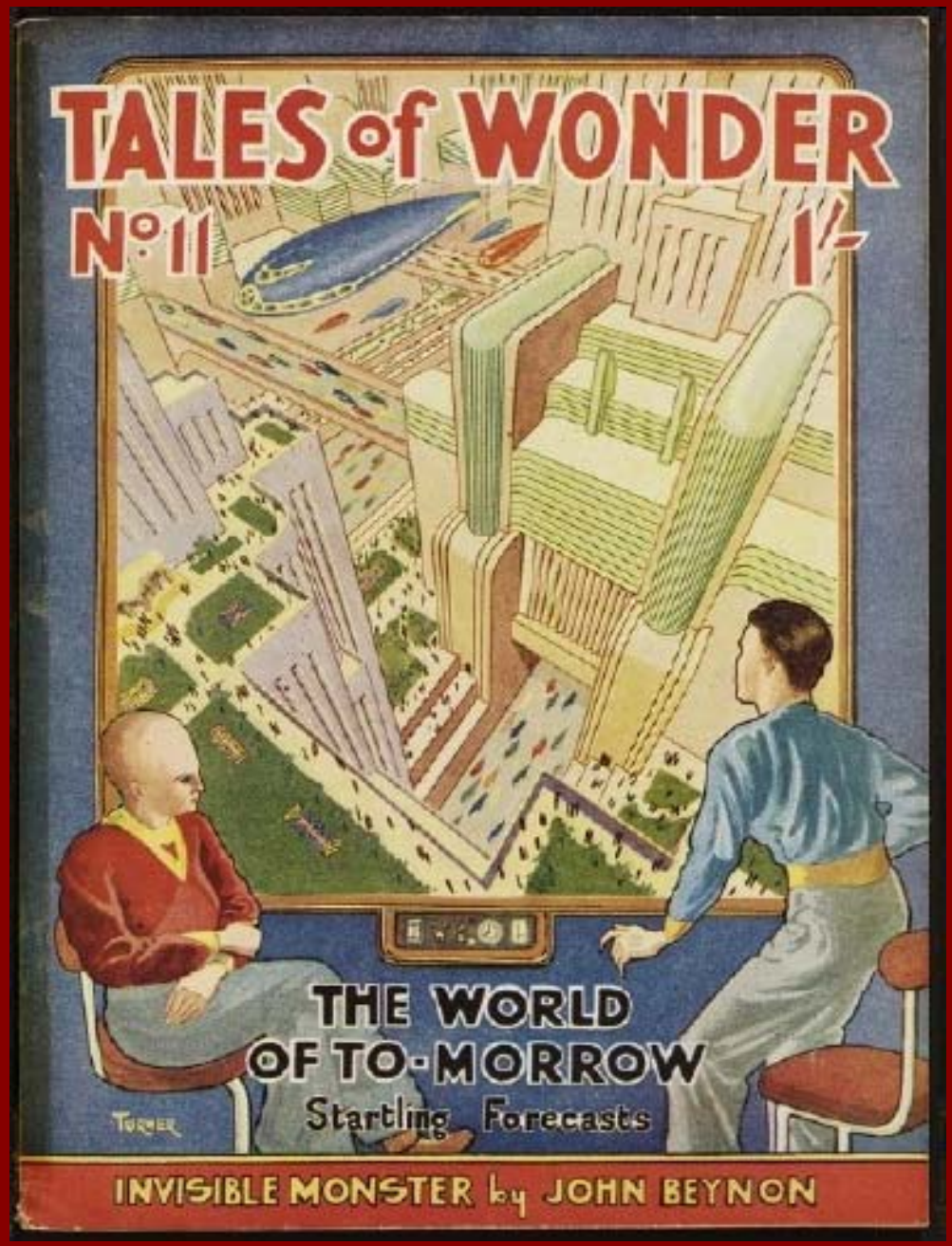
Find and Use:
New Standards Link Information to Provide New Answers

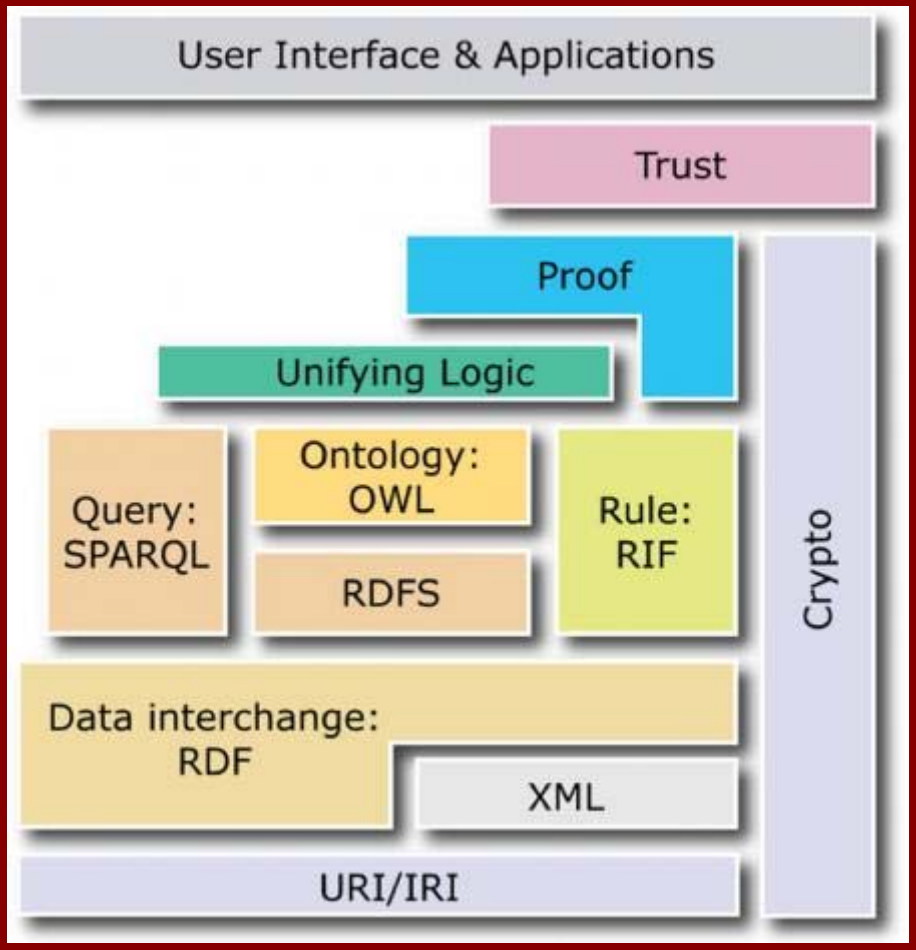
RSC Publishing

Richard Kidd | kiddr@rsc.org
19th October 2009



Are we at a tipping point when the tools and approaches common within bioinformatics finally cross over into chemistry and provide compelling use cases?

RSC Publishing



RSC Publishing

http://www.leirdal.net

Role:

- To fund development and support of the IUPAC InChI standard
- Working groups set up by IUPAC Subcommittee: reactions, organometallics, polymers, markush, business rules for structure input

RSC Publishing

Current members of the Trust:

ACD/Labs
ChemAxon
Elsevier
FIZ Chemie
Informa / Taylor & Francis
NPG
OpenEye
RSC
Symyx Technologies
Thomson-Reuters
Wiley-Blackwell

RSC Publishing

- Project Director: Steve Heller

- Board Chairman: Jason Wilde, NPG
- Secretary: Alan McNaught
- Treasurer: Richard Kidd, RSC

RSC Publishing

RInChI

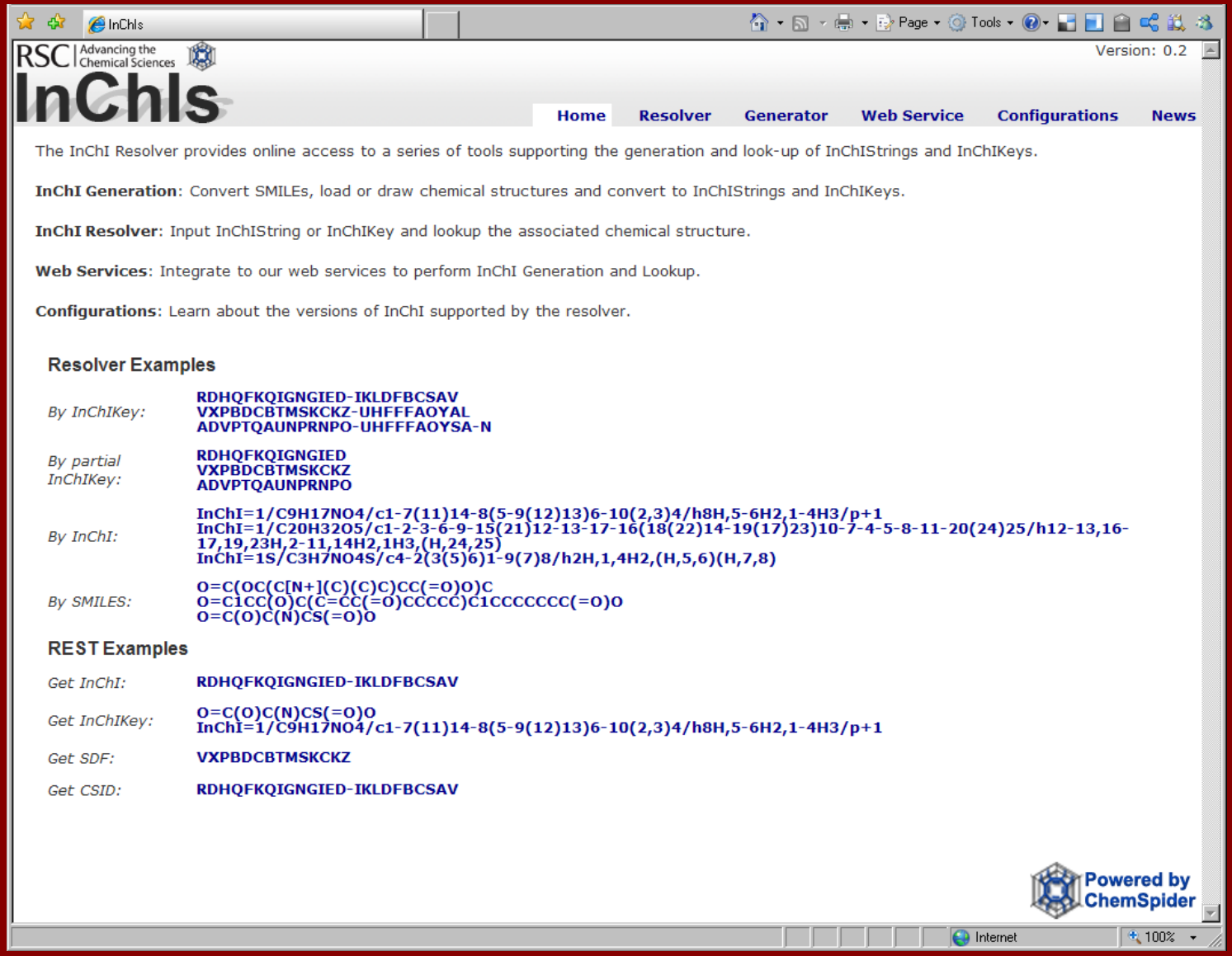
- Reactions

The RInChI Project

The aim of the RInChI project, in the same vain as InChI, is to create a unique data string to describe a reaction. Reaction InChIs, or RInChIs, are such data strings. They use the InChI software and from an rxn input file a RInChI can be created. The tools for doing this are below, and some helpful information is given in the help pages.

  - Jonathan Goodman
  - http://www-rinchi.ch.cam.ac.uk/

RSC Publishing



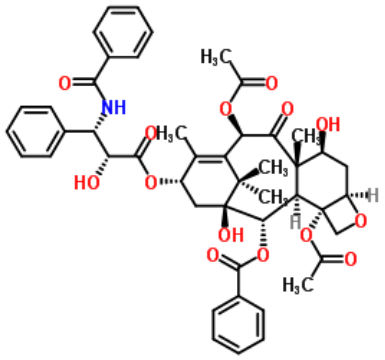
RSCPublishing

inchis.chemspider.com

InChIs

Version: 0.0.0.0

Home Resolver Generator Web Service Configurations



Download MOL Download SDF

ChemSpider ID: 10368587

SMILES: O=C(N[C@@H](c1ccccc1)[C@@H](O)C(=O)O[C@H]5C[C@@]6(O)[C@@H](OC(=O)c2ccccc2)[C@H]3[C@@](C)([C@@H](O)C[C@H]4OC[C@@]34OC(C)=O)C(=O)[C@H](OC(C)=O)\C(=C5/C)[C@]6(C)C)c7ccccc7

Molecular Formula: $C_{47}H_{51}NO_{14}$

Molecular Weight: 853.9061

InChI (v1.02b): InChI=1/C47H51NO14/c1-25-31(60-43(56)36(52)35(28-16-10-7-11-17-28)48-41(54)29-18-12-8-13-19-29)23-47(57)40(61-42(55)30-20-14-9-15-21-30)38-45(6,32(51)22-33-46(38,24-58-33)62-27(3)50)39(53)37(59-26(2)49)34(25)44(47,4)5/h7-21,31-33,35-38,40,51-52,57H,22-24H2,1-6H3,(H,48,54)/t31-,32-,33+,35-,36+,37+,38-,40-,45+,46-,47+/m0/s1

InChI Key (v1.02b): RCINICONZNJXQF-MZXODVADBJ

InChI (v1.02b, fixedH): InChI=1/C47H51NO14/c1-25-31(60-43(56)36(52)35(28-16-10-7-11-17-28)48-41(54)29-18-12-8-13-19-29)23-47(57)40(61-42(55)30-20-14-9-15-21-30)38-45(6,32(51)22-33-46(38,24-58-33)62-27(3)50)39(53)37(59-26(2)49)34(25)44(47,4)5/h7-21,31-33,35-38,40,51-52,57H,22-24H2,1-6H3,(H,48,54)/t31-,32-,33+,35-,36+,37+,38-,40-,45+,46-,47+/m0/s1/f/h48H

InChI Key (v1.02b, fixedH): RCINICONZNJXQF-GXKQXQCDDN

InChI (v1.02s): InChI=1S/C47H51NO14/c1-25-31(60-43(56)36(52)35(28-16-10-7-11-17-28)48-41(54)29-18-12-8-13-19-29)23-47(57)40(61-42(55)30-20-14-9-15-21-30)38-45(6,32(51)22-33-46(38,24-58-33)62-27(3)50)39(53)37(59-26(2)49)34(25)44(47,4)5/h7-21,31-33,35-38,40,51-52,57H,22-24H2,1-6H3,(H,48,54)/t31-,32-,33+,35-,36+,37+,38-,40-,45+,46-,47+/m0/s1

InChI Key (v1.02s): RCINICONZNJXQF-MZXODVADSA-N

Linked to literature

InChI Key (v1.02b, fixedH): HVYWMOMLDIMFJA-DPAQBDIFBB

InChI (v1.02s): InChI=1S/C27H46O/c1-18(2)7-6-8-19(3)23-11-12-24-22-10-9-20-17-21(28)13-15-26(20,4)25(22)14-16-27(23,24)5/h9,18-19,21-25,28H,6-8,10-17H2,1-5H3/t19-,21+,22+,23-,24+,25+,26+,27-/m1/s1

InChI Key (v1.02s): HVYWMOMLDIMFJA-DPAQBDIFSA-N

References:

Mével Mathieu. **Novel neutral imidazole-lipophosphoramides for transfection assays**, Chemical Communications, 2008
[DOI: **10.1039/b805226c**]

Aparicio Jesús F.. **Microbial cholesterol oxidases: bioconversion enzymes or signal proteins?**, Molecular BioSystems, 2008
[DOI: **10.1039/b717500k**]

Numata Munenori. **Creation of polynucleotide-assisted molecular assemblies in organic solvents: general strategy toward the creation of artificial DNA-like nanoarchitectures**, Organic & Biomolecular Chemistry, 2008
[DOI: **10.1039/b713354e**]

Gater Deborah L.. **Formation of the liquid-ordered phase in fully hydrated mixtures of cholesterol and lysopalmitoylphosphatidylcholine**, Soft Matter, 2008
[DOI: **10.1039/b710726a**]

Tawakol Ahmed. **Intravascular detection of inflamed atherosclerotic plaques using a fluorescent photosensitizer targeted to the scavenger receptor**, Photochemical & Photobiological Sciences, 2008
[DOI: **10.1039/b710746c**]

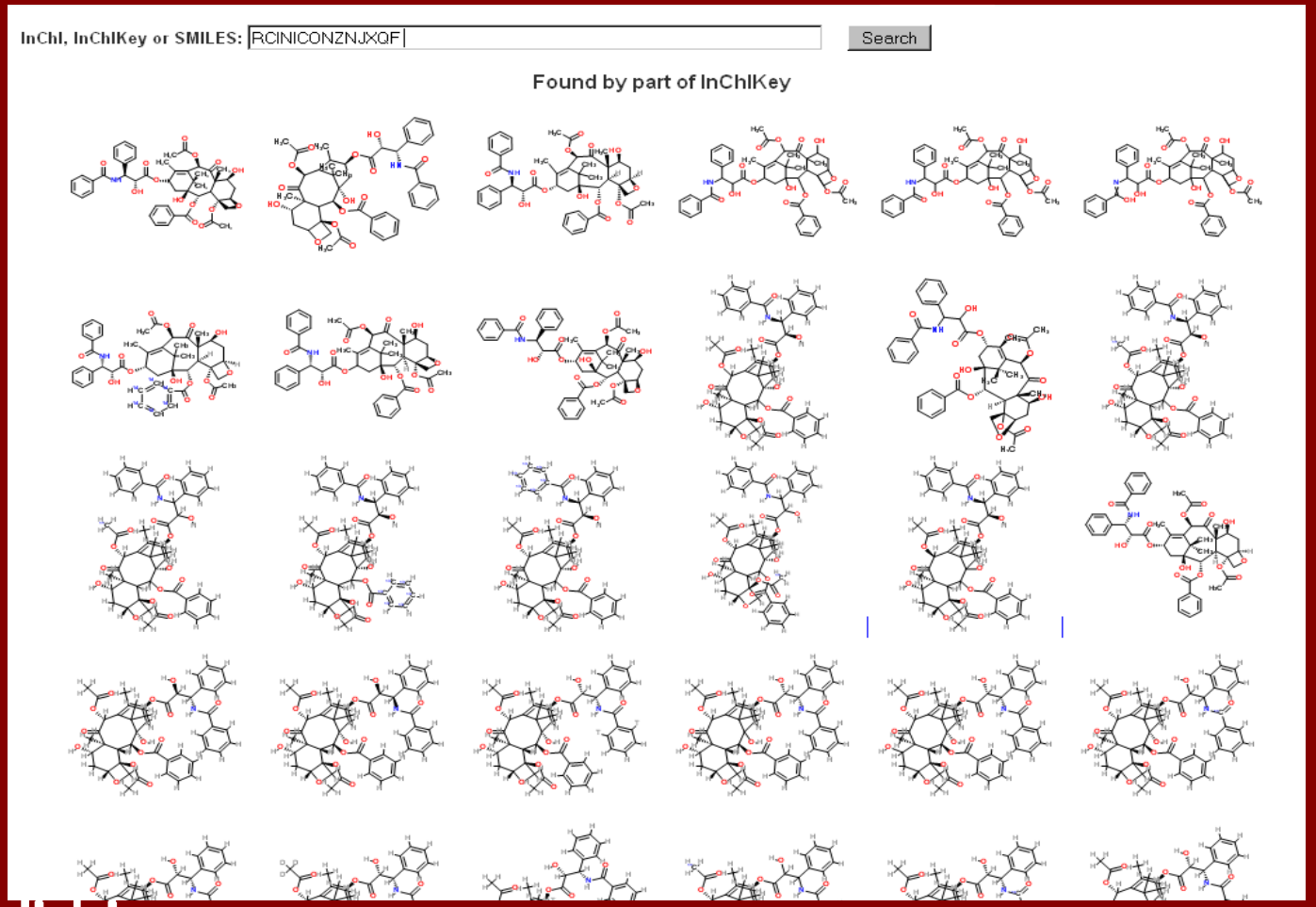
Tadashi Yoshida, Akira Honda, Hiroshi Miyazaki and Yasushi Matsuzaki. **Determination of Key Intermediates in Cholesterol and Bile Acid Biosynthesis by Stable Isotope Dilution Mass Spectrometry**, Analytical Chemistry Insights 2008:3 45-60
[DOI:]

Polozov et al.. **Progressive Ordering with Decreasing Temperature of the Phospholipids of Influenza Virus**, Nature Chemical Biology, doi: 10.1038/nchembio.77, published online 2 March 2008
[DOI: **10.1038/nchembio.77**]

Meloni et al.. **Metal swap between Zn7-metallothionein-3 and amyloid-beta Cu protects against amyloid-beta toxicity**, Nature Chemical Biology, doi: 10.1038/nchembio.89, published online 4 May 2008.
[DOI: **10.1038/nchembio.89**]

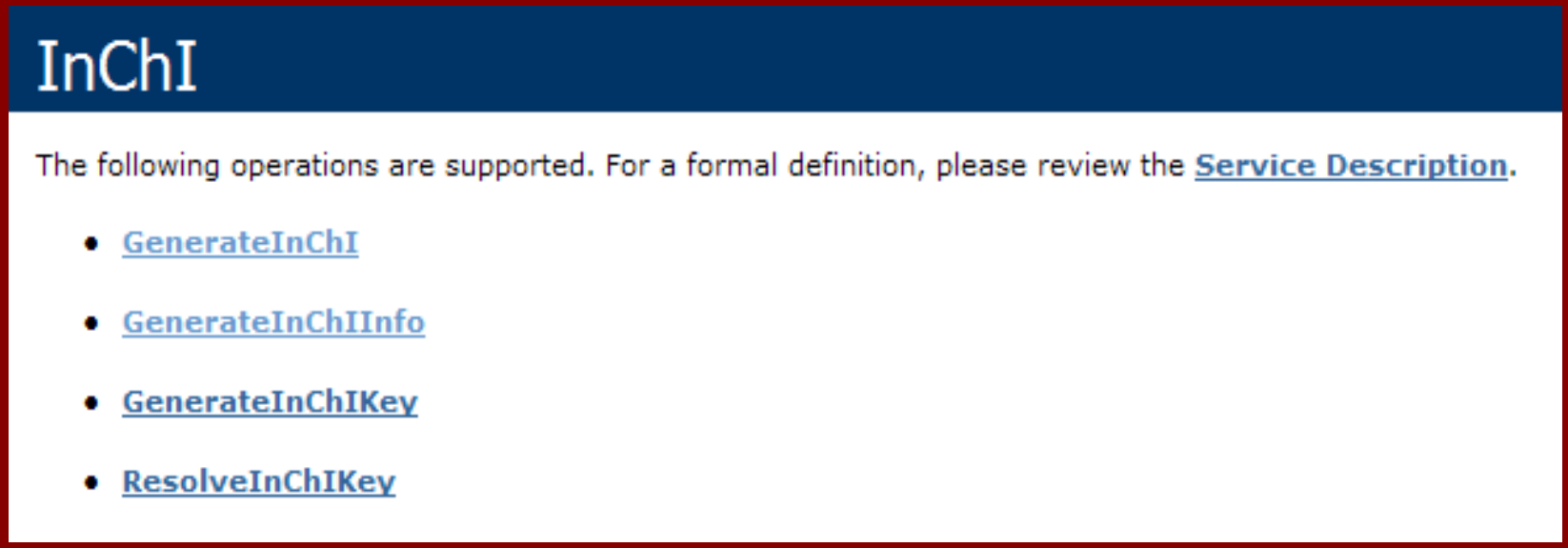
RSC Publishing

Resolve the skeleton

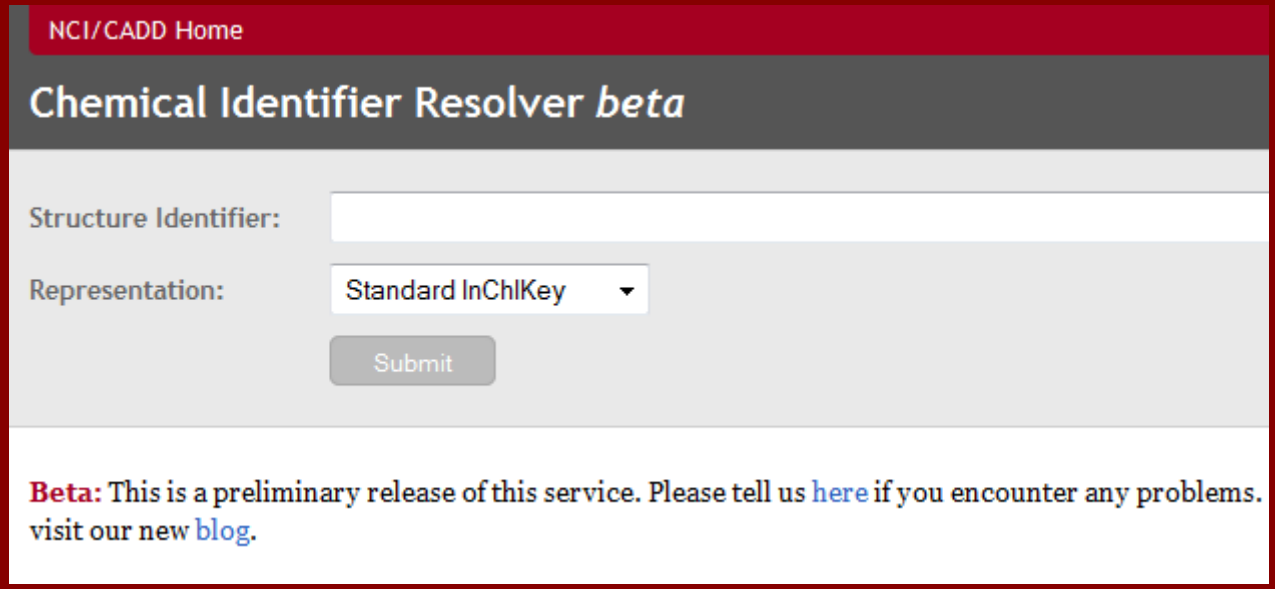


RSC Publishing

Serve Up Services



NCI Resolver



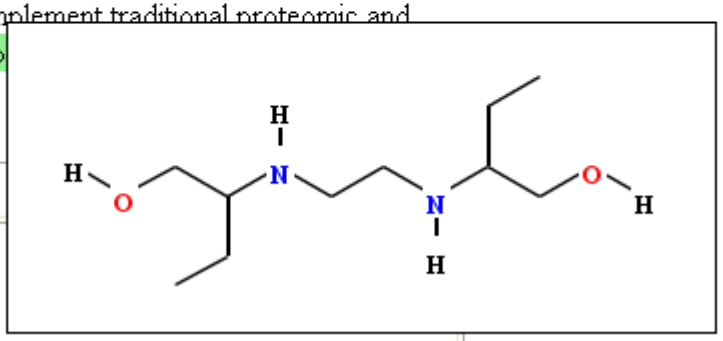
- So we need a Resolver Protocol

Publishers’ semantic markup

RSC Publishing

RSC and mining

Bacterial surface-associated proteins play essential roles in mediating pathogen–host interactions and represent privileged targets for anti-adhesion therapy. We used atomic force microscopy (AFM) to investigate, *in vivo*, the binding strength and surface distribution of fibronectin attachment proteins (FAPs) in *Mycobacterium bovis bacillus* Calmette-Guérin (BCG). We measured the specific binding forces of FAPs (~50 pN) and found that they increased with the loading rate, as observed earlier for other receptor–ligand systems. We also mapped the distribution of FAPs, revealing that the proteins are widely exposed on the mycobacterial surface. To demonstrate that the proteins are surface-associated, we showed that treatment of the cells with pullulanase, an enzyme possessing carbohydrate-degrading activities, or with protease, an enzyme that conducts proteolysis, led to a substantial reduction of the FAP surface density. A similar trend was also noted following treatment with ethambutol, an antibiotic which inhibits the synthesis of cell wall polysaccharides. The nanoscale analyses presented here complement traditional proteomic and molecular biology approaches for the functional analysis of surface-associated pro
novel anti-adhesive drugs.



Insight, innovation, integration

Studying the structure and function of bacterial cell adhesion proteins—referred to as adhesins—is essential given

RSC Publishing

Enhanced HTML

Proteolysis

Definition: The chemical reactions and pathways resulting in the breakdown of a protein by the destruction of the native, active configuration, with the hydrolysis of peptide bonds.

ID: GO:0006508

Synonyms:

- peptidolysis
- ATP-dependent proteolysis

Articles referencing this term

Microcins, gene-encoded antibacterial peptides from enterobacteria
Sophie Duquesne, Delphine Destoumieux-Garzón, Jean Peduzzi and Sylvie Rebuffat, *Nat. Prod. Rep.*, 2007 , **24** , 708
DOI: 10.1039/b516237h

The type I fatty acid and polyketide synthases: a tale of two megasynthases
Stuart Smith and Shiou-Chuan Tsai, *Nat. Prod. Rep.*, 2007 , **24** , 1041
DOI: 10.1039/b603600g

The ubiquitous carrier protein – a window to metabolite biosynthesis

RSC Publishing

Why are we doing this?

A solution looking for many problems

- Enhanced reader experience
- Current awareness
- Information retrieval (pre-indexing)

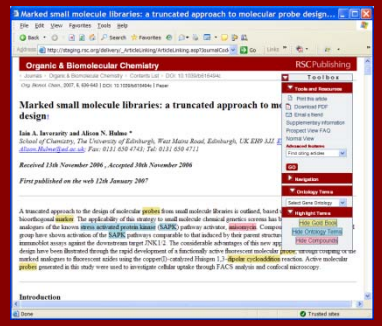
XML

Text mining (Oscar)

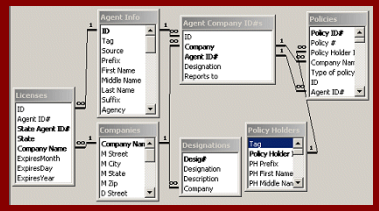
http://www.sciborg.org.uk/

http://oscar3-chem.sourceforge.net/

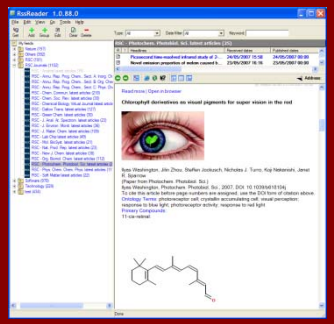
Manual QA



Enhanced HTML



Database



Enhanced RSS

RSC Publishing

Resources we use

Static

IUPAC Gold Book

Dynamic

OBO biomedical ontologies, especially ChEBI

RSC ontologies (http://www.rsc.org/ontologies)

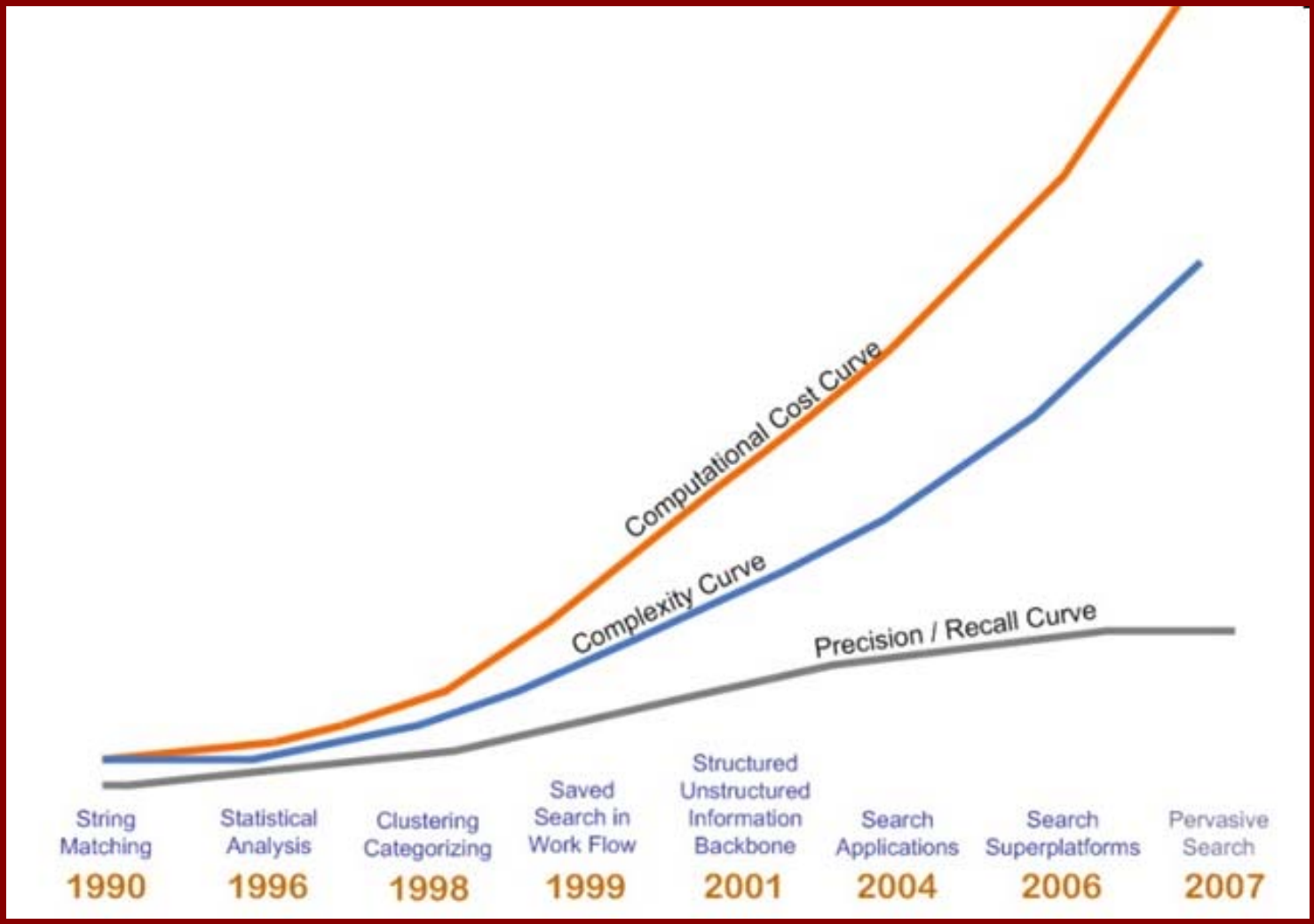
CMO, RXNO, MOP (and more to come)

RSC Ontology development

Annotations to a particular ontology are a moving target

And we can't guarantee completeness

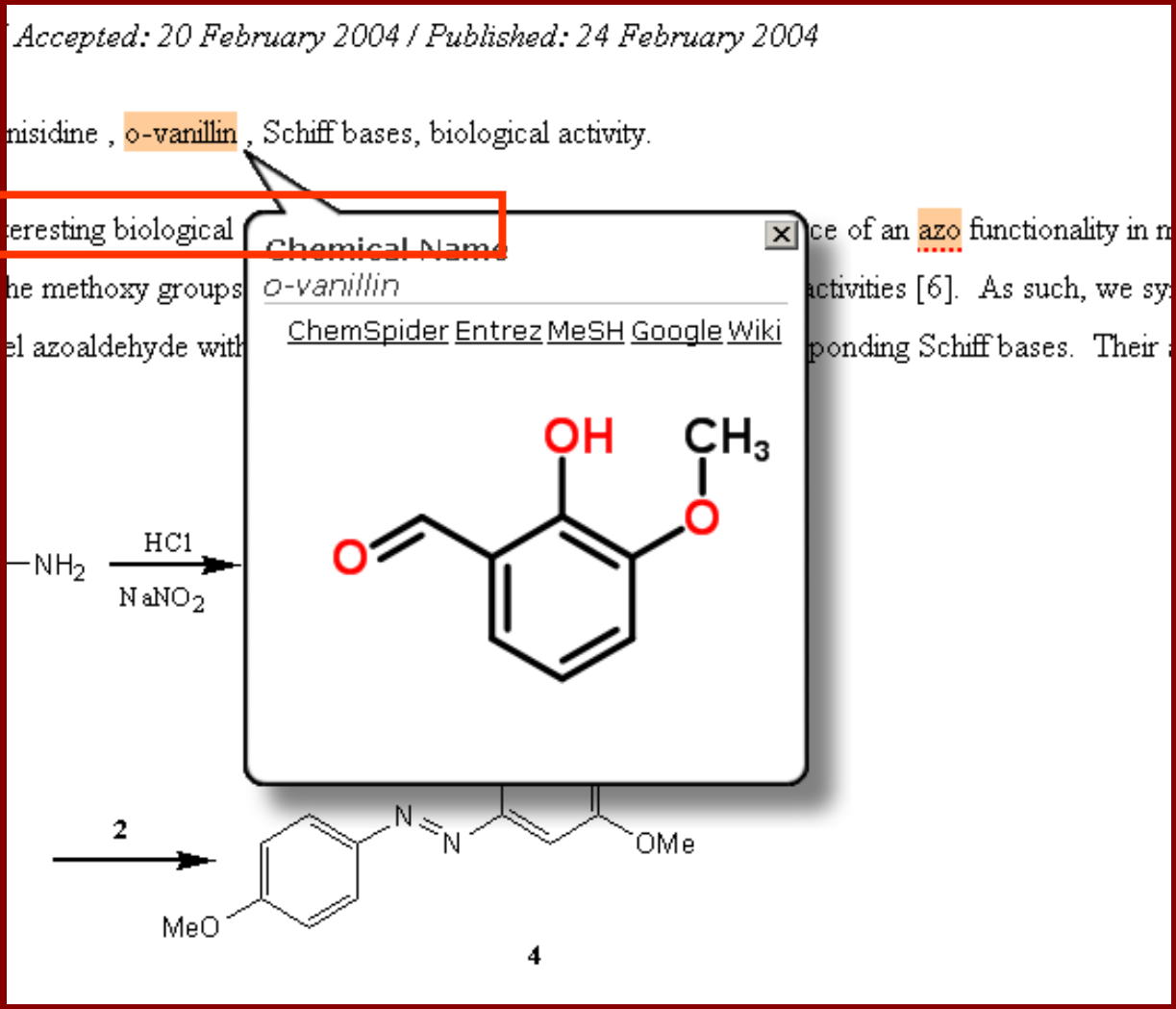
RSC Publishing



Stephen Arnold
Search: The Three Curves of Despair
March 2008

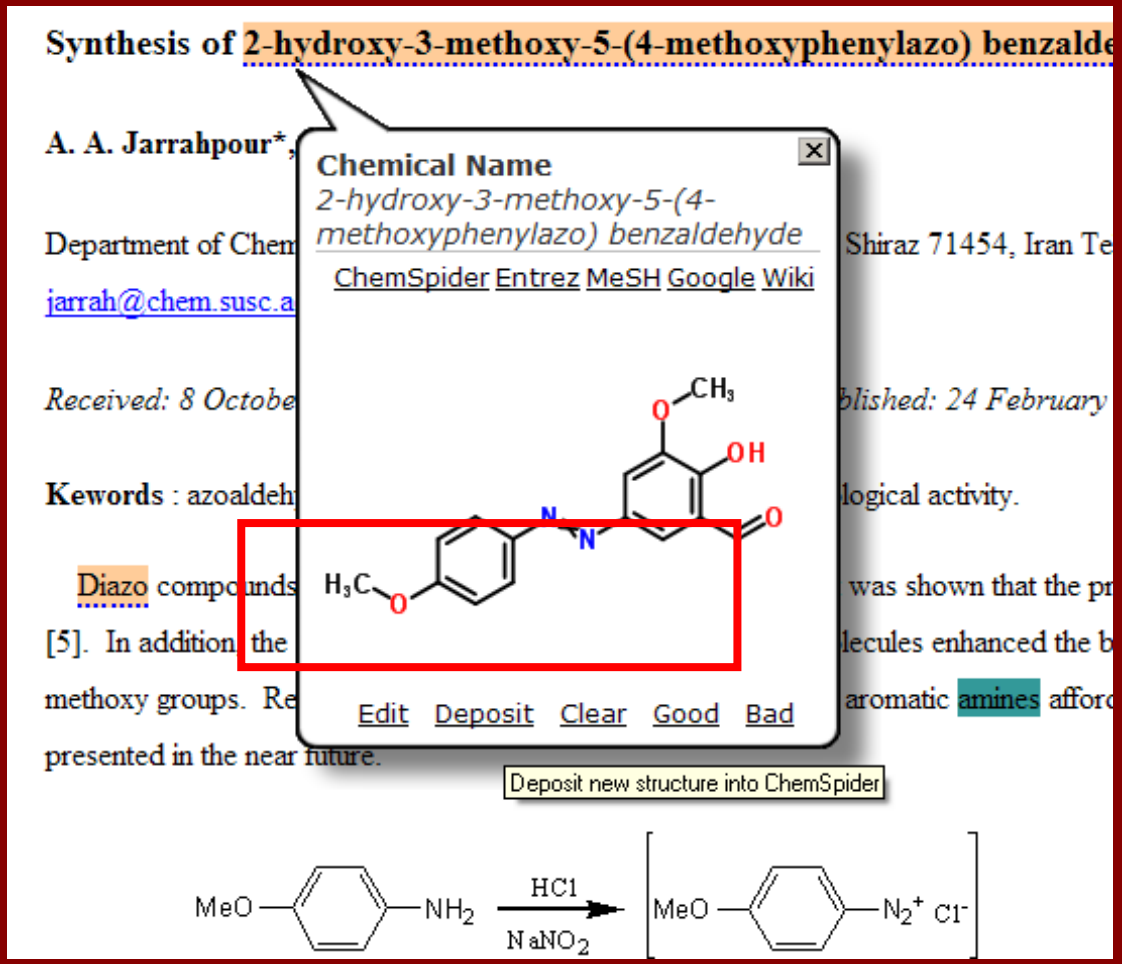
RSC Publishing

ChemMantis



RSC Publishing

Deposit Structures



RSC Publishing

Annotation: where and when?

Pre-publication?

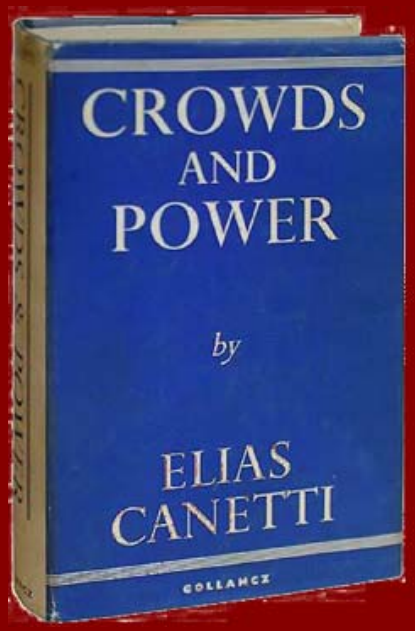
(by authors)

?

At publication?

(by editors)

Prospect



After publication?

(by the crowd)

ChemMantis

RSC Publishing

Challenges

Open problems

- Chemical structures from images
- Productive identifiers for productively-named entities

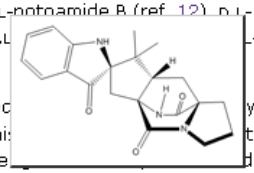
Putting ChemMantis and Prospect together

- Backfile (to 1841)
- Community curation

RSC Publishing

NPG

Work from our laboratory[4], as well as that from Sammes[9] a
revealed that these alkaloids are all derived from one or tw
isoprene units, tryptophan, and a cyclic amino acid such as
-methylproline or pipecolic acid. Significant experimental evi
the bicyclo[2.2.2]diazaoctane core common to all of these n
biosynthetically via an intramolecular hetero-Diels–Alder (IM
5-hydroxypyrazin-2(1*H*)-one[4,9]. Indeed, we have applied s
strategies to the total synthesis of several of these prenyla
including D,L-stephacidin A (refs 11,12), D,L-brevianamide B
D,L-marcfortine C (ref. 16), D,L-notoamide B (ref. 12), D,L- an
17,18), and, most recently, D,L-ma
19).



Within the family of prenylatec yclo
core, two distinct stereochemi th r
configuration at the C19 stere de r
Whereas the brevianamides **9** and **10** possess an *anti* relat
the members of the paraherquamide and notoamide family,
7, possess a *syn* relative configuration. To date, all of the st
also possess a *syn* configuration (C6, stephacidin numberin
notable exception, versicolamide B (**8**; ref. 20). Owing in pa
stereochemical anomaly, versicolamide B (**8**) attracted our ir
biogenetic and a synthetic perspective.

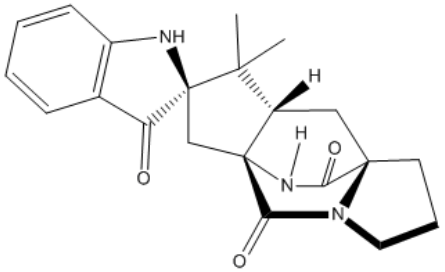
As mentined above, it has been hypothesized that the bicyc
core common to all of these natural products arises biosyntl
reaction, and evidence is mounting that this key transforma
enzyme-mediated. Enzyme-catalysed Diels–Alder reactions

RSC Publishing

Compound 9

From the following article
Asymmetric total syntheses of (+)- and (-)-versicolamide B and biosynthetic implications
Kenneth A. Miller, Sachiko Tsukamoto & Robert M. Williams
Nature Chemistry **1**, 63-68 (2009) Published online: 19 March 2009
doi:10.1038/nchem.110
Supplementary information

back to article



Compound 9
(-)-Brevianamide A

View in PubChem

View in 3D (5 KB) | Download CML file (4 KB) | Download ChemDraw file of structure (4 KB)

Synonyms

Spiro(5H,6H-5a,9a-(iminomethano)-1H-cyclopent(f)indolizine-7(8H),2'-(2H)indole)-3',5,10(1'H)-trione, 2,3,8a,9-tetrahydro-8,8-dimethyl-, (2'S,5aR,8aS,9aR)-

Chemical Formula: $C_{21}H_{23}N_3O_3$

Molecular Weight: 365.43

Elemental Analysis: C, 69.02; H, 6.34; N, 11.50; O, 13.13

InChI=1/C21H23N3O3/c1-18(2)14-10-19-8-5-9-24(19)17(27)20(14,23-16(19)26)11-21(18)15(25)12-6-3-4-7-13(12)22-21/h3-4,6-7,14,22H,5,8-11H2,1-2H3,(H,23,26)/t14-,19+,20+,21-/m0/s1

InChIKey: MWOFPQAPILIIPR-DJJZHVJBBC

Standard InChI=1S/C21H23N3O3/c1-18(2)14-10-19-8-5-9-24(19)17(27)20(14,23-16(19)26)11-21(18)15(25)12-6-3-4-7-13(12)22-21/h3-4,6-7,14,22H,5,8-11H2,1-2H3,(H,23,26)/t14-,19+,20+,21-/m0/s1

Standard InChIKey: MWOFPQAPILIIPR-DJJZHVJBSA-N

InfoChem

- Chemisches Zentralblatt
  - Digitised
  - Structure searchable
  - 98k unique names
  - 48k unique structures
  - Text searchable from FIZ Chemie

RSC Publishing

Microsoft External Research

- Organization within Microsoft Research that engages in strong **partnerships** with academia, industry and government to advance computer science, education, and research in fields that rely heavily upon advanced computing
- Initiatives that focus on the **research process** and its role in the innovation ecosystem, including support for open access, open tools, open technology, and interoperability
- Developers of **advanced technologies** and services to support every stage of the research process

RSC Publishing

Ontology Add-in for Word 2007

THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY

science commons

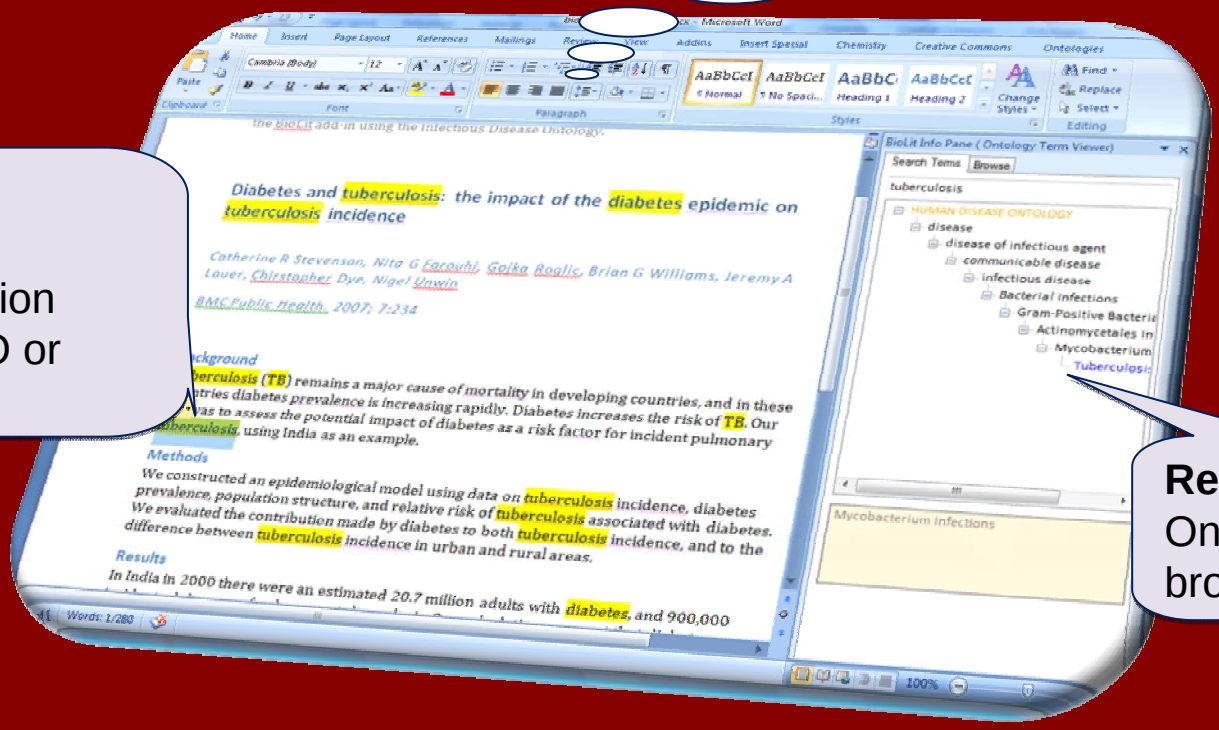
- John Wilbanks

University of California San Diego

- Phil Bourne
- Lynn Fink

Services: Ontology download web service

Intent: Term recognition & disambiguation based on OBO or OWL formats



Relationships: Ontology browser

Source code and binary:
http://research.microsoft.com/ontology/

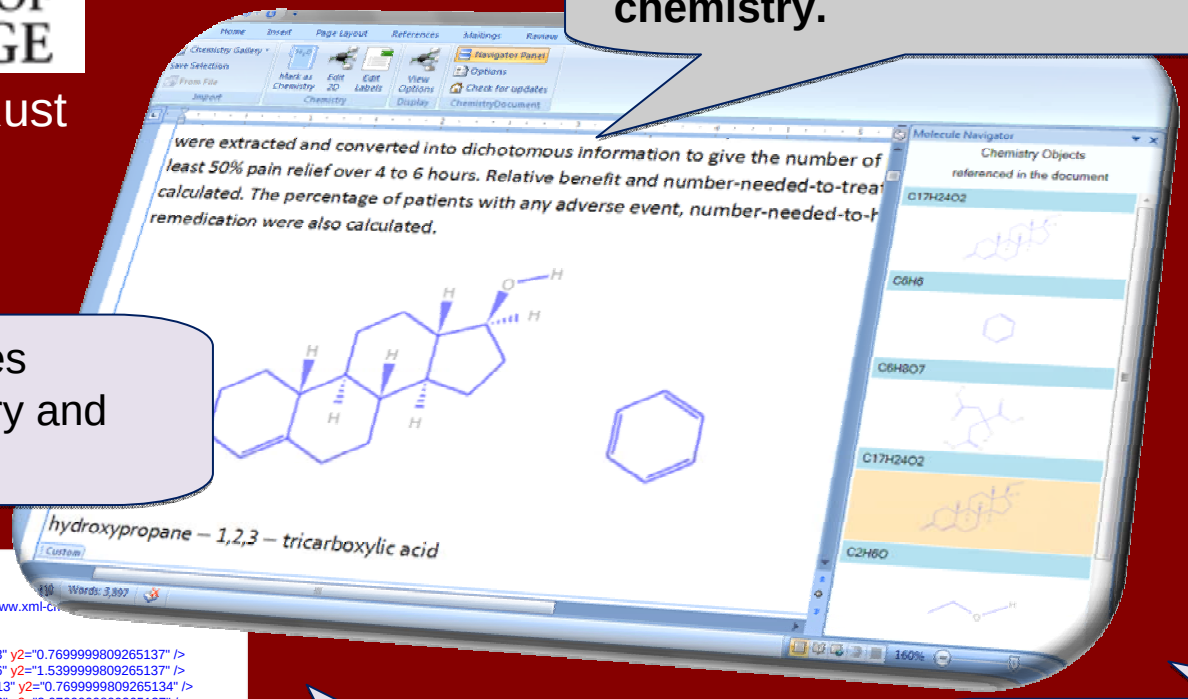
RSC Publishing

Chem4Word – Chemistry Drawing in Word

UNIVERSITY OF CAMBRIDGE

- Peter Murray-Rust
- Joe Townsend
- Jim Downing

Author and edit 1D and 2D chemistry.



Intent: Recognizes chemical dictionary and ontology terms

```
<?xml version="1.0" ?>
<cml version="3" convention="org-synth-report" xmlns="http://www.xml-cml
  <molecule id="m1">
    <atomArray>
      <atom id="a1" elementType="C" x2="-2.9149999618530273" y2="0.7699999809265137" />
      <atom id="a2" elementType="C" x2="-1.5813208400249916" y2="1.5399999809265137" />
      <atom id="a3" elementType="O" x2="-0.24764171819695613" y2="0.7699999809265134" />
      <atom id="a4" elementType="O" x2="-1.5813208400249912" y2="3.0799999809265137" />
      <atom id="a5" elementType="H" x2="-4.248679083681063" y2="1.5399999809265137" />
      <atom id="a6" elementType="H" x2="-2.9149999618530028" y2="-0.7700000190734864" />
      <atom id="a7" elementType="H" x2="-4.248679083681063" y2="-1.907348645691087E-8" />
      <atom id="a8" elementType="H" x2="1.0860374036310796" y2="1.5399999809265132" />
    </atomArray>
    <bondArray>
      <bond atomRefs2="a1 a2" order="1" />
      <bond atomRefs2="a2 a3" order="1" />
      <bond atomRefs2="a2 a4" order="2" />
      <bond
      <bon
      <bon
      <bon
    </bon
  </mole
</cml>
```

Data: Semantics stored in Chemistry Markup Language

Intelligence: Verifies validity of authored chemistry

Relationships: Navigate and link referenced chemistry

RSC Publishing

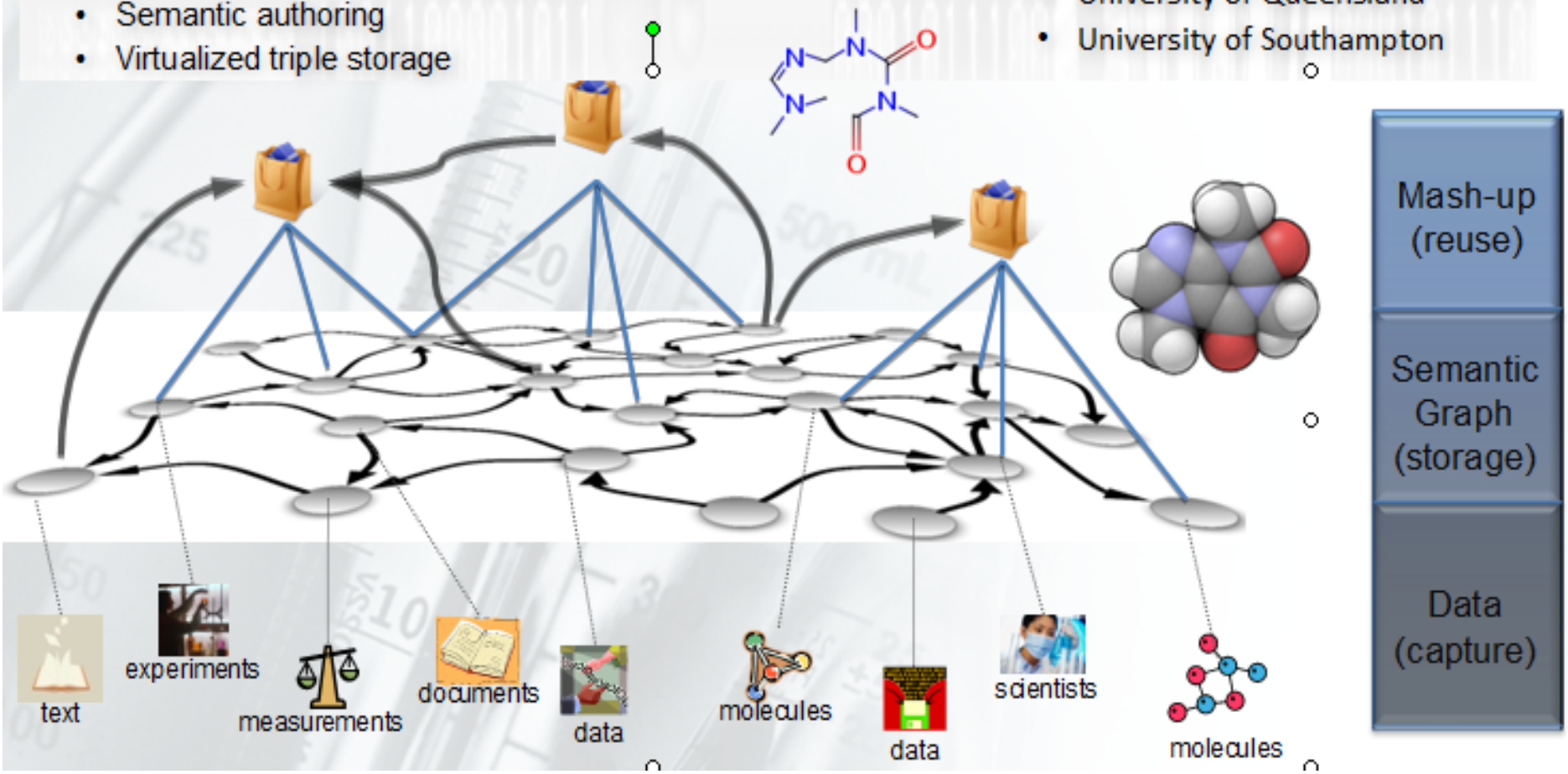
Available soon:
http://research.microsoft.com/chem4word/

OreChem – the Chemical Semantic Web (?)

- Large collaboration project focusing on interoperability
- At-source capture of chemistry data
- Chemical structure search
- Compound object authoring
- Retrospective harvesting of chemistry data
- Reuse through common ORE data model
- Semantic authoring
- Virtualized triple storage

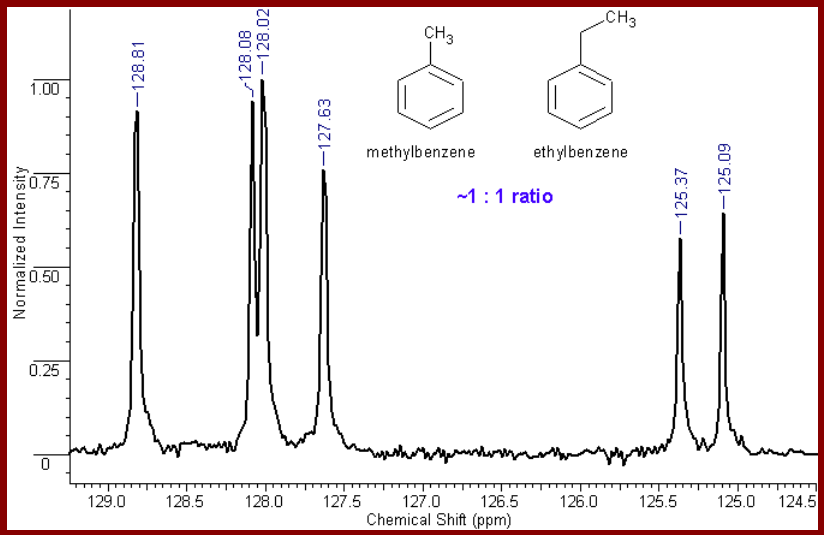
Participating Institutions:

- University of Cambridge
- Cornell University
- Indiana University
- Penn State University
- University of Queensland
- University of Southampton



Extraction of NMR data

- PDF
- Extraction of lines, polylines, text
- CML spectrum



- Mark Borkum, Southampton

RSC Publishing

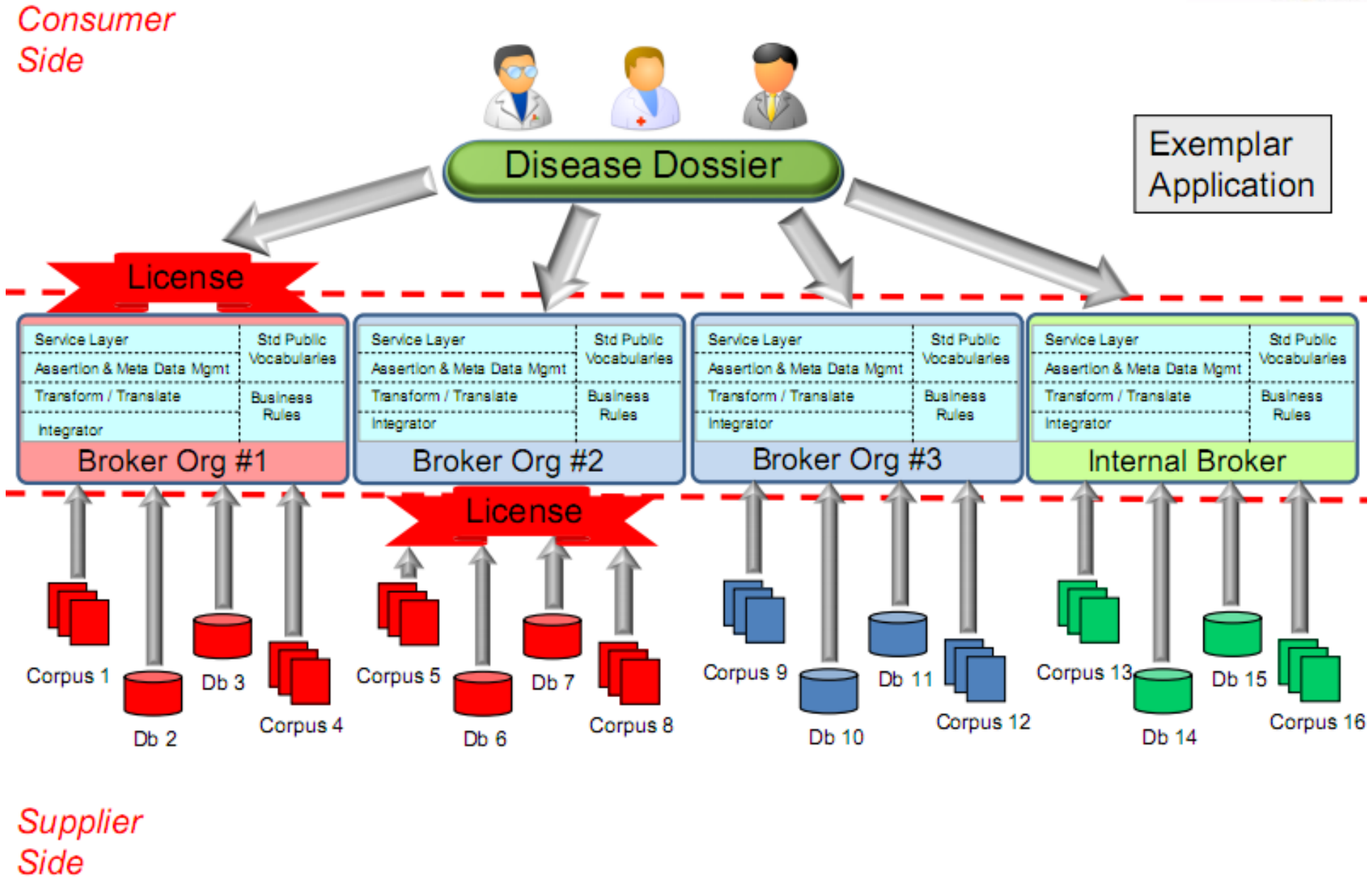
Semantic Enrichment of the Scientific Literature (SESL)

- Pistoia-funded
- EBI
- Elsevier, NPG, OUP, RSC

- Oct 2009 – Oct 2010

RSC Publishing

A Production Service ...



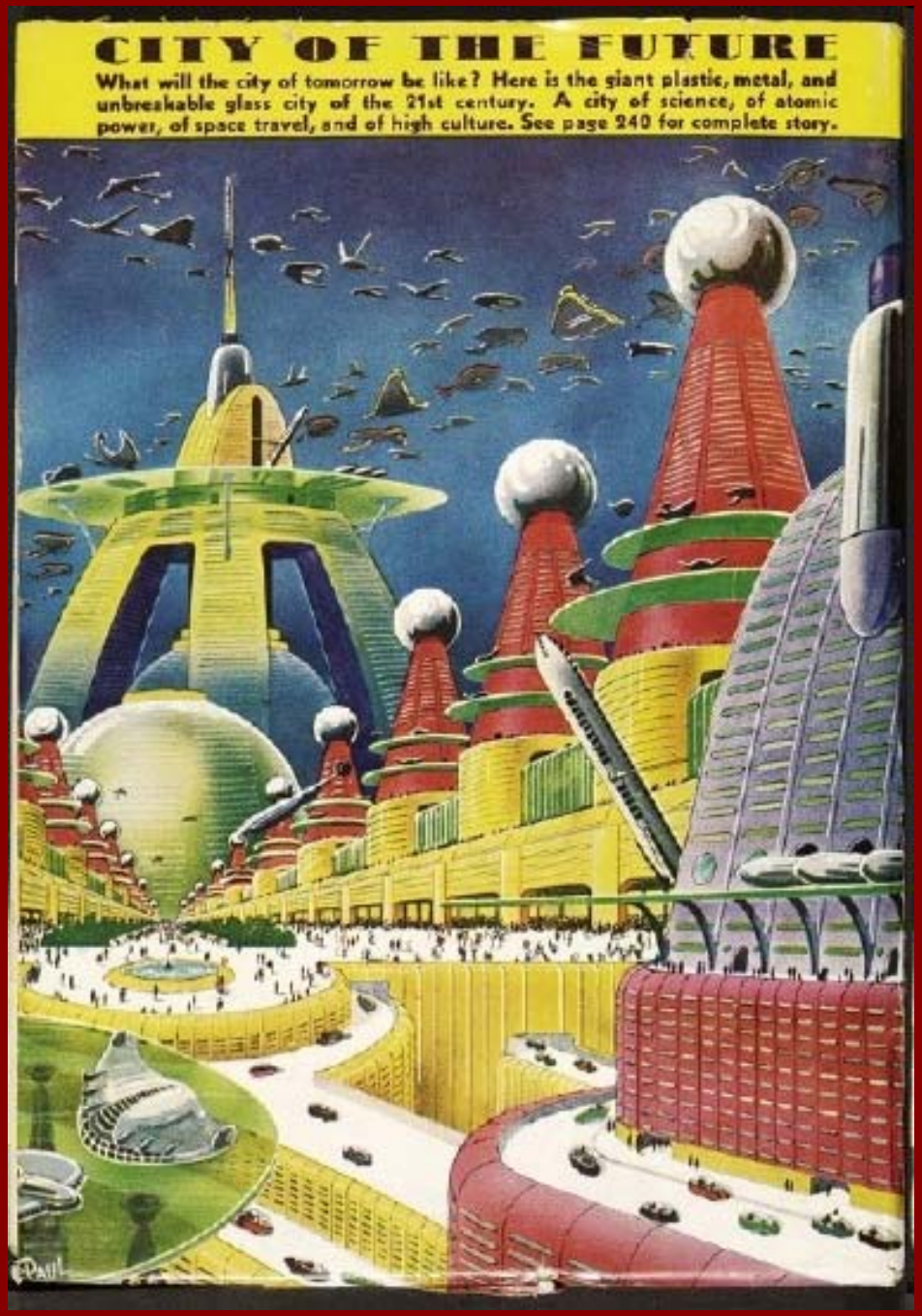
SESL deliverables

- Pilot to deliver target-disease assertions
- Publication of data, application and web service standards

- So: to deliver standards for semantic delivery

RSC Publishing

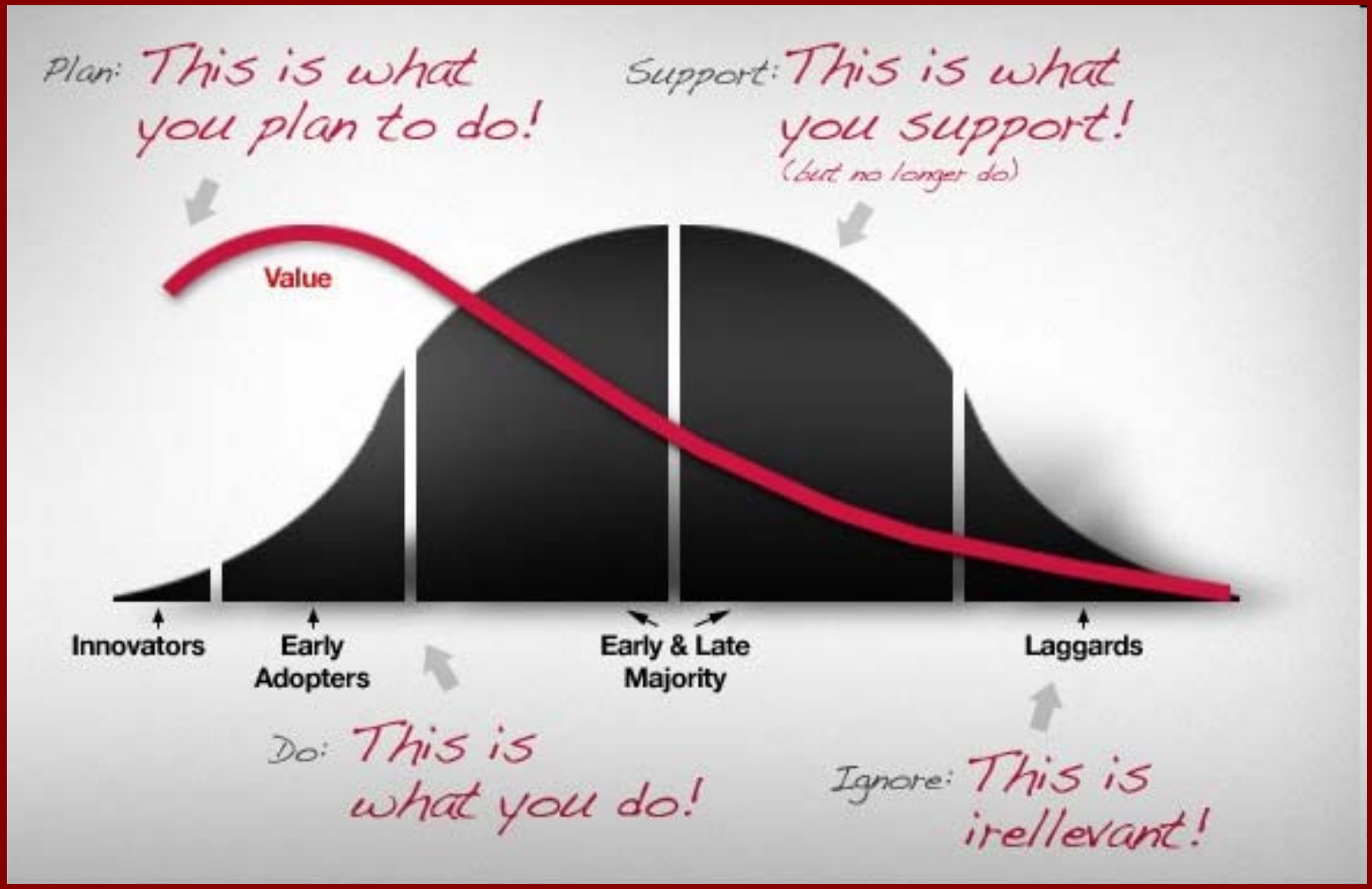
RSC Publishing



Are we there yet?

- InChI Trust
  - Compound standards
  - Reaction InChIs
  - Resolver Protocol
- Pistoia/EBI
  - Semantic standards for web services
- Microsoft/Academia
  - oreChem
  - Chem4Word
- Semantic markup by publishers

RSC Publishing



RSC Publishing

http://www.baekdal.com

Images

http://www.pantopicon.be

http://www.leirdal.net

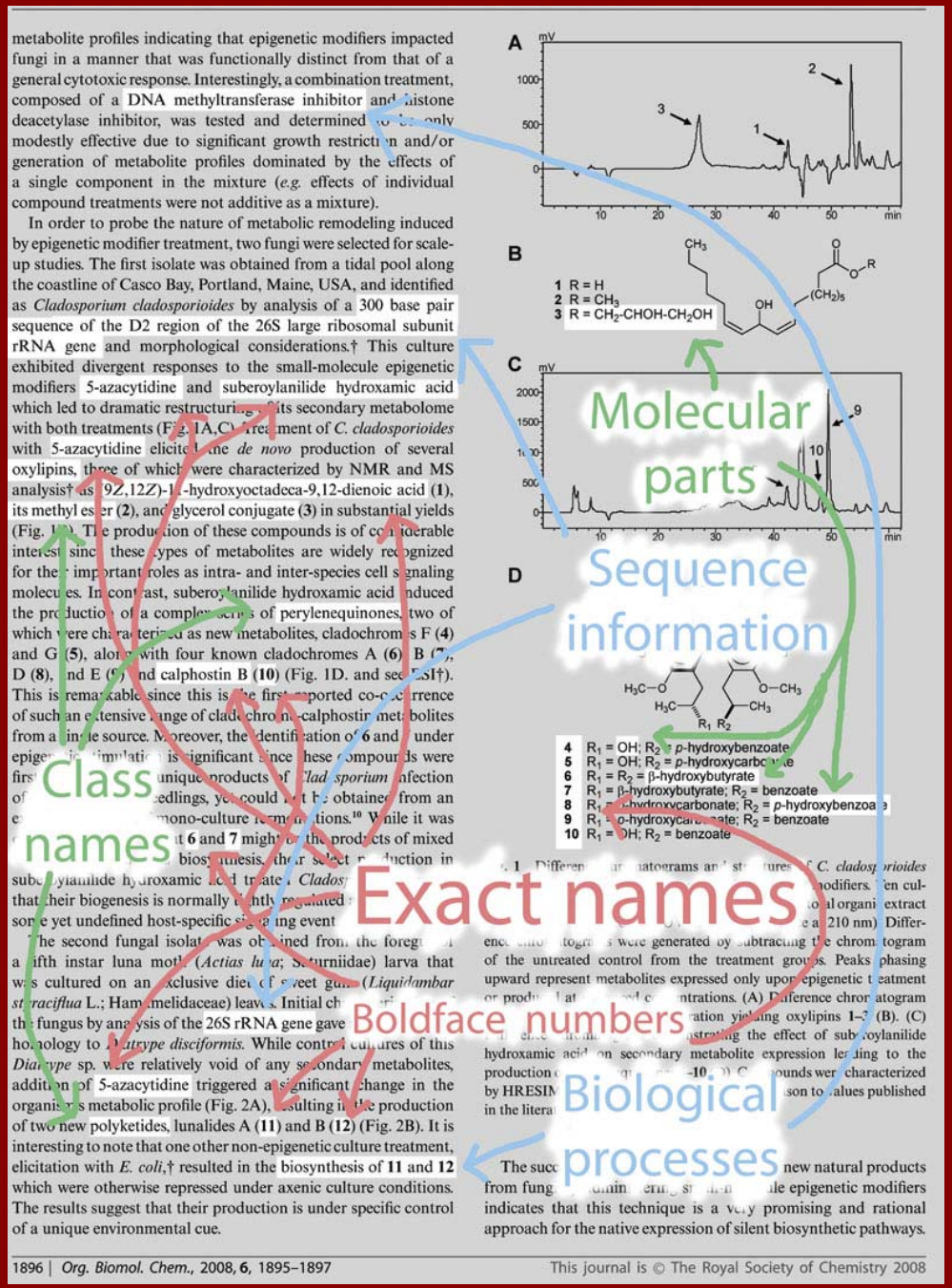
http://www.baekdal.com

Microsoft Research

Pistoia Alliance

Steve Arnold

Richard Kidd
kiddr@rsc.org



RSC Publishing