Semantics *vs.* statistics in chemistry

Colin Batchelor

Royal Society of Chemistry

2019-05-01

… or… Rationalism *vs.* empiricism in chemical data

Colin Batchelor

Royal Society of Chemistry

2019-05-01

Overview

- What are these two approaches?
- Five questions from 2011
- How can you tell whether something works?
- What have we done at the Royal Society of Chemistry?
- Some tentative conclusions

Two approaches

Language as a formal, logical system (Montague, Steedman, Wittgenstein).

Language as use (Firth, Spärck Jones, Masterman, also Wittgenstein).

“Words stand only for themselves” (Karen Spärck Jones)

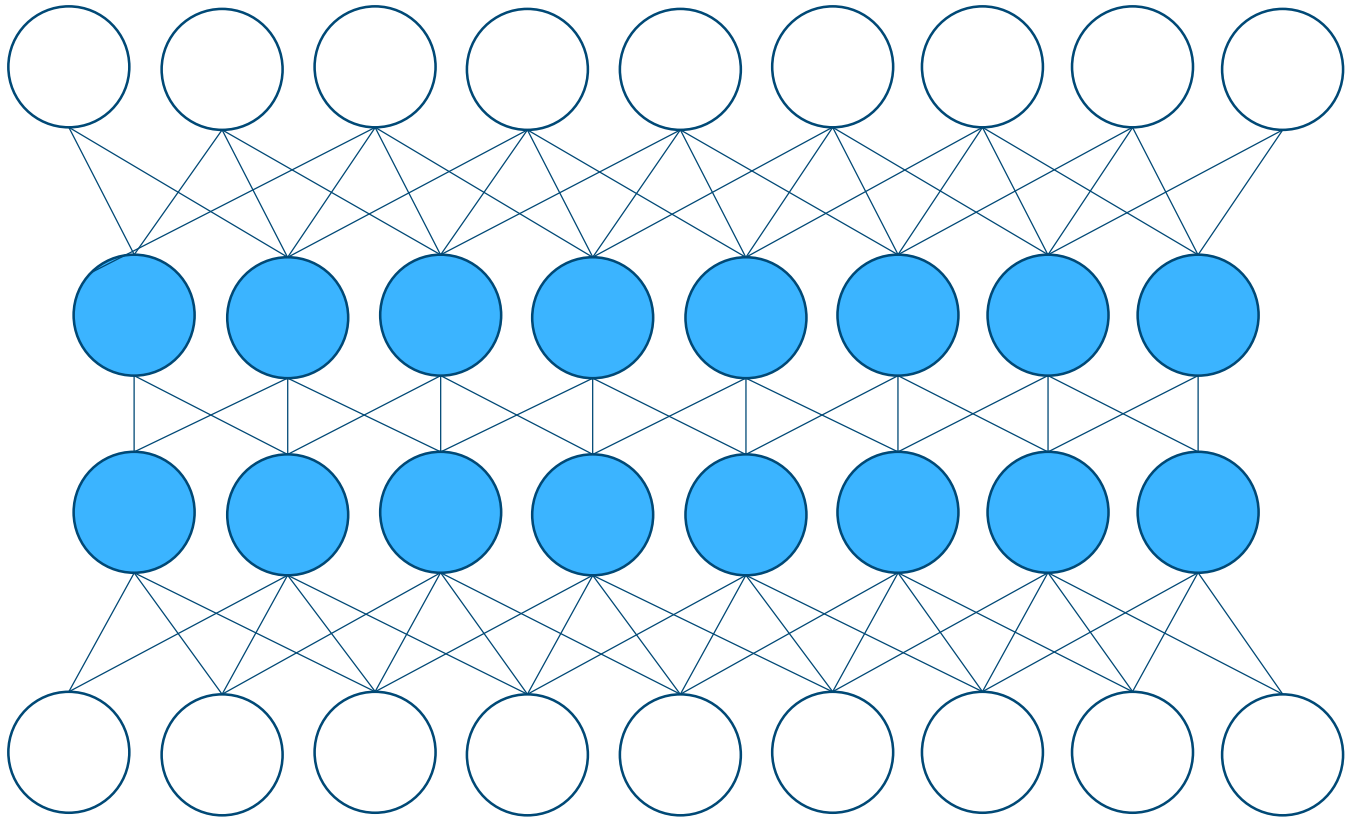
“You shall know a word by the company it keeps” (J. R. Firth)

Semantics in pictures

```
(p / picture-01 :polarity - :ARG0 (p2 / picture :ARG1-of (d / draw-01 :ARG0 (i / i))) :ARG1 (h / hat))

(p / picture-01 :ARG0 (i / it) :ARG1 (b2 / boa :mod (c / constrictor :ARG0-of (d / digest-01 :ARG1 (e / elephant))))

(c / contrast-01 :ARG2 (a2 / and :op1 (d3 / draw-01 :ARG0 (i / i) :ARG1 (p2 / picture :mod (a / another)) :ARG1-of (c3 / cause-01 :ARG0 (p3 / possible-01 :polarity - :ARG1 (u / understand-01 :ARG0 (g / grown-up) :ARG1 (i2 / it))))) :op2 (d / draw-01 :ARG0 I :ARG1 (i3 / inside :part-of (b2 / boa :mod (c4 / constrictor))) :purpose (p / possible-01 :ARG1 (s / see-01 :ARG0 g :ARG1 (i4 / it) :ARG1-of (c2 / clear-06))))))

(n / need-01 :ARG0 (t / they) :ARG1 (e / explain-01) :time (a / always))
```

https://amr.isi.edu/download/amr-bank-struct-v1.6-dev.txt

Deep learning



Input vector:
made with embeddings

Hidden layers hopefully “learn” something transferable.
“Feature engineering” happens here through training.

Output vector

ROYAL SOCIETY OF CHEMISTRY

Classification

Feature-driven with a metric – potentially hierarchical

- Clustering

Feature-driven, no metric – necessarily hierarchical

- Formal concept analysis

Aristotelian (an *x* is a *y* that *z*) – necessarily hierarchical

- Ontologies

Five questions (2011)

1. What can you achieve through “letting words (2019: or images) stand for themselves”?
2. What do you hope to achieve through semantic mark-up?
3. What kinds of non-textual (2019: or non-pictorial) internal data do you have access to?
4. What can you do with your non-textual (2019: or non-pictorial) data?
5. What external data do you have access to?

How can you tell whether something works?

Intrinsic evaluation *vs*. extrinsic evaluation.

https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-291.html (Galliers and Spärck Jones)

Intrinsic criteria relate to a system's objective. Examples: *F* score, BLEU score.

Extrinsic criteria relate to a system's function embedded in a given context. Examples: usage, usability, downloads, number of submissions, ease of integration, speed.

What have we done?

- Project Prospect (2006–2010)
- Subject categories for *RSC Advances* and *ChemComm* (2011 to present)
- Open PHACTS (2013)

Project Prospect

Idea: semantic annotation of chemistry articles.

Using: manual annotation, text-mining.

Outputs: enhanced HTML, RSS feed, open-source ontologies

Evaluation: partly on usage, partly on “impact”.

Project Prospect (and CVSP): The case for chemical markup

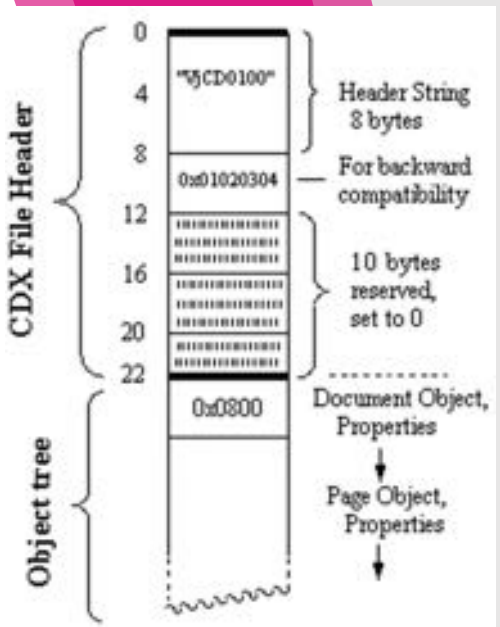
There are very well established and understood methods for indexing and searching chemical structures.

Chemical structures are the chief route into the literature for many chemists.

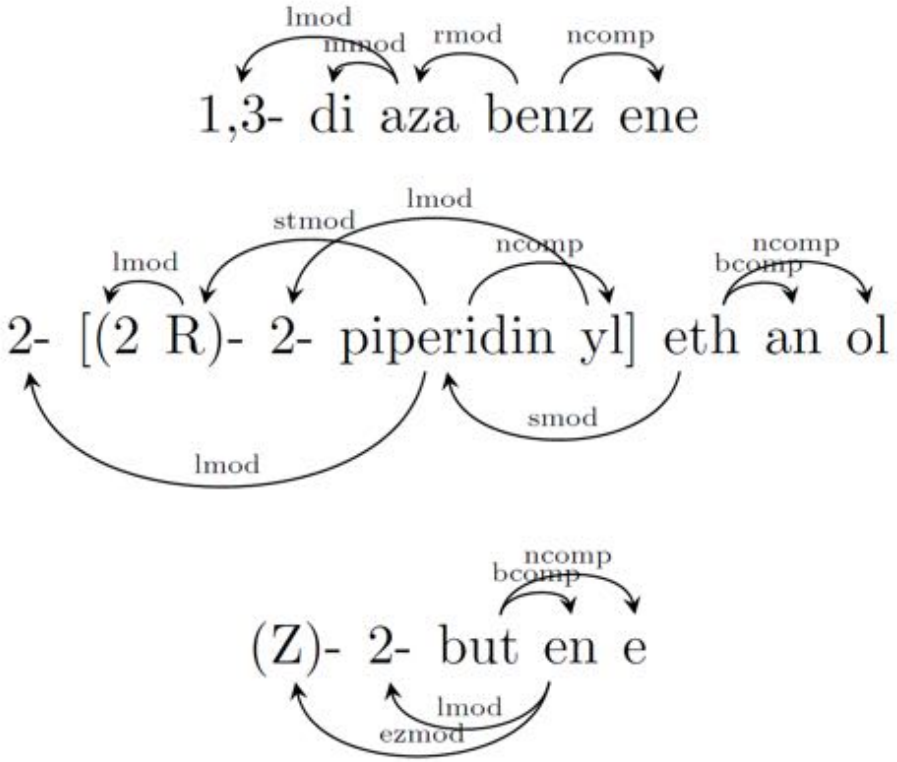
We can extract chemical structures at least semi-automatically from:

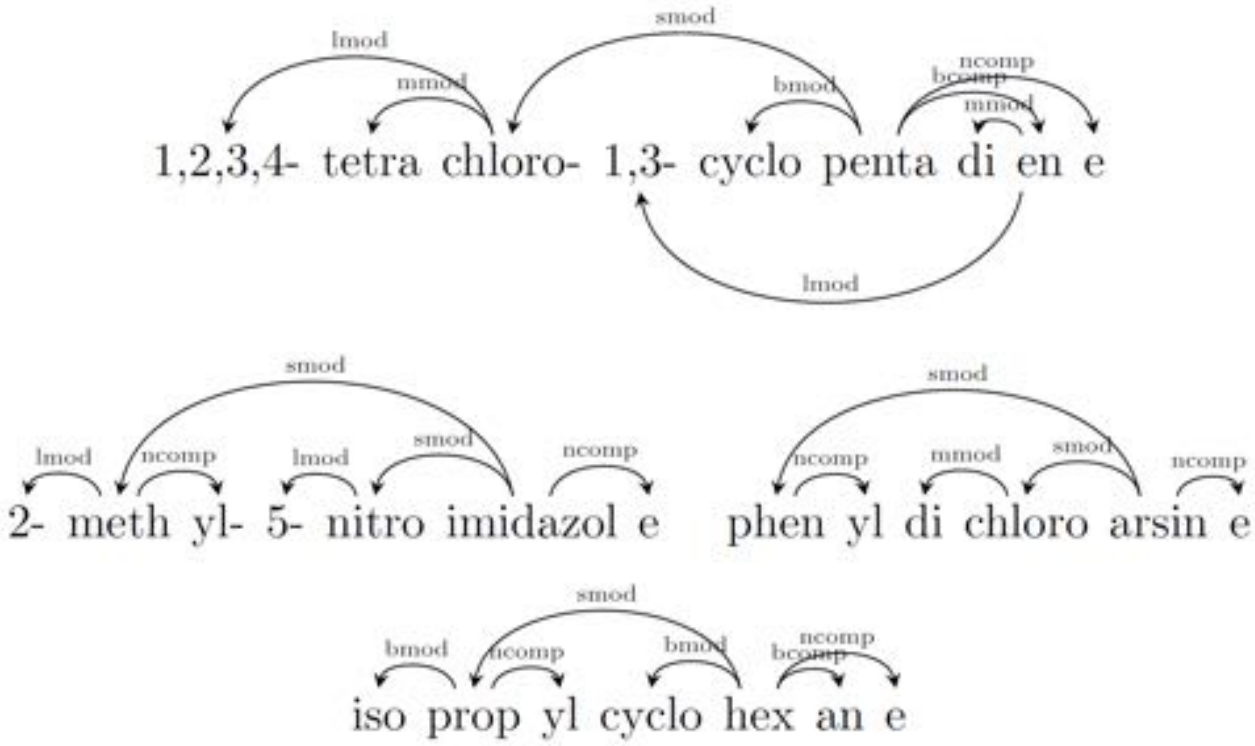
- Molecule names (https://opsin.ch.cam.ac.uk/)
- Author-supplied graphics(http://www.cambridgesoft.com/services/documentation/sdk/chemdraw/cdx/IntroCDX.htm)

But what can hand-written ontologies get us beyond chemical structures?



Aside: dependency grammar analysis of chemical names





ROYAL SOCIETY OF CHEMISTRY

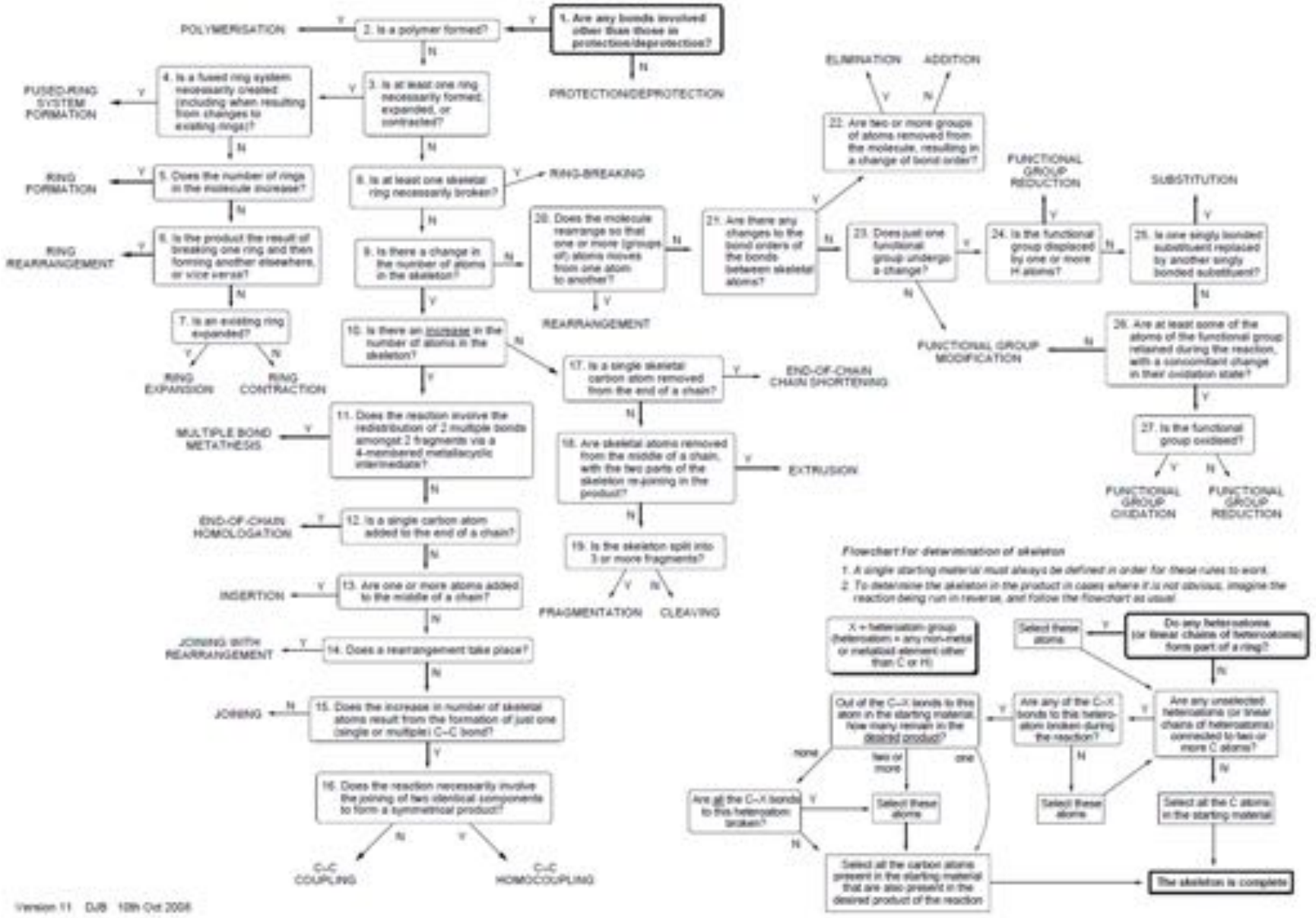
RXNO

Sandmeyer Gattermann Koerner–Contardi Heck Beckmann Cope Baeyer–Villiger Mannich Mitsunobu Stille Reformatsky Clemmensen Ullmann Birch Claisen Knoevenagel Friedel–Crafts Oppenauer Prins Wolff Bischler–Napieralski Curtius Vilsmeier–Haack Horner–Wadsworth–Emmons Horner Ritter Pictet–Spengler Arbuzov Mukaiyama Skraup Arndt–Eistert Fischer Dieckmann Staudinger Dimroth Edman Reimer–Tiemann Wurtz Baylis–Hillman Darzens Peterson Goldberg Paternò–Büchi Barbier Henry Barton Negishi Meerwein–Ponndorf–Verley Williamson Wittig Diels–Alder Ramberg–Backlund Fischer–Hepp Glaser Eglinton Cadiot–Chodkiewicz

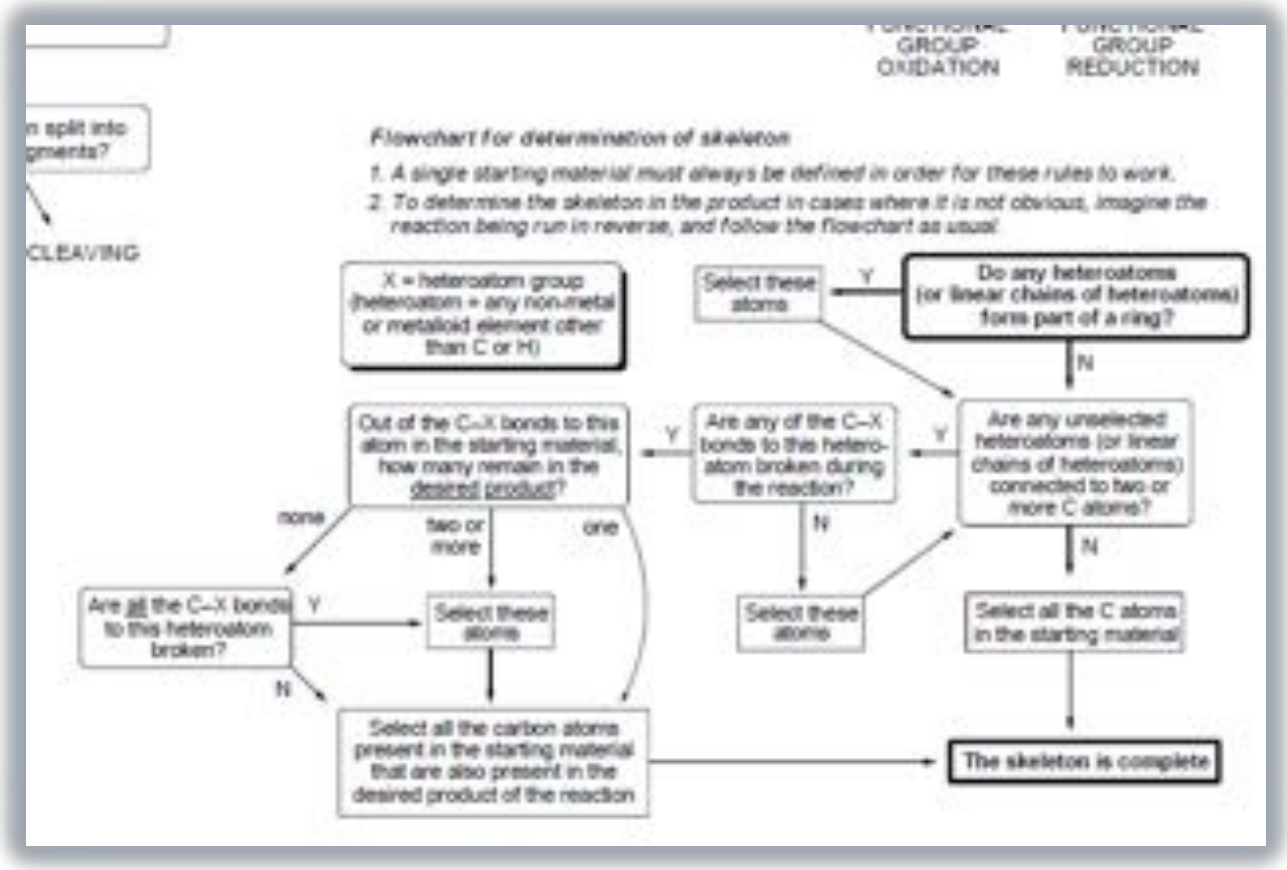
Methodology

- Get two annotators with synthetic organic chemistry PhDs together.
- Classify batches of reactions independently.
- Assess agreement.
- Update guidelines.

https://github.com/rsc-ontologies/rxno


1. Are any bonds involved other than those in protection/deprotection?
PROTECTION/DEPROTECTION
2. Is a polymer formed?
POLYMERISATION
3. Is at least one ring necessarily formed, expanded, or contracted?
4. Is a fused ring system necessarily created (including when resulting from changes to existing rings)?
FUSED-RING SYSTEM FORMATION
5. Does the number of rings in the molecule increase?
RING FORMATION
6. Is the product the result of breaking one ring and then forming another elsewhere, or vice versa?
RING REARRANGEMENT
7. Is an existing ring expanded?
RING EXPANSION
RING CONTRACTION
8. Is at least one skeletal ring necessarily broken?
RING-BREAKING
9. Is there a change in the number of atoms in the skeleton?
10. Is there an increase in the number of atoms in the skeleton?
11. Does the reaction involve the redistribution of 2 multiple bonds amongst 2 fragments via a 4-membered metallacyclic intermediate?
MULTIPLE BOND METATHESIS
12. Is a single carbon atom added to the end of a chain?
END-OF-CHAIN HOMOLOGATION
13. Are one or more atoms added to the middle of a chain?
INSERTION
14. Does a rearrangement take place?
JOINING WITH REARRANGEMENT
15. Does the increase in number of skeletal atoms result from the formation of just one (single or multiple) C–C bond?
JOINING
16. Does the reaction necessarily involve the joining of two identical components to form a symmetrical product?
C–C COUPLING
C=C HOMOCOUPLING
17. Is a single skeletal carbon atom removed from the end of a chain?
END-OF-CHAIN CHAIN SHORTENING
18. Are skeletal atoms removed from the middle of a chain, with the two parts of the skeleton re-joining in the product?
EXTRUSION
19. Is the skeleton split into 3 or more fragments?
FRAGMENTATION
CLEAVING
20. Does the molecule rearrange so that one or more (groups of) atoms moves from one atom to another?
REARRANGEMENT
21. Are there any changes to the bond orders of the bonds between skeletal atoms?
22. Are two or more groups of atoms removed from the molecule, resulting in a change of bond order?
ELIMINATION
ADDITION
23. Does just one functional group undergo a change?
FUNCTIONAL GROUP MODIFICATION
24. Is the functional group displaced by one or more H atoms?
FUNCTIONAL GROUP REDUCTION
25. Is one singly bonded substituent replaced by another singly bonded substituent?
SUBSTITUTION
26. Are at least some of the atoms of the functional group retained during the reaction, with a concomitant change in their oxidation state?
27. Is the functional group oxidised?
FUNCTIONAL GROUP OXIDATION
FUNCTIONAL GROUP REDUCTION
Version 11 DJB 18th Oct 2006
Flowchart for determination of skeleton
1. A single starting material must always be defined in order for these rules to work.
2. To determine the skeleton in the product in cases where it is not obvious, imagine the reaction being run in reverse, and follow the flowchart as usual.
X = heteroatom group (heteroatom = any non-metal or metalloid element other than C or H)
Out of the C–X bonds to this atom in the starting material, how many remain in the desired product?
none
two or more
one
Are all the C–X bonds to this heteroatom broken?
Select these atoms
Select all the carbon atoms present in the starting material that are also present in the desired product of the reaction
Select these atoms
Are any of the C–X bonds to this hetero-atom broken during the reaction?
Select these atoms
Do any heteroatoms (or linear chains of heteroatoms) form part of a ring?
Are any unselected heteroatoms (or linear chains of heteroatoms) connected to two or more C atoms?
Select all the C atoms in the starting material
The skeleton is complete

Defining the skeleton

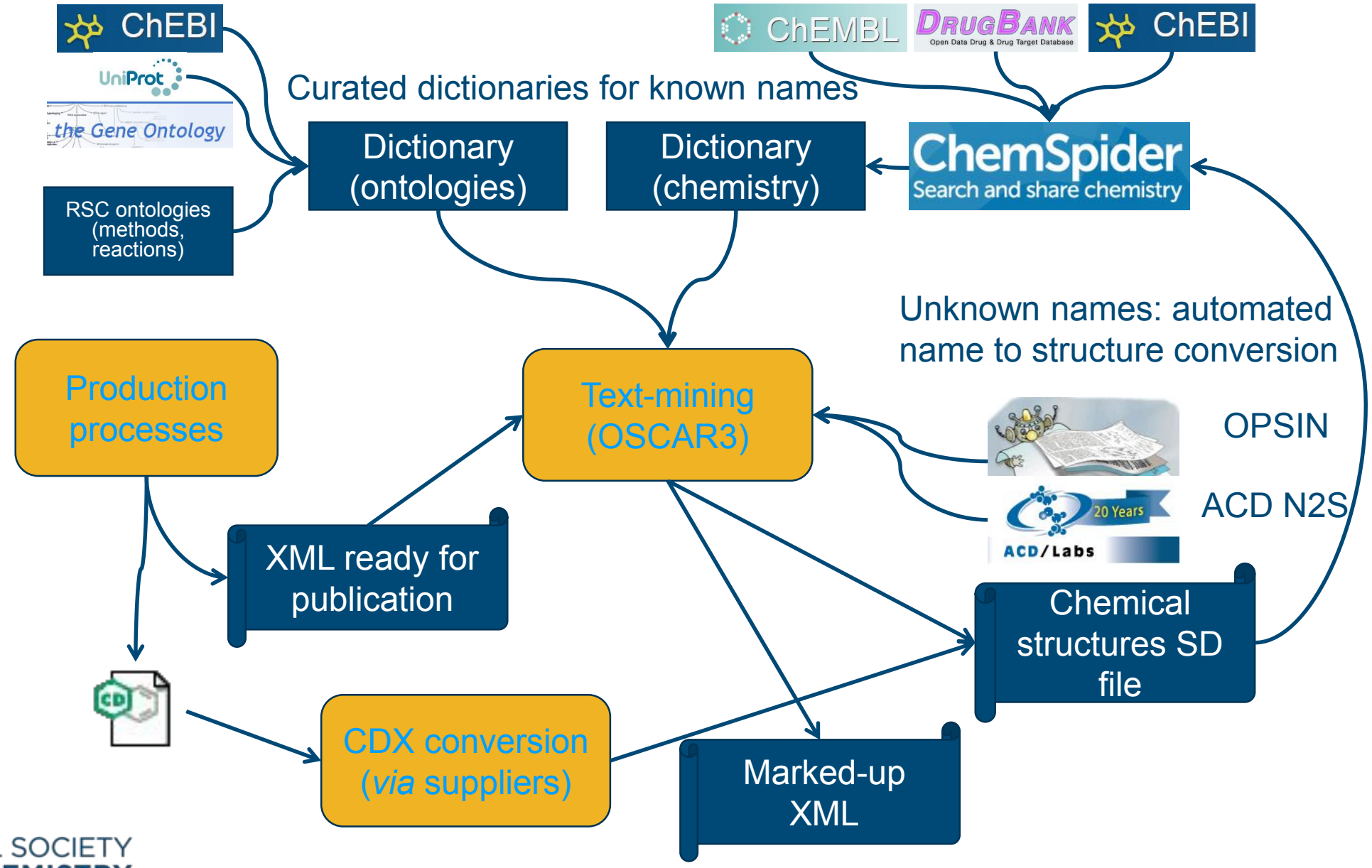


Limits of reaction classification

Much of RXNO is classified by hand.

Automatic reasoners cannot cope with cyclic graph systems.

Example: we can’t just define a cyclization as a reaction where a cyclic compound is formed. The Friedel–Crafts acylation produces a cyclic compound but is not a cyclization!



ROYAL SOCIETY OF CHEMISTRY

Subject categories

Idea: divide megajournal (*RSC Advances*) up into browseable chunks

Using: bag-of-words, bag-of-cited-journals

Output: contents pages

Evaluation: partly on usage, partly *via* governance.

Development: take suggestions from editors – determine whether the category is suitable – find “seed” articles. Iterate till satisfied.

Open PHACTS

Idea: use shared identifiers to describe disparate data of interest in drug discovery

Using: CHEMINF

Output: HUGE RDF files for input into industry-internal systems, OPS Explorer.

Evaluation: does it work at all? Then, take up by industry (difficult to measure and largely anecdotal.)

Characteristics of RDF

All predicates are binary.

There are no variables.

Everything is divided necessarily into predicates and individuals (and classes using RDFS/OWL).

Compare first-order logic

Quantifiers, arbitrary arity of predicates. There are no classes, only predicates.

Unary predicate: $\exists x,\ \exists y,\ \mathrm{cat}(x),\ \mathrm{dog}(y)$

Binary predicate, “Some cat chases some dog”: $\exists x,\ \exists y,\ \mathrm{chases}(x,y)$

Ternary predicate, “Some cat chases some dog at some time”: $\exists x,\ \exists y,\ \exists t,\ \mathrm{chases}(x,y,t)$

Quaternary predicate, “Some cat chases some dog some time somewhere”: $\exists x,\ \exists y,\ \exists t,\ \exists l,\ \mathrm{chases}(x,y,t,l)$

5-ary predicate, “Some cat chases some dog some time somewhere at some temperature”: $\exists x,\ \exists y,\ \exists t,\ \exists l,\ \exists T,\ \mathrm{chases}(x,y,t,l,T)$

6-ary predicate, “Some cat chases some dog some time somewhere at some temperature and Julius Caesar was killed by Brutus on the Appian Way”: $\exists x,\ \exists y,\ \exists t,\ \exists l,\ \exists T,\ \exists S,\ \mathrm{chases}(x,y,t,l,T,S)$

Characteristics of OWL

- Sophisticated handling of classes
- (Relatively) less-sophisticated handling of predicates (= roles, = relations)
- Distinction between TBox (what holds in general) and ABox (what holds in particular)

- Suggests that RDF subjects and objects, not predicates, should do the heavy lifting.

Semantic representations: options

Dublin Core-style semantics: hard work done by predicates

```
http://dx.doi.org/10.1039/b311404j dc:creator "Christine A. Williams"
```

***Ad hoc* semantics**: no clear division of labour between predicates and subjects/objects

mosquitoes **transmit** malaria

Event semantics: minimize work done by predicates

Event semantics in FOL: Davidson (1967)

- “Jones did it slowly, deliberately, in the bathroom, with a knife, at midnight. What he did was butter a piece of toast.”

As n-ary predicate:

- buttering(Jones, piece of toast, knife, midnight, bathroom)

Davidson introduces an event variable and separate predicates for most of the details:

- exists(e) [buttering(Jones,piece of toast,e) & with(knife,e) & at(e,midnight) & in(e,bathroom)]

Event semantics in RDF: the neo-Davidsonian approach

Jones did not butter all toast with all knives in all bathrooms using all butter.

$$\exists e\ [\text{instance\_of}(e,\text{buttering}) \ \&\ \text{agent}(e,\text{Jones}) \ \&\ \text{patient}(e,x) \ \&\ \text{instance\_of}(x,\text{piece of toast}) \ \&\ \text{instrument}(e,y) \ \&\ \text{instance\_of}(y,\text{knife}) \ \&\ \text{location}(e,z) \ \&\ \text{instance\_of}(z,\text{bathroom})]$$

RDF

```
e     instance_of      House:Buttering
e     has_agent        Jones
e     has_patient      x
x     instance_of      House:PieceOfToast
e     has_instrument   y
y     instance_of      House:Knife
e     has_location     z
z     instance_of      House:Bathroom
```

Worked example

Ad hoc approach:

```
http://www.chemspider.com/236 has_melting_point "279 K"
```

Brief, simple, but no provenance information.

Event semantics approach:

```
e instance_of CHMO:meltingpointmeasurement
e has_participant s
s has_granular_part http://www.chemspider.com/236
e has_output b
b instance_of CHMO:meltingpoint
b has_value "279 K"
```

More complicated, but provenance obligatory (is *e* a measurement in the lab, a calculation, or scraped off a website?), and clearer relation to workflows.

Physicochemical properties

log P log D (at pH 5.5, at pH 7.4) bioconcentration factor K_{OC} (at pH 5.5, at pH 7.4) index of refraction polar surface area molar refractivity molar volume polarizability

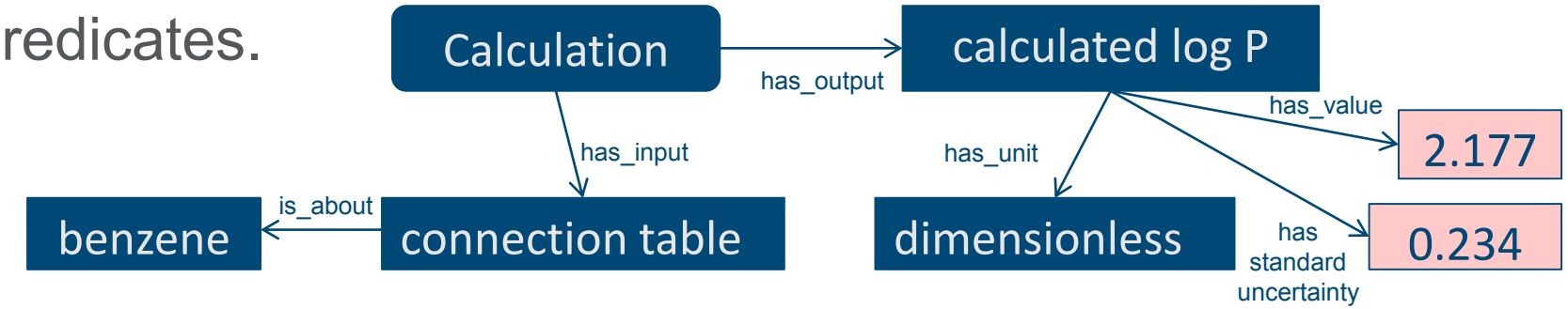
surface tension density at STP flash point at 1 atm boiling point at 1 atm enthalpy of vaporization at STP vapour pressure at STP

Open PHACTS RDF schema

Two dozen calculated properties for >10^6 molecules.

- CHEMINF ontology for cheminformatics.
- QUDT for units and numeric values
- ChemSpider IDs for molecules.

Event-based (rather than a predicate-based) schema: avoids an explosion in the number of predicates.

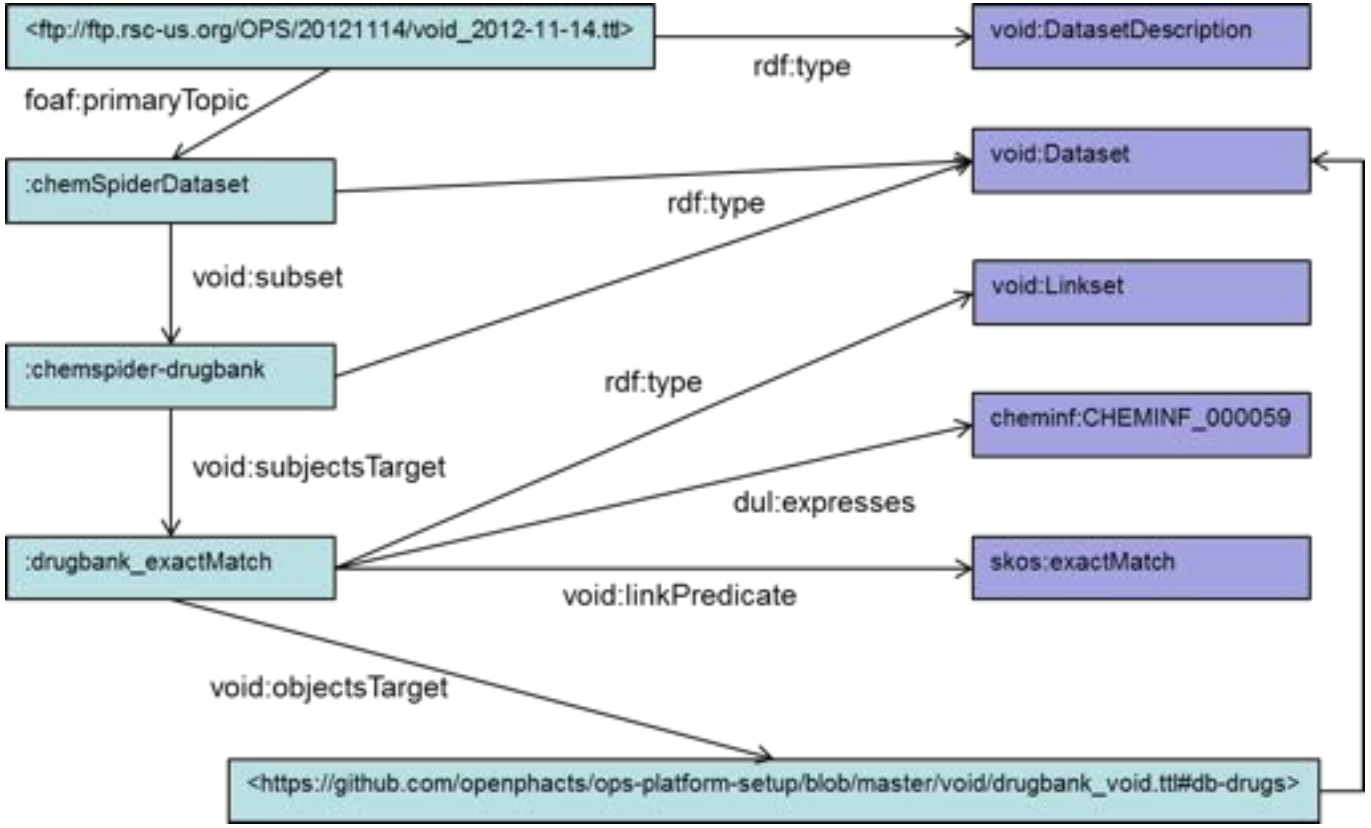


Vocabulary of Interlinked Datasets

Metadata describing the RDF

- Can be used to build a directory of the RDF available
- Find what's there without having to download all of it first
- Describes how Datasets are linked using Linksets
- `skos:exactMatch` (Simple Knowledge Organisation System)
- `skos:relatedMatch`
- `dul:expresses` (DOLCE+DnS Ultralite)
- Recommendations for how to create the VoID has been specified by Manchester here: http://www.openphacts.org/specs/2013/WD-datadesc-20130912/

RDF/VoID – DrugBank example



ROYAL SOCIETY OF CHEMISTRY

Synonyms and identifiers: On OWL annotation properties 1

Most relations in scientific ontologies (parthood, participation, mutual manifestation) express **ontological dependence**.

It is impossible, for example, for something to be a **benzene molecule** without having as part a **benzene ring**.

The **hangover of the conference participant** cannot exist without the existence of the **conference participant**.

Synonyms and identifiers: On OWL annotation properties 2

Names and identifiers are not like this.

It is neither necessary to the word “benzene” nor to the benzene molecule that the other exist; *cf.* wyvern, polywater.

(It is necessary to the successful act of reference that both exist, but we are not representing speech acts of that kind here.)

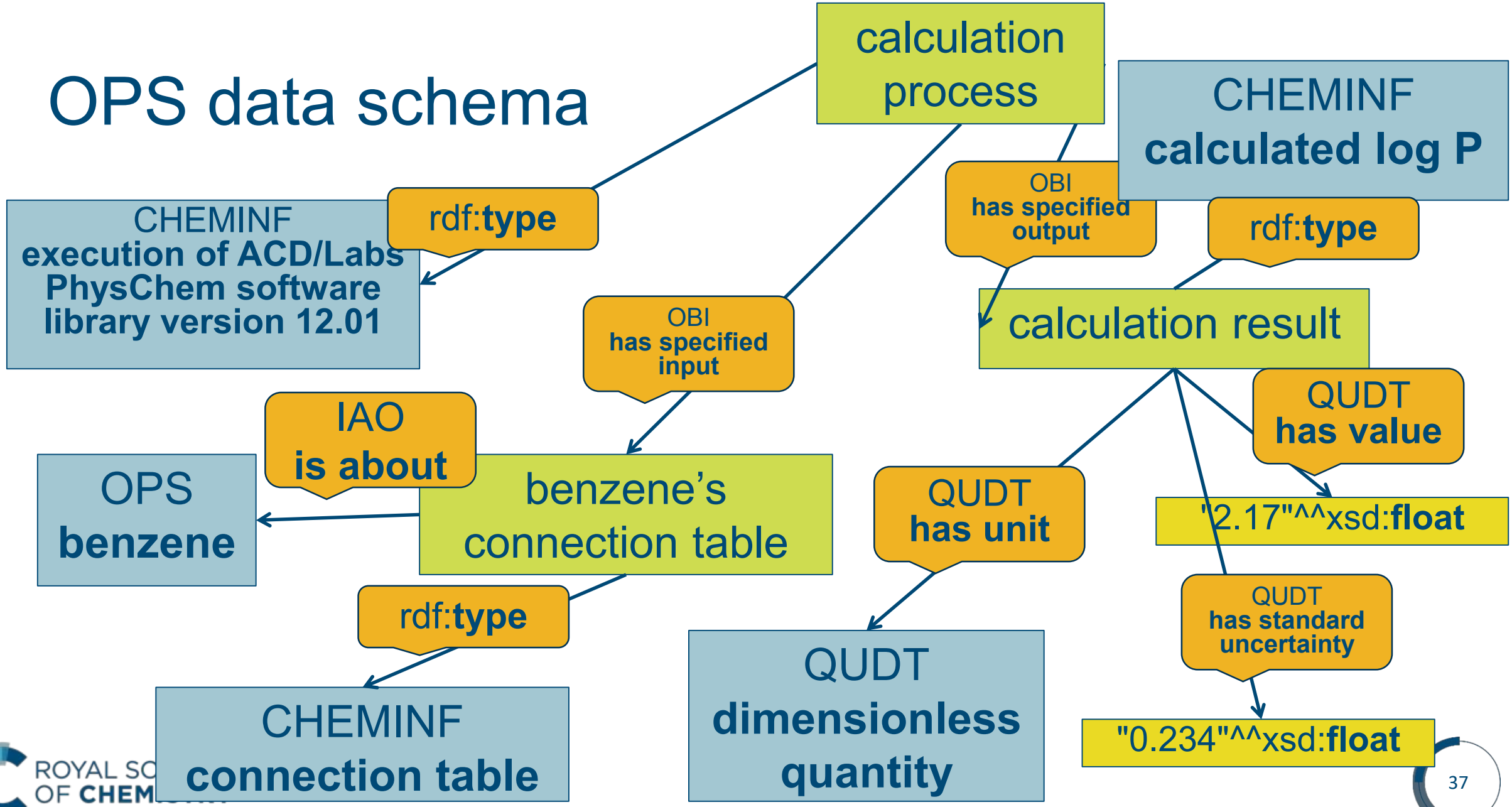
Synonyms and identifiers: On OWL annotation properties 3

And yet there are things we want to say in OWL about identifiers; an InChI is necessarily interpretable by the InChI algorithm, a SMILES is necessarily a chemical identifier and so on.

We use **OWL annotation properties**, which are not used for inference by reasoners.

Validated ChemSpider synonyms **Unvalidated ChemSpider synonyms**
Validated database identifiers **Unvalidated database identifiers** **InChI**
InChIKey **SMILES** **Preferred ChemSpider name**

OPS data schema



Conclusions

“Semantics” can mean all manner of things and pluralism about semantics can yield practical insights.

All sorts of data come in non-textual form: tables in articles, microarray data, experimental protocols, crystallographic data, chemical structures, and formal semantics can help.

You need a fair amount of infrastructure for this.

BUT initial success of applications with the community (librarians, board members) may not be borne out by actual usage.

Thank you.
Any questions?

ROYAL SOCIETY OF CHEMISTRY