PubChemRDF: Towards a semantic (web) description of PubChem

Evan Bolton, Gang Fu, Bo Yu
Scientific Information

Presentation → Human understanding → Data
What about computer understanding?
Presentation vs. Data

CC(=O)OC1=CC=CC=C1C(=O)O

PNG Image of a chemical structure

CC(=O)OC1=CC=CC=C1C(=O)O

SMILES of a chemical structure
What you see... presentation layer

Aspirin - Compound Summary (CID 2244)

Also known as: ACETYSALICYLIC ACID, 2-Acetoxybenzoic acid, Acypyrin, Ecotrin, Acenterine, Polopiryna, Acetosal, Colfarit, Enterosarein

Molecular Formula: C₉H₈O₄  Molecular Weight: 180.15742  InChIKey: BSYNRYMUTXBXSQ-UHFFFAOYSA-N

The prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. (From Martindale, The Extra Pharmacopoeia, 30th ed, p5)  From: MeSH

Table of Contents

Identification
Related Records
Use and Manufacturing
Pharmacology
Biomedical Effects and Toxicity
Safety and Handling
Environmental Fate and Exposure Potential
Exposure Standards and Regulations

2D Structure  3D Conformer

\[
\begin{align*}
\text{\includegraphics[width=0.4\textwidth]{aspirin_molecule.png}}
\end{align*}
\]
What computer sees... presentation layer

*(NOT the DATA you see)*

**Aspirin - Compound Summary (CID 2244)**

**Also known as:** ACETYLSALICYLIC ACID, 2-Acetoxybenzoic acid, Acylypin, Ecotrin, Acenterine, Polopiryna, Acetosal, Colfarit, Enterosarein

**Molecular Formula:** C₉H₈O₄  **Molecular Weight:** 180.15742  **InChIKey:** BSYNRYMUTXBXSQ-UHFFFAOYSA-N

The prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. *(From Martindale, The Extra Pharmacopoeia, 30th ed, p5) From: MeSH*
Presentation is human specific

- PDF, HTML, CSS, XML
- Display is for human interpretation but the computer cannot pull out the relevant data (well not easily)
  - Molecular Weight
  - Units
  - Value
Imagine...

• You ‘view’ a web page and get data presentation
• Computer ‘views’ the web page and gets data
• Each with a custom view of the same information
  – Nice layout of text, images, and tables for you
  – Interpretable data contents for the computer
• While you view the page, the computer pulls, organizes, and combines information
• Browser ‘interprets’ the data while you view the page and maybe ‘learns’ useful things to know relative to your interests or ‘remembers’ things
How do we allow the computer to get the same (or better) understanding as you?

- OCR and natural language processing?
  - Ugh... good luck

- Need a common language for computers to interpret the relevant information

- We have HTML/PDF for display-centric layout... why not markup language and vocabularies for data markup?
Welcome to the semantic web....
Enable users to find, share, and combine information more easily.

What is RDF?

• **Resource Description Framework**

• Family of World Wide Web Consortium (W3C) specifications for data exchange on the Web
  
  http://www.w3.org/RDF/

• Machine-readable statements (for computers)

• Emphasis on data exchange (via the Web)
How does RDF work?

• Data model employs the concept of triples

  “subject-predicate-object”
  “atorvastatin may treat hypercholesterolemia”

• Models statements as directed graphs

• Uses Uniform Resource Identifiers (URI) for subject, predicate, and object
  – Object can also be a constant value (literal)
**URIs and RDF**

- **Uniform Resource Identifier (URI)** is
  - a general identifier for anything
  - can be independently created
  - Is a Uniform Resource Locator (URL) or Uniform Resource Name (URN) or both


- **URN** identifies something (e.g., CID2244)

- **RDF** uses **URI references**
  - **URI reference** is a URI with an optional fragment identifier at the end (e.g., URL#frag_identifier)
What is PubChem doing with semantic web technologies and Why?
PubChem has a fair bit of data
Why semantic web?
... *improve data access* *(not just presentation)*

- **PubChem** has a large corpus of information
- **Human web interfaces** are powerful and useful, but lack computer data interpretation
- **Programmatic interfaces** help to automate tasks but large analyses face usage limitations
- Providing a PubChem SQL database is beyond the limits of PubChem human resources but import RDF in a triple store and use SPARQL to query away.... a schema-less database
Aspirin - Compound Summary (CID 2244)

Also known as: ACETYLSALICYLIC ACID, 2-Acetoxybenzoic acid, Acilpyrin, Ecotrin, Acenterine, Polopiryna, Acetosal, Colfarit, Enterosarein

Molecular Formula: C₉H₈O₄  Molecular Weight: 180.15742  InChIKey: BSYNRMUTXBSQ-UHFFFAOYSA-N

The prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. (From Martindale, The Extra Pharmacopoeia, 30th ed, p5)  From: MeSH

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix sio: <http://semanticscience.org/resource/> .
chemical-descriptor:CID2244_Molecular_Weight
  sio:has-unit obo:UO_0000055 ;
  sio:has-value 180.15742 ;
  sio:is-attribute-of compound:CID2244 ;
  a sio:CHEMINF_000334 .
Introducing *PubChemRDF*

- provides **RDF formatted information**
- is a **subset of PubChem**
- includes a **REST-ful interface** and bulk **downloadable data** on the PubChem FTP site

- leverages existing ontology frameworks
- aims to help facilitate data sharing, analysis, and integration with external resources
What does PubChemRDF (initially) cover?

- **Substance**
  - Standardized CID
  - Depositor Identifier
  - Data source information
  - Bioactivity links

- **Assay**
  - Data source information
  - Assay type
  - Assay title
  - Bioactivity links

- **Compound**
  - Parent, CIG links
  - 2-D and 3-D similarity information
  - Computed property and descriptors

- **Target**
  - Type, Title
  - CDD, neighbors
  - Encoding gene
  - Bioactivity links
What does PubChemRDF look like?

• For molecular weight of CID2244: (in RDF-XML format)

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:sio="http://semanticscience.org/resource/"
  <rdf:Description rdf:about="chemical-descriptor/CID2244_Molecular_Weight">
    <sio:is-attribute-of rdf:resource="compound/CID2244"/>
  </rdf:Description>
  <rdf:Description rdf:about="chemical-descriptor/CID2244_Molecular_Weight">
    <sio:has-value rdf:datatype="http://www.w3.org/2001/XMLSchema#double">180.15742</sio:has-value>
  </rdf:Description>
  <rdf:Description rdf:about="chemical-descriptor/CID2244_Molecular_Weight">
    <sio:has-unit rdf:resource="http://purl.obolibrary.org/obo/UO_0000055"/>
  </rdf:Description>
  <rdf:Description rdf:about="chemical-descriptor/CID2244_Molecular_Weight">
  </rdf:Description>
</rdf:RDF>

:: in Turtle format ::

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix sio: <http://semanticscience.org/resource/> .

chemical-descriptor:CID2244_Molecular_Weight
  sio:has-unit obo:UO_0000055 ;
  sio:has-value 180.15742 ;
  sio:is-attribute-of compound:CID2244 ;
  a sio:CHEMINF_000334 .
What can you do with PubChemRDF?

• Import it and use locally for search/analysis
  – Use PubChemRDF from FTP site
  – Many RDF-enabled analysis/query packages
    • RDF-triple store
      – E.g., Apache Jena, OpenLink Virtuoso
    • SPARQL query engine (like SQL)
• REST-ful interface allows linking to PubChem from other RDF/Semantically-aware resources
Summary and parting thoughts...

• **PubChemRDF** released **this fall** as a **beta**
  – Provides a semantic description of PubChem, bulk data dump of FTP site, RESTful interface

• More to come...
  – Expose more information

• Questions to be answered
  – Is it what folks are expecting? Will it be adopted?

• Community feedback desired
Acknowledgements

- **Gang Fu**
  - Did most of the RDF implementation work

- **Bo Yu**
  - Help in making RDF production ready

- **PubChem** staff
  - Special thanks to Yanli Wang

- External collaborators for helping to extend ontologies and map PubChem data
  - Michel Dumontier (Carleton U.), Egon Willighagen (Maastricht U.), Janna Hastings (EBI & U. of Geneva), Colin Batchelor (RSC), Stephan Schurer, Uma Vempati, Hande Küçük (U. of Miami)