With the amount of published research, patents, white papers and other written knowledge out there, it’s hard to be even reasonably sure you’re aware of the goings-on around a certain topic or field. Omnity is a search engine made to make it easier by extracting the gist of documents you give it and finding related ones from a library of millions — and now supports more than a hundred languages.

The process is simple and free, at least for the public-facing databases Omnity has assembled, comprising U.S. patents, SEC filings, PubMed papers, clinical trials, Library of Congress collections and more.

You upload a document or text snippet and the system scans it, looking for the least common words and phrases — which generally indicate things like topic, experiment type, equipment used, that sort of thing. It then looks through its own libraries to find documents with similar or related phrases that appear in a manner that suggests relevance.

For example, say you put in the results of your clinical trial testing a food additive on a certain strain of mice, and found it resulted in a certain condition. Omnity would return documents describing other tests of that additive, on mice or other animals, or unrelated tests that produced that condition, with no need for you to specify the important aspects or drill down. The similarities and connections between your document and the results are presented in a nice pretty graph, as well.

For example, say you put in the results of your clinical trial testing a food additive on a certain strain of mice, and found it resulted in a certain condition. Omnity would return documents describing other tests of that additive, on mice or other animals, or unrelated tests that produced that condition, with no need for you to specify the important aspects or drill down. The similarities and connections between your document and the results are presented in a nice pretty graph, as well.

This portion of Omnity has been operational for some time, but yesterday the company announced that it was expanding its system to encompass more than a hundred languages. So you can put in research papers or filings in Chinese, Russian, Arabic, etc. — and it will conduct the same process in a cross-lingual way and return relevant results.

The process works the same for documents in other languages, but Omnity knows that a phrase in French is the equivalent of a phrase in English, even though it may not grasp the subtleties of the translation process. It still knows, and it still comes back with the right docs.

For now the database is focused on English-language repositories, but CEO Brian Sanger told TechCrunch in an email that the company is “in the process of international expansion. Enabling individual foreign language documents is the first step in that process, and we will be adding non-English documents over time.”
The service is free, so you may be wondering how these people make money. As with so many other companies, private customers pay the bills. The public website is the lure, showing that the system works with a large — 15 terabytes, at present — database. But the system can also be deployed internally at a company, for example at a law firm that must track millions of documents and court cases and have them ready for instant recall.

“A single enterprise customer may have hundreds of terabytes to tens of petabytes, a scale 1000X or 10000X greater than all public documents combined,” Sanger wrote. “This is where our revenue model is focused, cross-connecting and auto-classifying custom ingested documents at scale.”

The method is similar in some ways to Semantic Scholar, another machine learning-powered search engine that extracts meaning from text to make documents more easily searchable and categorizable, but Omnity's approach is a little more abstract.

“Omnity makes mathematical equations that describe statistical patterns of rare words in discrete documents, without regard to work proximity, and without regard to grammar or grammatical content,” wrote Sanger. Semantic Scholar, on the other hand, understands grammar the way we do and uses it to draw meaning out of the text.

It’s an interesting juxtaposition: two AI-powered search engines that, despite similarities, are still very distinct. Perhaps it’s a look into the next generation of search.