Semantic Internet Searching Using Active Structure

Hit and miss is not the way to do it

Powered by ORION

Four Pieces

“Yet the human processing of information makes use of a much deeper understanding of text than these systems exhibit. Is there a way to combine the computing power of machines with the sophisticated processing of humans?” – HCSNet NGS09

People may have billions of neurons in their heads, but they also have a limit of no more than four or so pieces of information in play at once. More than that, and “they can’t get their head around it” (Greenspan). How about we examine our own limitations before inflicting our methods of processing on machines – four pieces is pretty crummy

We Do the Clever Stuff

Do we take the view that the search tool should show us all the right documents, but we do any synthesising?

That is obvious nonsense, given our limitations and the difficulty in finding the right documents

So, the machine has to do some synthesising of its own, to find the right documents

We need to cede some cognitive activity

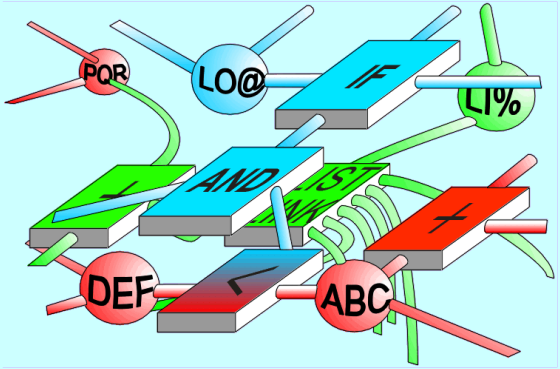
Not Just Bigger, Different

When we want to do more, we don't just scale up what we can do, we make something conceptually different

Do More Differently

If we want to synthesise across a thousand documents, four pieces is a barrier we need to break. Like the truck, we should look at doing something different.

Just as the truck has wheels, we can use undirected operators and wires to go beyond the limitations of a living cell. Then we can have as many things in play as need be.



Not Just Diagrams

Don’t turn words into a representation, use machinery to turn them into machinery – an active structure. Then the machinery can turn more words into more machinery, and the same machinery can do the searching for us – it is already doing searching for itself as it builds its structure.

We blow away the four pieces limit

A Systems Approach

Let's assume reasonably complex text – legal, scientific, medical, technical - something worth researching, not "What's on at the flicks?"

So what bits do we need, and how do we get them to work together?

A Dictionary

The system will already know the meanings for a lot of words, so we have a dictionary to find them. But a lot of the text we read will introduce new words or new definitions of existing words, so the dictionary will need to be dynamic.

We could have a stable global dictionary and a local dictionary for the particular document – two documents could define a term differently. Multiple definitions will pose a problem for spanning documents.

Mission failure is defined as…
Fibroblast growth factor (FGF)…

Logics

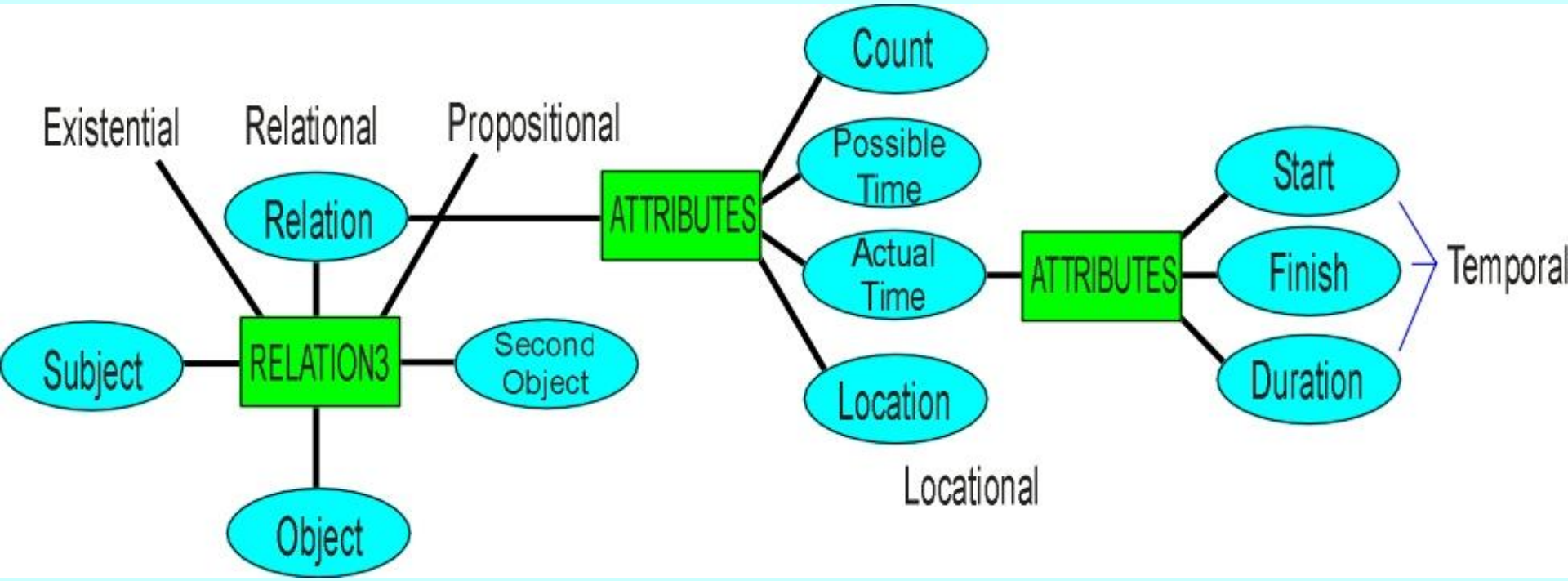
What about all the logics we will find in the text?

- Propositional and Existential
- Relational
- Temporal and Locational
- States and Transitions
- Sets, Groups and Numbers

This is part of our human four pieces limit – we can handle them all in one spot, but we try to keep them separate in general, because to synthesise across all of them over a document would overwhelm us.

It is easier for the machine if they are synthesised into a common form and it doesn’t have to treat them separately

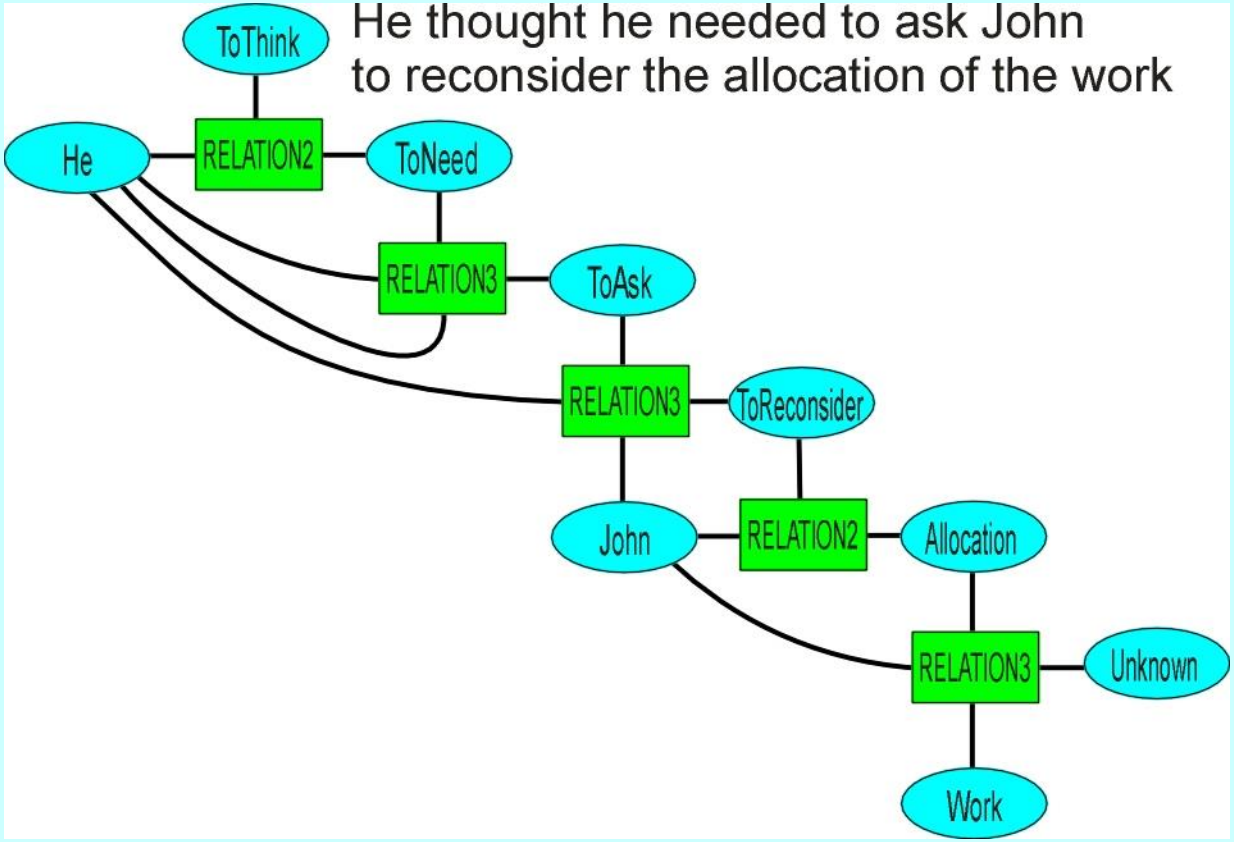
Common Form



A relation combines relational, propositional and existential logic. Attributes handle temporal and locational logic with linkable objects

The connections behave as wires – they have no implicit direction, and carry states and objects in any direction

Layering without Limit

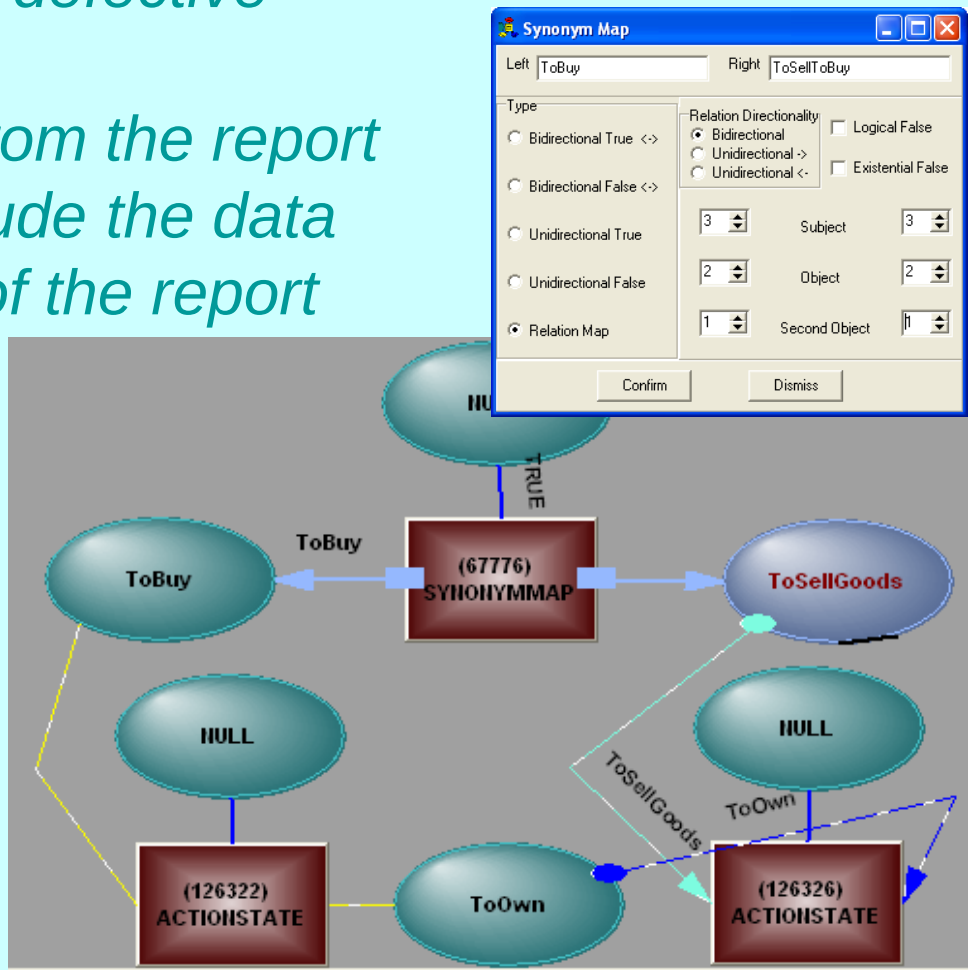


Structures layer without limit – intent, possibility, actuality

Mapping of Structure

John sold Fred a car
Fred bought a car from John
The car sold Fred was defective

Joe omitted the data from the report
The report did not include the data
The data was left out of the report



We need to do a lot better at mapping of relation structures, not just simple synonyms/antonyms

Complex Text

Complex text refers to itself

> The restrictions on trade in section 3.2 are relevant

It recycles a prologue to a structured list

> …in one of the following states:
> (a)…;(b) Student logged on;(c)…;

It turns itself on and off

> Clause 4 is void if…

It creates alternative streams

> Treatment may be either by radiotherapy or chemotherapy…

Part of it survives itself

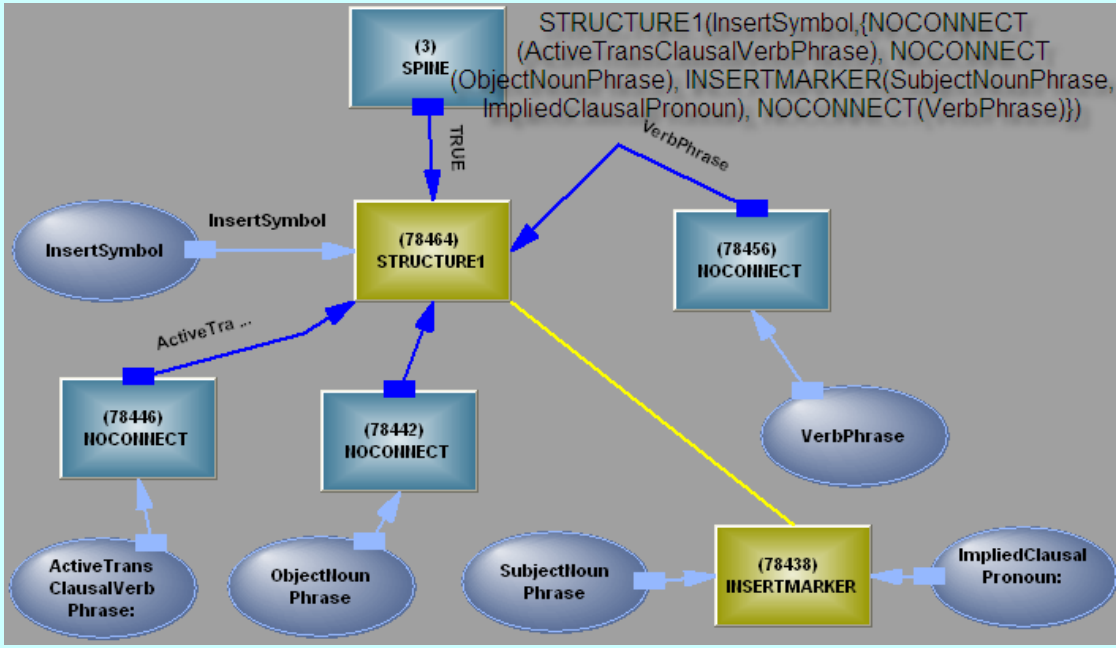
> Clause 15 shall survive termination of this contract

It is complex, in other words, and we had better be capable of handling these complexities, otherwise our searching is only kidstakes, and not for research

Some Grammar

Grammar patterns to build

- Collocations
- Noun phrases
- Embedded lists
- Main and subordinate clauses



About 5000 patterns, some heavily tinged with semantics
Some patterns rearrange the parse chain or insert implied objects – building the grammar structure is highly dynamic, because our real aim is to build the object/relation structure

Grammar Is Not Enough

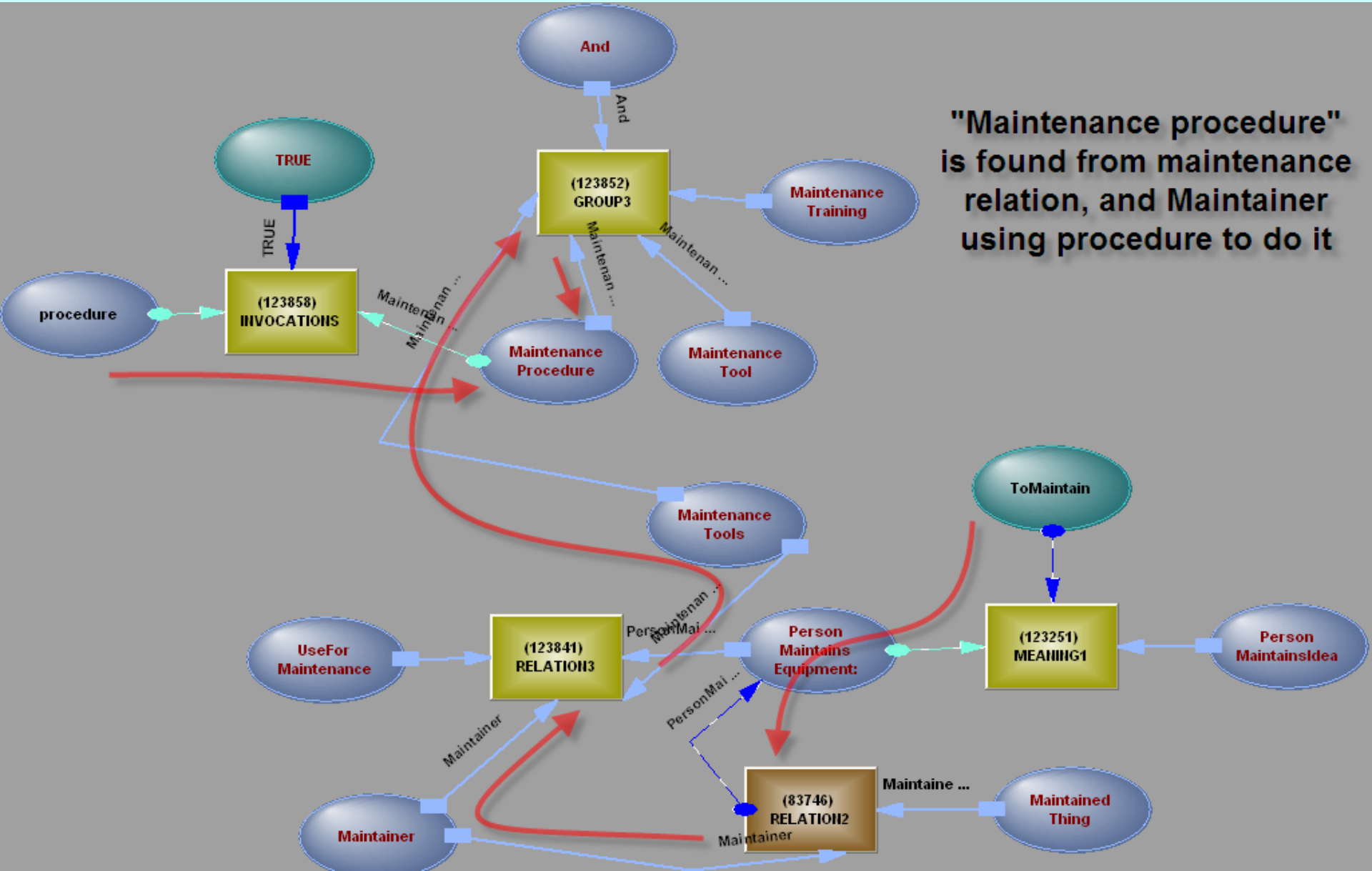
During A, B of C and D of E, F and G shall be required….

Is the subject -

- B
- B and D
- F and G
- B, F and G
- B and D, F and G

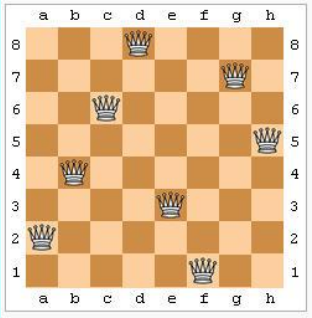
We can't tell without semantics
Grammar can only be a helper in dense texts

Some Semantic Modelling



When It Gets Desperate

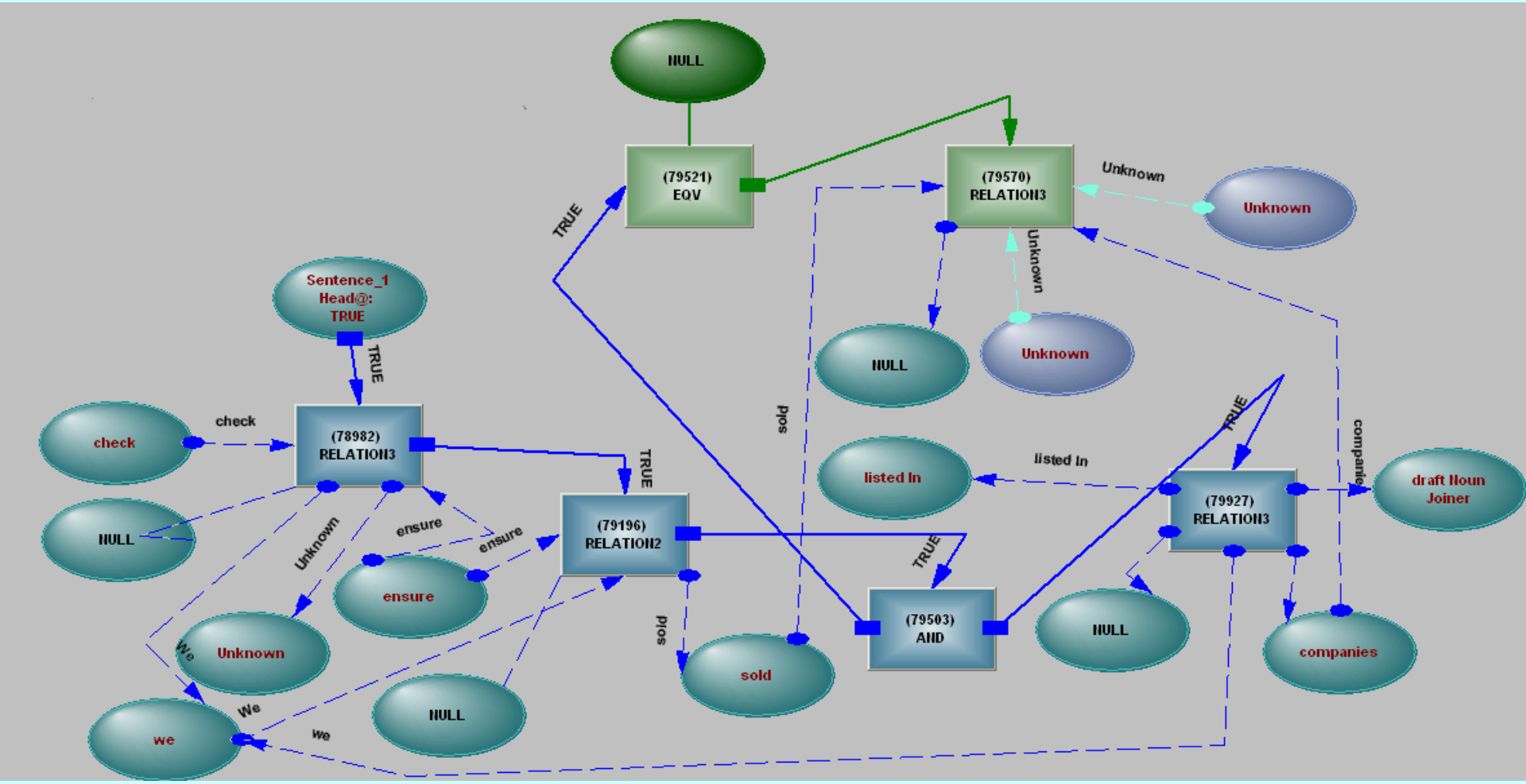
Where analytic use of grammar and semantics fails, the system resorts to hypothesising.



This is the hypothesising of constraint reasoning, except here we are building the structure, propagating states in it, observing the consequences, then tearing it down and building something else – dynamic structural backtrack

The maintenance impacts on the schedule we have chosen for next week in several ways that are undesirable

A Sentence



A sentence has a logical spine linking all the relations built from verbs, noun phrases, prepositional chains

No Longer a Bag of Words

A document is turned into a complete semantic structure – the structure extends itself and uses what is already built to disambiguate what it reads later.

It fixes up all the anaphora, forward references and ambiguities that can't be resolved immediately

This is computationally expensive to do,
so it has to be worth doing

Search Queries

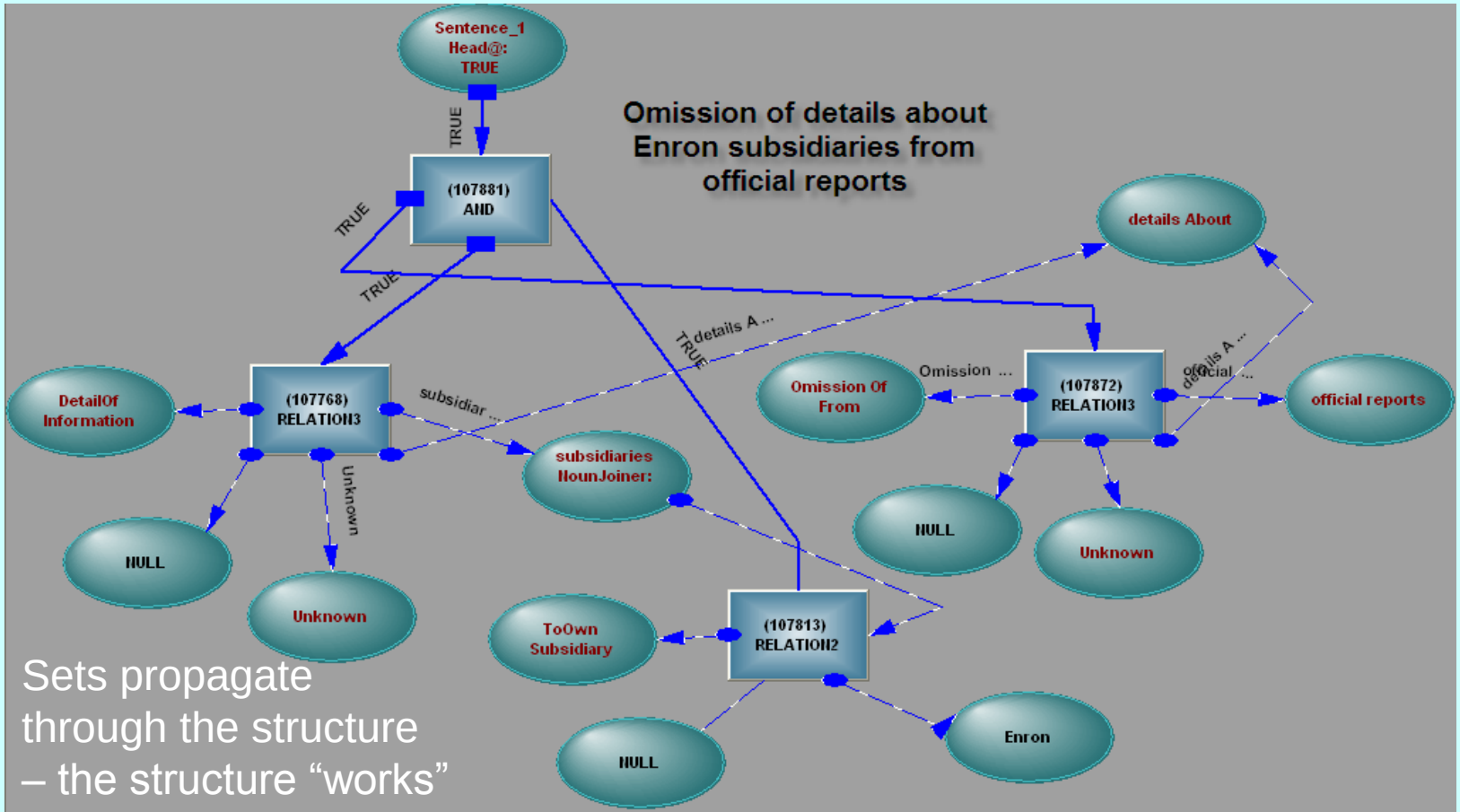
If we have broken the four pieces limit, then our queries can be more complex.

Queries can be posed in free text, and can be multi-sentence using pronouns – the same transformation mechanism is used on them as is used on the original text.

In operation, queries can become active – generating alternatives, shuttling them through the structure, and pruning them

Searching

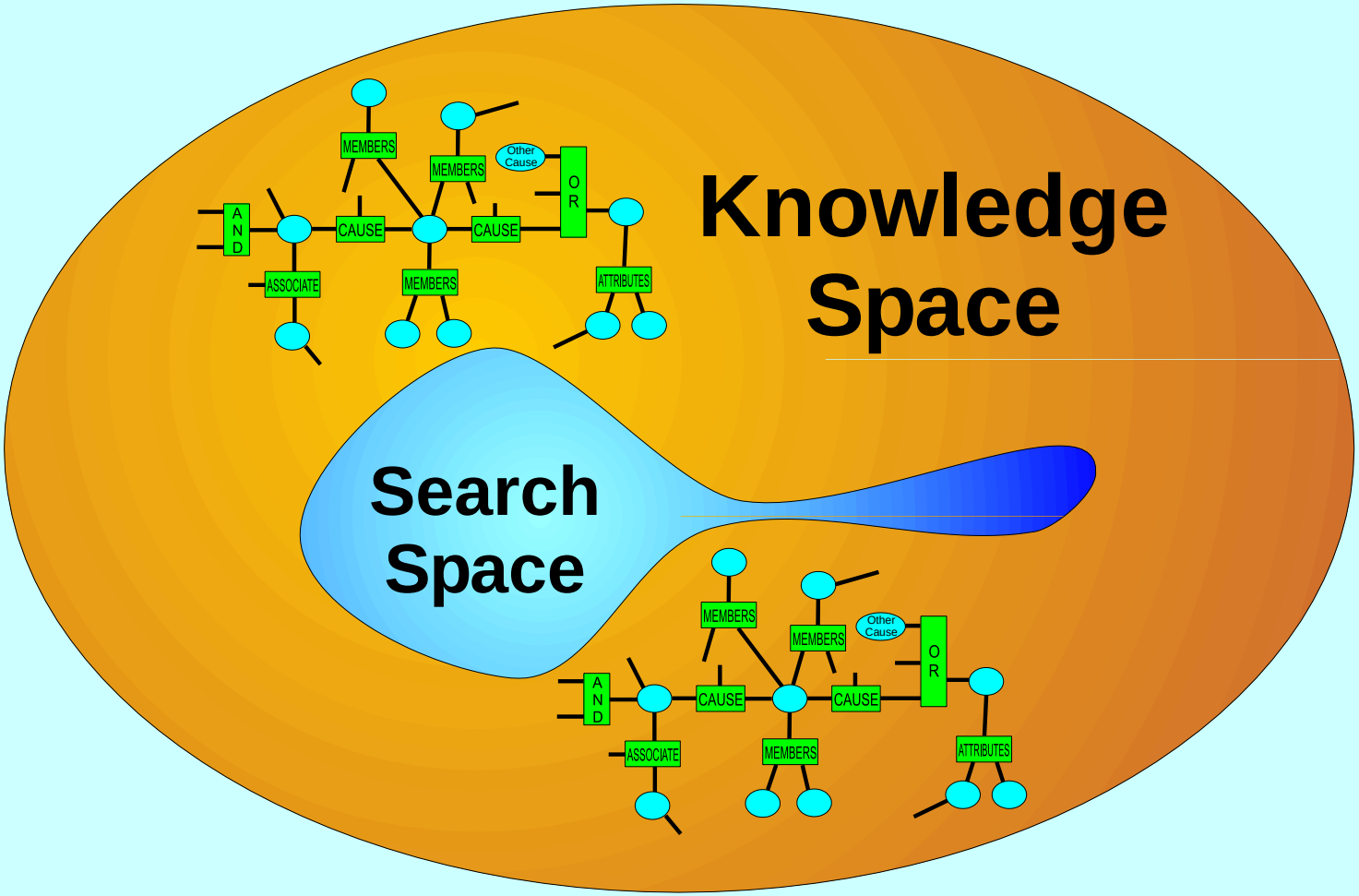
A search structure is created, and it is then matched to the structure built from the text.



Problems for Internet Extension

- Slow – about the speed of an attentive human reader
- Intended to be accurate and complete on heavily packed noun phrases and long prepositional chains
- The modelling required for so many knowledge domains could be daunting

Moldable Search Space



When there is an overwhelming amount of information, you need a powerful and flexible way of deciding what to ignore - held in the same structure

Machine Power

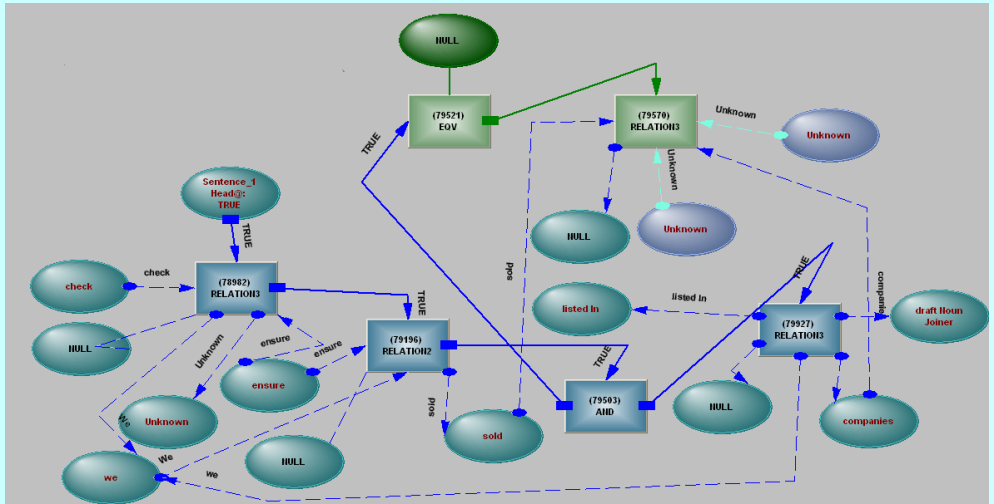
It is estimated that Google has a million servers. If we compare finding key words with searching on meaning – maybe 1000 times the effort.

Is it worth 1000 times the value – for valuable stuff yes, for simple stuff no.

Can we search without understanding?

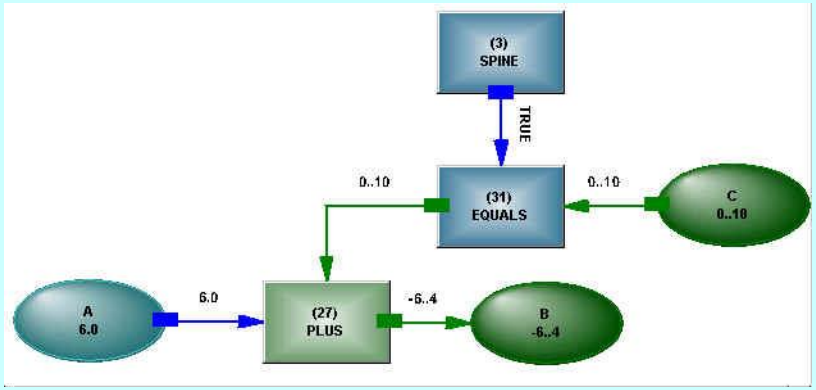
What do we expect to find, other than lots of misses and extra hits?

Enough with the fudges!



No Fudges

Relations conceptualised as objects
No pipelining
Parsing of objects between words
Building of undirected structure as goal
New structure seamless with old
Synthesis of all logics



A Thoroughness Not Hitherto Attempted