

BASIC CONCEPTS FOR DATABASE SEARCHING

By Marvin Hunn

There are two common approaches to learning about database searching. One approach stresses software features (e.g., “how can I search for an exact phrase or limit results to one language?”). But search engines change, and the AI revolution is bringing major change that requires deeper understanding. Another approach to learning is to focus on basic concepts that are likely to retain their importance as software changes. But that leaves the student unfamiliar with interfaces. We think both approaches are needed.

This basic introduction stresses concepts but sometimes deals with interface details. We begin with a section on database content, follow with a section on traditional database search engines, and finish with a section on AI-powered search engines.

Documents and Databases

Source document. For our purposes, this means a recorded intellectual work that can be a source of information. The work can be any length and can use any recording technology. For example, a source might be a print book, a digital journal article, a streamed video, a microfilm, a hand-written manuscript, or an ancient clay tablet.

Metadata. Bibliographic metadata is information that describes a source, such as author, title, date, subject, and link-to-source. So we distinguish data (sources) and metadata (descriptive information about the sources). Metadata is often organized into records, one record per source, and the records organized into a database. Here is a simplified metadata record with typical fields labeled.

Author: Bibfeldt, Franz
Title: John Knox and the British Reformations
Publication info: Dallas : Nonesuch Press, c2001.
Series: Studies in Reformation history
Subject: Reformation—Great Britain
Subject: Great Britain—Church history—16th century
Subject: Knox, John, ca. 1514-1572
Link: <https://nonesuch.com?id=123456>

Search engine. Search engines are computing systems that attempt to find source documents whose contents satisfy the conditions specified in the search statement. Google and EBSCO are examples.

Searching metadata vs searching full-text documents. Many databases allow you to search metadata records; other databases allow you to search all the words in the documents (“full-text”) or to search both metadata and the full-text.

Paywalls. A paywall is a mechanism that requires a user to pay in order to obtain access to content. Internet search engines like Google and Bing can’t provide access to sources behind paywalls. (Well, that’s not quite true; an enormous amount of illegally pirated content is available on the dark web.) Commercial databases legally and ethically provide access to content behind paywalls.

Keyword. In the 1800s librarianship, “keyword” often meant a significant or important word used for indexing and subject access, but in the modern context it means any searchable word. It might appear anywhere in a record or source. It need not be an important word.

Controlled vocabulary database. This is a database which uses standardized terminology in metadata records to describe sources. Standardization is meant to guarantee a name or concept is always expressed in a consistent way. Standardization lessens problems caused by variant spelling (e.g., Koran or Quran or Qur'an) and variant forms (e.g., J Smith vs John Smith) and synonyms (e.g., anger vs wrath). So controlled vocabulary supports consistency and cross references from non-standard terminology to standard terminology.

Controlled vocabulary subject headings. Descriptors and subject headings are terms (words and phrases) assigned to sources to indicate the subject/topic. These terms are drawn from a thesaurus which lists standardized headings.

Tagged text. A source is “tagged” if controlled vocabulary metadata identifiers have been inserted into the text to provide information about portions of that text. For example, every word might be grammatically analyzed (e.g. this word is a verb, first person singular, and the lemma is xzy) or geographically analyzed (e.g., this word refers to a city in Israel and here is a pointer to its location on a map). Syntactical and literary structure might be tagged (e.g. this is a verb phrase; this is an independent

clause; this is a paragraph; this is a pericope within a longer work). Greek and Hebrew texts of the bible are available as tagged texts. You may also see the word “**treebank**” in linguistic contexts. A treebank is a corpus (collection) of tagged sources. The tagging typically includes information about phrase and clause structure, and is displayed as a tree structure of linked elements. So the “tree” part of treebank refers to the display of the tagged structure of a source, and the “bank” part of treebank refers to the collection of tagged sources. There are treebanks of ancient Greek text available for linguistic study, and there are some standards that govern how the texts are tagged. One example is the Diorisis Ancient Greek Corpus. The corpus is searchable.

Rank order. Most search engines rank sources in an attempt to show the best sources first. Rank order depends on a combination of factors like relevance, authority/accuracy, availability, etc. Sorting a retrieved set by a field value like date or author is not called ranking, but both ranking and sorting assign a display order.

Relevance of search results. The concept of search relevance is slippery, context dependent, and changes with time and personal perspective. Most often we search for sources “about” a certain topic and a searcher might say relevance is aboutness. Sometimes a searcher might say a retrieved source is relevant (pertinent, germane, apposite) if it provides information useful to the searcher. So relevance is usefulness. Sources may be considered relevant because they are helpful to the problem at hand even though they are not about the originally specified topic. For example, suppose a student is studying how a specific OT law is interpreted in the NT. He stumbles on an article about US constitutional interpretation which makes a point about “originalism” in modern legal hermeneutics. This makes him ask new questions about how the NT is interpreting the OT law. The student calls that useful article “relevant.” But it never mentions the bible. It is not about biblical hermeneutics.

Search precision and recall. Don’t expect a perfect search result. Search success is often assessed in terms of precision and recall. Precision refers to accuracy, and recall refers to completeness or thoroughness. Both are percentages. Precision and recall are usually defined as follows. If d = number of sources retrieved in a search, and R = number of relevant sources in the database, and r = number of relevant sources retrieved by the search, then precision of that search = r/d and recall = r/R . For example, suppose you search for information about women in the gospel of John. You retrieve 50 sources but only 25 are relevant. So search precision is $25/50 = 50\%$ (good). However, the database

actually contains 250 sources truly on topic, so the recall is $25/250 = 10\%$. Precision and recall are inversely related; a high precision search is usually a low recall search. Often you can design a search to be high precision or high recall, but it is usually difficult to execute a perfect search (high precision and high recall).

Traditional Search Engines

Word search versus concept search. Traditional keyword search engines search for words, not concepts. This is important because one word may have many meanings (polysemy), and two different words may have approximately the same meaning (synonymy). Knowing how to discover good search terms is the key to successfully searching a traditional database. See “Search Tips 1” for an explanation of how to discover good search terms.

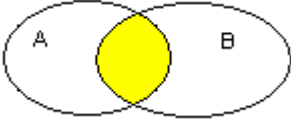
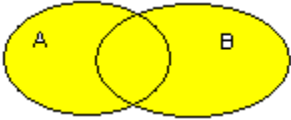
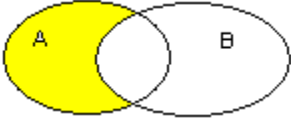
Field-specific searches, and field codes. Most database interfaces provide search forms with some means of selecting the fields to search.

calvin	⊗	Author - AU	∨	
AND ∨	institutes	⊗	Title - TI	∨
OR ∨			Subject Terms - SU	∨

It is also possible to use field codes. << AU calvin AND TI institutes >> means search the author field for the word ‘calvin,’ and the title field for the word ‘institutes.’ Field codes are usually two letter abbreviations. Field codes can be used to formulate complex search statements that can’t be formulated with a simple form like we have here.

Search operators. Many search engines require/allow you to specify not only search terms, but also information about how to relate those terms to each other. Operators are commands that tell the search software how to relate terms. Logical operators and proximity operators are the most common. See following description.

Logical search operators. Most search engines support the logical operators AND, OR, NOT (also called Boolean operators after George Boole, the mathematician who popularized their use in set operations).

Operator/ function	Common symbol	Example	Explanation	
Intersec- tion	AND	A AND B infant AND baptism		'AND' retrieves records containing both terms. You will commonly AND concepts to narrow a search.
union	OR	A OR B clergy OR pastor		'OR' retrieves records containing either term. You will commonly OR synonyms to broaden a search.
exclusion	NOT	A NOT B spirit NOT holy		'NOT' excludes records containing the second term. It is easy to accidently exclude desired material.

Proximity search operators. Many search engines support proximity operators. Proximity operators specify how far apart matching words can be. They may also specify word order. Details vary considerably system to system. Proximity operators tend to be more precise than logical operators, and more appropriate for full-text searching. Examples follow.

Operator/ function	Common symbol	Example	Explanation
proximity	NEAR + number N + number	spirit NEAR4 filled spirit N4 filled	'NEAR' specifies maximum distance between words in the same field, in any word order. The example specifies a maximum distance of four words (NEAR4). It matches "Spirit filled" as well as "filled with the Spirit". Systems count distance differently. Some say the distance between spirit and filled in the exact phrase "spirit filled" is zero , but others say the distance from spirit to filled is one word.
Exact phrase	" . . . "	"spirit filled"	Most systems will interpret quotation marks as an exact phrase operator. Use straight double quotes ("spirit filled"), not curly quotes ("spirit filled"), not single quotes ('spirit filled').
Word order specific proximity	W + number PRE/ + number	big W4 dog big PRE/4 dog	This operator is similar to NEAR: it specifies maximum distance between words in the same field, but it also specifies word order. Use it to allow intervening modifiers. For example, <<big W4 dog >> matches 'big brown dog' as well as 'big dog.'

Truncation operator. This operator allows partial word matches by truncating (cutting off) part of the word, usually the final part. Truncation is usually indicated by * (asterisk). For example, in some systems bapt* matches any word that starts with the four letters "bapt"(such as baptism and baptist). It is easy to accidentally include undesired material when you truncate. Some systems have "wildcard" or character masking operators that can be used in the middle of a word (e.g., wom#n to match woman or women).

Stemming and lemmatization. Linguistic stemming attempts to identify the root or stem of a word. For example, search for the word "baptize," and the system determines the stem is "bapt-" so it matches baptism, baptized, baptist etc. It may also match Anabaptist and rebaptize, depending on how stemming is performed in the particular system. Since English words are mostly inflected with suffixes, stemming is often equivalent to truncating the end of a word. (But this is not so in Hebrew, for example, which regularly inflects with prefixes and suffixes and infixes). Lemmatization attempts to identify the base or dictionary form of a word, which is known as the lemma. For example, "be" is the lemma for all of the following: are, am, is, was, were. Stemming and lemmatization are features of some special purpose software. For example, Accordance and Logos support lemma searches of the Greek and Hebrew texts of the Bible.

Default operator. The search engine will use the default operator if the search statement does not explicitly indicate how to relate multiple search terms. For example, if you type a two word search statement like << word1 word2 >>, the software might AND the terms or look for an exact phrase or do something else. The default operator for EBSCO is n5 (near five). The default operator for WorldCat is AND.

Filters. Database filters acts like a sieve, limiting search results to the selected categories. For example, it is common to have filters for language of publication, date range, and format (book, movie, whatever).

Grouping and nesting. Many retrieval systems allow the searcher to use parentheses to group terms and specify the order in which search operators are to be executed. For example, consider this search:

<< brown AND dog OR cat >> It is pretty clear that the dogs must be brown. But what about the cats? Most systems would retrieve cats of any color. To specify both cats and

dogs must be brown we would use parentheses like this: brown AND (dog OR cat) This specifies the OR operator is to be executed first, creating a set that holds the results of the dog OR cat. Then that intermediate result is to be AND-ed with brown to yield a final result. Nesting refers to embedding one set of parentheses within another set like this: A OR (B AND (C OR D)).

Search statements. A search statement is a combination of terms, operators and options that constitute a search. For example, (law OR covenant) AND (romans OR galatians) is a search statement.

Mis-coordination is an undesired semantic relation between matched words in a multi-word search. Consider the guy who wanted to buy car polish (a compound used to preserve car paint). He searched for car AND polish. He found information about a Polish car (car manufactured in Poland). Mis-coordination is very common when logical AND is used to combine words. It is less common when exact phrases and proximity operators are used to combine words. In this example we could use an exact phrase operator to improve our search results: "car polish." Even when words are syntactically related the way we want, they may still fail to be semantically related the way we want. In this example we have two different words spelled the same way (polish and Polish). This is a good example of why searching requires more than matching character strings; it requires matching meaning/concepts. (Mis-coordination is not a standard term; I am using it to cover what the IR people call "false coordination" and "incorrect term relation" and some similar retrieval problems.)

Stopwords or noise words. These are specific words that a system does not search for. To save time and storage space, some retrieval systems ignore some very common words that carry little meaning such as conjunctions, prepositions, and articles. The assumption is that you do not need to search for "a" or "in" or "the." But you do. ("A" is necessary in a search for Vitamin A; "in" is necessary for a search for the phrase "in christ." "Thé" is French for tea, and there was once a band named "the the.") Some systems will search for a stopword if you put it in quotes.

Browsing vs Searching. Some databases support two primary ways of finding material: 1) searching for words or 2) browsing a list of headings. Searching retrieves records by matching combinations of words. Looking for all records with the phrase "Jesus Christ" is a search. Searching is what you normally do. Browsing is fundamentally different. It

is a two-step process. First, you begin by supplying a word or phrase you expect to match the start of a field. The database then displays a sorted list of headings that start with those words. Second, you pick one or more specific headings from the list, and matching records are displayed. For example, start with Jesus Christ. The system responds with a long list that begins like this:

Jesus Christ -- Ascension

Jesus Christ -- Authority

Jesus Christ -- Baptism

Jesus Christ -- Betrayal

Jesus Christ -- Burial

You then pick an entry from this browse display.

Auto-suggest. This software feature automatically displays a list of terms (phrases) that are similar to the terms the searcher is typing. For example, in Google start to type << charitable income tax deduction >> and before you finish typing the system suggests the following.

charitable income tax deduction limits

charitable income tax deduction carry forward

charitable income tax deduction calculator

Auto-suggest is similar to browsing. Both involve a two-step process and both display a sorted list to help you select terms. But the lists are not created in the same way. Auto-suggest is usually based on a history of search statements supplied by other people, expanded by common phrases in the database. Browsing is based on fields that exist in records in the database. It assumes structured metadata records with fields and controlled vocabulary. So it is not possible to browse unstructured web pages.

AI-Powered Search Engines

Most database vendors are gradually adding Artificial Intelligence (AI) features to their traditional keyword databases. This includes EBSCO, ProQuest, JSTOR and others. We are writing separate documents about search engines from those vendors. Here is a general introduction to basic concepts.

Anthropomorphic language. AI systems are not intelligent. They don't think about anything. They don't understand anything. They don't intend anything. They are just algorithmic systems. But many people use anthropomorphic language in reference to AI

systems because the algorithms produce human-like results. Don't be deceived by the anthropomorphic language or the amazing results.

Answer Bots and AI-powered search engines. Most people associate AI with the Big Four general purpose Answer Bots: OpenAI's ChatGPT, Google's Gemini, Anthropic's Claude, and Microsoft's Copilot. These question-answering systems are used by billions of people. They are popular for two main reasons.

1. The Answer Bots seem to know everything about everything.
2. The Answer Bots use a natural language interface, so communication is easy. In addition, they can dialog (chat). This back and forth format allows progressive refinement of questions and answers.

We are interested in AI-powered search engines. Often the Answer Bots cite a few sources to support their answers. But they are not optimized for searching. Perplexity is a key exception; it is optimized for searching. And traditional database search engines are rapidly adding AI capabilities.

What AI-powered database search engines can do. At present, traditional library database search engines are gradually implementing the following AI-powered features.

- Some library search engines will accept your natural language description of what you want instead of forcing you to formulate a traditional search statement.
- A few search engines use semantic search technology to search for concepts, not just words.
- Many AI-powered search engines can summarize a source to help you determine if it contains relevant information.
- Some can translate a source for you. Copyright controls the right to distribute translations, so this feature is limited by copyright license arrangements. So you might be able to translate one article but not another.

Natural language queries vs traditional search statements. Natural language queries allow the searcher to use ordinary English (or other natural language) to tell the system

what the searcher wants to retrieve. For example, you might give the machine this prompt: << why do clergy quit the ministry? >>. Contrast that with traditional keyword search engines which require the searcher to use special syntax and operators to express a search statement, e.g., << (pastor OR clergy OR minister OR priest) NEAR10 (leave OR resign* OR quit OR attrition OR depart*) >>.

During the present transitional period, we are seeing **hybrids** which use AI-powered natural language interfaces to aid traditional keyword search engines. The hybrids convert a natural language query into a traditional search statement and then execute a traditional search (as with the clergy attrition example above). So the hybrids are not using semantic search; they are using traditional keyword search. As of 2025, both EBSCO and ProQuest provide hybrid systems. See [Natural Language Searching in EBSCO Databases](#) for practical tips about how to deal with hybrids.

Semantic search technology includes a variety of techniques to search for concepts, not mere words. It is a big advance over traditional keyword searching, but it is not a total replacement; keyword is still sometimes the right tool to use.

At the present time, semantic search usually uses a **vector space model of semantic relations**. If you want a deep understanding, you must deal with math and some advanced linguistic concepts. Check the footnote.¹ But here is a simple description. The vector space model represents semantic similarity as geometric distance. The vector space model represents the contextual meaning of a word, phrase or larger portion of text as a vector in multidimensional linguistic space. . The vectors indicate direction and

¹ Here are some advanced sources about the Vector Space Model; these sources are in the DTS library. Lenci and Sahlgren (*Distributional Semantics*, 2023) explain how meaning can be modeled through usage patterns. Widdows (*Geometry and Meaning*, 2004) discusses the mathematical basis of information retrieval and semantic search using vector spaces. Kornai (*Vector Semantics*, 2022) links vector theory and cognitive linguistic theory. Gärdenfors (*Conceptual Spaces: The Geometry of Thought*, 2000) and (*The Geometry of Meaning: Semantics Based on Conceptual Spaces*, 2014) discusses a model of how human knowledge can be represented in the human brain as geometric structures in high-dimensional spaces, just like the vector space model AI uses for semantic search and other AI magic.

distance in the space. Vectors with similar meanings are neighbors in the same region of space, while unrelated vectors will be far apart. The model creates a multi-dimensional map of how words are related to each other. A model might use 500 or 1,000 dimensions; that combined with a large vocabulary might result in billions of parameters to distinguish meaning.

Knowledge graphs. While vector space is the key to modern semantic search, knowledge graphs and other tools also contribute to semantic search. A knowledge graph is a network of real-world entities (people, places, things) and their interconnected relationships. It is organized to support inferences. For example, if John is the father of Mary, and Mary is the sister of Fred, the graph might infer that John is the father of Fred. (Of course casual language might not distinguish half-siblings from full-siblings, in which case John might not be the father of Fred. Knowledge graphs that deal with casual language sometimes make false inferences.) Knowledge graphs may work with other tools. Thus knowledge graphs can be organized based on ontologies which can be derived from patterns observed in data. So that is three tools working together.

Problems with AI-powered search engines

- AI systems have access to everything on the open internet, but they typically do not have access to copyrighted digital content sequestered behind paywalls. Likewise they typically do not have access to content that only exists in print. That is important because much scholarly material is behind paywalls or available only in print. So it no surprise that AI-powered systems sometimes cite biased sources like press releases or company blogs. This goes back to their limited access to scholarly sources. (However, they all now purchase access to some scholarly sources at least for training purposes. They are making progress.)
- AI systems are trained to answer based on probability. If an AI has access (during training) to many sources that answer a given question, then it will usually generate an answer consistent with those sources. But when there are few or no relevant sources that answer a question, the bot will still answer because it has been trained

to guess rather than admit “I don’t know.”² In such cases, the bot will depend on very low probability patterns, and this can result in false or nonsensical claims. It can result in references to people and books that do not exist. Such false claims are called "hallucinations."

- General purpose Answer Bots learn during a training period. They do not keep track of what sources provided what information during training. Once training is complete, and they are used for production, they are sometimes able to answer but unable to find sources to support what they learned. When a source they can access cites a source they cannot access, they often provide incomplete citations. The bots tend to produce short bibliographies.
- AI tools designed to function as search engines (like Perplexity and Elicit) have some special virtues. In particular, they usually cite more scholarly sources, they cite accurately, and they follow a transparent and reproducible process. Google is a special case. The AI mode of Google is based on their general purpose Answer Bot (Gemini) but it has been adapted to searching, and we think it is more helpful for a search of open access sources than other general purpose Answer Bots. This could change at any time.

Problems related to the vector space model

- The vector space model is based on symbols (words) organized according to language patterns to approximate a network of concepts. The vector space model treats the entire linguistic system as independent and self-contained. The meaning of a word in a particular context is defined in terms of its relations to other words (“distributions”). Symbols (words) define other symbols. There is no grounding in the real world. This is called the "symbol grounding problem." Grounding is particularly important for inferences.
- The vector space model claims semantic similarity is like geometric distance. There is some theoretical and empirical grounding for this claim, and empirical testing shows the model can be configured to work well, but the model is incomplete. For example, there are many different ways of calculating distance/similarity between

² See Adam Kalai, et al., "Why Language Models Hallucinate" (arXiv preprint, Sept. 4, 2025, <https://doi.org/10.48550/arXiv.2509.04664>) for more detail.

words in vector space (e.g., naïve Euclidian distance, cosine similarity, dot product, etc.) In practice, each method works better with some kinds or collections of data than with other kinds or collections of data. Theory does not specify one particular similarity function. It takes trial and error testing to find a similarity function that works well enough.

- The vector space model considers all words used in a prompt and in source documents. It also considers co-occurring words. But it does not consider “order-sensitive compositionality” (syntax or basic word order), or the distinction between reference and truth, or complex logical structure or rhetorical argumentation.
- The vector space model is just a way to organize data and relations; it is not inherently about meaning. For example, it can be used to organize non-linguistic patterns. Many face-recognition and general-image AI systems use vector space models to represent and process visual data. We associate words with semantic meaning but we don’t associate visual data with semantic meaning. This helps us realize the vector space model is just a mathematical model of patterns, and it can be applied to all sorts of things, but it is not inherently semantic.

Prompt engineering is the process of perfecting the instructions you give the AI. Here we assume you are using AI for some serious non-trivial research. In such cases, prompt engineering can help you improve you understand your topic. Here are a few suggestions.

- **Iterate.** Begin with a simple prompt, evaluate the AI response, and craft an improved prompt. This gives you an opportunity to read, learn, and shape the factors the AI considers. The process of refining the prompt is educational and it helps you develop clear and precise terminology to express key ideas.
- **Push back gently.** If you don’t like what the AI is saying, remember you might be wrong, the AI might be wrong, both of you might be wrong . . . In such a situation, you need to probe without biasing the AI. One way to do that is to quote a claim the AI makes, and ask if there are alternative views about the issue. (Don’t insist there are alternatives; ask if there are alternatives.) After it lists alternatives, tell it to evaluate each view. If you imply a certain view is right or

wrong, the AI might just accept what you say and supply what you want, not what is true.

What you can do to minimize hallucination. Hallucination is an especially important problem. AI hallucinations are most likely to occur when the AI doesn't know enough to answer the question properly. Here are two ways to reduce hallucination.

- 1. Upload sources.** Begin by uploading a source to the AI. It should be highly relevant to your topic, and trustworthy. Tell the AI this is a good quality source. The source may not answer the specific question you have, but it can provide context as the AI then tries to answer the specific question. Some AIs always begin by searching for trusted sources to provide context for your question. This is called **RAG (retrieval augmented generation)**.
- 2. Include anti-hallucination instructions in your prompt**
 - To reduce hallucinations, tell the AI to say "I don't know" when it is uncertain.
 - To detect hallucinations, make the AI self-audit. Ask the AI to generate a response, then prompt it to create questions to verify that response. Then ask it to check for errors and generate final corrected answers.

What software engineers are doing to minimize hallucination

The companies that design and build AI products use reinforcement learning (RL). This is a form of trial and error with rewards for good results. The process acts similarly to a multiple-choice test with no penalty for wrong answers. In this scenario, the best strategy to maximize the score is to guess, because a correct guess earns points, while a false or blank answer earns zero. A side effect of RL is that it can teach AI software to guess when it does not know the answer. This behavior is probably the primary cause of hallucinations. When asked about topics outside its training data, the AI attempts to fill in the gaps based on the patterns it has learned, fabricating details, links, or facts.

Companies are working to fix this by altering the reward structure. Training will heavily penalize the AI for being confidently wrong, and will offer partial credit for answering "I don't know" when unsure. In addition, companies are improving use of Retrieval-Augmented Generation (RAG). They will work to provide verified data

sources to provide context. This will reduce the need to rely on guessing. Note traditional library databases and full-text collections could become the main source of RAG content for AI.