External Resources

The internet connects us to sources of knowledge in a way that has never before been possible - a reader can quickly cross reference a word or phrase to clarify meaning, usage or context.

Often the clarification requires no more than a scan of search results to find the answer within the first few results, in other cases it becomes necessary to follow one or more results and read (or at least browse) the content to discover the information we need.

If the reader happens to be a machine then scanning whatever results are presented might not be sufficient - a human reader can subconsciously apply a filter that disregards irrelevant or nonsensical results. A particular resource might be disregarded because of dubious advertising or other visual cues that discredit the content from being worthy of further reading, whereas a machine sees only the words as presented.

What is needed is a plan, where to begin, how to proceed, how to maintain consistency and coherence. Start with a highly regarded and authoritative source such as the Oxford Dictionary to determine if the word is immediately known.

Perhaps the word is common but has simply been misspelt, try Bing Spellcheck, a definitive response could take us back to the Oxford Dictionary to retrieve more meaning or if there are several alternative correct words they may need to be checked against the context of the original text to pinpoint the best choice (we may not have enough context yet, and the decision needs to be put off until we have more).

If the word is not misspelt, we move on to other resources:

> Google or Bing search, these could lead almost anywhere but some results can be given higher priority, Wikipedia isn't authoritative but does have editorial control and peer review. Generic searches can benefit from additional context, using the word within a phrase to narrow the results and avoid clutter.
>
> Medical dictionary or medical information providers, either as a matter of course or because it is relevant to the subject matter.
>
> Drug lists and other resources from the National Library of Medicine and the US FDA in case the word might be a pharmaceutical product or ingredient.
>
> Other domain specific resources can be included, the internet is almost boundless and nearly every subject is represented somewhere.
>
> Many words can be either a noun or a verb – “costs”. If the current use lacks enough context to decide, the system may need to look at similar documents.

The plan is not to know everything about a word, just enough to provide meaning in the current context and to give a word its place in the overall structure. Different purposes can require different plans with authority given to resources that have more focus on the subject domain.

Dictionaries do much more than single words, a list of synonyms for the word “channel”, collocations with “take”, the meaning of a phrase like “cash flow” or “tennis ball”. Just some of the rich collection of content accessible to a connected machine.

New words being accepted into Orion are also checked to see if they include prefixes and suffixes:

multichannel

inelastic

The word may need to be decomposed for use around a conjunction – “a single or multichannel receiver” – or to allow the structure to accommodate use in different forms:

subchannel

elasticated.

Particularly in medical text, a word may be fabricated out of prefixes and suffixes, but appear in no medical dictionary. If so, Orion has to retreat to its own resources, and make sense of the word by assembling the prefixes and suffixes itself, just as a medical reader would do.

Many times a verb will be contained within a collocation to invoke a precise usage:

call off

open out

take after

A dictionary check can provide the intended meaning where it differs from the verb alone. Collocations can have multiple meanings:

The plane took off.

He took off his jumper.

and we need to be sure we have the right meaning in context when we are dealing with a new collocation, or the collocation may allow separation:

He took his jumper off.

or the collocating preposition may be a long way away:

He waited with bated breath for the operation to start (waited for).

If the phrase contains a key idiom word like “stitch” when used for “stitch in time”, the idiom must be recognised to avoid simplistic interpretations. Except for a select few in the 1960's no-one was ever literally “over the moon”.

Search engines throw away capitalisation even when it gives implicit meaning – “Brown syndrome” which is very different to “brown syndrome” – capitalisation needs to be retained and respected where it imparts more information and ignored where it doesn't – “john asked emily NOT TO SHOUT when sending Emails”.

Like any student trying to learn, Orion will make mistakes. Its human overseers can examine the new structure it has built based on external sources, and make corrections.

Exploring the vast ocean of information that is provided by the internet presents enormous opportunities for automated learning and domain analysis, as long as the knowledge can be interpreted, understood and built upon rather than just harvested and stored.

Luckily, Orion is designed to be both plastic (everything can change) and a fast learner.