

Linked Open Data

Egon Willighagen (@egonwillighagen)

May 2012, Maastricht, HARMONY 2012

Open Data

- **Use**
- **Mix** (or convert!)
- **Reshare**

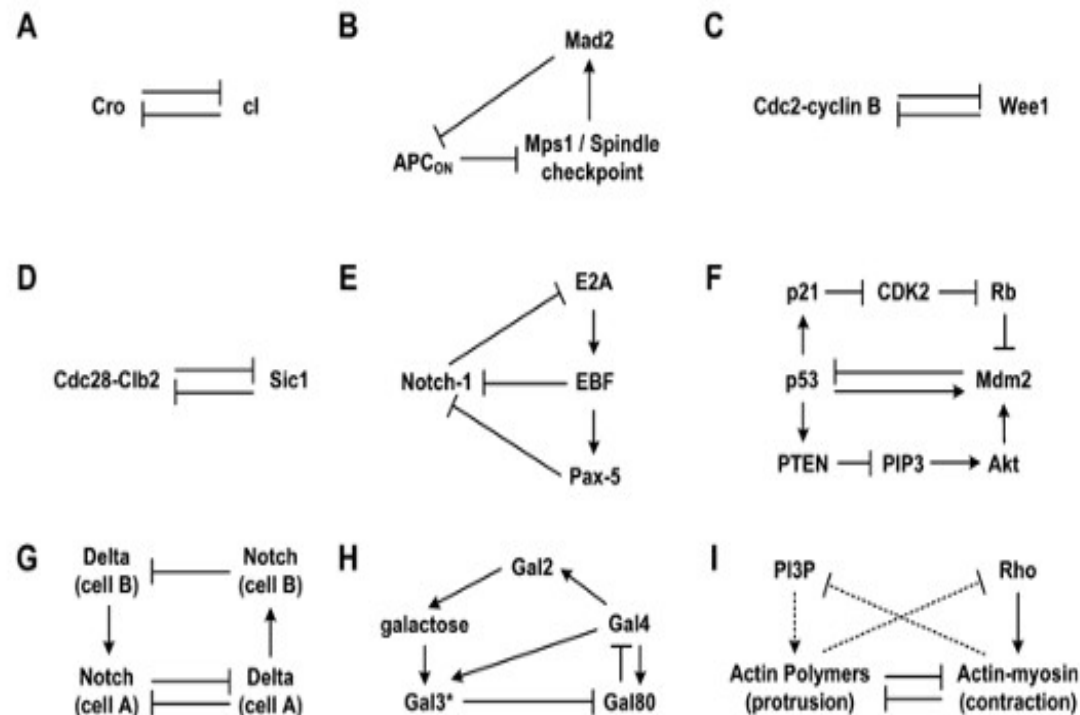
- Licenses
 - Creative Commons flavors, ...
- Waivers
 - CCZero

Linked Data: meaning

- Why Semantic Web?
 - **Unique identifiers:** URLs (registries are redundant)
 - **Meaning:** what is glucose? What are Mad2, CDK2, Pax-5 (BioModel of the month 12/05, CCZero)?

Ontologies, ...

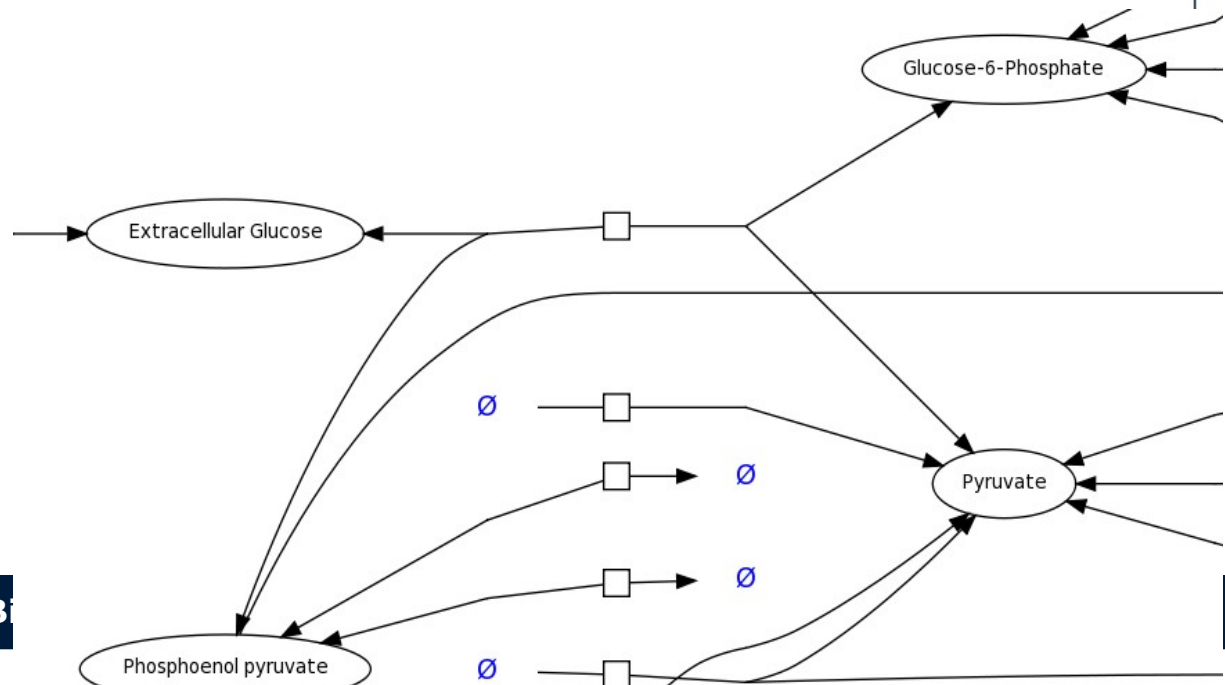
- Gene Ontology
- BioPortal, OLS, ...
- ...



Linked Data: linking

- Why Semantic Web?
 - **Web:** stand-up scientist
 - **Linking:** “Give me all models that predict something that predict glucose concentration”
 - “What are the chemical properties of glucose”

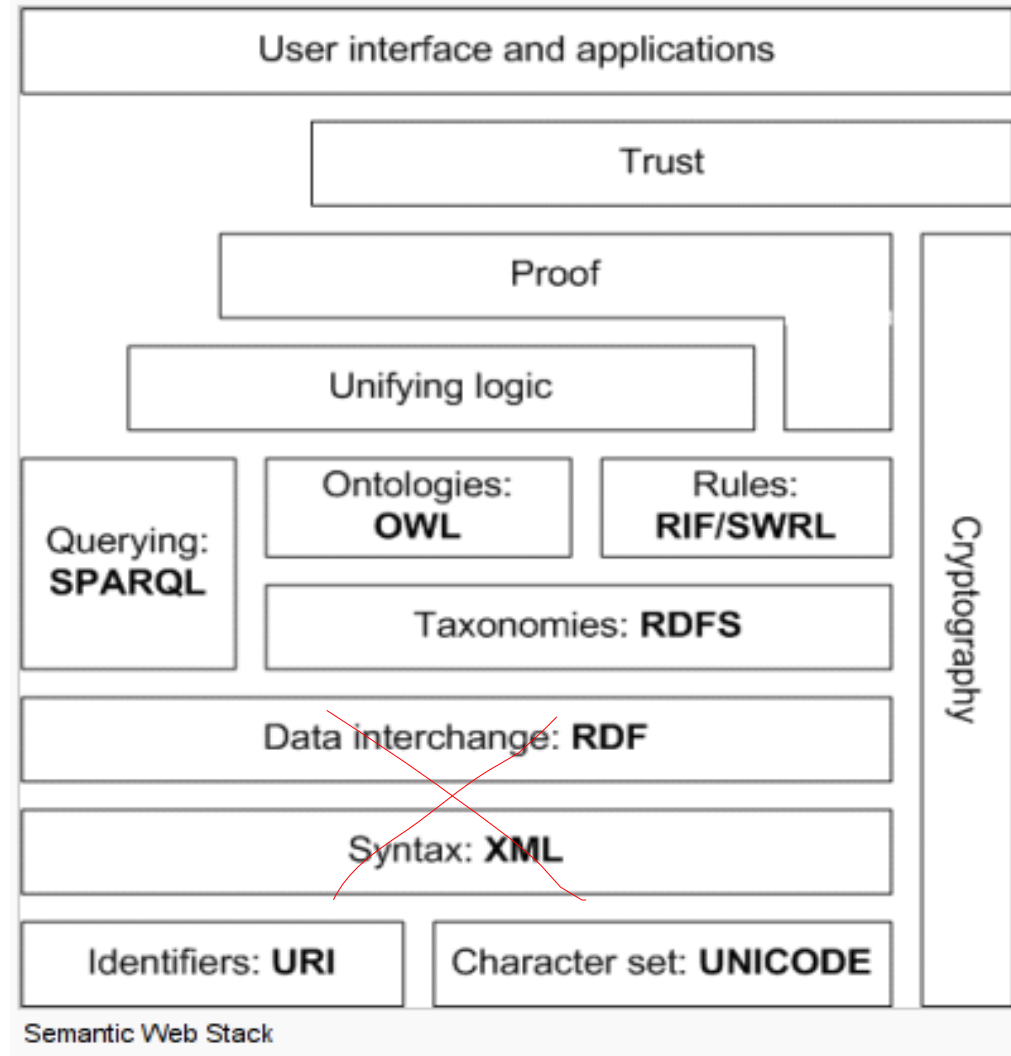
BIOMD0000000051
 CCZero



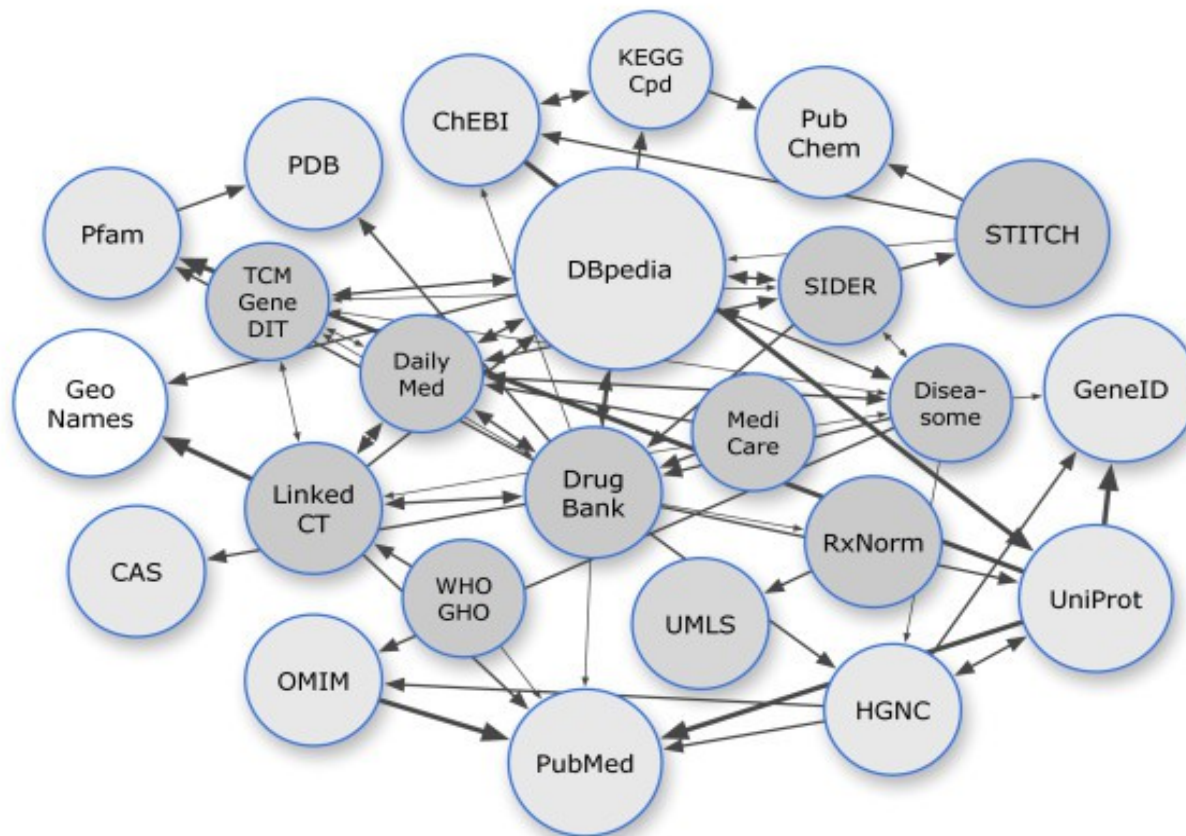
Resource Description Framework

RDF != RDF/XML

Wikipedia

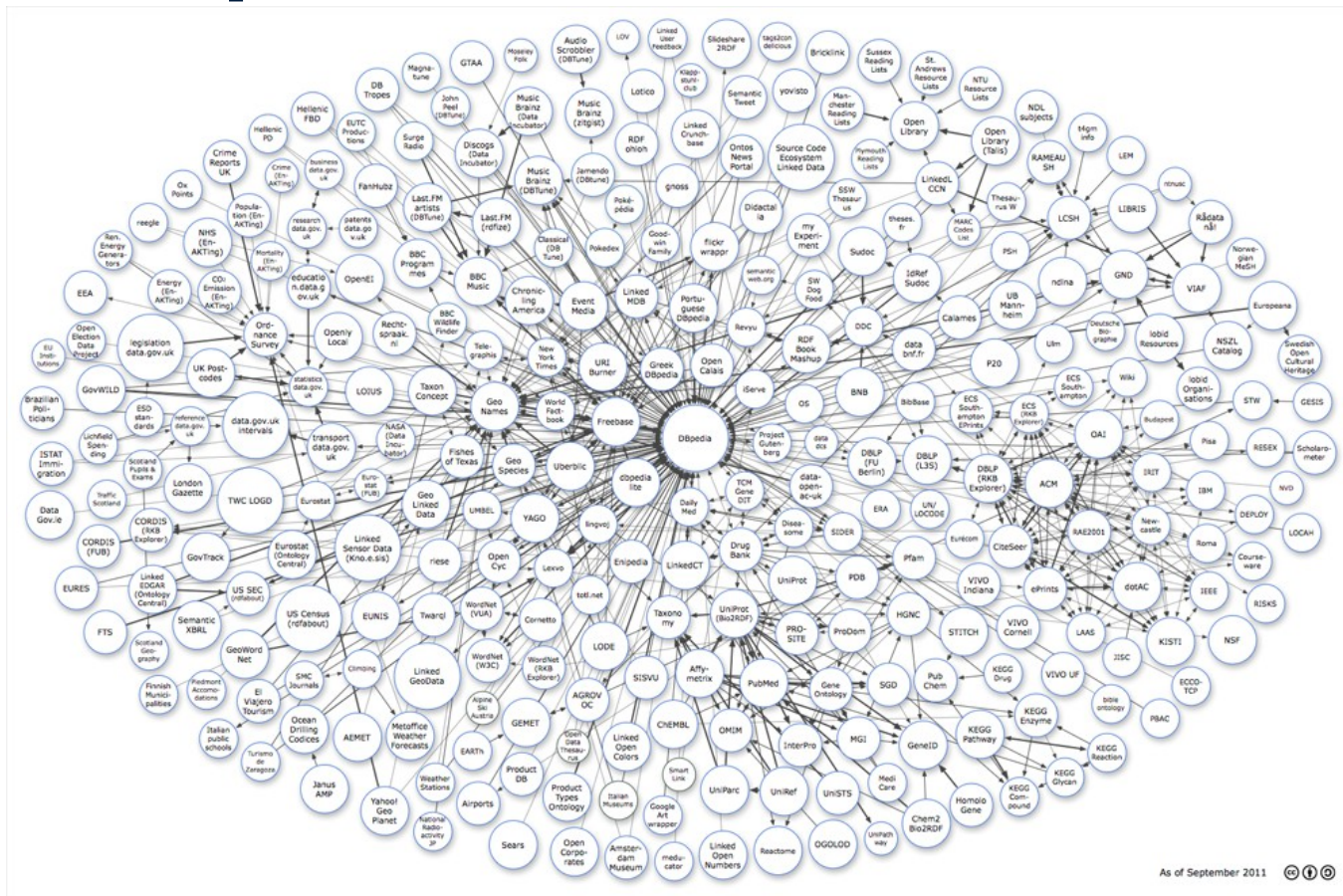


Linked Open Drug Data



M. SamWald, et al, Linked open drug data for pharmaceutical research and development, 2011, JChemInf.

Linked Open Data



Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch.
<http://lod-cloud.net/> Sept 2011, CC-BY-SA.

Linked Open Data in the Life Sciences



Google ...

glucose



40 personal results. 62,000,000 other results (0.43 seconds)

[Glucose - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Glucose

Glucose (/ˈɡluːkoʊs/ or /ˈkoʊz/; C₆H₁₂O₆, also known as D-**glucose**, dextrose, or grape sugar) is a simple sugar (monosaccharide) and an important ...

[Function](#) - [Structure and nomenclature](#) - [Physical properties](#) - [Production](#)

[Blood Glucose Levels: Testing and Normal Range](#)

diabetes.webmd.com/blood-glucose

A blood **glucose** test measures the amount of a type of sugar, called **glucose**, in your blood. **Glucose** comes from carbohydrate foods.

[What Is Glucose?](#)

www.wisegeek.com/what-is-glucose.htm

6 days ago – **Glucose** is a type of sugar the body uses for energy. Diabetics, who have poorly regulated **glucose** levels, must be sure to...

[Carbohydrates - Glucose](#)

www.elmhurst.edu/~chm/vchembook/543glucose.html

Glucose is by far the most common carbohydrate and classified as a monosaccharide, an aldose, a hexose, and is a reducing sugar. It is also known as dextrose ...

[Glucose Tests: The Test](#)

labtestsonline.org/understanding/analytes/glucose/tab/test

23 Mar 2012 – Describes how **alucose** tests are used. when a **alucose** test is ordered.

Glucose

Glucose is a simple sugar and an important carbohydrate in biology. Cells use it as the primary source of energy and a metabolic intermediate. Glucose is one of the main products of photosynthesis and fuels for cellular respiration. Wikipedia

Formula: C₆H₁₂O₆

Molar mass: 180.16 g/mol

Melting point: 146° C (294.8° F)

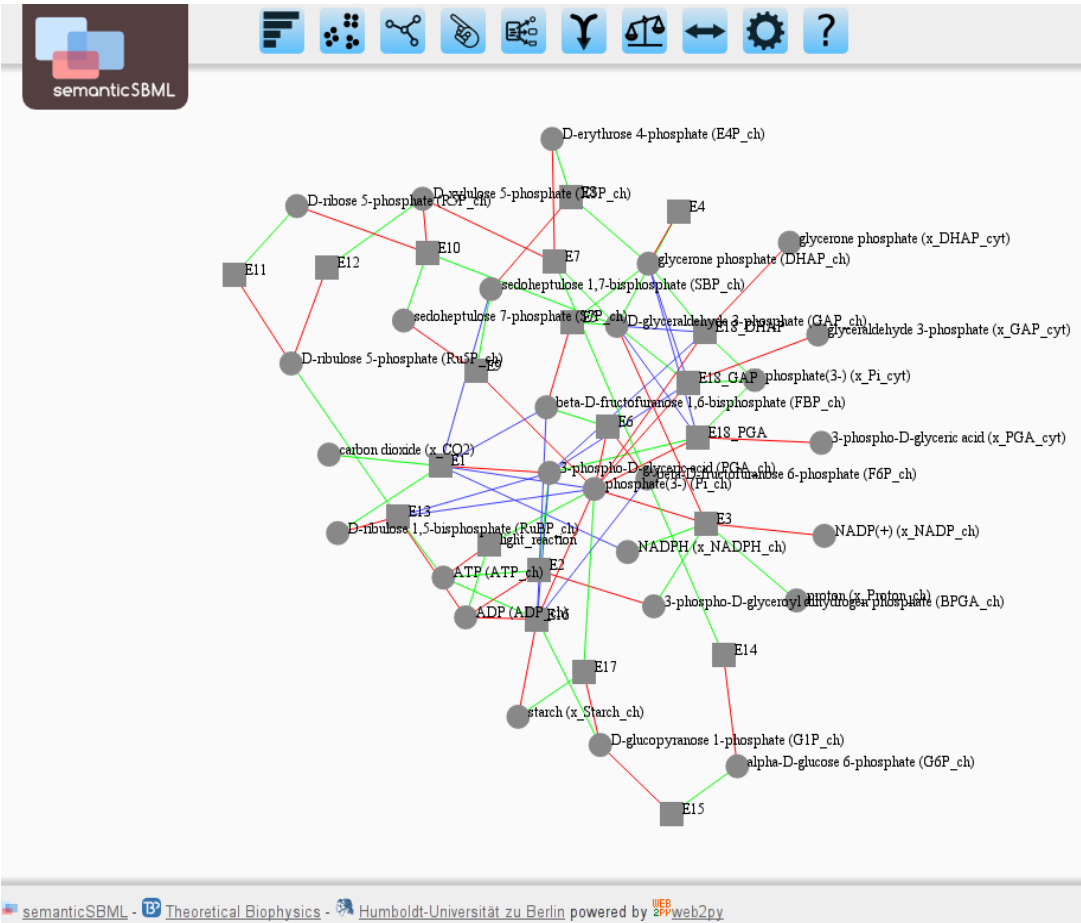
IUPAC ID: D-glucose

Density: 1.54 g/cm³

Soluble in: [Water](#)

[Report a problem](#)

W3C Health Care and Life Sciences interest group



Semantic Systems Biology

Protein-protein, protein-drug and gene-gene interaction manipulation, it only takes into consideration one layer

Semantic web technologies enables the representation

The systems biology task force will therefore identify enabling the discovery of adverse effects when a patient

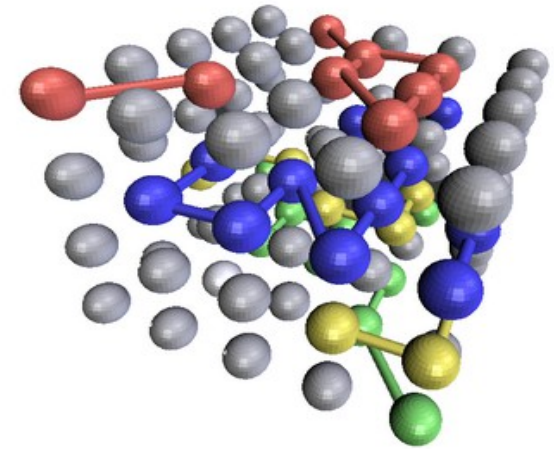


Figure 1. Different layers of information encode for data based on cellular component (e.g. both gene product and others still may include information about co-diseases)

Initial Goals and Structure of the Task Force

1. In this task force, we will pursue a pathway-based approach (e.g. MapK signalling pathway). This starts by 1) identifying the pathway

2. The approach will be that of collecting the experimental datasets that we know of such as a) TCGA; b) ICGC

W3C Health Care and Life Sciences interest group



Web Semantics: Science, Services and Agents on the World Wide Web

Available online 3 April 2012

In Press, Corrected Proof — [Note to users](#)



Emerging practices for mapping and linking life sciences data using RDF — A case series

M. Scott Marshall^{a, b, c, d, e, f}, Richard Boyce^c, Helena F. Deus^d, Jun Zhao^e, Egon L. Willighagen^f, Matthias Samwald^{g, h}, Elgar Pichlerⁱ, Janos Hajagos^j, Eric Prud'hommeaux^k, Susie Stephens^l

^a Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

^b Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands

^c Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

^d Digital Enterprise Research Institute, National University of Ireland at Galway, Ireland

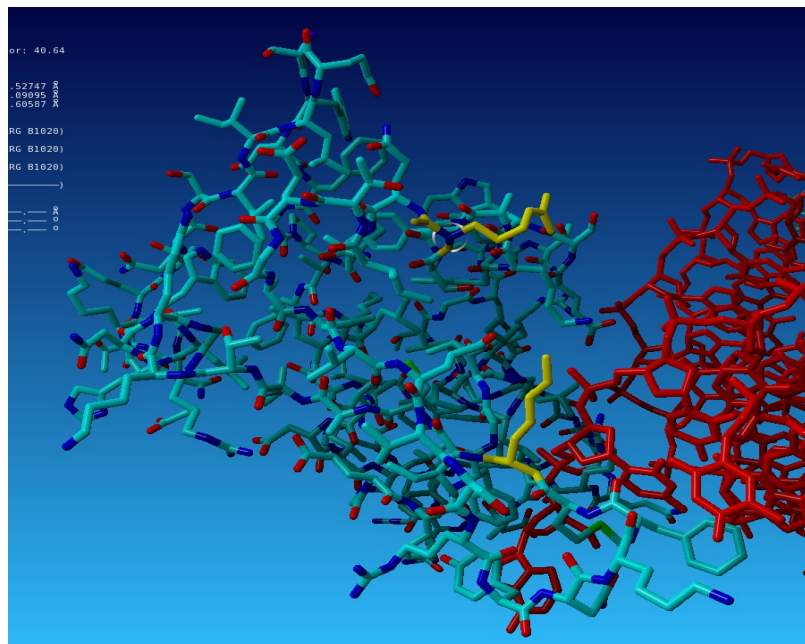
^e Department of Zoology, University of Oxford, Oxford, UK

^f Division of Molecular Toxicology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

^g <http://www.costofknowledge.com/>

<http://thecostofknowledge.com/>

Why does it matter again?



2D-Structure

Decision Support

- 145-49-3 [tanimoto=0.75]
- 116-85-8 [tanimoto=0.75]
- 641-90-7 [tanimoto=0.75]
- OpenTox
 - ECOSAR LC50 fish
 - L(expt_provided in ECOSAR manual) = 1.299674391746521
 - Lipinski Rule of Five
 - LipinskiFailures = 0.0
 - MolecularWeight
 - MW = 240.2100067138672
 - QSAR SRC KOWWIN fingerprints AD
 - AppDomain_Tanimoto = 0.0
 - Tanimoto = 0.0830889567732811
 - START biodegradation and persistence plug-in
 - START Biodegradability = Class 2 (persistent chemical)
 - START Biodegradability#explanation = Q1.Terminal tert-butyl bra
 - SmartCYP: Cytochrome P450-Mediated Drug Metabolism
 - SMARTCyp.Rank1.Accessibility = 0.8571428656578064
 - SMARTCyp.Rank1.Energy = 74.0999984741211
 - SMARTCyp.Rank1.Score = 67.24285888671875
 - SMARTCyp.Rank1.sites = C5,C7
 - SMARTCyp.Rank1.sites#explanation = <a href="#" onClick="chang
 - SMARTCyp.Rank2.Accessibility = 1.0
 - SMARTCyp.Rank2.Energy = 77.19999694824219
 - SMARTCyp.Rank2.Score = 69.19999694824219
 - SMARTCyp.Rank2.sites = C14,C17
 - SMARTCyp.Rank3.Accessibility = 1.0
 - SMARTCyp.Rank3.Energy = 80.80000305175781
 - SMARTCyp.Rank3.Score = 72.80000305175781
 - SMARTCyp.Rank3.sites = C9,C12
 - SMARTCyp.Rank>=4.sites =
 - ToxTree: Benigni/Bossa rules for carcinogenicity and mutagenicity
 - Error when applying the decision tree = NO
 - For a better assessment a QSAR calculation could be applied. = N
 - Negative for genotoxic carcinogenicity = NO
 - Negative for nongenotoxic carcinogenicity = YES

Conclusion

- Open Data
 - Use, mix, reshare
- Linked Data
 - Open standard (RDF, ...)
 - ... thus domain independent
 - Meaningful
 - via ontologies and resolvable, unique identifiers