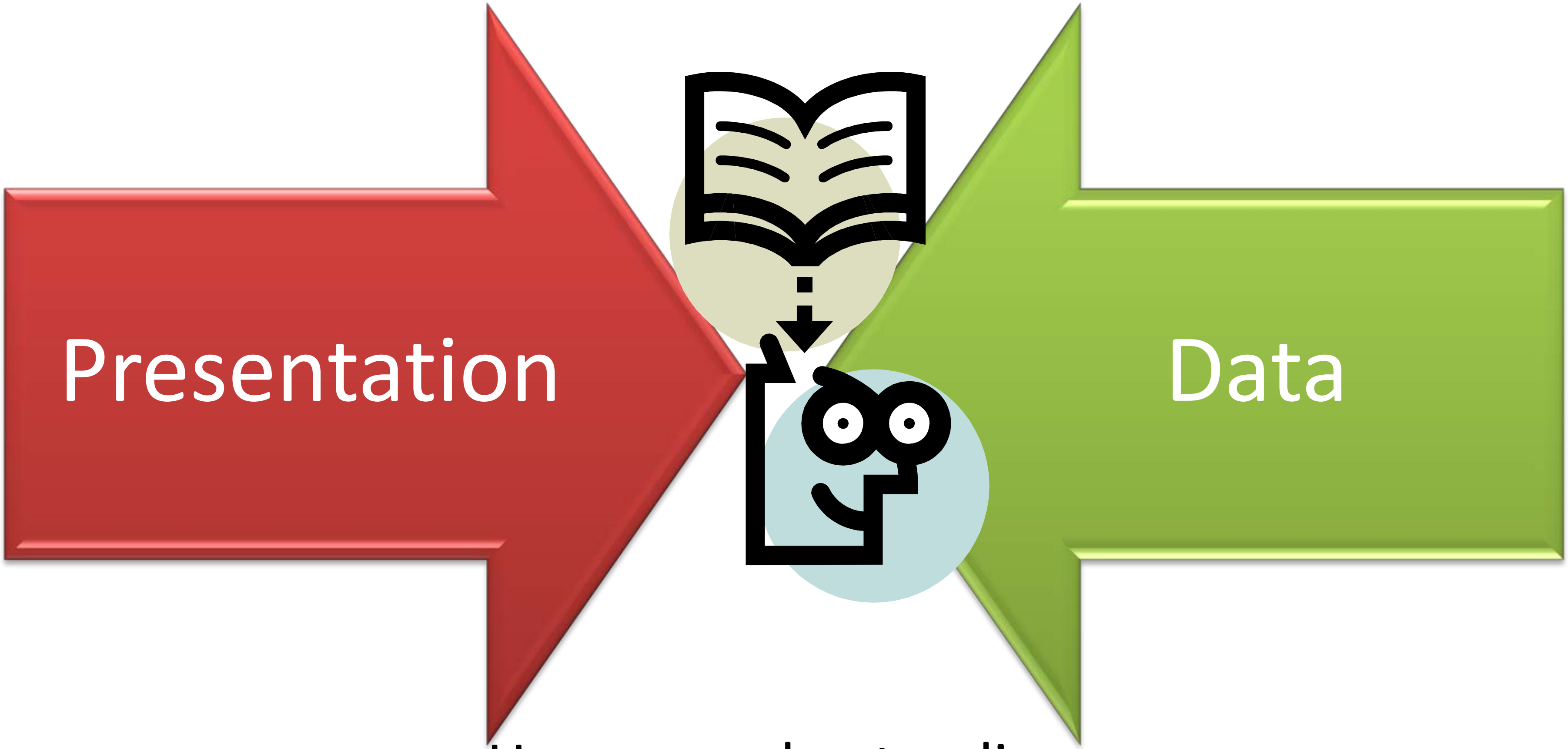
PubChemRDF: Towards a semantic (web) description of PubChem

Evan Bolton, Gang Fu, Bo Yu

Scientific Information

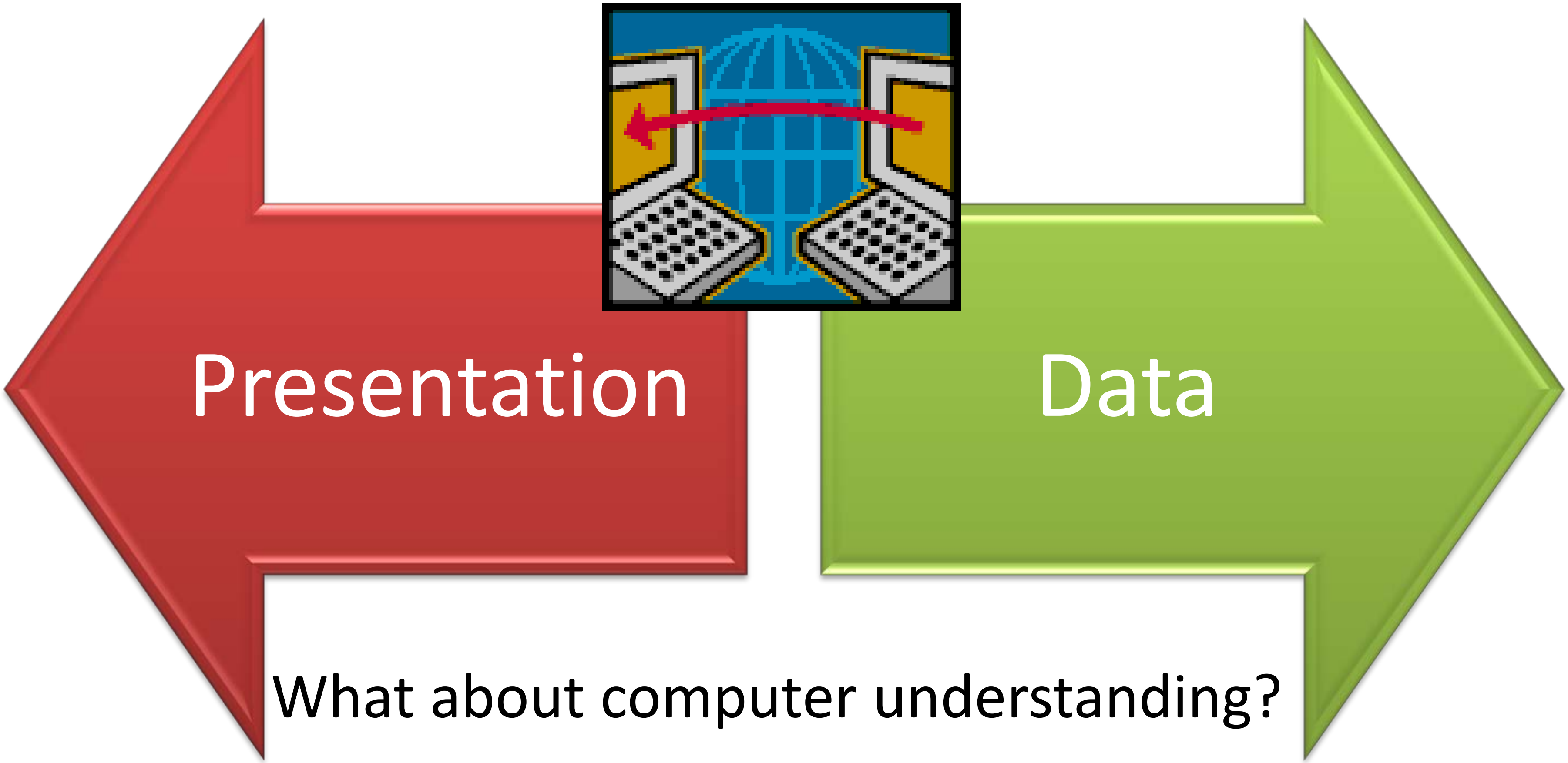


Presentation

Data

Human understanding

Scientific Information

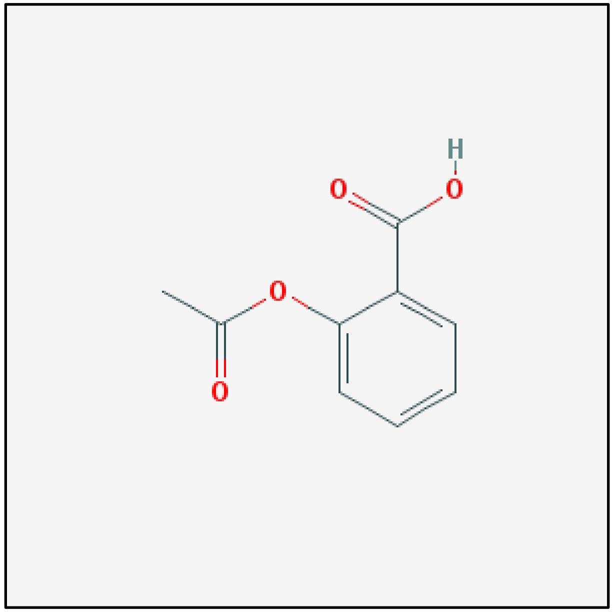


Presentation

Data

What about computer understanding?

Presentation vs. Data

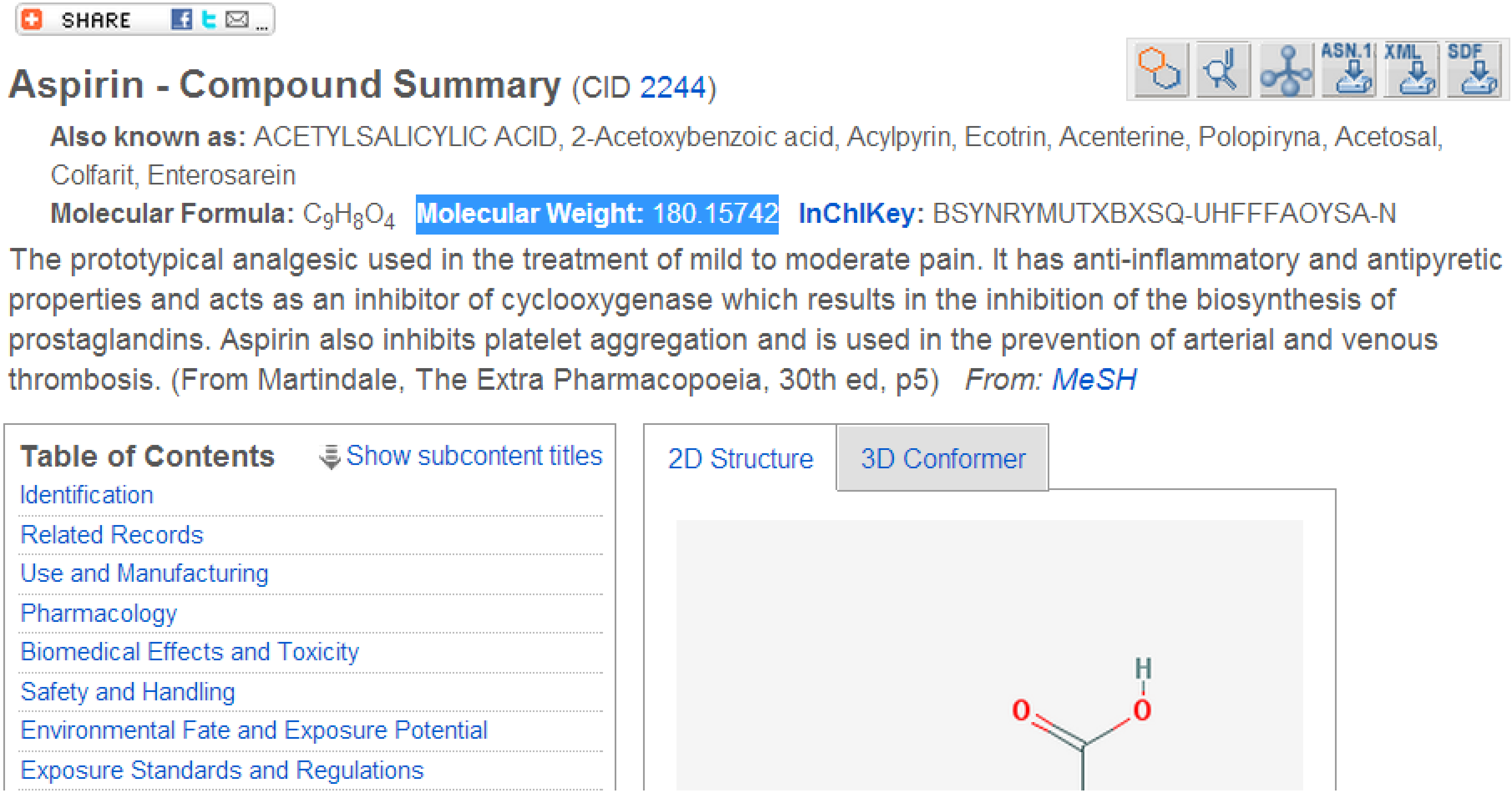


PNG Image of a chemical structure

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: ACETYLSALICYLIC ACID, 2-Acetoxybenzoic acid, Acylpyrin, Ecotrin, Acenterine, Polopiryna, Acetosal, Colfarit, Enterosarein

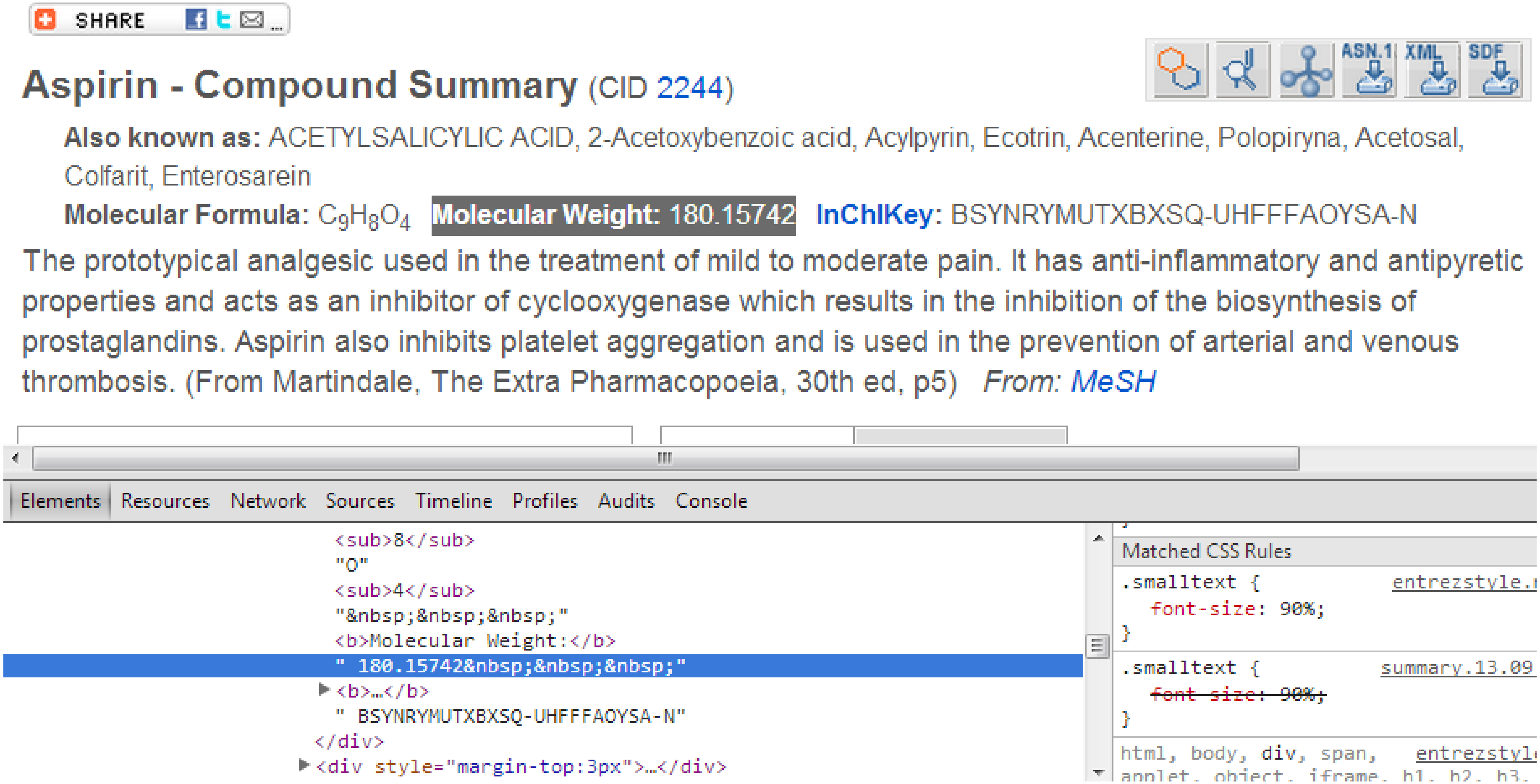
Molecular Formula: $C_9H_8O_4$ **Molecular Weight:** 180.15742 **InChIKey:** BSYNRYMUTXBXSQ-UHFFFAOYSA-N

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 Show subcontent titles

What computer sees… presentation layer

(NOT the DATA you see)



Aspirin - Compound Summary (CID 2244)

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

Molecular Formula: $C_9H_8O_4$ **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*

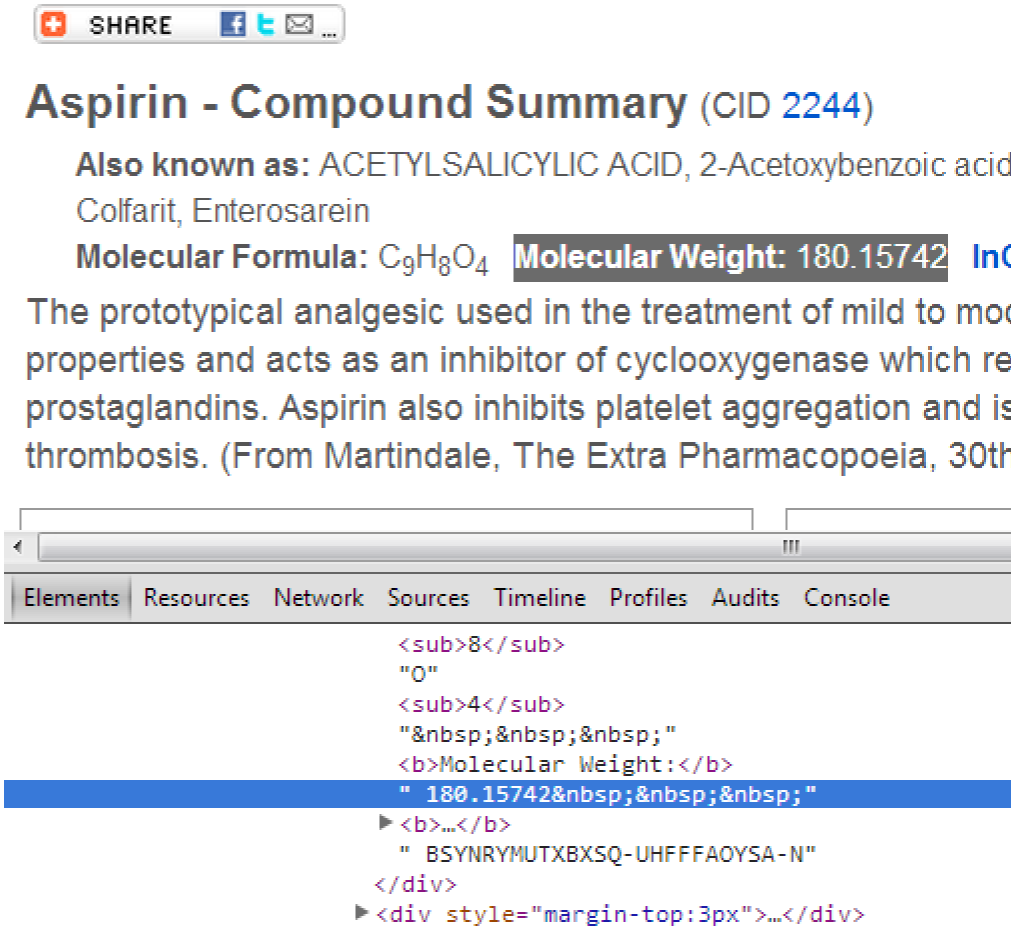
```
      <sub>8</sub>
      "O"
      <sub>4</sub>
      "&nbsp;&nbsp;&nbsp;"
      <b>Molecular Weight:</b>
      " 180.15742&nbsp;&nbsp;&nbsp;"
    ▶ <b>…</b>
      " BSYNRYMUTXBXSQ-UHFFFAOYSA-N"
    </div>
  ▶ <div style="margin-top:3px">…</div>
```

Matched CSS Rules

```
.smalltext {                    entrezstyle.
   font-size: 90%;
}
.smalltext {                    summary.13.09
   ~~font-size: 90%;~~
}
html, body, div, span,         entrezstyl
applet, object, iframe, h1, h2, h3,
```

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



Aspirin - Compound Summary (CID 2244)

Also known as: ACETYLSALICYLIC ACID, 2-Acetoxybenzoic acid
Colfarit, Enterosarein
Molecular Formula: $C_9H_8O_4$ **Molecular Weight:** 180.15742 In

The prototypical analgesic used in the treatment of mild to mo
properties and acts as an inhibitor of cyclooxygenase which re
prostaglandins. Aspirin also inhibits platelet aggregation and is
thrombosis. (From Martindale, The Extra Pharmacopoeia, 30th

```
      <sub>8</sub>
      "O"
      <sub>4</sub>
      "&nbsp;&nbsp;&nbsp;"
      <b>Molecular Weight:</b>
      " 180.15742&nbsp;&nbsp;&nbsp;"
    ▶ <b>…</b>
      " BSYNRYMUTXBXSQ-UHFFFAOYSA-N"
    </div>
  ▶ <div style="margin-top:3px">…</div>
```

Imagine…

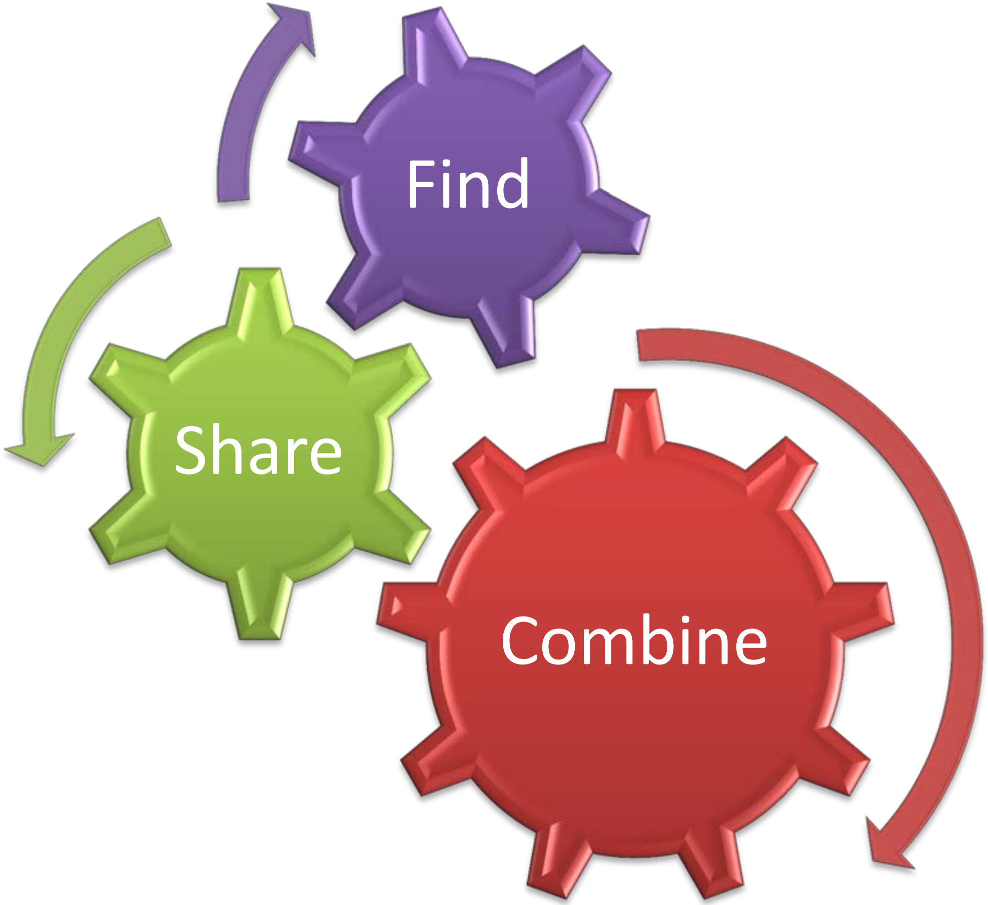
- You ‘view’ a web page and get data presentation
- Computer ‘views’ the web page and gets data
- Each with a custom view of the same information
  - **Nice layout of text, images, and tables for you**
  - **Interpretable data contents for the computer**
- While you view the page, the computer pulls, 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….

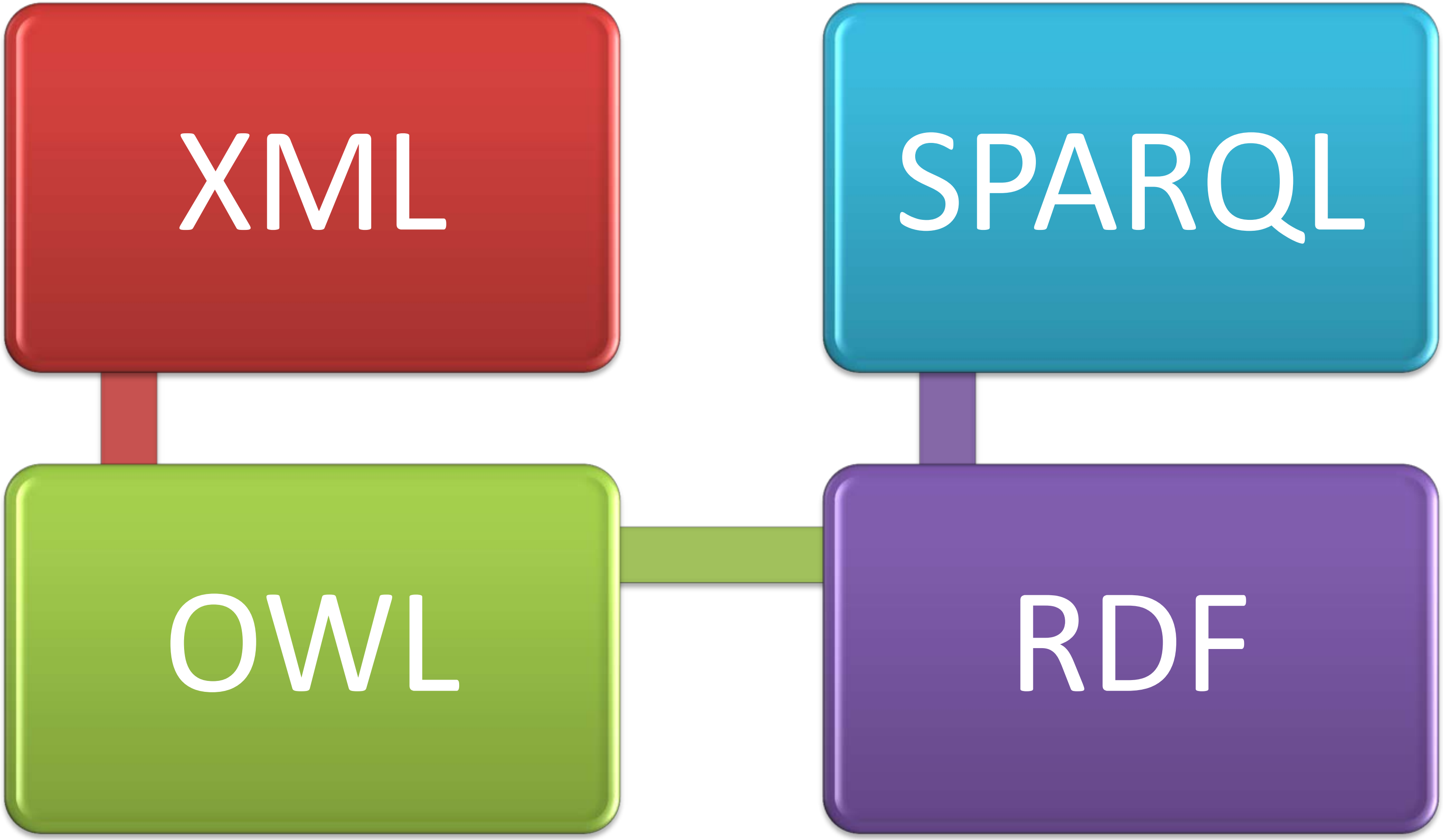
Semantic Web Purpose



Enable users to find, share, and combine information more easily.

http://en.wikipedia.org/wiki/Semantic_Web#Standards

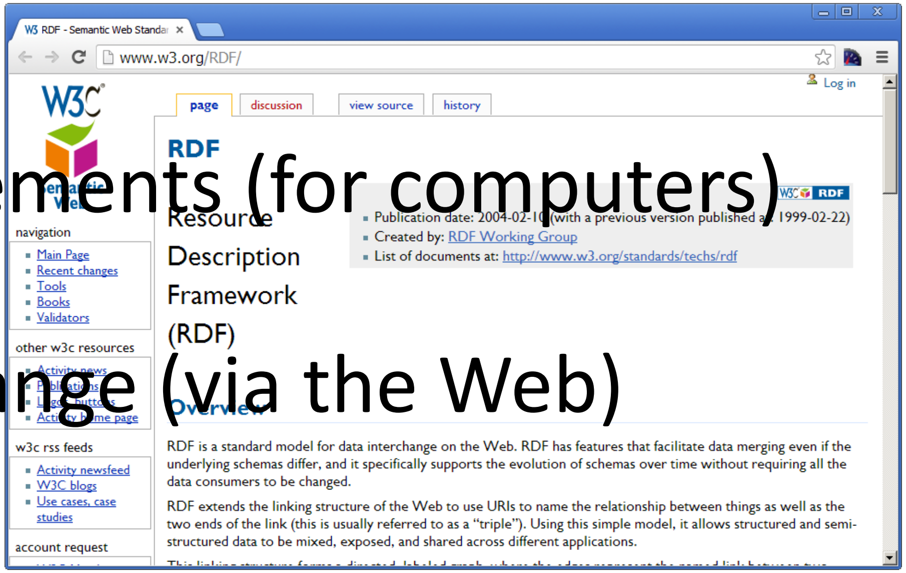
Semantic Web Technologies



RDF = Resource Description Framework **OWL** = Web Ontology Language **XML** = Extensible Markup Language
SPARQL = SPARQL Protocol and RDF Query Language

What is RDF?

- **R**esource **D**escription **F**ramework
- 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)



RDF

Resource Description Framework (RDF)

- Publication date: 2004-02-10 (with a previous version published at 1999-02-22)
- Created by: RDF Working Group
- List of documents at: http://www.w3.org/standards/techs/rdf

Overview

RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed.

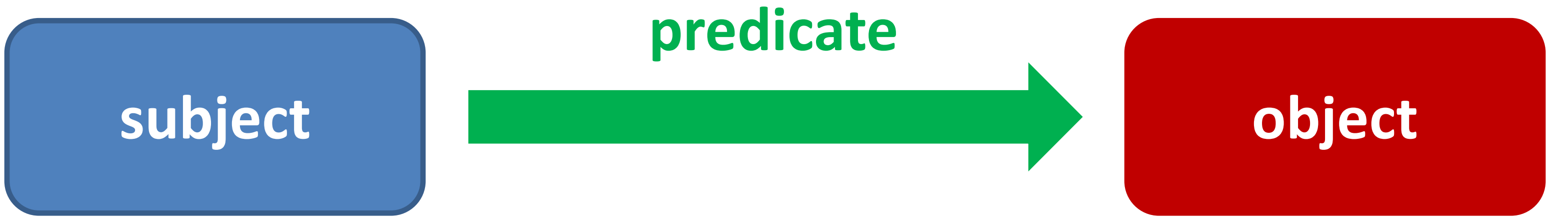
RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications.

How does RDF work?

- Data model employs the concept of triples

“**subject**-***predicate***-**object**”

“**atorvastatin** ***may treat*** **hypercholesterolemia**”



subject

predicate

object

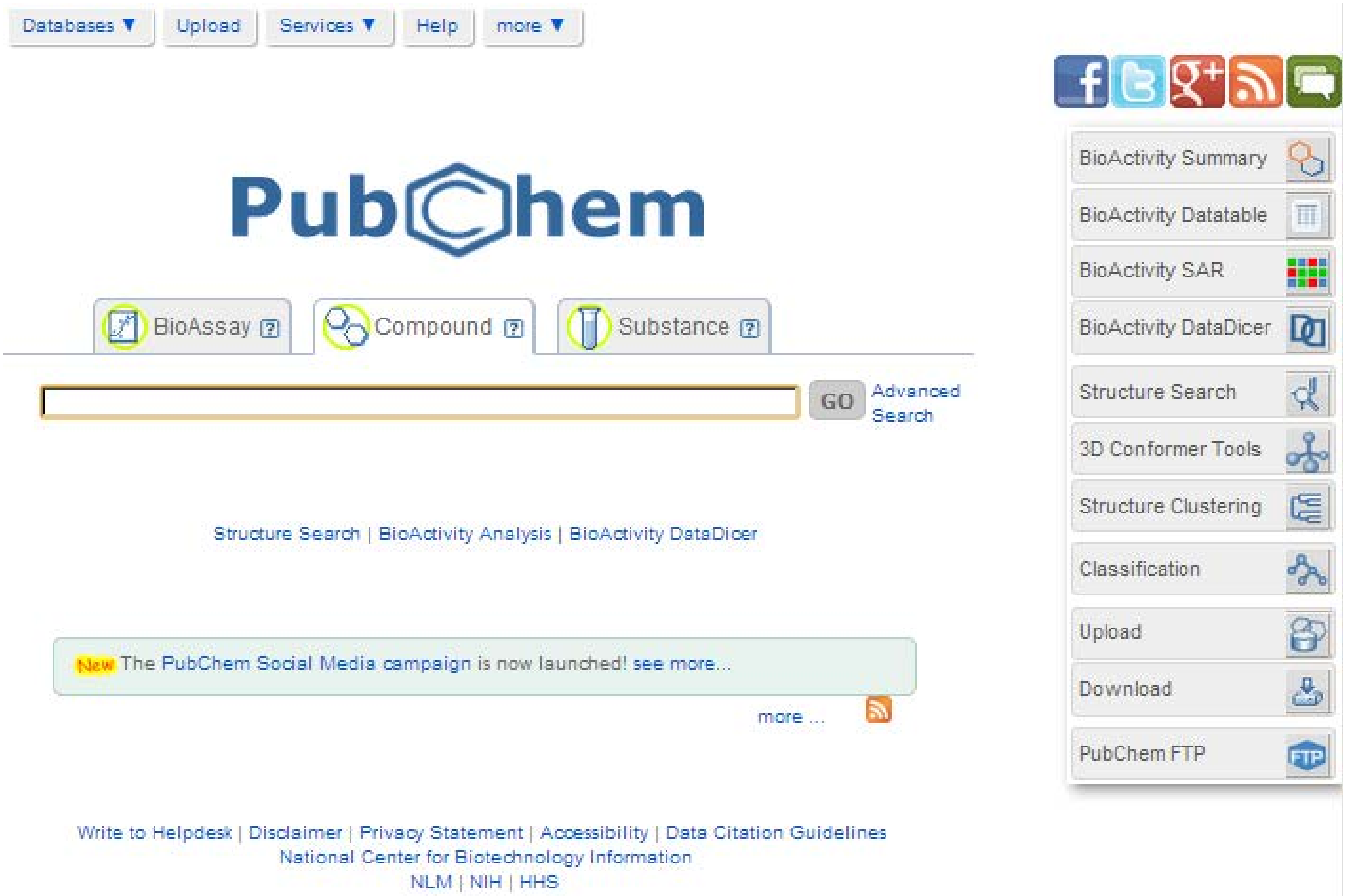
- Models statements as directed graphs

- 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
- ***URL*** is a web address (http://pubchem.ncbi.nlm.nih.gov)
- ***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., URI#frag_identifier)

What is ***PubChem*** doing with ***semantic web technologies*** and ***Why***?



New The PubChem Social Media campaign is now launched! see more...

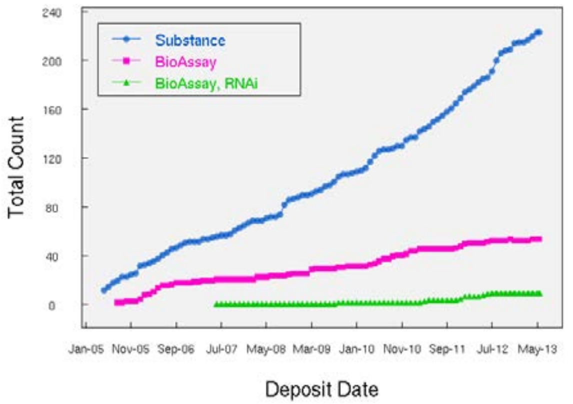
more ...

Write to Helpdesk | Disclaimer | Privacy Statement | Accessibility | Data Citation Guidelines
National Center for Biotechnology Information
NLM | NIH | HHS

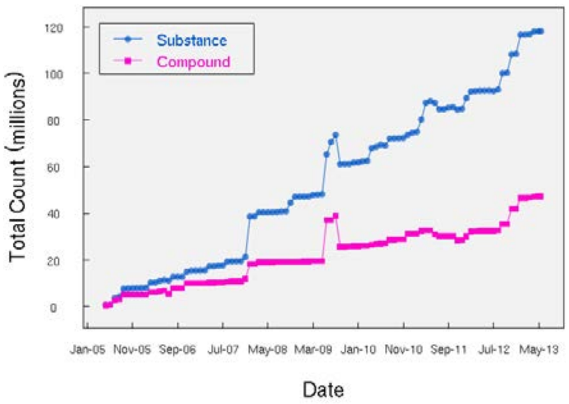
http://pubchem.ncbi.nlm.nih.gov

PubChem has a fair bit of data

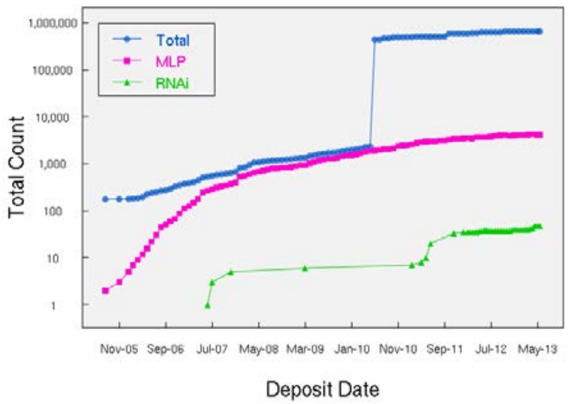
Contributors



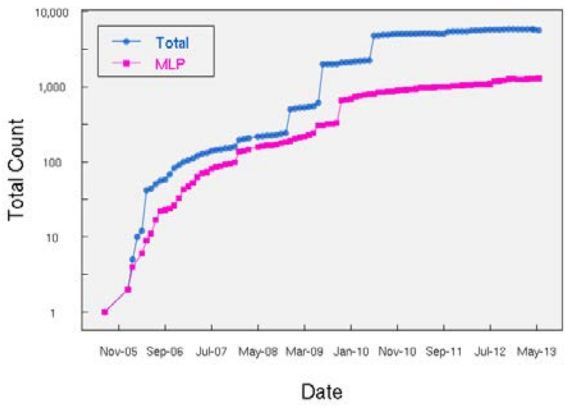
Chemicals



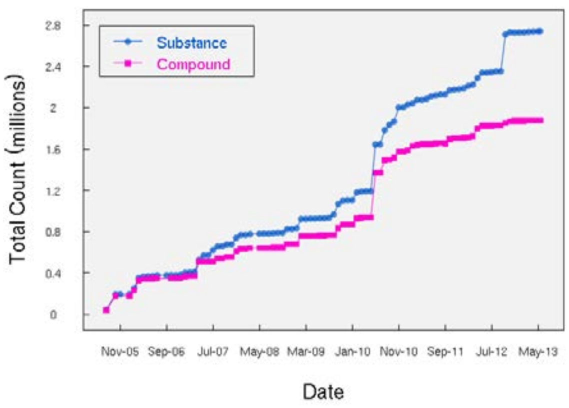
Biological Assays



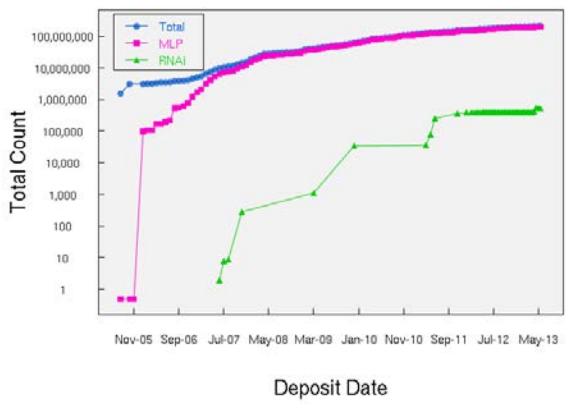
Protein Targets



Tested Chemicals



Bioactivity Results

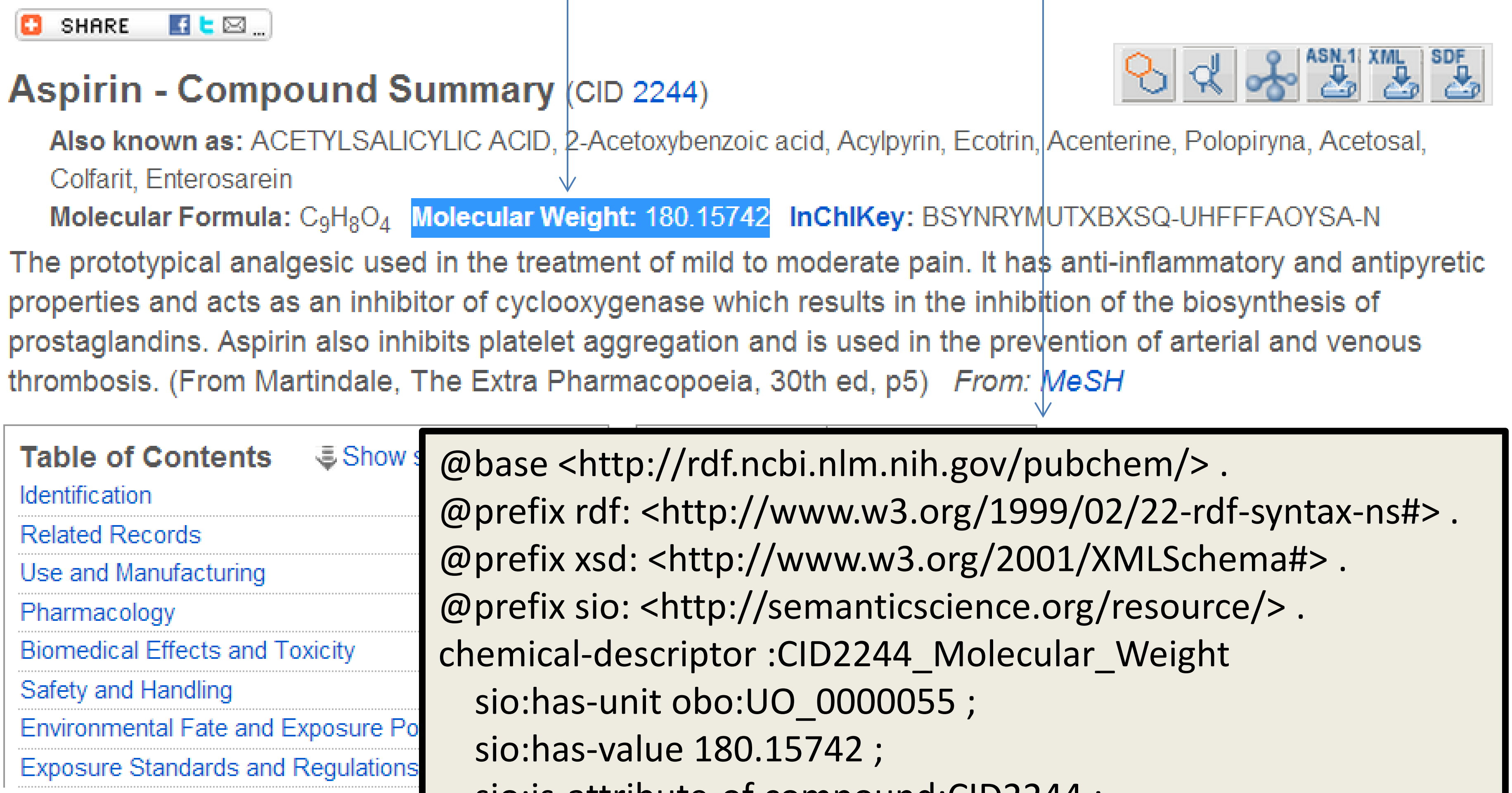


Why semantic web?

… *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**

Presentation and Data



SHARE

Aspirin - Compound Summary (CID 2244)

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

Molecular Formula: $C_9H_8O_4$ **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 Show s

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

```
@base <http://rdf.ncbi.nlm.nih.gov/pubchem/> .
@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
  - Target 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)

http://pubchem.ncbi.nlm.nih.gov/rest/rdf/chemical-descriptor/CID2244_Molecular_Weight.xml

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:sio="http://semanticscience.org/resource/"
        xml:base="http://rdf.ncbi.nlm.nih.gov/pubchem/">
  <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-descripto
    <sio:has-unit rdf:resource="http://purl.oboli
  </rdf:Description>
  <rdf:Description rdf:about="chemical-descripto
    <rdf:type rdf:resource="http://semanticscien
  </rdf:Description>
</rdf:RDF>
```

:: in Turtle format ::

```
@base <http://rdf.ncbi.nlm.nih.gov/pubchem/> .
@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)