



ReleMed, GoPubMed, and Cuil

I have chosen two variations of the PubMed database that have smart interface, search, and display features. They bring out the best from this rich database through their intelligent software that simplifies and, at the same time, enhances the search and output options. These sites also spare researchers the complexities of looking up and navigating the MeSH vocabulary to formulate a query that anticipates, for example, the British/American spelling and the pairing of MeSH terms with appropriate subheadings. The pan is the Cuil database (expected to be pronounced as “cool”), which has been anything but cool. In the database/search engine industry, it was possibly the biggest disappointment of 2008, if not the millennium.

“They bring out the best from this rich database through their intelligent software that simplifies and, at the same time, enhances the search and output options.”



the picks

RELEMED

The PubMed (www.pubmed.gov) database, with its 18 million records, is one of the best digital creations of the U.S. government. But its native interface is not for the faint of heart. ReleMed (www.relemed.com), an alternative interface, makes searching the deeply structured, content-rich file very effective, even pleasurable. It was developed by Intelligent Search Technologies, Inc., and the product, indeed, shows intelligent technologies that lead users to the most relevant articles, with one exception.

For starters, in the results list, ReleMed shows the matching elements of the query in the appropriate fields, something that is similar to (but more compact than) what a traditional KWIC format offers. It also highlights the query elements in every matching extract, which makes it easy to quickly scan through the match list. All the sentences or segments of the bibliographic records where at least two of the query elements (and the elements in the translated/enhanced query occur together) are included. This proximity operation is a powerful way to increase precision, and it is a rarity even in the search software of subscription-based information services.

Similarly compact and efficient is the listing of the target sites that contain the full text of the document. This is derived from the LinkOut option of the PubMed records. The only oddity I found in this excellent interface is that the list does not include the PubMed Central linkouts, where the full text would be available free of charge (depending on the possible moratorium chosen by the publisher within the limits of the current mandate set by Congress).

It is odd because this is immediately visible even in the PubMed interface using unambiguous icons for the free availability of the primary document. Learning about an open access version of the source document is so important that it should be included as a filter, both in the native and third-party interfaces. It is a nice touch that the links to the consumer-oriented primary documents are highlighted in yellow.

The other appealing feature of ReleMed is the detailed, informative, and well-presented display of how the original query was translated and automatically enhanced. This is far more user-friendly and more complete than in the native interface. It shows the terms taken from the UMLS (the Unified Medical Language System). ReleMed's intelligent search automatically maps terms, adding synonyms and related words so that **elevated** also retrieves **high** and **dose** retrieves **dosage**. ReleMed calls this "query translation" and "query expansion." This can increase recall considerably, without a sharp drop in precision.



Results list with highlighted matches and outlinks



Details of query modifications and enhancement by synonyms and lexical variants

GOPUBMED

Accessing PubMed through the GoPubMed (www.gopubmed.org) interface also has a few shortcomings (although primarily because of easy-to-correct programming glitches or questionable filtering preferences). I still picked it for its innovative and smart features. One of the limitations is the same missed opportunity (as in ReleMed) to offer an option to limit the search to records for open access items available through PubMed Central (www.pubmedcentral.nih.gov) or to just alert users about the availability of a free version.

I am stressing this issue because PubMed Central, which was my pick in my last column, keeps growing at an impressive pace, and it offers free access to more than 1.5 million papers from many of the best life science journals. Interestingly, the developers of GoPubMed at Transinsight GmbH, in Germany, thought about it. The software offers a filter in the cluster pane labeled PubMed Central, but it erroneously, and irritatingly, sends users back to the GoPubMed basic query page instead of running the search in PubMed Central or just decoding/converting behind the scenes the icon/linkout options for a PMC/open access filter.

The other shortcoming is that the theoretically useful filter option that limits the search to items published in the highest impact journals is neither fair nor practical. It is not fair because, although the journals picked by GoPubMed are indeed the highest impact factor journals, they cross all the disciplines rather than showing the top ones in the Web of Science disciplinary category/categories that fit the query best. In the case for a topic on **acquired ocular toxoplasmosis**, the highest impact factor journals in parasitology and ophthalmology would qualify for such a filter rather than *Science*, *NEJM*, *CA*, *The Lancet*, and *Nature*. Not surprisingly, applying this filter to the query eliminated 238 hits of the original result list of 240—quite a "hitside."

The most productive journals for this topic do appear on a separate filter list, applaudably with hit counts. However, only one journal at a time can be chosen. It would be a good compromise to allow users to choose several journals instead of shooting in the dark with these high-impact journals.



Clusters by four of the five classic Ws: who, what, where, and when

The clusters rhyme with four of the five basic Ws of journalism: who (authorship), what (topical categories), where (journal names and geographic author affiliation), and when (publication year). Getting an answer to the fifth W (*why*) was the given research done and the paper published) still requires reading the entire article. However, quite often, records with the best structured abstracts also answer that question in the PURPOSE section of abstracts to help the users in filtering the primary result set. In the long run, these talented developers certainly will come up with a technology to create an answer to the why question from parsing the nonstructured abstracts and the MeSH subheadings assigned to the records.



the pan

CUIL

Everything was ready for the July 2008 launch of the purported Google-killer Cuil (www.cuil.com) search service. I imagined the developers—including a couple of former super-thinkers from the Googleplex—popping the corks of a few bottles of Himalayan goji juice, reading the “message” in unison, and holding hands, while playing Enya’s lovely *Paint the Sky With Stars* album, before they turned on the master switch for their boxes.

Everything was ready on the PR side, that is, because neither the hardware nor the software were ready. The press corps got the scoop, and the blogosphere went to high gear just to witness a giant fiasco when the moment of truth came. Cuil’s hardware system crashed under the workload right at the beginning. The software was in prealpha stage (and still is) in my opinion. Cuil may still be a candidate for Microsoft to acquire, after that company luckily failed with the acquisition of Yahoo!.

I could finally access and test the service a couple of days after Cuil’s debut, and I was mightily unimpressed for reasons explained in my long review (www.gale.cengage.com/reference/peter).

Suffice it to say here that the music that kept popping up in my head during the first test was the title song from Deana Carter’s quadruple platinum debut album, *Did I Shave My Legs for This?* Cuil’s message (parroted by many starstruck journalists and bloggers) promised to have collected about three times as many pages as any other search engine (meaning Google) and 10 times as many pages as Windows Live (which is not exactly a good benchmark). Noticeably, they did not mention Yahoo!, which produced the highest number of hits for my cross-search-engine test. Yahoo! and Google often produced (or at least reported) orders of a magnitude of more hits for my test queries than Cuil (except for my name, where Cuil had more hits reported than Google).

YAHOO!	Google	cuil	Test queries
6,680,000,000	2,940,000,000	546,032,790	Yahoo
4,430,000,000	2,740,000,000	516,386,388	Google
37,900,000	5,690,000	8,512	Cuil
4,130,000	9,830,000	494	Cengage
33,800	5,330	6,925	"peter/jasco"
300	88	35	Scopus calculating h-index

Hit Rates			
YAHOO!	Google	cuil	Test queries
12.2	5.4	1	Yahoo
8.6	5.3	1	Google
4,452.5	668.5	1	Cuil
8,360.3	19,898.8	1	Cengage
4.9	0.8	1	"peter/jasco"
8.6	2.5	1	Scopus calculating h-index

Hit counts reported for some test queries, and the hit rates

Almost a month later, I retested Cuil for this column. There were no hardware problems this time, and the hit counts increased. But in terms of hit rates, the picture remained the same for most of the queries (except for the test query word Cuil) because the other systems also kept growing in the past 2 months.

Interestingly, the number of hits for the last test query dropped. The reason is that Cuil eliminated most (but not all) of the triplicate, quadruplicate, and quintuplicate hits that popped up in the original test (www.jasco.info/cuil), in which 35 hits were promised, 16 were presented, and merely three were unique out of those.

The ratio is also better now for Cuil, as out of the eight hits promised and six delivered, four are unique, and “only” one is a triplicate. The first hit (from Wikipedia) may not look immediately identical to the other two items (that I marked with a border) because for that first hit Cuil extracted a different passage from the page, but all three marked hits are functionally—if not URL-wise—identical, from mirror sites of Wikipedia.

Eight hits reported, six presented—with one triplicate

Of course, hit counts reported by most, if not all, search engines are worth as much as your stock shares in the miserable financial institutions whose nauseatingly overpaid and incompetent managers managed to drive into bankruptcy. The gurus at the search engine companies claim that the discrepancy between reported and actual hits is due to the fact that the reported counts are only estimates.

Interestingly, these are always overestimations, never underestimations. Cuil cannot use this excuse, as it manages to overestimate the hit counts even when it finds only a single page, and by Pavlovian reflex, it reports three. Quite tellingly, Google reports 686 hits and shows 293, and Yahoo! reports 2,680 hits and shows 352 for the query `jacso scopus`, but their confabulation is not as obvious as in Cuil.

One cannot stop wondering how Cuil could be so much larger than any other search engine when its hit counts are vastly smaller than those of Yahoo! and Google for almost all the test queries, even 2 months after its debut.

Beyond its claim about its size (rhyming with the mantra of the dot-com bubble era of “get large or get lost”), the claim from the developers that, “Rather than rely on superficial popularity metrics, Cuil searches for and ranks pages based on their content and relevance,” also seems to be like a page from the tales of Baron Munchausen.

The presentation of the hits does not reflect functioning relevance ranking, as the duplicate pages are scattered throughout the results. If relevance ranking really worked, such records would line up as do passengers at a bus stop in London, where it is decent to form a queue when there are two or more people.

These problems are dwarfed by other, essential shortcomings in the software. Cuil only recognizes the Boolean AND; there’s no OR, NOT, or phrase search. Limiting the query to the title field, a domain, document type, or language is also impossible. The insertion of images with the text snippet is often absurd, as the image doesn’t match the snippet. Beyond being laughable, it may also represent liability issues when someone will not be as “understanding” as professor Grattage, whose bio was paired up with a pornographic image by Cuil’s pathetic matchmaker.

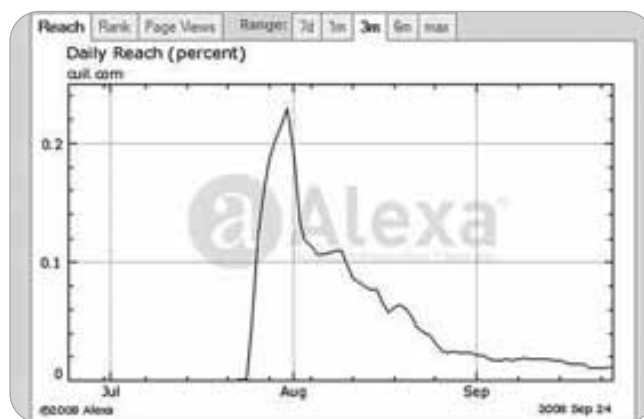


Three hits reported—one was actually found.



Cuil only recognizes the Boolean AND; there’s no OR, NOT, or phrase search. Limiting the query to the title field, a domain, document type, or language is also impossible.

Cuil developers and PR specialists may dismiss the “superficial popularity metrics” that makes Google so efficient, but they have to face harsh reality outside the fantasy world of the Googleplex. They had the fastest rate of burning money, to the tune of \$33 million. And the venture capitalists who financed this exercise will certainly wonder how and why could Cuil’s “popularity” plummet so steadily from its debut, to hit bottom and stay there. As of the end of September, 2 months after the launch, Alexa paints a very dark picture with its graph for Cuil usage (www.alexa.com/data/details/traffic_details/cuil.com).



A graph is worth a thousand words about Cuil.

The developers of worthy services such as ReleMed, GoPubMed, and many others with no venture capital but only their own money, time, talent, brains, and labor of love can feel pride. The Cuil people can sing “Money can’t buy me love” where “love” is replaced by “popularity.”

Péter Jacsó (jacso@hawaii.edu) is professor of library & information science at the University of Hawaii’s department of information and computer sciences.

Comments? Send email to the editor (marydee@xmission.com).