<moiety>
<moiety>
Part-whole relationship
<partMoiety>
Defining characteristics
<partMoiety>
Reference to universal definition
Instance particulars

Moiety role	NCIt code	Defining characteristic/representation type	Part-whole relationship	Instance particula
Simple chemical	-[a]	Chemical structure/ MOLFILE, InChI, InChIKey Stereochemistry Type/CV	<quantity>	<id>
Protein subunit	C118424	Chemical structure/ amino acid letter sequence	<quantity>	<id>
Polymeric subunit	???	Chemical structure/ MOLFILE, InChI, InChIKey Stereochemistry Type/CV	<quantity>	<id>
Mixture component	C103243	Variable[b]		
Structural modification	C118425	Chemica InChIKey Stereoch		
Amino acid connection points	C118427	-		
Linear SRU connection points	???	-		

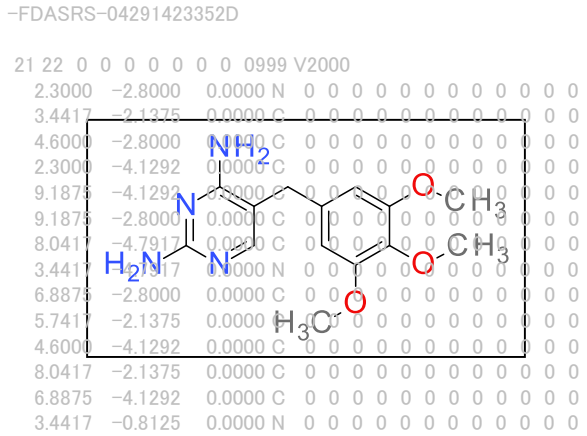
Type	MIME Media Type
Molfile	application/x-mdl-molfile
InChI	application/x-inchi
InChIkey	application/x-inchi-key
Amino acid sequence	application/x-aa-seq
DNA Sequence	application/x-dna-seq
RNA Sequence	application/x-rna-seq

Letter code	Amino acid
A (a)	Alanine
R (r)	Arginine
N (n)	Asparagine
D (d)	Aspartic acid
B (b)	Asparagine or aspartic acid
C (c)	Cysteine
E (e)	Glutamic acid
Q (q)	Glutamine
Z (z)	Glutamine or glutamic acid
G	Glycine
H (h)	Histidine
I (i)	Isoleucine
L (l)	Leucine
K (k)	Lysine
M (m)	Methionine
F (f)	Phenylalanine
P (p)	Proline
S (s)	Serine
T (t)	Threonine
W (w)	Tryptophan
Y (y)	Tyrosine
V (v)	Valine
X	a non-standard amino acid

Stereochemistry type	NCIt code
Square Planar 1 Molecular Geometry	C103211
Square Planar 2 Molecular Geometry	C103212
Square Planar 3 Molecular Geometry	C103213
Square Planar 4 Molecular Geometry	C103214
Tetrahedral Molecular Geometry	C103215
Octahedral 12 Molecular Geometry	C103216
Octahedral 22 Molecular Geometry	C103217
Octahedral 21 Molecular Geometry	C103218
Cahn-Ingold-Prelog Priority System	C103219
Axial R	C103220
Axial S	C103221

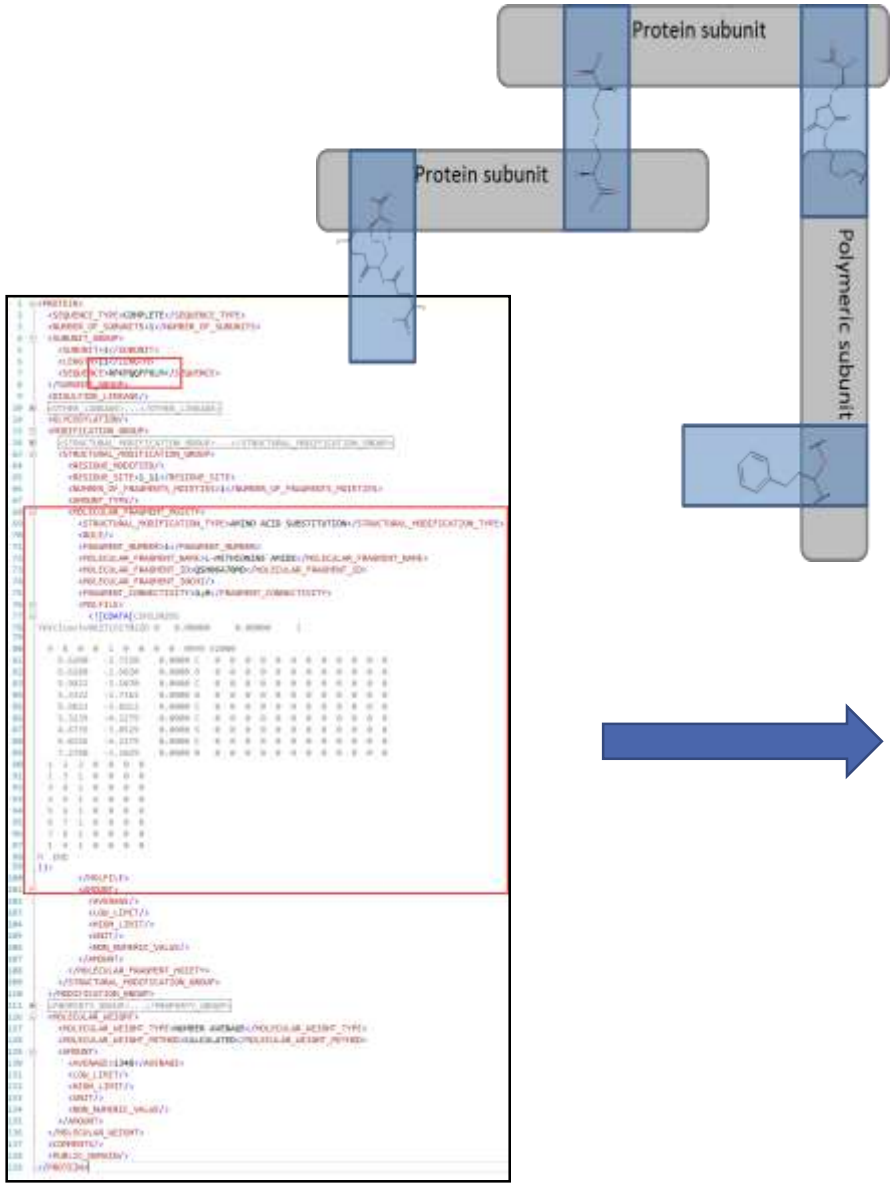
Representation of chemical substance in SPL standard

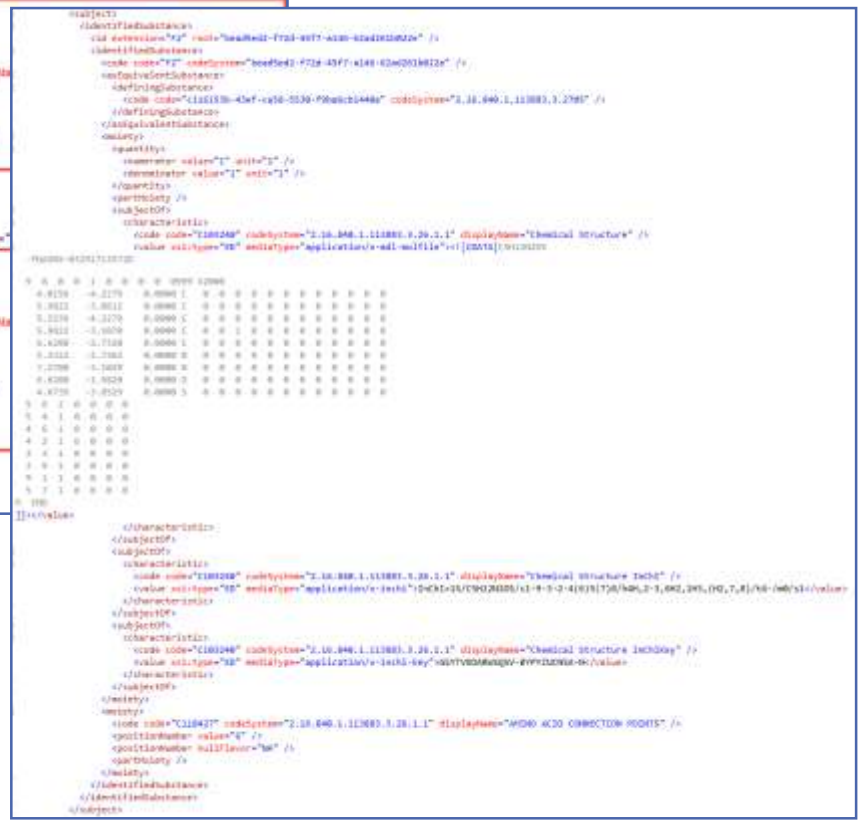
- Chemical substance
  - Chemical structure (MOLFILE)



  - InChI=1S/C14H18N4O3/c1-19-10-5-8(6-11(20-2)12(10)21-3)4-9-7-17-14(16)18-13(9)15/h5-7H,4H2,1-3H3,(H4,15,16,17,18)
  - IEDVJHCEMCRBQM-UHFFFAOYSA-N

Modified Proteins





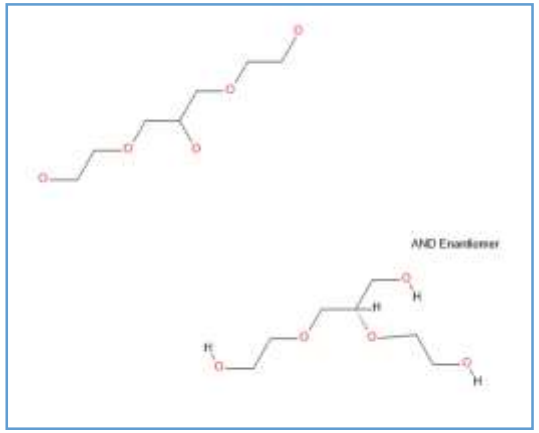
Mixtures

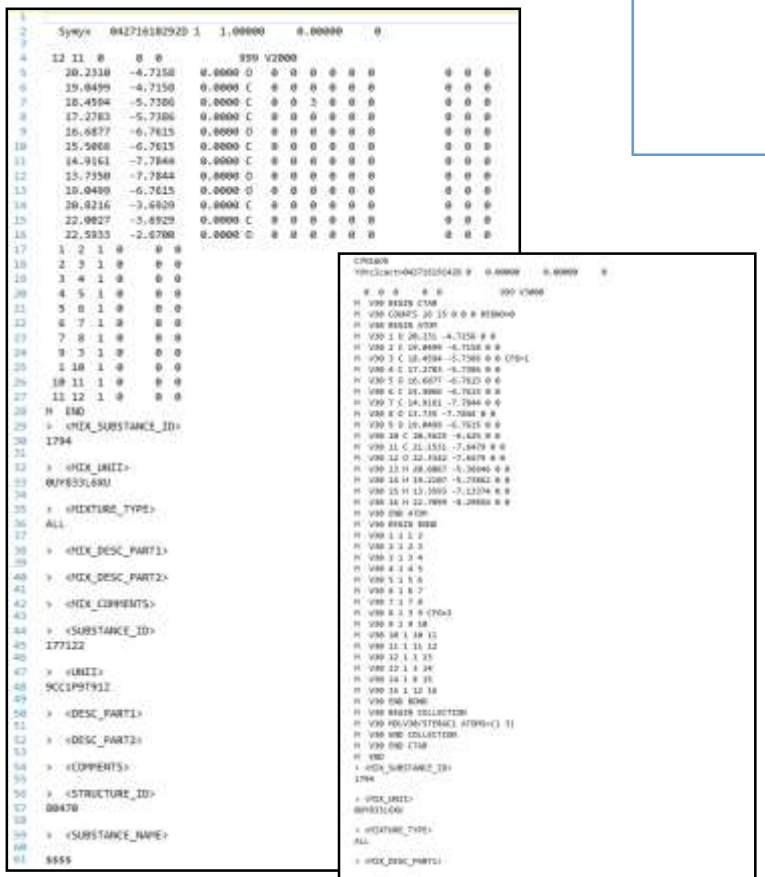
IDMP:

- Mixture substances shall be described as simple combinations of single substances that are either isolated together or are the result of the same synthetic process.

- Mixture substances shall not be combinations of diverse material brought together to form a product.

Mixtures





Example Complex Mixtures

- “Aroclors” are complex mixtures of polychlorinated biphenyls (PCBs). There are 209 possible PCBs and different Aroclors are combinations of a series of these 209 variants and at specific ranges of concentrations.
- Ideally SPL will carry information about the individual components and the concentration of each component for a specific Aroclor
- Work in progress and looking promising!

Substances in products

- Small molecules
- Proteins
- Nucleic acids
- Polymers
- Organisms
- Parts of organisms
- Mixtures

FAIR Data Principles

To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

To be Accessible:

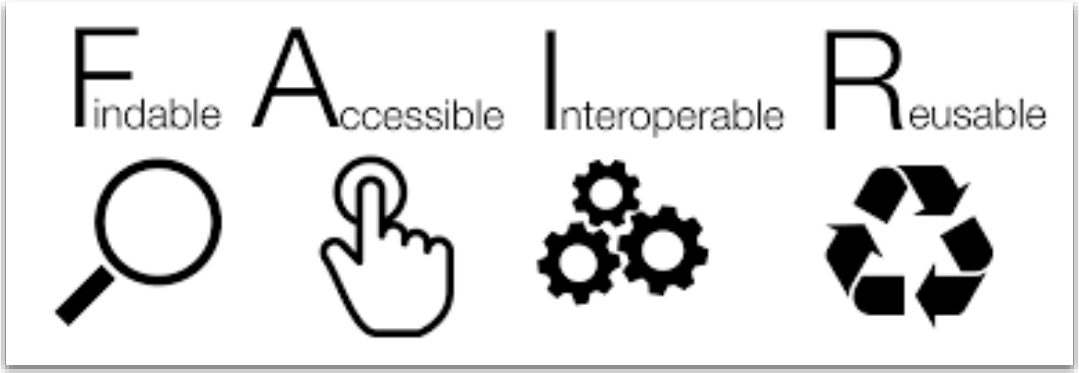
A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

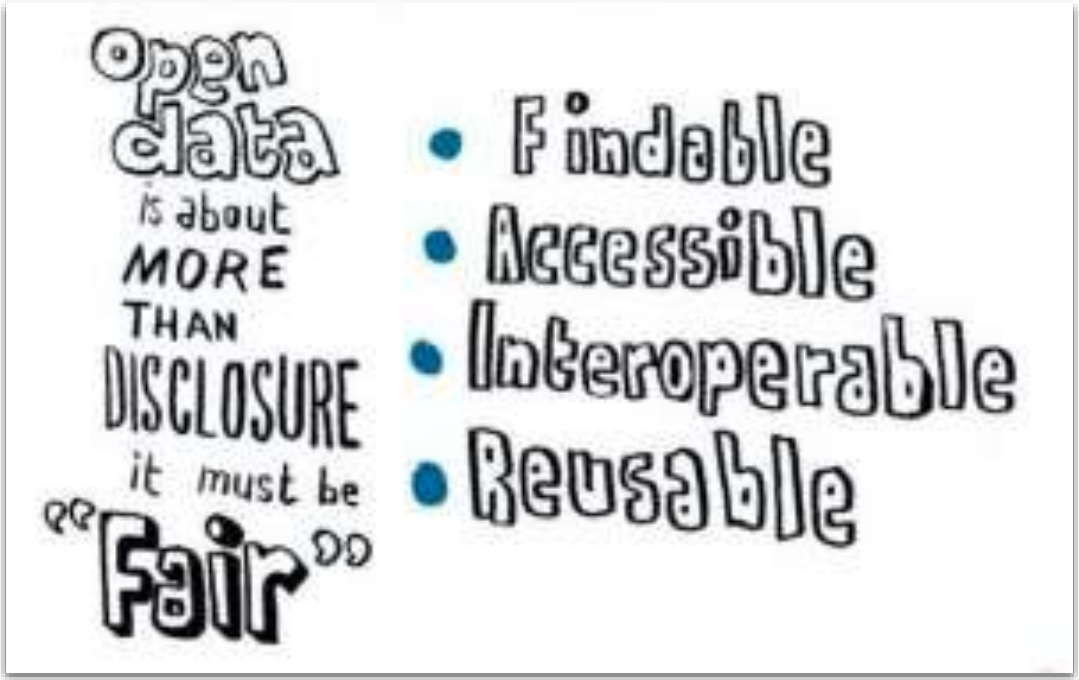
To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

To be Re-usable:

R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

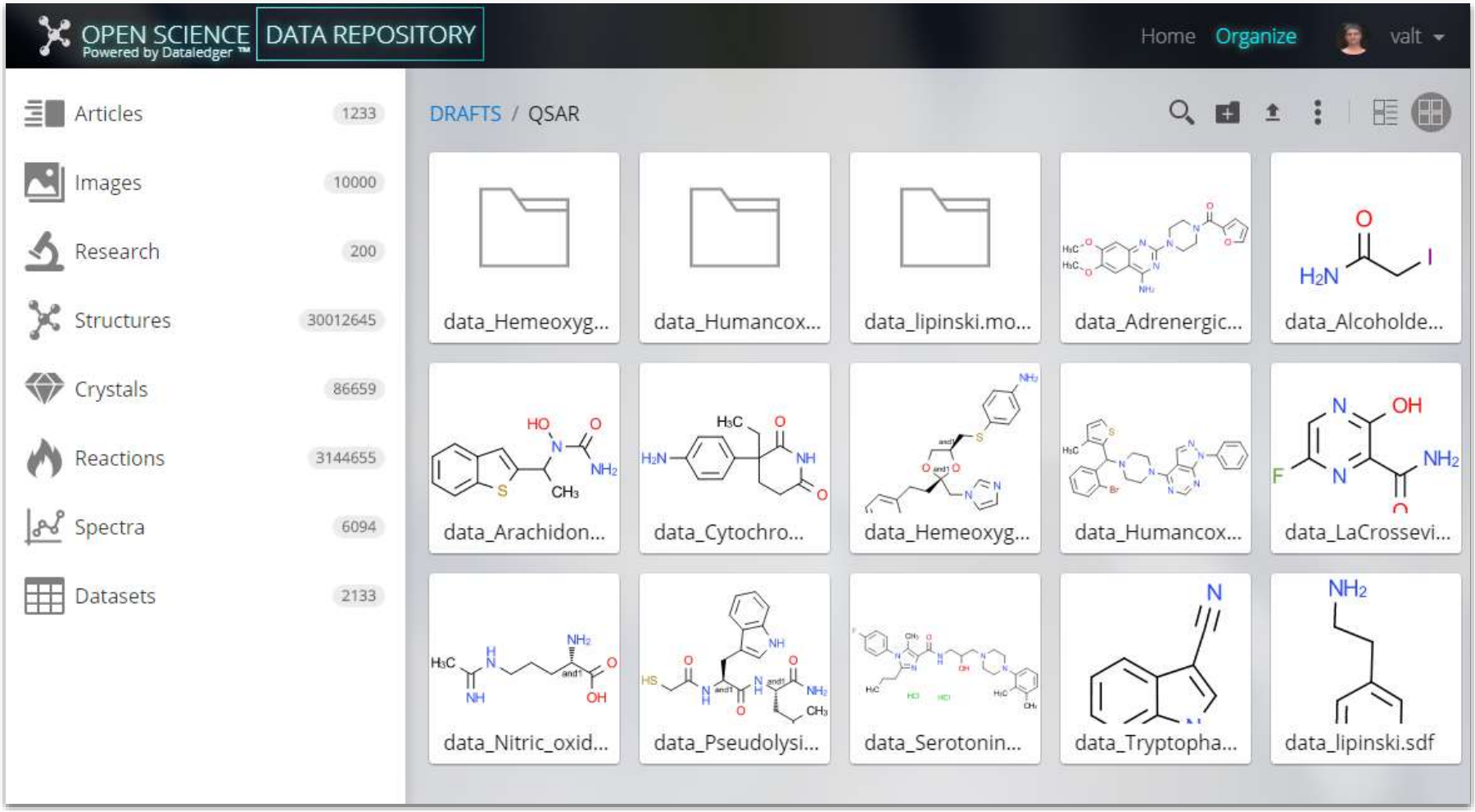




Open Science Data Repository powered by Dataledger™

Chemical processing

- Support for chemical formats
- Chemistry validation and standardization
- Automatic processing and visualization



Possible solution

- Agreed and machine-readable (digital) standards
- Open-source (and therefore fully transparent) solution
- Organizational AND community support and involvement
- Accessible solution
- Data triaging at data repositories level
- Real-time validation/standardization (API, library, “docker”, etc)

Thank you!

On Web:
scidatasoft.com

Contact us:
info@scidatasoft.com

Slides:
https://www.slideshare.net/valerytkachenko16