Using the Structured Product Labeling format to index versatile chemical data

Valery Tkachenko[1], Yulia Borodina[2] and Antony Williams[3]

[1]Science Data Software, Rockville, MD

[2]FDA, Silver Spring, MD

[3]National Center for Computational Toxicology, US-EPA

ACS Spring 2017
San Francisco, April 1-6th 2017

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 is a universal exchange standard

- Reusable data types
- Coded data elements
- Data specific validation procedures

To provide machine readable data

- Extract from text or legacy databases
- Harmonize data according to the standard
- Code in a machine readable format

Scope of data that can be represented by SPL

- Health informatics
- Cheminformatics
- Bioinformatics

SPL applications

- Human and animal approved and unapproved Rx and OTC drug and biologic product labeling
- Electronic drug establishment registration & drug listing
- Electronic content of labeling for medical devices
- Compounded drug facility and product reporting
- Data needed for the importation of drugs and biologics
- Risk Evaluation & mitigation Strategies
- Maximum residue level in pesticides
- Identification of Medicinal Products (IDMP)
- Pharmacologic class, ***substances***, indications, biosimilar non-proprietary names identification
- Etc.

SPL product data exchange

Drug manufacturers

Product SPL file

Drug distributers

Product SPL file

Product SPL file

SPL Substance index file
SPL Pharm Class index file
SPL Billing Unit index file
SPL Product Concept index file

Labels.fda.gov

DAILYMED

openFDA
open.fda.gov

Substances in products

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

IDMP standard defines “WHAT”

Example:

- Proteins shall be defined by the final expressed sequence
- The description of modified proteins shall capture structural changes that result from the modification when a definitive structure is known

SPL standard defines “HOW”

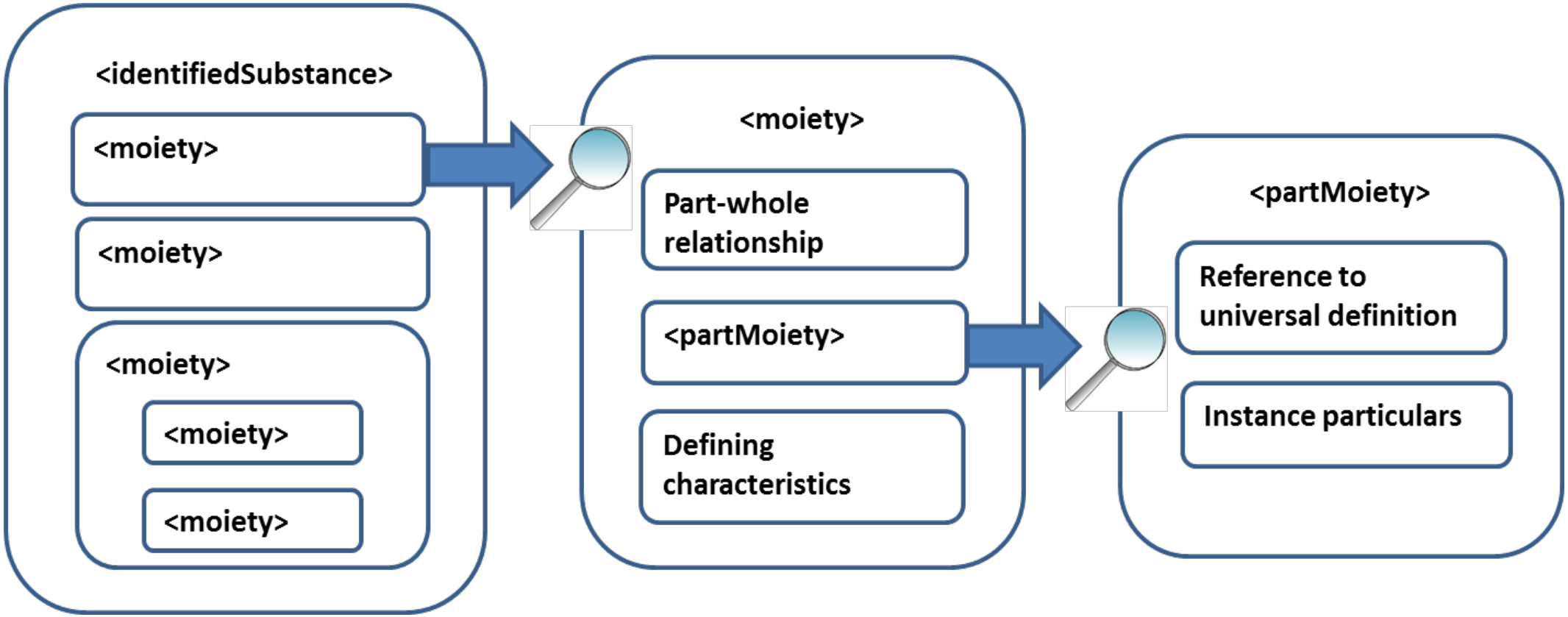
Example:

- Structural Representation Type

Type	MIME Media Type
MOLFILE	application/x-mdl-molfile
Amino Acid Letter Sequence	application/x-aa-seq

- All post-translational modifications and other variances from the closest amino acid sequence are specified by means of amino acid substitutions.
- Amino acid substitutions are specified as moieties of code “structural modification” (C118425)

SPL model


<identifiedSubstance>
<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	V		
Structural modification	C118425	C I S		
Amino acid connection points	C118427	-		
Linear SRU connection points	???	-		

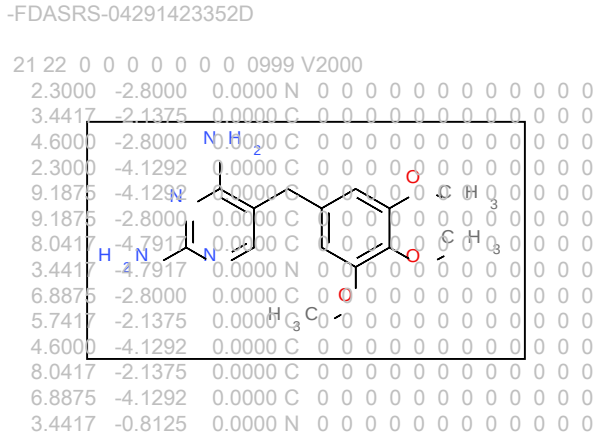
Type	MIME Media Type
Molfile	application/x-mdl-molfile
InChI	application/x-inchi
InChIkey	application/x-inchi-key
Amino acid sequence	application/x-aa-seq
DNA Sequence	application/x-dna-seq
RNA Sequence	application/x-rna-seq

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

Representation of structurally diverse substance in SPL standard

- Biological substance (plant)
  - Bibliographic reference: Cichorium intybus L.

Structurally diverse substances

```
> <SUBSTANCE_ID>
79785

> <UNII>
00338G7O3S

> <REF_INFO_PART1>
<TAXON_AUTHOR>BLOCH & SCHNEIDER</TAXON_AUTHOR>

> <ITIS_TSN>
169276

> <DESC_PART1>
<STRUCTURALLY_DIVERSE>
<SOURCE_MATERIAL>
<SOURCE_MATERIAL_CLASS>ORGANISM</SOURCE_MATERIAL_CLASS>
<SOURCE_MATERIAL_TYPE>BONY FISH</SOURCE_MATERIAL_TYPE>
<SOURCE_MATERIAL_STATE/>
<PARENT_SUBSTANCE_ID/>
<PARENT_SUBSTANCE_NAME/>
<DEVELOPMENTAL_STAGE/>
<PART_GROUP>
<PART>WHOLE</PART>
<PART_LOCATION/>
</PART_GROUP>
<FRACTION>
<MATERIAL_TYPE/>
<FRACTION/>
</FRACTION>
<ORGANISM>
<KINGDOM>ANIMALIA</KINGDOM>
<PHYLUM>CHORDATA</PHYLUM>
<CLASS>ACTINOPTERYGII</CLASS>
<ORDER>PERCIFORMES</ORDER>
<FAMILY>SCIAENIDAE</FAMILY>
<GENUS>MENTICIRRHUS</GENUS>
<SPECIES>SAXATILIS</SPECIES>
<HYBRID_SPECIES_MATERNAL_ORGANISM_ID/>
<HYBRID_SPECIES_PATERNAL_ORGANISM_ID/>
<HYBRID_TYPE/>
<INFRASPECIFIC_TYPE/>
```

```
<component>
  <structuredBody>
    <component>
      <section>
        <id root="f66254a0-7bc6-416e-8de5-a9894b505ede" />
        <code code="48779-3" codeSystem="2.16.840.1.113883.6.1" displayName="SPL indexing data elements section" />
        <effectiveTime value="20161003" />
        <subject>
          <identifiedSubstance>
            <id extension="00338G7O3S" root="2.16.840.1.113883.4.9" />
            <identifiedSubstance>
              <code code="00338G7O3S" codeSystem="2.16.840.1.113883.4.9" />
              <asEquivalentSubstance>
                <definingSubstance>
                  <code code="dcbaa56c-8cb0-eacd-b932-637517989caf" codeSystem="2.16.840.1.113883.3.2705" />
                </definingSubstance>
              </asEquivalentSubstance>
            </identifiedSubstance>
            <subjectOf>
              <document>
                <bibliographicDesignationText>Menticirrhus saxatilis Bloch &amp;amp; Schneider</bibliographicDesignationText>
              </document>
            </subjectOf>
          </identifiedSubstance>
        </subject>
      </section>
    </component>
  </structuredBody>
</component>
</document>
```

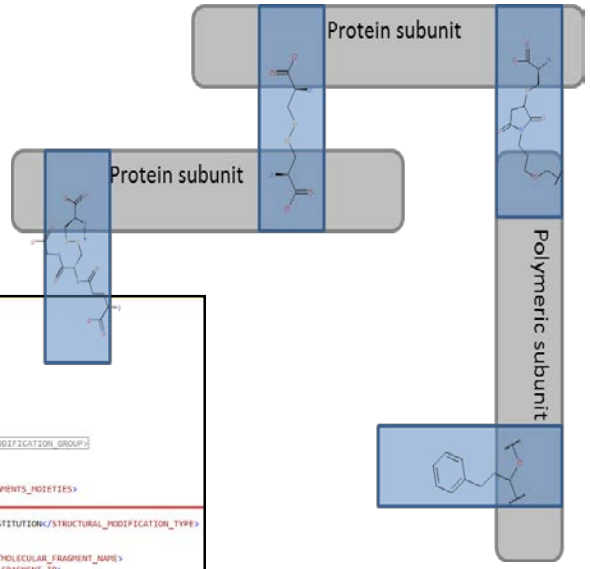
Representation of protein substance in SPL standard

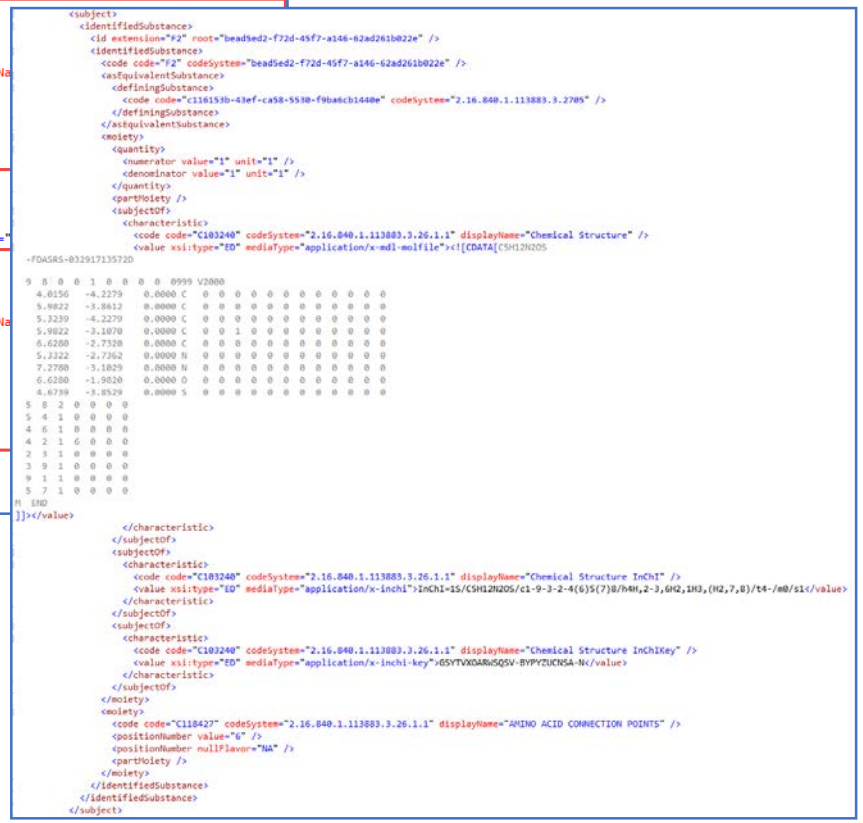
- Protein substance
  - Chemical structure (SEQUENCE)

QVQLQQSGSELKKPGASVKVSCKASGYTFTNYGMNWVKQAPGQGLKWMGWINTYTGEPTYT
DDFKGRFAFSLDTSVSTAYLQISSLKADDTAVYFCARGGFGSSYWYFDVWGQGSLVTVSSASTKGP
SVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPS
SSLGTQTYICNVNHKPSNTKVDKRVEPKSCDKTHTCPPCPAPELLGGPSVFLFPPKPKDTLMISRT
PEVTCVVVDVSHEDPEVKFNWYVDGVEVHNAKTKPREEQYNSTYRVVSVLTVLHQDWLNGKE
YKCKVSNKALPAPIEKTISKAKGQPREPQVYTLPPSREEMTKNQVSLTCLVKGFYPSDIAVEWESN
GQPENNYKTTPPVLDSDGSFFLYSKLTVDKSRWQQGNVFSCSVMHEALHNHYTQKSLSLSPGK

  - Other characteristics

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.

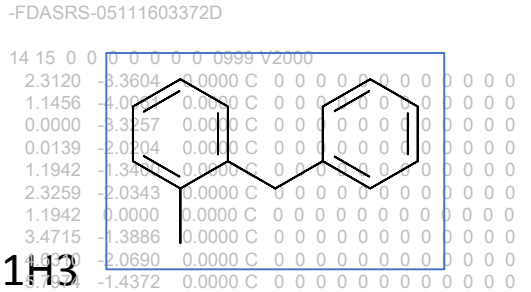
- Mixture substances shall not be combinations of diverse material brought together to form a product.

Representation of mixtures in SPL

- Chemical mixture
  - Moiety (mixture component)
  - Quantity: Undefined not 0

    InChI=1S/C14H14/c1-12-7-5-6-10-14(12)11-13-8-3-2-4-9-13/h2-10H,11H2,1H3

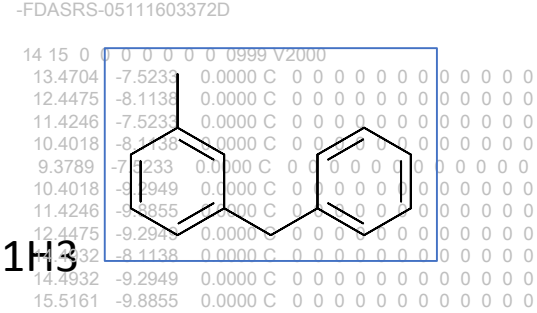
    PQTAUFTUHHRKSS-UHFFFAOYSA-N



  - Moiety (mixture component)
  - Quantity: Undefined not 0

    InChI=1S/C14H14/c1-12-6-5-9-14(10-12)11-13-7-3-2-4-8-13/h2-10H,11H2,1H3

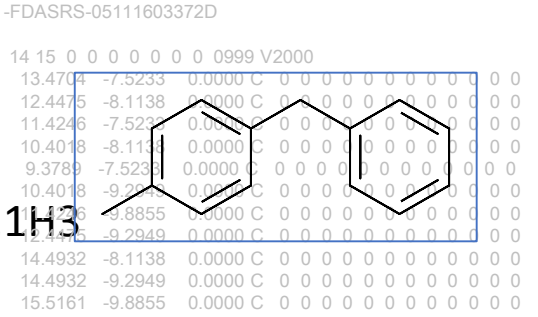
    KSYQGOYOIKQFNA-UHFFFAOYSA-N



  - Moiety (mixture component)
  - Quantity: Undefined not 0

    InChI=1S/C14H14/c1-12-7-9-14(10-8-12)11-13-5-3-2-4-6-13/h2-10H,11H2,1H3

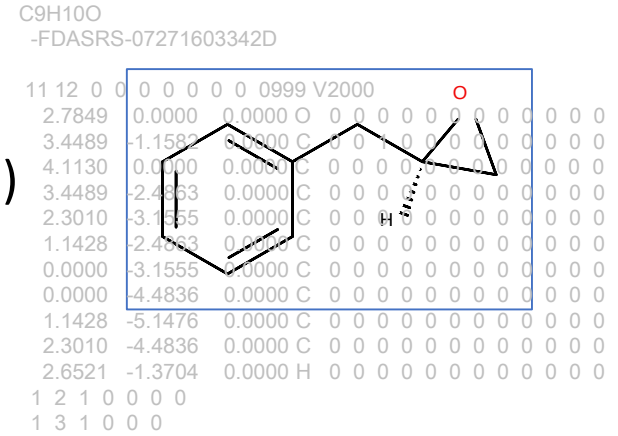
    SIYISNUJKMAQBV-UHFFFAOYSA-N



Representation of mixtures in SPL

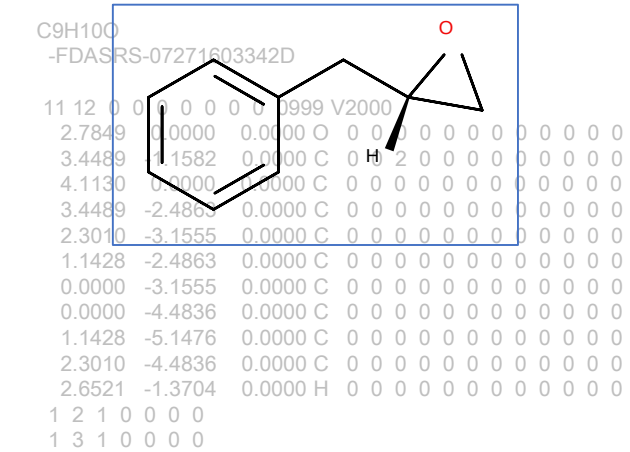
- Racemic mixture
  - Moiety (mixture component)
  - Quantity: 1/2

InChI=1S/C9H10O/c1-2-4-8(5-3-1)6-9-7-10-9/h1-5,9H,6-7H2/t9-/m1/s1)
JFDMLXYWGLECEY-SECBINFHSA-N

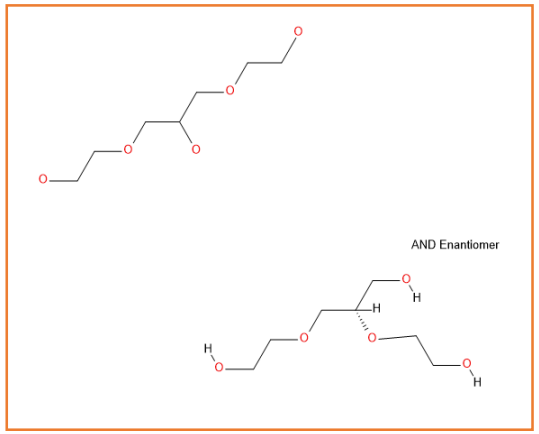


  - Moiety (mixture component)
  - Quantity: 1/2

InChI=1S/C9H10O/c1-2-4-8(5-3-1)6-9-7-10-9/h1-5,9H,6-7H2/t9-/m0/s1
JFDMLXYWGLECEY-VIFPVBQESA-N



Mixtures




```
  Symyx   04271610292D 1   1.00000     0.00000     0

 12 11  0     0  0            999 V2000
   20.2310   -4.7158    0.0000 O   0  0  0  0  0  0           0  0  0
   19.0499   -4.7158    0.0000 C   0  0  0  0  0  0           0  0  0
   18.4594   -5.7386    0.0000 C   0  0  3  0  0  0           0  0  0
   17.2783   -5.7386    0.0000 C   0  0  0  0  0  0           0  0  0
   16.6877   -6.7615    0.0000 O   0  0  0  0  0  0           0  0  0
   15.5066   -6.7615    0.0000 C   0  0  0  0  0  0           0  0  0
   14.9161   -7.7844    0.0000 C   0  0  0  0  0  0           0  0  0
   13.7350   -7.7844    0.0000 O   0  0  0  0  0  0           0  0  0
   19.0499   -6.7615    0.0000 O   0  0  0  0  0  0           0  0  0
   20.8216   -3.6929    0.0000 C   0  0  0  0  0  0           0  0  0
   22.0027   -3.6929    0.0000 C   0  0  0  0  0  0           0  0  0
   22.5933   -2.6700    0.0000 O   0  0  0  0  0  0           0  0  0
  1  2  1  0     0  0
  2  3  1  0     0  0
  3  4  1  0     0  0
  4  5  1  0     0  0
  5  6  1  0     0  0
  6  7  1  0     0  0
  7  8  1  0     0  0
  9  3  1  0     0  0
  1 10  1  0     0  0
 10 11  1  0     0  0
 11 12  1  0     0  0
M  END
>  <MIX_SUBSTANCE_ID>
1794

>  <MIX_UNII>
0UY833L6XU

>  <MIXTURE_TYPE>
ALL

>  <MIX_DESC_PART1>

>  <MIX_DESC_PART2>

>  <MIX_COMMENTS>

>  <SUBSTANCE_ID>
177122

>  <UNII>
9CC1P9T91Z

>  <DESC_PART1>

>  <DESC_PART2>

>  <COMMENTS>

>  <STRUCTURE_ID>
80470

>  <SUBSTANCE_NAME>

$$$$
```

```
C7H16O5
YUtclcactv04271619142D 0   0.00000     0.00000     0

  0  0  0     0  0            999 V3000
M  V30 BEGIN CTAB
M  V30 COUNTS 16 15 0 0 0 REGNO=0
M  V30 BEGIN ATOM
M  V30 1 O 20.231 -4.7158 0 0
M  V30 2 C 19.0499 -4.7158 0 0
M  V30 3 C 18.4594 -5.7386 0 0 CFG=1
M  V30 4 C 17.2783 -5.7386 0 0
M  V30 5 O 16.6877 -6.7615 0 0
M  V30 6 C 15.5066 -6.7615 0 0
M  V30 7 C 14.9161 -7.7844 0 0
M  V30 8 O 13.735 -7.7844 0 0
M  V30 9 O 19.0499 -6.7615 0 0
M  V30 10 C 20.5625 -6.625 0 0
M  V30 11 C 21.1531 -7.6479 0 0
M  V30 12 O 22.3342 -7.6479 0 0
M  V30 13 H 20.6467 -5.36646 0 0
M  V30 14 H 19.2107 -5.73862 0 0
M  V30 15 H 13.3593 -7.13374 0 0
M  V30 16 H 22.7099 -8.29856 0 0
M  V30 END ATOM
M  V30 BEGIN BOND
M  V30 1 1 1 2
M  V30 2 1 2 3
M  V30 3 1 3 4
M  V30 4 1 4 5
M  V30 5 1 5 6
M  V30 6 1 6 7
M  V30 7 1 7 8
M  V30 8 1 3 9 CFG=3
M  V30 9 1 9 10
M  V30 10 1 10 11
M  V30 11 1 11 12
M  V30 12 1 1 13
M  V30 13 1 3 14
M  V30 14 1 8 15
M  V30 15 1 12 16
M  V30 END BOND
M  V30 BEGIN COLLECTION
M  V30 MDLV30/STEREA1 ATOMS=(1 3)
M  V30 END COLLECTION
M  V30 END CTAB
M  END
> <MIX_SUBSTANCE_ID>
1794

> <MIX_UNII>
0UY833L6XU

> <MIXTURE_TYPE>
ALL

> <MIX_DESC_PART1>


> <MIX_DESC_PART2>


> <MIX_COMMENTS>


> <SUBSTANCE_ID>
177124

> <UNII>
C0096KF291

> <DESC_PART1>
<STEREOCHEMISTRY><OPT_ACT>(+/-)</OPT_ACT><TYPE>RACEMIC</TYPE><COMMENTS></COMMENTS></STEREOCHEMISTRY>
```

```
 9 10  1  0  0  0  0
10 11  1  0  0  0  0
11 12  1  0  0  0  0
 1 13  1  0  0  0  0
 3 14  1  6  0  0  0
 8 15  1  0  0  0  0
12 16  1  0  0  0  0
M  END
]]></value>
                </characteristic>
              </subjectOf>
              <subjectOf>
                <characteristic>
                  <code displayName="Chemical Structure" codeSystem="2.16.840.1.113883.3.26.1.1" code="C103240" />
                  <value p6:type="ED" mediaType="application/x-inchi" xmlns:p6="http://www.w3.org/2001/XMLSchema-instance">InChI=1S/C7H16O5/c8-1-3-11-6-7(5-10)12-4-2-9/h7-10H,1-6H2/t7-/m1/s1</value>
                </characteristic>
              </subjectOf>
              <subjectOf>
                <characteristic>
                  <code displayName="Chemical Structure" codeSystem="2.16.840.1.113883.3.26.1.1" code="C103240" />
                  <value p6:type="ED" mediaType="application/x-inchi-key" xmlns:p6="http://www.w3.org/2001/XMLSchema-instance">RLVWMVMTMYXMBZ-SSDOTTSWSA-N</value>
                </characteristic>
              </subjectOf>
```

```
 1 13  1  0  0  0  0
 3 14  1  0  0  0  0
 8 15  1  0  0  0  0
12 16  1  0  0  0  0
M  END
]]></value>
                </characteristic>
              </subjectOf>
              <subjectOf>
                <characteristic>
                  <code displayName="Chemical Structure" codeSystem="2.16.840.1.113883.3.26.1.1" code="C103240" />
                  <value p6:type="ED" mediaType="application/x-inchi" xmlns:p6="http://www.w3.org/2001/XMLSchema-instance">InChI=1S/C7H16O5/c8-1-3-11-6-7(5-10)12-4-2-9/h7-10H,1-6H2/t7-/m0/s1</value>
                </characteristic>
              </subjectOf>
              <subjectOf>
                <characteristic>
                  <code displayName="Chemical Structure" codeSystem="2.16.840.1.113883.3.26.1.1" code="C103240" />
                  <value p6:type="ED" mediaType="application/x-inchi-key" xmlns:p6="http://www.w3.org/2001/XMLSchema-instance">RLVWMVMTMYXMBZ-ZETCQYMHSA-N</value>
                </characteristic>
              </subjectOf>
            </moiety>
```

FDA Substance Registration System (SRS)

- Agency-Wide Substance Database
- 100,000 substances
- Small molecules, mixtures, proteins, polymers, biopolymers, plant parts, tissue parts, vaccines, etc.
- Highly curated information (chemical structures, names, protein and nucleic acid sequence, taxonomic information)
- Unique Ingredient Identifiers (UNIIs)

FDA open substance data in SPL format

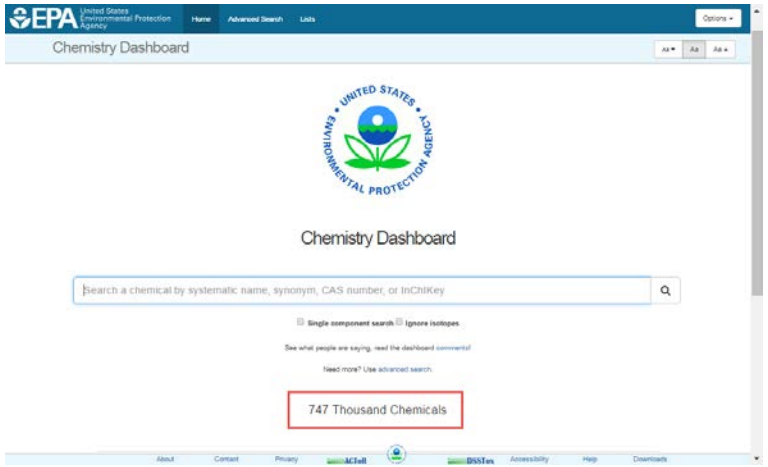
- Definitions of non-confidential substances from FDA Substance Registration System (**SRS**)
- UNique Ingredient Identifier (**UNII**)
- Compliant with ISO IDMP 11238 standard (**IDMP**)
- Over **60,000** chemical substances
- Over **10,000** structurally diverse substances
- Started publishing proteins
- Available on DailyMed for direct download
- Available in openFDA s3 bucket
- Regularly updated and versioned
- Openly documented
  http://www.fda.gov/downloads/ForIndustry/DataStandards/StructuredProductLabeling/UCM321876.pdf

Inter-agency database interoperability

- Chemical structure linking and connectivity is “fairly well” solved using Standard InChIs – not perfect but very useful!
- Challenges remain with enhanced stereochemistry, polymers, organometallics and Markush structures but InChI is moving towards solving these challenging problems – iteratively and slowly
- FDA has solved a lot of the problems with SPL so can it be applied to non-FDA databases and potentially enhance interoperability?
- Looking for a good source of high-quality “controlled” but open data

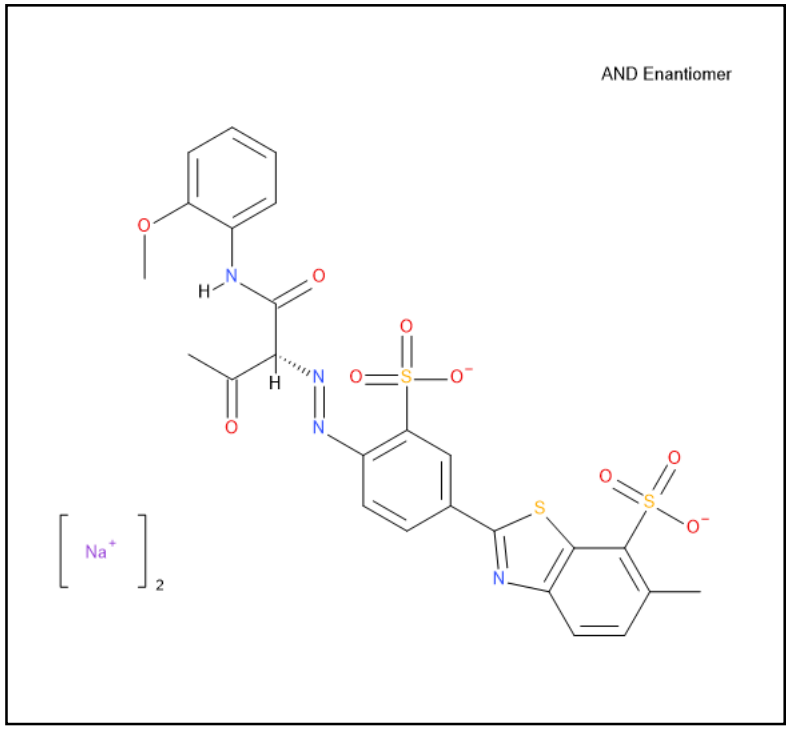
The Comptox Chemistry Dashboard data

- 747,000 chemical substances and ca. 727k chemical structures!
- Diverse structures – organics, inorganics, organometallics
- Fully open data available for download, reuse and repurposing: https://comptox.epa.gov/dashboard/downloads
- Increasingly an integration hub for EPA chemical structure data
- Strong team with cheminformatics experience
- Manual curators to provide feedback on CVSP results and processing
- Lots of complex mixtures and “UVCB” chemicals -**U**nknown or **V**ariable **C**omposition, Complex Reaction Products and **B**iological Materials



Example Complex Stereochemistry v3000

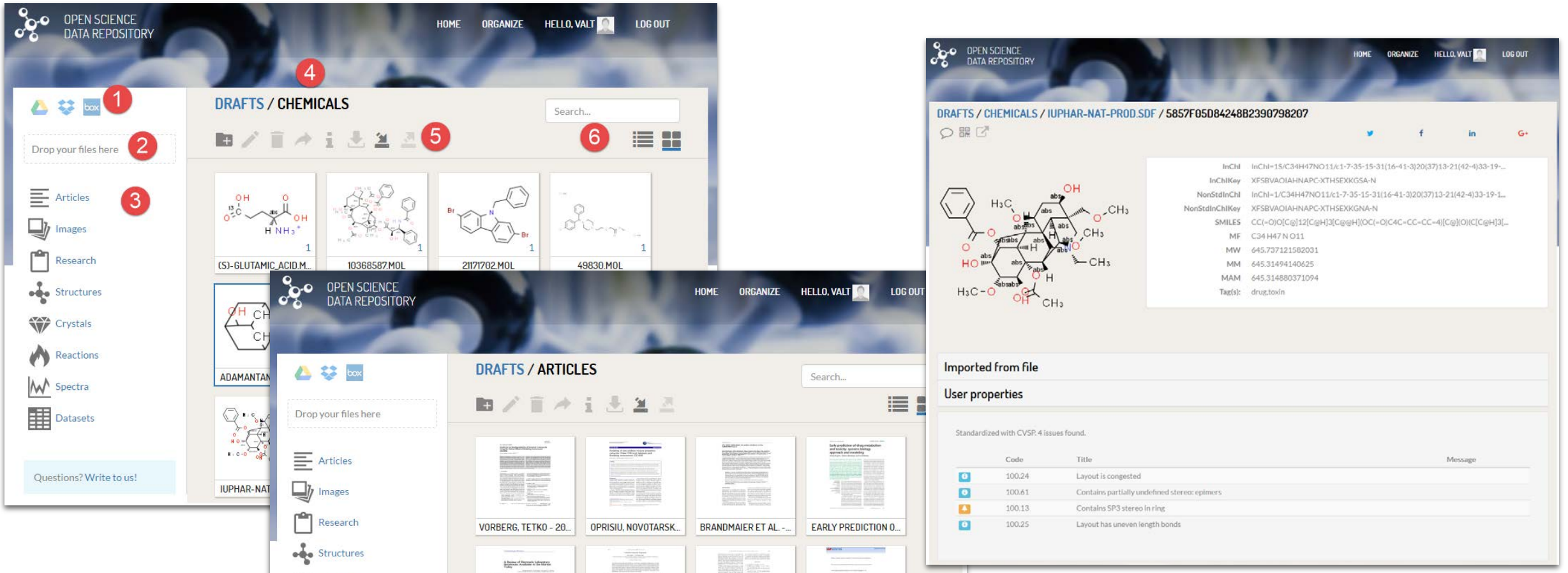
- Insert structure and result?



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!

Open Science Data Repository



Summary

- CompTox Dashboard content was converted to SPL files
- Stoichiometric mixtures (e.g. racemates) were converted as is
- Complex mixtures required additional changes
- SPL as a container is suitable for storing much more complex chemical and biological substances
- SPL is machine readable
- Open Science Data Repository provides on-the-fly SPL capabilities
- Work in progress and definitely iterative in nature
- Watch this space for further progress

Acknowledgements

- Open source cheminformatics toolkits
  - CDK
  - Indigo
  - RDKit

- RDA and IUPAC for organizing workshop on chemical structure representation