All aboard the semantic bandwagon…
… hold on a moment

Colin Batchelor
20101202

What do people mean by semantics?

Semantic web

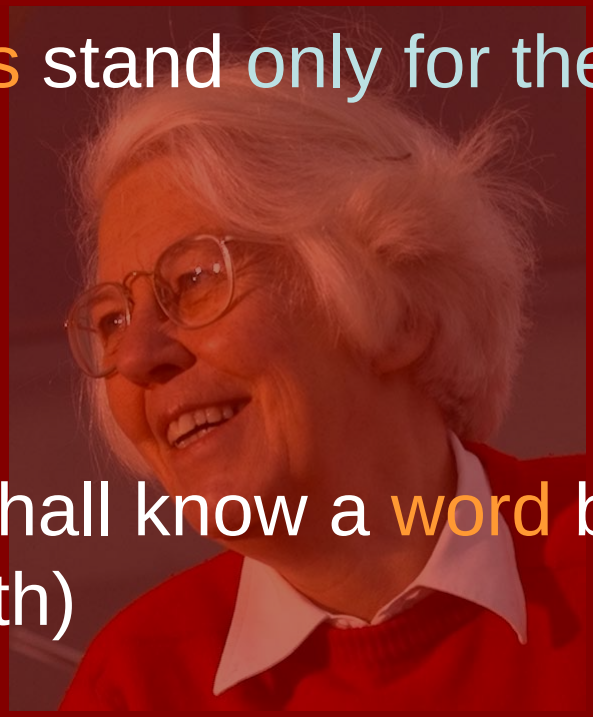
- Specifying how web service shall speak unto web service

Natural-language semantics

- Lexical semantics
- Compositional semantics

What do search engines do?

Words stand only for themselves (Spärck-Jones)



You shall know a word by the company it keeps (Firth)

Search engines don’t use semantic methods.

Statistics *versus* semantics

A counterblast:

Alon Halevy, Peter Norvig and Fernando Pereira, "The Unreasonable Effectiveness of Data", *Intelligent Systems*, 2009, **24**, 8.

But STM publishers don't just publish text.

Summary of preferred values of parameters for uptake on ice surfaces

No.	Species	α/γ	$\pm\Delta\alpha_s$	K_{linC} cm	N_{max} molecule cm^{-2}	$\Delta(E_{ads}/R)$ $\Delta\ln N_{max}$	Te
	O	$7 \times 10^{-6} + 2.6 \times 10^{-24}$ exp(1370/T) [O_2]	± 0.3				110-1
	O_3	$<1 \times 10^{-5}$	0.7($\Delta\log\gamma$)				220-2
	OH						
	HO_2			No recommendation			
	H_2O_2	0.22	0.5($\Delta\log\gamma$)	1.6	3×10^{14}	±0.5 ($\Delta\log K_{linC}$)	228-2
	H						
	NO			No recommendation			
	NO_2	$\leq 3 \times 10^{-4}$	1.0($\Delta\log\gamma$)				
	NO_3			3.07×10^{-09} exp(2646/T)		±100	
	N_2O						170-2
	N_2O_5			No recommendation			190
	HONO	0.02	± 0.01	1.0×10^{-05} exp(3843/T)	3×10^{14}	±50	180-2
	HNO_3			2.5×10^{-5} exp(4585/T)	2.7×10^{14}	± 700	190-2
	HO_2NO_2	0.15	± 0.10	No recommendation			190-2
	N_2O_5	0.02	± 0.01				190-2
	SO_2	0.9×10^{-6}	$\pm 0.5 \times 10^{-6}$	See data sheet			210-2
	HCHO			0.7	2.7×10^{14}	± 0.3	198-2
	HCOOH			5.8×10^{-11} exp(6500/T)	2.2×10^{14}	± 300	187-2
	CH_3CHO			No recommendation			
	CH_3COOH			1.9×10^{-11} exp(6660/T)	2.4×10^{14}	± 300 0.1	195-2

Beyond text

Data which doesn't come in sentences:

- Tables in articles
- Microarray data
- Experimental protocols
- Crystallographic data
- Chemical structures

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

Author-supplied graphics

5 questions for you

1. What could you achieve through “letting words stand for themselves”?
2. What do you hope to achieve through semantic markup?
3. What kinds of non-textual data do you have access to?
4. What can you do with your non-textual data?
5. What external resources could be opened to your readers?

Any questions for me?