Copyright Clearance Center | RightsDirect
A Copyright Clearance Center Subsidiary

# The Limitations of Keyword Search

*Using Semantic Search to Uncover Scientific Meaning*

## KEYWORD SEARCH IS THE NORM

R&D and information management professionals routinely employ simple keyword searches or more complex Boolean queries when using databases such as PubMed and Ovid and search engines like Google and Google Scholar to find the information they need. While satisfying the basic needs of the researcher, keyword search has limitations which can negatively affect both precision and recall — reducing productivity and hindering researchers' ability to discover new insights.

## WITH KEYWORD SEARCH, YOU GET WHAT YOU ASK FOR

Keyword searches are quite literal in the sense that computers find terms wherever they appear—even if part of a larger phrase or used in a different context. This approach is effective if the researcher knows exactly what they are looking for. The problem is words often have multiple meanings, so keyword searches often return irrelevant results (false positives), failing to disambiguate unstructured text. For example, a keyword search on the term "AIDS" might include not only references to "Acquired Immune Deficiency Syndrome" but also hearing aids.

Keyword searches may also fail to turn up related materials that don't specifically use the search term (false negatives). Under these conditions, researchers can miss pertinent information. There is also the danger of making business decisions based on a less than comprehensive set of search results.

# KEYWORD VERSUS SEMANTIC SEARCH

Red represents keyword search hits for the phrase "adverse events." Green represents missed hits for this keyword search that a semantic search for the *concept* adverse event would produce, improving recall (comprehensiveness) without requiring the searcher to explicitly include search terms in their query for every adverse event of interest.

complication rates over 6 times higher (>28%). Patients who had complications were significantly older then patients who did not. The most common adverse events were dysphagia and cardiac complications. The most severe morbid complications, in terms of increased treatment needs and hospital stay, were paraparesis and seizure. Conclusions: Perioperative complication rates in cervical spine surgery are significantly lower in younger patients, surgery performed through an anterior approach (compared with a posterior or combined approach), with fewer levels involved (particularly in prosterior surgery), and in primary (compared with revision) procedures.

In the example below, green represents a valid hit that a semantic search would provide. Red represents false positives that a keyword search for the concept "AIDS" would provide reducing precision and resulting in more irrelevant hits.

| |
|---|
| Methods: The study enrolled patients with chronic, severe pain of nonmalignant etiology with a demostrable neurological basis, or pain related to cancer or AIDS or their treatment. |
| Additionally, dye study employing live fluoroscopy or DSA can aid in planning of surgical revision by pinpointing the location of the leak, potentially eliminating other possibilities of leak (retrograde CSF leak into the pocket from a catheter fracture, or dural leak). |
| He was unable to walk more than 5 minutes without pain and used a walking stick aid as he could not bend his leg at the knee. |
| At 3 months post implant, pain was reported as 0 VAS for knee pain, with continued improvements to sleep and function, including the cessation of the walking stick aide. |
| Our method can be further developed to reliably differentiate between effective and non-effective contacts and aid DBS programming. |

While keyword search may recognize plurals, variations and stemming (connecting a text string to other related text strings, as in *fish, fishing, fished*), thorough queries must still account for every term and permutation. Researchers often maintain highly-complex queries that require constant refinement. Compounding the issue, keyword searches get more complicated when users want to go beyond co-occurrence to identify potential causal relationships. Searchers are all too familiar with sifting through lists of document results that contain all required search terms but lack a clear conceptual connection between them. These are a special kind of false positive, and a case where semantic search has the clear advantage.

Vocabularies that exploit linguistic relationships — such as verbs indicating that a particular biomedical entity is related to another, as when one upregulates, downregulates, inhibits, or disinhibits the other — can be applied in semantic search effectively to further limit results and achieve greater precision. This not only increases accuracy and saves time, it also builds user trust.

## SEMANTIC SEARCH LOOKS AT MEANING

Unlike keyword search, semantic search takes into consideration the researcher's intent to get at the contextual meaning of terms. Semantic search pushes beyond the boundaries of the organization's collective base of understanding to get at information and concepts that haven't been explicitly written into the query. Semantic technology deciphers concepts and meaning by associating search inputs with clarifying terms such as related synonyms that have been built into the system. For example, a search for the common drug brand *Lipitor* would surface any documents also mentioning *atorvastatin*; a search for *cancer* would yield documents discussing types of cancer such as lung, breast, or brain.

This is possible because of the process of semantic enrichment which helps researchers find new ideas and concepts by tagging unstructured text with information about its meaning. To distinguish concepts, semantic technology references vocabularies that contain all known terms for the same thing, relates these entities to each other in hierarchical relationships, and employs algorithms to analyze the context within which those terms appear.

Semantic search greatly improves precision and recall, giving you and your colleagues in R&D the most comprehensive and relevant set of results to help extract new insights, accelerate discovery and guide business decisions.

## APPLYING SEMANTIC SEARCH ACROSS THE ORGANIZATION

In R&D-intensive industries such as the life sciences and chemical manufacturing, semantic search delivers value in several areas:

### Early Phase Research

- Researchers can discover interesting potential biomarkers and drug targets they hadn't known to look for in advance. These initial results can be linked to supporting source content for further review prior to wet lab.

### Competitive Intelligence

- Competitor patent filings, often intended to hinder discovery, can be explored alongside non-patent literature (NPL) to provide a full picture of competitor strategy, claims, and prior art for patent landscaping or other purposes.

### Pharmacovigilance

- Literature monitoring for pharmacovigilance can become both more comprehensive and more precise through semantic searches that suggest links between adverse events and pharmacological substances, increasing the efficiency of these vital monitoring workflows.

### IDMP (Identification of Medicinal Products) Compliance

- Semantically enriched internal and external content can provide a fuller view of medicinal product attributes, supporting IDMP compliance.

### Discovery of Chemical Compounds

- Researchers can take advantage of well-established chemical ontologies to conduct more efficient semantic search for chemicals, more easily identifying relevant chemical compounds and their properties and relationships.





Copyright Clearance Center

Copyright Clearance Center (CCC) is a global leader in content management, discovery and document delivery solutions. Through its relationships with those who use and create content, CCC drives market-based solutions that accelerate knowledge, power publishing and advance copyright. With its subsidiaries RightsDirect and Ixxus, CCC provides solutions for millions of people from the world's largest companies and academic institutions around the world.

RightsDirect

A Copyright Clearance Center Subsidiary

RightsDirect provides content workflow, document delivery and rights licensing solutions that allow companies around the world to use, share and store content while simplifying copyright compliance. Working with partners around the world, RightsDirect offers sophisticated solutions tailored to the needs of national and global organizations. RightsDirect is a wholly-owned subsidiary of Copyright Clearance Center (CCC) based in Amsterdam with a presence in Tokyo.

### LEARN MORE

Learn how the integration of RightFind® XML for Mining and SciBite DOCstore can help improve the results of semantic enrichment initiatives, reduce costs and simplify copyright compliance.

For U.S. inquiries:

**info@copyright.com**

**+1.978.750.8400 (option 3)**

**www.copyright.com/SciBiteDOCstore**

For inquiries outside the U.S.:

**info@rightsdirect.com**

**+31-20-312-0437**