Communication and data curation

V3. The solution to the data crisis is . . . *more* data

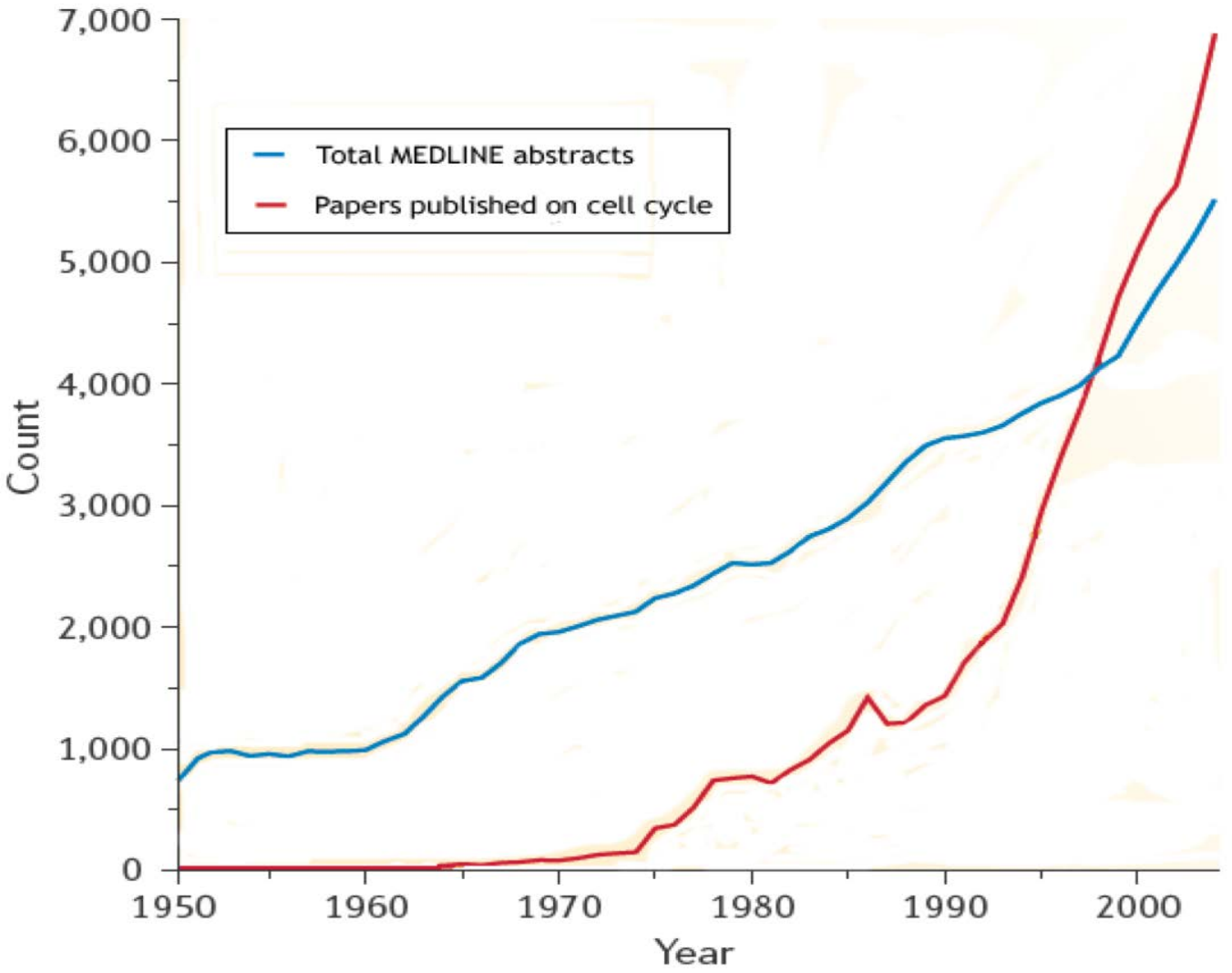
Recall the problem: more relevant articles than any specialist could possibly read

Background: How *do* scientists and other professionals read these days?

How could we help them read better?

Examples, obstacles, the future

Are you kidding me???



[Axis is x10^{-2} for total Medline abstracts]

Adapted from Jensen, Saric, & Bork; *Nature* (2006).

Some big numbers

- 25,000 publishers
- 1.5 million scholarly articles a year
- 600,000 abstracts added to Medline each year

So, how do scientists and other professionals engage the literature

User behavior during (scientific) search

in, e.g., PubMed, SCOPUS, Google Scholar, Web of Science, ADS, etc.

The *search trance*.

Researchers engage with the literature as if playing a video game

They rapidly, almost subconsciously…

- develop queries likely to find known items, or retrieve subject or topic results
- track references backward and citations forward,
- make rapid relevance judgments: assessments of impact, and quality
- locate and compare key terms, equations, definitions, protocols, findings

This is almost sub-cognitive, kinaesthetic, even trance-like,

sessions are often considered successful

— even though no article to read was found and read!!

Their goal appears to be *not* finding an article to read. . .

but *avoiding reading*

Now, this is nothing new…

- indexing and citation analysis help decide whether articles are relevant… *without reading them*.
- abstracts and literature reviews help us take advantage of articles … *without reading them.*
- the articles we do read, in their analyses and summaries help us take advantage of other articles … *without reading them*.
- friends, colleagues, and, best of all, graduate students, help us take advantage of articles … *without reading them*.

Paper based strategic reading

… engineers describe a common pattern for utilizing document components by zooming … and filtering information …

[they] first read the abstract, then skim section headings.
Next … lists, summary statements, definitions, and illustrations.

… they disaggregate and re-aggregate article components for use in their own work … perhaps by using a marker to highlight … perhaps by creating a mental register…

— Bruce Schatz et al.
"Federated Search of Scientific Literature"
IEEE Computer, 1999.

Online strategic reading

[*informant*] …

I used the sections of the papers for the equations….

I even wouldn't read all the other parts of the article …

I look for specific surface tensions, experimental measurements …

I sometimes need to look specifically at other methods and theories.

— Ann Bishop
"Document Structure and Digital Libraries:
How Researchers Mobilize Information in Journal Articles"
Information Processing and Management, 1999.

Longstanding behaviors, sure: but newly urgent

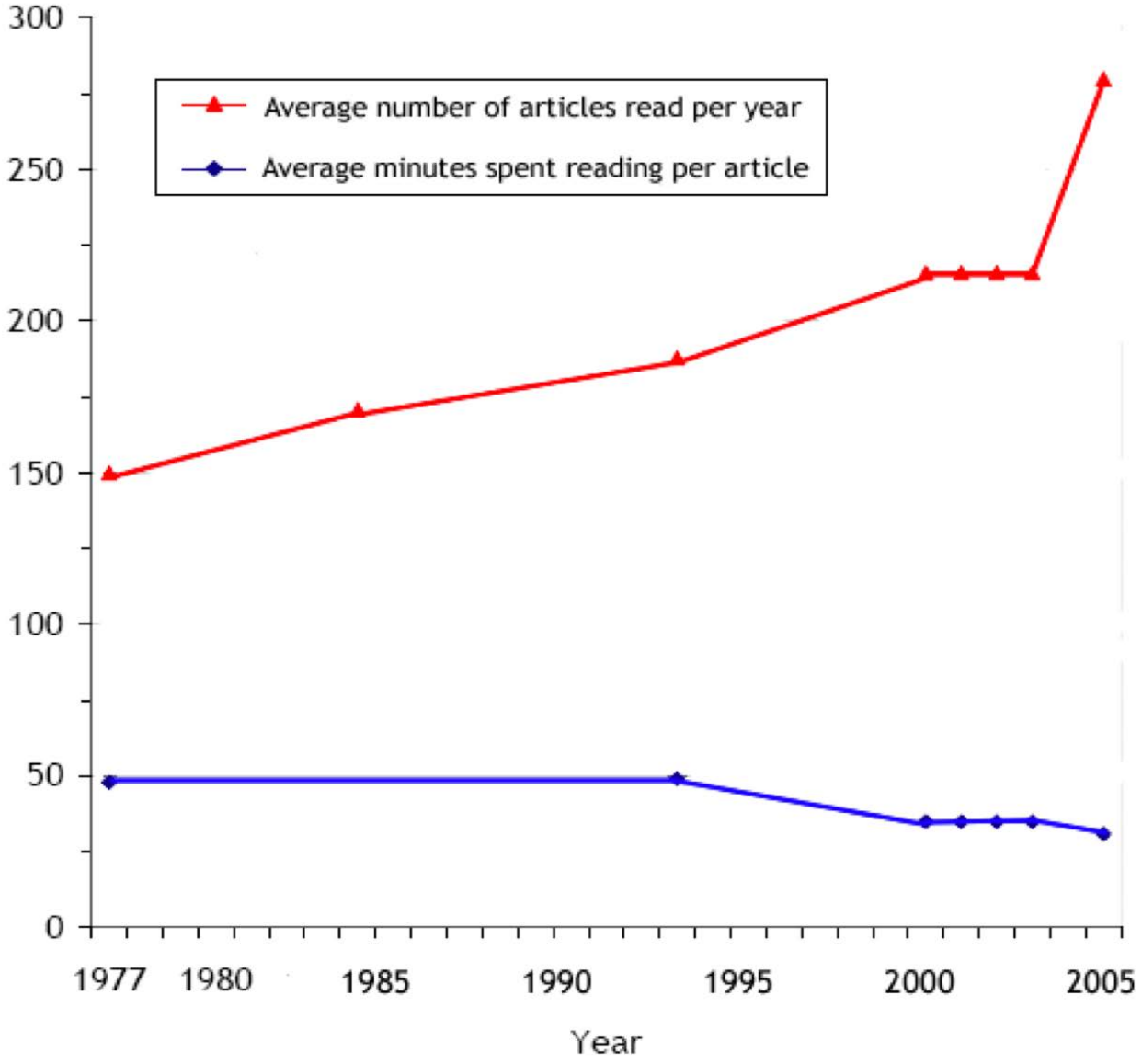
Key studies from C. Tenopir and D. King (2000-2007) show …

- time *searching* and *browsing* has been rising rapidly from 1984 to 2000.
- until the mid-1990s the number of articles "read" was more or less steady.
- but since then the number "read" (*not "browsed"*) has been climbing
- and so *reading time per article is dropping,* in some fields fairly fast (< 24 minutes!)

In addition studies from C. Palmer et al (2000-2007) show …

- researchers use sophisticated techniques to mobilize information according to varied research needs.
- these information needs vary with discipline, research life cycle (local and global), and research strategies, as well as with varying affordances of current technology
- the pace of evolution and innovation in these techniques is increasing, and increasingly driven by researchers themselves and not information specialists or publishers

Faster, faster, faster, more more more



Tenopir et al. 1977-2005

So what to do??

Human reading of natural language prose is uniquely valuable, providing nuance, clarity, and insight.

So reading cannot be totally replaced by text mining

So let's provide tools that support *strategic reading*

For example

— computationally available data items accessible with discipline-specific tools
(chemical formulae, proteins, equations, etc.)

— advanced navigation and viewing
optimized for domain-specific and objective-specific browsing and analysis,

— typed hypertext linking with links as first class objects,

— data-driven interactive diagrams and graphics

— computable equations,

— supportive ontological inferencing

— thoroughgoing interoperability with other tools

… and so on, and on, and on.

the grand old dream of radical new functionality as envisioned by
Paul Otlet, Vannevar Bush, Douglas Engelbart, Ted Nelson et al.

More of what scientists want. . .

The *datument* is a hypermedia document accessible to robots and humans

… for transmitting "complete" information including content and behaviour.

… the machine is … semantically aware of the document content [through] domain-specific XML components...

We argue that a cultural change in our approach to information is needed.

P. Murray-Rust and H.S. Rzepa[*]
"The Next Big Thing: From Hypermedia to Datuments,"
Journal of Digital Information, 5:1 2004

[*]Chemistry faculty at Cambridge University and Imperial College London, respectively.

Imagine what could be achieved if articles, rather than consisting entirely of free-form natural languages, contained explicit assertions about biological knowledge in unambiguous machine readable form … some progress is being made…

Mathew Cockerill, Editorial,
BMC Bioinfomatics, 6:140 2005

In a nutshell

As scientific ontologies are integrated into the publishing workflow
many enhancements to scientific communication will become possible

…support for text mining, information extraction, and literature-based discovery.

And one is not so obvious,

… support for the long-standing practice of *strategic reading*

"Strategic Reading and the Future of Scientific Publishing"
Allen H. Renear, Carole L. Palmer,
Science, August 16 2009.

Necessary data standards are now, finally, in place

Character encoding interoperability
Unicode/UTF-xx [Adoption: nearly total]

Data structure serialization interoperability
XML, JSON [Adoption: nearly total]

Syntactic interoperability
i.e. RDF(S), OWL [Adoption: underway]

Semantic interoperability
RDF/OWL ontologies; linked data.
[Adoption: substantial]

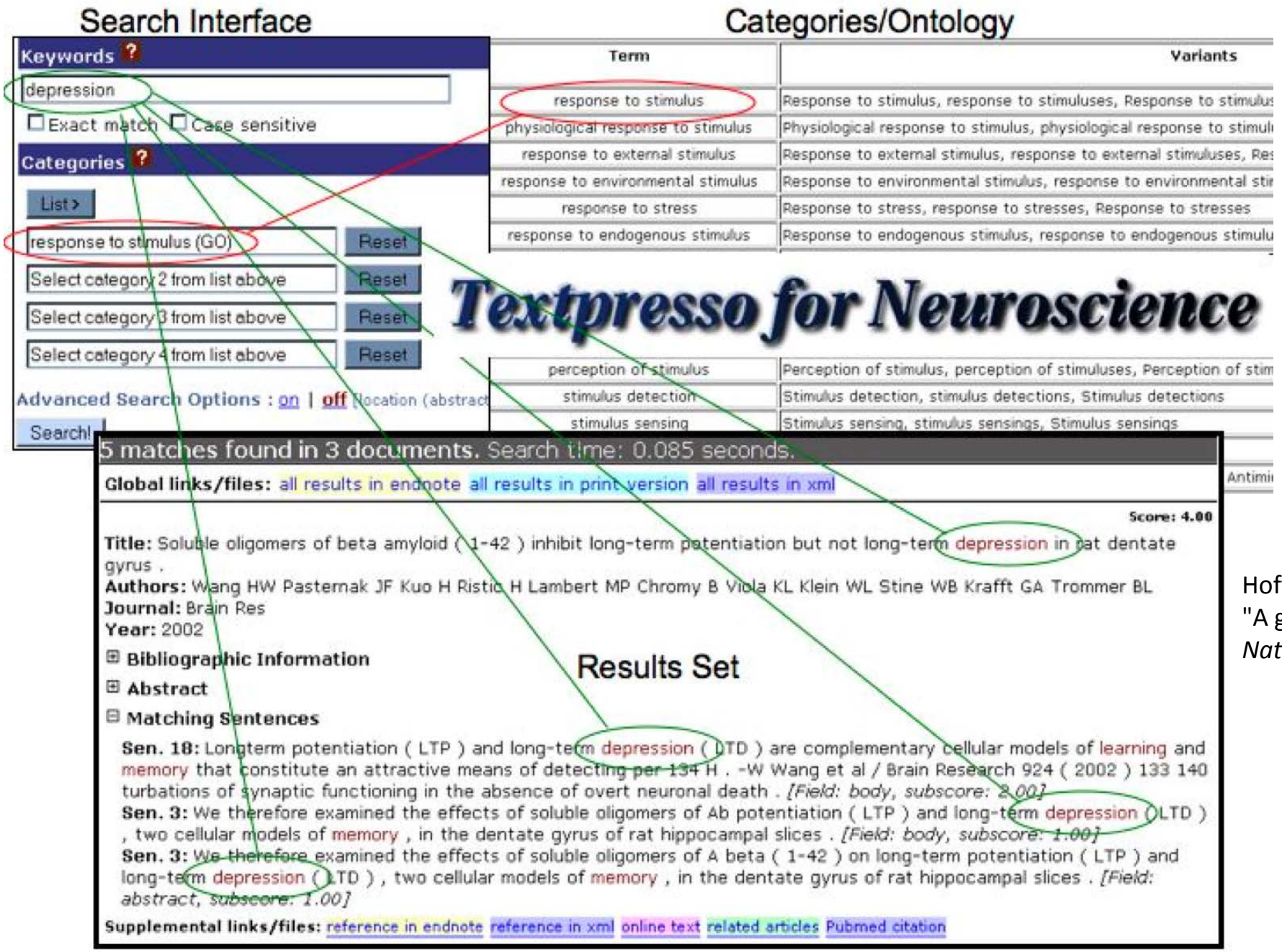
Document markup meta-languages
XML [Adoption: nearly total]

Document markup languages
e.g, NLM/DTD, XHTML, TEI, DocBook, DITA
[Adoption: widely adopted]

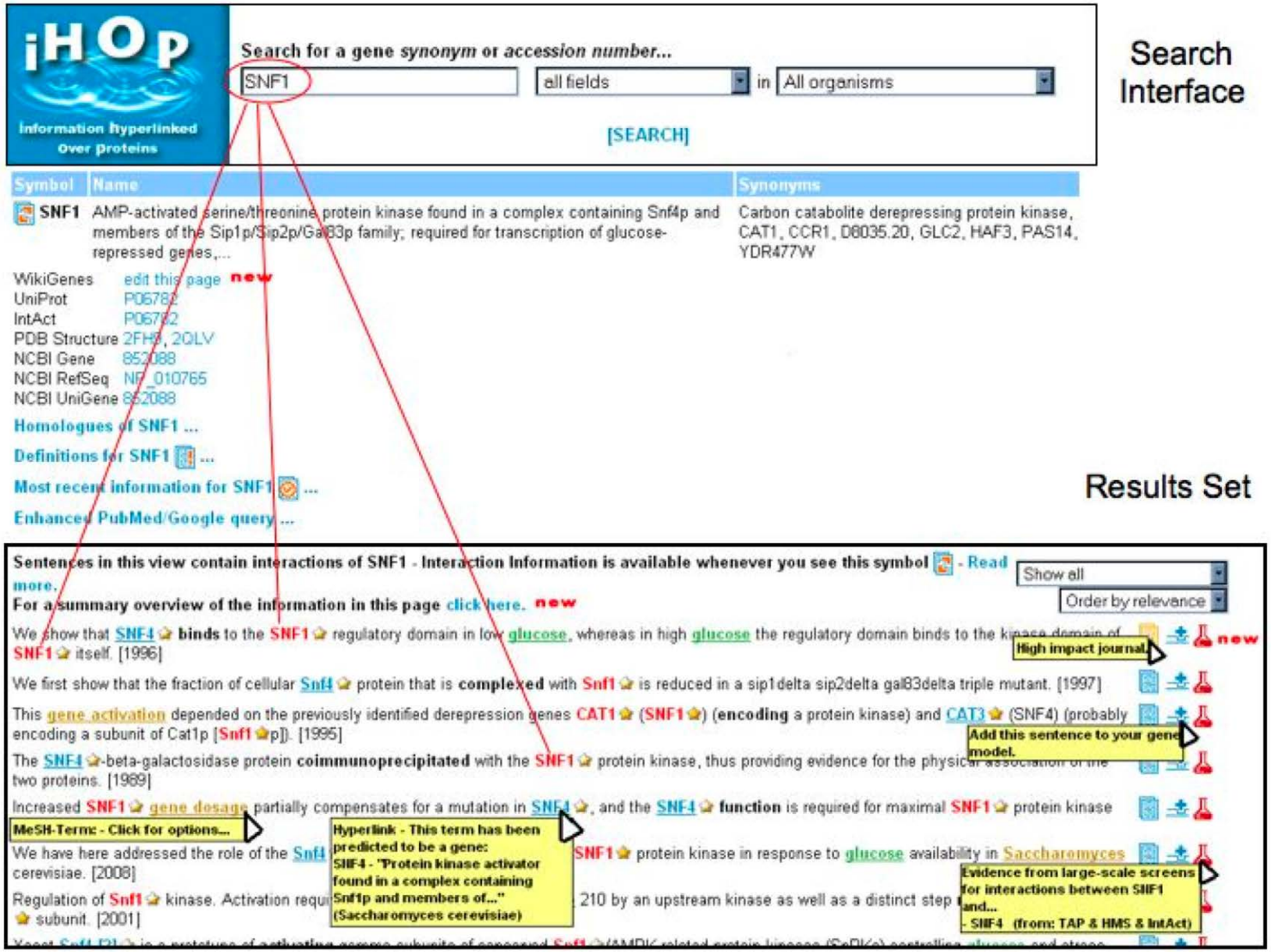
Metaphysical interoperability
"upper" ontologies [Adoption: (hard to say)]

Domain ontologies and terminologies [Adoption: steady improvements]
hundreds

Some examples . . .



Hoffmann, R; Valencia, A (Jul 2004).
"A gene network for navigating the literature.".
Nature Genetics. **36** (7): 664.



Muller HM, Kenny EE, Sternberg PW "Textpresso: an ontology-based information retrieval and extraction system for biological literature" *PLoS Biol*. 2004 Nov;2(11)..