

The Semantic Web in Ten Passages

Harold Boley

NRC-IIT Fredericton

University of New Brunswick

Short Research Presentations
For New Graduate Students

2 October 2002

(Revised: 10 September 2003)

1. Meaningful Search

Conventional Web expanded into a **Semantic Web**:

Search engines in future should ‘understand’ meaning of Web pages far enough to enable ‘sensible’ queries

At the moment **semantic search engines** only exist for specialized areas of knowledge

Shall use **conceptual representation** of Web pages:

- Help *people* as direct users
- Help ‘*agent systems*’ of AI: based on core technology of Semantic Web, offer higher Web services such as info comparison, integration, abstraction, or trading

2. The Search Engine and its Crawler

Crawler: A program that periodically navigates across Web pages and for every page analyses text, entering central words into ‘address book’

Every word in ‘address book’ refers to a list of all the pages *in which this word was discovered by the crawler*

You get 'hit list' of pages after you type in that (compound) word

===== Google-Search: "wonder drug" =====
24,000 pages, too low in “precision”, e.g. ambiguous



3. Precision and Recall – Conflicting Measures for Search Results (I)

Sample goal: Check Aspirin remedy for head pain



==== Google-Search: Aspirin ====
625,000 pages, too low in “precision”, although unambiguous

Crawler enters *all important words of analysed page* into ‘address book’, so you can now narrow down search by typing combination of words in the search line.

Then receive a page only if crawler has discovered in it *all of these search words*

==== Google-Search: Aspirin “head pain” ====
79,900 pages, “precision” improved

3. Precision and Recall – Conflicting Measures for Search Results (II)

But wait: Have we perhaps cut out pages because only wrote “head pain” but not “head hurt” which means same?

Indeed: In improving the **precision measure** we have seriously forfeited the **recall measure**



=== Google-Search: Aspirin “head hurt” OR “head pain” ===
84,300 pages, “recall” improved

4. Semantics – From Common Words to Standard Concepts

“Semantically”, i.e. ‘with respect to meaning’, we look for the **concept** that can be named in pages by “head pain” OR “head hurt” OR another such **word**

‘Semantic Search Engine’ could use **one** semantic **standard concept** for whole group of such words, named, e.g., by capitalized English term “Headache”

‘Address book’ internally only uses “Headache”. But this standard concept refers to all pages in which crawler found “head pain” OR “head hurt” OR another such common word

===== Semantic Search: Aspirin Headache =====
“recall” complete! – “precision” perfect?

5. Semantic Relationships Between Standard Concepts & ... (I)

Wanted pages claiming that Aspirin *cures* head pain –
not pages claiming that Aspirin *causes* head pain

Semantic relationships: ‘address book’...►“knowledge base”
Contains so-called ‘facts’ like “Aspirin CURES Headache”
(here triple of the form “Subject PREDICATE Object”)

All-capitalized term “CURES” is **standard predicate**,
standing for common relational words in pages such as
“remedies”, “heals”, etc.

==== Semantic Search: Aspirin CURES Headache ====
“precision” perfect!

5. Semantic Relationships Between Standard Concepts & ... (II)

Some pages claim both semantic relationships, the curing *and* the causing one

=== Semantic Search: Aspirin CURES Headache AND Aspirin CAUSES Headache ===

Describe such pages with a further standard predicate “AMB”, even if they do not contain a corresponding common word such as “ambivalent”, “conflicting” etc.

== Semantic Search: Aspirin AMB Headache ==

Store “Aspirin AMB Headache” as a **fact** in ‘address book’?

5. ... & Knowledge Derivation

Better: Logic languages, e.g. RuleML, instead allow this triple to be *derived* from the two stored facts with a so-called **rule**

A special ‘If-then’ derivation like

IF Aspirin CURES Headache AND Aspirin CAUSES Headache THEN Aspirin AMB Headache

is performed with the general ‘IF-THEN’ rule (‘?’ for variables)

IF ?Pharm CURES ?Sick AND ?Pharm CAUSES ?Sick THEN ?Pharm AMB ?Sick

via ‘bindings’ like ‘?Pharm = Aspirin’ and ‘?Sick = Headache’

- Rules explicitly deduce knowledge (here on ‘ambivalence’) implicitly hidden in the facts (here in ‘cures’ plus ‘causes’)
- In parallel, they can find every page that fulfills the ‘IF’ part, hence also the ‘THEN’ part (here each “AMB” page)

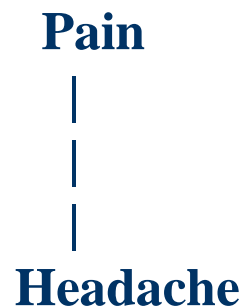
6. Where Do the Standard Concepts and Standard Predicates Come from?

Experts of field, in this case medicine, have to agree on standard definitions of connected concepts and predicates

Hierarchical, superconcept-subconcept connection is most important

Example:

Pain-Headache connection puts Headache below Pain:



For such **shared explicit concept catalogues** AI often borrows the expression “ontologies” from philosophy

7. How Does One Assign the Standard Concepts/Predicates to Common Words?

Ideally, crawler would navigate through pages for important common words and assign right standard concepts and standard predicates to them fully automatically

But such full automation is very difficult

Interactive classification of pages together with experts:

- 1) The crawler for a given page proposes standard concepts, some with semantic relationships via standard predicates
- 2) At least for unclear cases these will then be corrected and if necessary completed by experts

8. Where Will the Assignments be Stored as Metadata?

Two principal possibilities for storing these metadata:

“EXTERNAL”: **‘Address book’** can store standard concept/relationship together with its assignment to all pages with the corresponding common words

“INTERNAL”: **Pages themselves** can store their own descriptive standard concepts/relationships (“annotations”)

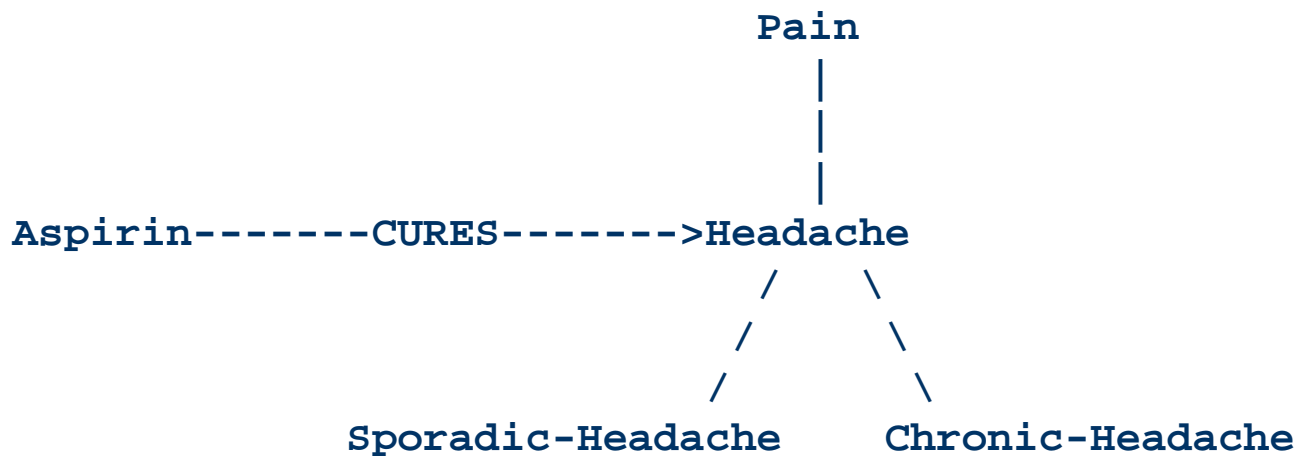
Advantage of “EXTERNAL”, disadvantage of “INTERNAL”:
Only by separating metadata from pages themselves is it possible to **describe pages one does not own**

Advantage of “INTERNAL”, disadvantage of “EXTERNAL”:
For every page change **affected annotations can be immediately updated** as well without searching for metadata

9. Refined Standard Concepts Inherit Refined Semantic Relationships (I)

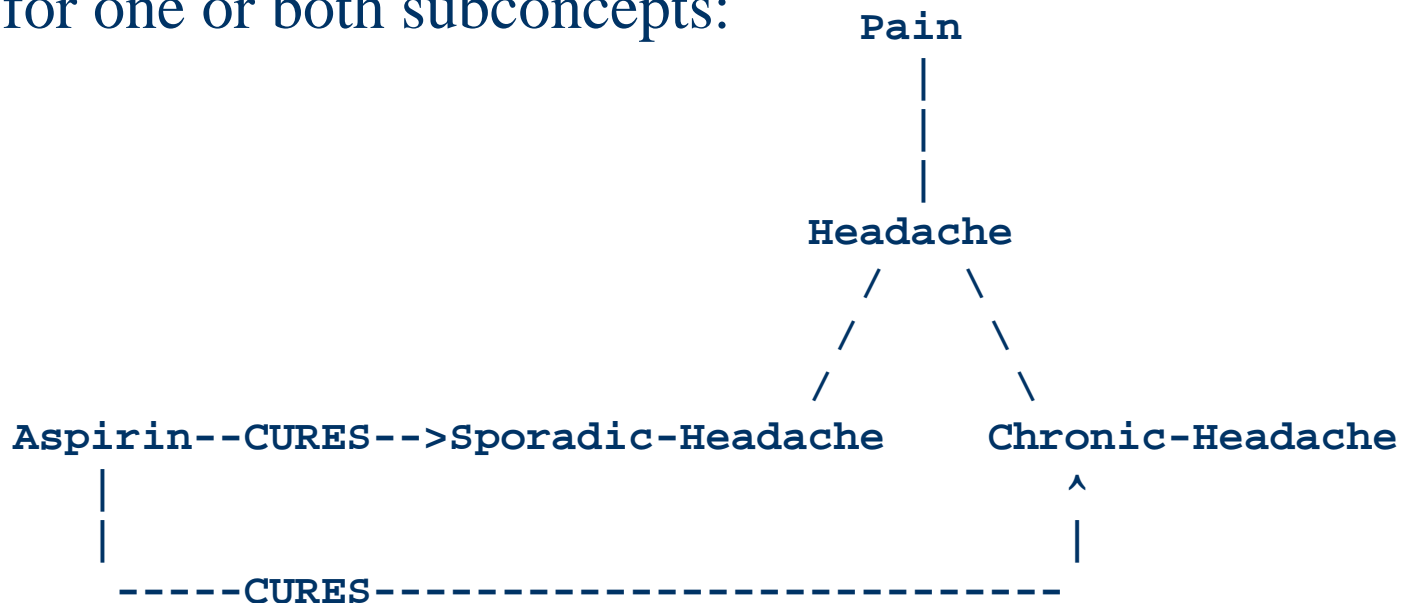
What happens when standard concepts or semantic relationships change, e.g. through **concept refinements** following new scientific discoveries?

E.g., split our sample standard concept **Headache** into subconcepts such as **Sporadic/Chronic-Headache**; we agree on the following ontology, including our semantic relationship “Aspirin CURES Headache”:



9. Refined Standard Concepts Inherit Refined Semantic Relationships (II)

Relationship “Aspirin CURES Headache” may be refined for one or both subconcepts:



If relation holds for *all* subconcepts (here: Sporadic/Chronic-Headache), it can also be left at the superconcept (Headache), from where it is automatically **‘inherited’** to the subconcepts on demand only (similarly as in OO programming)

9. Refined Standard Concepts Inherit Refined Semantic Relationships (III)

As a result of such concept refinements two principal possibilities arise for pages classified by them:

“UPDATE”: Try corresponding **retroactive updates to metadata of all affected ‘old’ pages.**

Domain experts should decide whether one or more subconcept such as Sporadic-Headache and Chronic-Headache were ‘meant’ or whether their old common superconcept Headache remains correct

“SWITCH”: Switch metadata ontology at certain points in time, **continue to access ‘old’ pages via ‘old’ metadata, and only for ‘new’ pages use ‘new’ metadata.**

Headache would stay unrefined as a standard concept for an old page, even if domain experts would immediately notice that it were, e.g., only about Sporadic-Headache

10. Library Catalogues as Metadata Ontologies

“UPDATE” would be ‘nicer’ solution, but many libraries have chosen solution “SWITCH”, i.e. put up with users having to search in two or more catalogues sometimes

The Semantic Web will not solve *this* problem either, but both possible solutions, “UPDATE” and “SWITCH”, will be supported by software tools of the Semantic Web

Conversely, the Semantic Web can learn a lot from Library Sciences. Initiatives – e.g. within Math-Net and CISTI – attempt to bring both together

The Semantic Web, on the basis of AI, is a new subfield of computer science with various further interdisciplinary relations, e.g. to logic, linguistics, and cognitive science