

The Law Is Where You Find It

by David Post

On-line research is not just a fancy new tool. By changing how and where we find legal information, it is changing the very substance of the law.

On-line research is not just a fancy new tool. By changing how and where we find legal information, it is changing the very substance of the law.

mentally new ways of finding legal information may be changing the very substance of the law, and perhaps even the notion of "the law."

CONTROLLED BY CATEGORIES

Tennyson was correct: in 1793 law was indeed a "lawless science" because the information that constituted "the law" was, in his day, largely inaccessible. In a decentralized common law system, which relies upon precedent for its decisions, information about what the decision makers have done in particular cases is not available to others, it doesn't exist: it's the tree that falls in the forest when no one is around to see it. And that precisely the situation prior to the advent of a system for distributing reported decisions—a system that really took shape, in this country, during the middle and latter parts of the nineteenth century, with the development of the West National Reporter System, the first comprehensive collection in standardized format of all of the decisions of all of the nation's courts. As that database grew (and it grew rapidly during the late 1800s), no one could possibly keep track of the information that was growing there: some way to locate information within the database was required, a need soon filled, of course, by the West Digest system. When I describe the West Digest system to some of my current students, I am met with puzzled looks. For those of you who, like them, have never had the exquisite pleasure of doing legal research with the digests, a brief description is in order. The digests were nothing less than a comprehensive index of every proposition of law in every reported case. The original classification scheme grew from what West called the seven "grand divisions"—persons, property, contracts, torts, crimes, remedies, and government—and encompassed over 400 main topics ("criminal law," "copyright law," and the like) and thousands of subdivisions (often layered three or four layers deep). Each case from the National Reporter System was divided into individual propositions of law (the "Headnotes"), and each proposition of law in every case decided by each tribunal was then assigned to one, or more, of these subdivisions. The digests then collected all propositions from all cases under each of the subdivisions, so that, in theory, one could find what every court had ever said about the subject at hand.

The development of the West Digest was largely the work of a small number of men (Benjamin Abbott, John Hudson, G. Fred Williams, and John Mallory) prior to the turn of the century, and it was surely a grand scheme, as grand in its way as the Linnaean system for classifying the animal and plant kingdoms (and the result of much the same intellectual impulse). Without this remarkable information retrieval system, the Reporters themselves—the comprehensive database of cases—would have been almost unusable, and the system's influence on the structure of the law was profound. But it was far more than just a convenient way to find the law; in many ways it was the law. West referred to it, and many of us thought of it, as a kind of map of the landscape of the reported decisions, but if it was a map it was a very curious map indeed, for unlike most maps it actually helped to shape the very features of the landscape that it was supposedly mapping. An ordinary map can be "wrong"—it can incorrectly describe the features of this physical landscape by putting, say, something in the east quadrant when it is really in the west. But the West Digest had a kind of self-perpetuating quality; if it put something in the east, it pretty much stayed there forever. As Robert Berring, professor of law and law librarian at Boalt Hall, has written:

It is commonplace among information scientists that an unindexed or misindexed piece of information is forever altered or lost... If cases are only available through the West Digest system, then only that intellectual structure for understanding the interrelationships of law can prevail. Like it or not, practitioners and researchers internalized the West structure, and it became the skeleton upon which the rest of the system was built. To illustrate, in 1986 I clerked for then-judge Ruth Bader Ginsburg on the U.S. Court of Appeals for the D.C. Circuit, and worked with her on a case, Marvin v. D.C. Metropolitan Police Department, 812 F.2d 1425, over which we struggled mightily. The case involved a "Bivens" claim of unconstitutional conduct lodged against a member of the D.C. police force, and raised difficult questions about the way that courts would evaluate such claims on the pleadings. About a year after the decision was handed down, I had occasion to look for the decision in the West Digest, and found, to my surprise, that the key propositions of law for which the case stood did not concern "Civil Rights/Federal Remedies," or "Constitutional Law/ Remedies for Injuries," or "Constitutional Law/Personal, Civil,
and Political Rights/Personal Liberty and Security”—all perfectly plausible categories for this case—but rather “United States/Government in General/Liability of Officers and Agents.” If that was the only way you could find the case, then others would find it (and cite to it) only if they were looking for the law of the liability of U.S. government officers and agents, not civil rights law or constitutional law, and it would over time come to mean something about law of the liability of U.S. government officers and agents, not civil rights law or constitutional law.

ON-LINE LIBERATION
The advent of on-line searching has, of course, changed all of that. Now you enter "Bivens claim" and "summary judgment" and "immunity" and "government official"—or, even more powerfully, you enter a simple question: "What standards of specificity should a court apply to a Bivens claim when evaluating a motion for summary judgment based on absolute or qualified immunity?"—and you’ll find Martin, even if you have no idea what category your question falls into. No longer, in short, does an indexer stand between you and the proposition of law for which you may be searching. No longer does any single entity—West Publishing or anyone else—decide what is or is not "constitutional law" or "civil rights law" or "governmental immunity" law. Where once the legions of indexers compiling the West digests determined the meaning of individual cases, now we have literally an infinite number of ways to find any particular case and make of it what we will.

This is likely to have a profound impact on the very existence of legal categories like "constitutional law" or "civil rights law" or "law of governmental official immunity." We have in this country, for instance, no coherent body of "information law," and we are only now beginning to see people describe themselves as "information lawyers." We have, to be sure, courses at law schools on information law, but such courses will usually cover a little bit of "privacy law," some "intellectual property law," some "communications law," and the like. Information law is seen, in other words, as a kind of cross-disciplinary study, cutting across the "ordinary" categories within which legal information is organized. But where did those categories come from? It is something of a stretch, I admit, to claim that we don't have a coherent body of information law because Metes, Abbot, Hudson, Williams, and Mallory did not include such a category in their grand indexing scheme—but it is, I would suggest, not too much of a stretch. There was, in effect, no place within the digest structure for information law to live and grow, while it was not impossible to bring together legal propositions from these disparate categories to see if common principles bound them together, it was awfully difficult. But those categories are crumbling, because they are in large part a product of an information retrieval system whose day is past.
1. Organizing Information

You should be generally mistrustful of statements that begin “the world of (some complex concept) can be divided into two: …” We all know that almost every complex idea has a myriad of subdivisions and anything so simplistic as a binary division almost can’t, by definition, express the nuances necessary to capture the idea. I’ll hope you’ll suspend your disbelief here, though, and agree with me that the world of information organization can be divided into two parts: pre-indexed and self-indexed information.

Perhaps that’s even too broad a statement. Perhaps the reality is that information is either pre-indexed or is not organized at all—a jumble of random concepts until we impose some order on it by the type of search we conduct. But since chaos doesn’t lend itself too well to the imagination, let’s stick with my pre-indexed and self-indexed approach for now. At least that gives us two terms we can understand and work with.

A. Indexing

It’s probably important to note at this point that “indexing” is one of those terms that has more power for the older generations and less power—maybe less relevance—for your generation. When computers were unavailable for research (or for any purpose), and all research was conducted using books, the use of indexes was learned and indexes were used by students as second-nature; you couldn’t conduct research without them. Nor was it possible, except under unusual circumstances, to self-index material: everything was pre-indexed.

Computers, and especially the internet, have changed this. But almost as profound as the change to search capability is the change in the way we think about research. Whereas self-indexing was impossible until a decade ago, it’s now the dominant model of research in the non-scientific world; every time you search on Google or some other internet search engine you’re self-indexing the internet for a term or terms you’ve selected. Pre-indexes have fallen largely, if not completely—and very swiftly—out of favor.

We could debate the merits of these changes for a long time and not reach a resolution. But that’s not the purpose of this discussion. Rather, the point here is a simple but important one: as lawyers, you’re going to have to come to terms with both types of research—pre-indexed and self-indexed—and you’ll need to understand the benefits and detriments of both. This is new, and it will be more difficult for you than it was for lawyers of earlier generations to learn about self-indexing because the concept of “indexing” itself was familiar to them. In order for you to understand legal research, though, you will first have to become secure with the concept of indexing. Only then should you be confident in your research skills. Because older lawyers can’t imagine a research world without familiarity with indexing, many will have difficulty understanding any problems you might have in wrestling with the concept of indexing in the law.

Sadly, this is one of those generational problems you will experience from time to time, brought about by the rapid and permanent changes in the way we do things in this digital age. Rather than the generations getting frustrated with each other, it’s better to understand the cause of the issue, and it’s probably going to fall to your generation to understand and forgive the older generations for their confusion.
B. Pre-Indexed Information

Pre-indexed information is simply information that someone has indexed before the information reached you. If you go to any large reference work—one of your textbooks, for example—and look in the back, you'll see an index that lists the major concepts discussed in the book and the corresponding pages on which those concepts are discussed. Some large reference works, like paper encyclopedias, might not have indexes, but the information in them is organized in relation to a specific organizational principle—usually the alphabet—that allows the reader to find specific information quickly and easily. Those works are pre-indexed also, just in a slightly less obvious manner.1

In the legal context, as in almost all other fields, pre-indexing was the only meaningful way of conducting research in the twentieth century before the advent of computer-stored information. Imagine the alternative: you're asked to research an issue—let's say the statute of limitations for personal injury cases in New York; you go to the library and stand in front of the reporters containing cases from New York state; you pick a reporter (probably the most recent one, since you're researching current law), and start reading; and if you don't find the answer to your question in that reporter, you move on to the next one, then the next, and then the next, and so on until you find a case that discusses your issue.

That's analog self-indexing, and just describing this approach explains why it's impossible to put into practice. As with all sets of complex information, you need an index to help you cut quickly to the information you need. Thankfully, such an index exists. More than one, actually, but let's stick with the most commonly known and used one: the West Key Number system.

Some quick history here. Long before the corporate behemoth now known as West, or Thomson Reuters, or whatever its name is when you read this, came into being, and before its most famous product—Westlaw—was invented, there really was a Mr. West. His name was John B. West, and he was from Massachusetts, although he settled in St. Paul, Minnesota, in the late nineteenth century. This was a time at which more courts were writing more opinions than ever before, making it difficult for lawyers to keep up with the flow of information. West was a businessman who sensed an opportunity when it presented itself, and in 1879 he compiled together opinions from courts in Iowa, Minnesota, Michigan, Nebraska, Wisconsin, and the Dakota Territory. Within ten years, West was publishing opinions from all state and federal jurisdictions, and publishing them in volumes he called "reporters." But, of course, the information organizational problem remained; the opinions were gathered together in one place, certainly, but the amount of information in each reporter was still daunting and couldn't reasonably be self-indexed by anyone.

The answer came in 1897 with a series of volumes called "digesting sets." These contained brief synopses of each case, organized under West's Key Number system. This

---

1. This differs, of course, from Wikipedia, the encyclopedia with which you might best be familiar. The way in which you search Wikipedia—entering search terms—is a perfect example of self-indexing; you identify the term or terms you're interested in and the search engine pulls up the articles in Wikipedia that correspond to those terms. There is no physical form of Wikipedia (unless someone has undertaken the monumental and incomprehensible task of printing the whole thing out on paper), and so no need for alphabetical or pre-indexed organization.
is the indexing tool that has become familiar to every lawyer since the turn of the last century and which is still the dominant model for book-based legal research today.

We discussed the Key Number process in the previous chapter, but it’s worth reminding ourselves how simple and powerful an idea it is. The “unlocking” metaphor implicit in the use of “Key” in the name is accurate; once the reduction of cases to a collection of doctrinal slices was complete, every case—from the most recently published one back to the earliest decision—could be used to support a legal argument. Instead of facing an opaque shelf full of reporters, the researcher now has a transparent index to opinions that allowed for complex, accurate, and speedy research. West’s innovation was powerful, and it remains so today; the matrix that defines the nature of each subdivision has been refined over time, and new subdivisions are added as the law becomes more nuanced and concepts are refined by the courts, but the fundamental idea of Key Numbers and digests remains today and is the preferred mode of research by many lawyers.

There are many things to be said in favor of this approach. Most importantly, it reduces the possibility of making a mistake based on linguistic incompatibility. If you conduct an online search for cases involving “wife beating,” for instance, your search will not include cases that employ alternative phrases like “spousal abuse.” The concepts are the same (or closely related), but the words themselves are different. That happens all the time in the law. But West editors have read the cases, and they have likely made the adjustment necessary to slot cases that discuss similar legal concepts, but use different words to discuss and describe those concepts, into the same box, making them readily retrievable by the researcher.

That advantage to the system, though, is also the basis for one of the main criticisms of it. In order for the process to work, the researcher must make the same set of analytical assumptions as the editor. These often involve narrow distinctions between different subdivisions of the law in a particular area. If the editor decides a case should go in one box, and you search for cases in a different box, you won’t find some cases that might be relevant to you. And sometimes, despite all the safeguards and protections, editors make mistakes. Careful researchers can adjust for this problem, of course, by conducting searches in multiple boxes that appear to be related. Nonetheless, some cases likely remain undiscovered by researchers who aren’t sufficiently able to imagine the same path for allocating Key Numbers as are the West editors.

The larger criticism of the Key Number approach, though, is more conceptual. Imagine you have a large, multi-volume history of the Civil War, and you are interested in a comparison of how the Union and Confederate armies were able to supply their troops with food. You go to the index at the back of the final volume, only to discover that the history has not been indexed with the concept of food supply in mind. You read the entire history, and you discover that it does, indeed, contain numerous mentions of food supply issues, complete with detailed statistics. But these references are scattered throughout the multiple volumes, and the only way to gain meaningful information about your topic is to read every word in all volumes and to take careful notes of where and what you find, in essence, self-indexing the work for a specific concept. As you can imagine, this would be a time-consuming and frustrating experience.

So it is with cases indexed using the Key Number system. If the concept for which you are searching is not one for which the cases were coded during the editing process,
that information will be virtually invisible to you: it can be found in plain sight when you read a case that contains the information, of course, but the only way for you to know that is for you to read all volumes that might possibly contain that information and that, as we already know, is a practical impossibility. So while pre-indexing renders transparent the information that is indexed, it renders opaque the information that is not indexed.

One of the most significant pieces of information that is not contained within the West Key Number system is the name of the judge. Were you, for instance, to be working on a federal case that had been specifically assigned to a particular judge, you would have no way of researching how that judge had ruled previously on issues that might come up as part of the case you are working on, were you limited to using the Key Number system. From a doctrinal perspective, this might not be a significant omission: the law is the law, after all, and even if a judge’s opinions suggest that the judge might be opposed to ruling in a particular way, the judge will have little choice but to rule in that way if the law compels the result, and the law can still be searched using the Key Number system. As any practicing litigator knows, however, knowing the judge allows the lawyer to frame arguments in a way that might make it easier for a judge to rule the way you want the judge to rule. It hardly ever hurts to know what your judge thinks about an issue.

More subtly, and yet perhaps even more importantly, the Key Number system—any pre-indexed system, in fact—sets boundaries for our thinking. It is easy to fall into a way of thinking about the law that is bounded by the divisions and subdivisions of the Key Number system and to forget that there are ways of thinking about law that have little to do with those divisions. It is impossible, for example, to use the Key Number system to examine the relationship between two concepts—the way in which courts consider the idea of “ numerosity” in class action cases and the way in which the federal Multi-District Litigation Panel thinks about numerosity when deciding on which cases to refer for treatment as multi-district cases, for example. Pre-index systems like the Key Number approach are inherently singular, not relational.

Over time, lawyers grew more and more frustrated with the practical and theoretical limitations of pre-indexed searching in the law, and sought alternatives that would allow them more flexibility and freedom to search in ways that made sense to them at the time they were conducting the research, not to the editors at West. The computer finally allowed for the flexibility lawyers were looking for.

C. Self-Indexed Information

Commercial computer-assisted legal research came to the law in the 1960s, first with Lexis—developed initially by a group of Ohio lawyers who had grown frustrated with the limitations of pre-indexed research—and then, a few years later, with Westlaw. After Westlaw caught up with Lexis, the two services have offered essentially complete databases of primary law, although they differ in secondary source coverage and in the types of finding aids they offer. Before we consider the specifics of these services, let’s consider their approach to searching.

But even before we do that, we should remember that both services—at the time of writing, at least—offer the ability to conduct pre-indexed searching as well. In fact,
although lawyers and law professors often use the concept of “computer-assisted legal research” as a way of distinguishing computers from “book-based legal research,” the dichotomy is a false one. The true distinction is between pre- and self-indexed research approaches, because both LexisNexis and Westlaw offer the researcher the opportunity to conduct pre-indexed research.

Westlaw does this by incorporating the Key Number system, allowing researchers to interrogate the various primary law databases with queries that use the same doctrinal segmentation as they would use in the books. There are no digest volumes online, but a researcher can create a unique “digest” for each research topic, simply by entering in the Key Number information and the relevant sub-database or databases within the complete Westlaw database for the search under consideration. Westlaw will return the results, allowing the researcher to analyze the synopses of cases just as would be possible using the books.

LexisNexis can’t use the Key Number system, because that’s copyrighted by West. Instead, it uses a feature currently known as2 LexisNexis Headnotes, which also provides a synopsis of various parts of a court’s opinion, although unlike West’s Key Number headnotes, LexisNexis’ headnotes are taken directly from the language of the case, rather than being specially written by a team of editors. In practice, though, you can search both LexisNexis and Westlaw by use of topic indexes. There may be differences in operation, result, language, and appearance, but both services allow for this pre-indexed type of search, meaning that the more useful distinction is between pre- and self-indexed research, not between books and computers.

Any time you run a computer search using any form of search engine—the ones that power the old or new forms of Westlaw or LexisNexis, the ones that power Google, Yahoo, Bing, or any other search engine you can name—you are, in effect, indexing the contents of the database you’re searching for the terms you enter. If, for instance, you enter the terms “Casablanca,” “My Town,” and “times” into Google, you are asking the Google search engine to look for all examples of those terms and to return the ones that rank most highly (using Google’s complex ranking algorithm) first, and the ones that rank less highly further down. The results appear very quickly, so it’s difficult to appreciate how much information has been sifted in order to generate the results, but an indexer would generate similar results in time, given the search and ranking parameters under which Google was operating.

Speed is not the only difference, of course. For one thing, a self-indexed computer search allows for one concept to be searched in relation to another concept. So, in my previous example, you’re not just searching for “Casablanca,” you’re searching for a movie in relation to a place and in relation to the times the movie is being shown. That’s more than most pre-indexes will allow, although you will see some books in which the

---

2. If this book sounds cagey about what these services contain, or the names for these various features, it’s with good reason. At the time of writing, the legal information world is experiencing a period of volatility that is almost unprecedented. The only comparable time might be the period when Lexis and Westlaw were first introduced, and even then, there were many fewer users or services. Part of the changes we’re seeing at the moment have to do with the way both LexisNexis and Westlaw (or WestlawNext, as West is currently calling its new incarnation and LexisAdvance, as the newer Lexis product is known) present information. With rapid change comes insecurity, hence an unwillingness to be definitive about what features these services offer or what they’re called.
primary search term is segmented by subdivisions that allow some semblance of relational searching: in the multi-volume Civil War history we hypothesized earlier, for example, you might find an entry for “Lincoln, Abraham” and a series of page references, or you might find something like “Lincoln, Abraham, ... assassination of, ... Booth, John Wilkes,” and so on. Each of these segmentations allows you to search for one concept and its relation to the principal entry. You might find later entries in which “Booth” is the principal entry, and segmentations like “escape,” “medical treatment of,” and “death of” allow for similar relational searches.

West's Key Number system doesn't allow for this sort of relational searching, as we discussed earlier. But self-indexed research, available on Westlaw and LexisNexis, allows the researcher to create highly sophisticated relational searches that allow the researcher to search for multiple concepts and to define the closeness of the relationship between those concepts. Until now, that type of searching was undertaken using Boolean logic, a term honoring the mathematician George Boole.

2. Boolean Searching

Even if it were easy to explain to you how Boolean logic works, I wouldn't try, because the attempt would likely finish any chance I might have that you would continue to read this chapter. Suffice it to say that Boolean logic allows the researcher to construct a search that allows the computer to search a given database or databases for a term or concept, including variants of the words used to describe the term or concept, and, if the researcher chooses, to search for additional terms or concepts and — and this is the crucial part — to define the nature of the relationship between the various search components.

So, for example, you could search LexisNexis or Westlaw for the term “statute of limitations,” and limit your search to the New York state law database, or you could search for that term in two or more different state databases — New York and California, say, or all state court databases, or all court databases, including the federal courts as well. And you could look for a relationship between “statute of limitations” and your judge, or you could look for a relationship between “statute of limitations” and the concept “slip and fall.” Or “personal injury.” Or anything else you chose. If you're concerned that the court might have said something like “this court will limit all actions based on a theory of personal injury to those filed within three years of the date of injury” — a wordy way of saying something very similar to “statute of limitations” but not an inconceivable way for a court to write — you can search for variations of “limit” that would include the word itself, but also forms like “limitations” within a certain number of words of the term “personal injury,” or you could define the relationship as the same sentence. Or the same paragraph. Or the same case.

You get the idea. Self-indexed research opens up a huge world of possibilities that pre-indexed research foreclosed. Once meaningful self-indexed research was possible (and it took a while for that to happen, but it's certainly here now), the researcher had almost complete control over every aspect of what was to be indexed, and how those results would be presented. It was a remarkable leap forward in legal research, and it allowed lawyers to research in ways that were almost literally inconceivable before computer-assisted legal research came along.
3. Problems with Self-Indexing

But along with these possibilities came some problems that hadn't existed before, and we need to consider those as well. These problems are true for all computer-assisted searches and they're particularly relevant to lawyers who conduct searches on the large, commercial databases like Westlaw and LexisNexis.

The first, and biggest, problem is easy to describe and perhaps harder to grasp. Put simply, when we find a lot of information, we lose precision over what it is we're finding. And when we increase the precision of our search, we don't find a lot of documents that might have been relevant to our searches. If we use Boolean-based search engines, it seems that there's nothing we can do to change this effect; it happens regardless of how experienced the researcher might be, even those who are relatively familiar with the database they're searching.

An example might make this problem easier to understand. Suppose you're asked to search a database consisting of millions of documents. The broader your search, the more documents you'll recover, but many of those documents will only have tangential relevance to your search. You read through your results but quickly realize that a lot of what you're reading doesn't help your research. You probably have all the relevant documents to assist you, but you won't find them because they're buried in a sea of irrelevant information.

Frustrated, you enter a much more precise search request into the search engine, and you end up with very few documents, all of them highly relevant. As it turns out, though, you will have excluded from your search result a number of relevant documents as well, and you won't know what you haven't seen. And because you won't know what you haven't seen, you have no way of knowing if your search was too precise or whether it was about right.

A lot of lawyers didn't understand this problem at first. But this relevance/precision dilemma can cause real issues for lawyers if they forget it. If you're asked to do a discrete search with a readily identifiable answer — what the statute of limitations for slip and fall cases in New York state might be, for example — the relevance/precision problem might not show itself; you might only recover five percent of the relevant cases that give you the answer, but for this query, that might be enough. You answer the question quickly and correctly and believe that computer-assisted research is infinitely superior to the older, book-based approach.

If you're asked a complex question that requires analysis, interpretation, and the ability to analogize and distinguish cases from each other, the absence of relevant cases that are not recovered in an overly precise search, or the mass of information recovered in an overly broad search, could have serious ramifications. You might fail to identify relevant cases, or even relevant lines of analysis, that might help or hurt your case. And there's nothing that can be done to improve your results.
4. Ranked Searching

Or is there? The large legal database providers never fully accepted the implications of the relevance/precision problem. They argued that their programs had made refinements that lessened the effects of the problem, and many lawyers believed that their particular knowledge of the subjects they were searching meant that they could tell if a large chunk of information was missing, so the problem wasn’t as serious as information scientists were telling us.

The commercial database providers went further when they introduced “Natural Language” searching. Using this technique, a legal researcher could enter a query written in English—“statute of limitations for personal injury,” for example—and the computer would produce a set of results. How it did that is very important for us to consider.

For those being introduced to Natural Language searching when it was first available on Westlaw and LexisNexis, it looked like magic. The computer looked like it was able to read the researcher’s mind and come up with answers to their questions, rather than making them contort their search into a mathematical equation. But that, of course, was exactly what was happening; it was just that the computer was making the decisions about the form that equation would take, not the researcher. In fact, our Natural Language search was being reduced to a Boolean search by operation of an algorithm written into the search engine, the search was being run, and the results presented were being ranked and then displayed. That meant that control was being taken away from the researcher and placed in the hands of the programmers who wrote the algorithm at a couple of crucial points along the way; first, when the search was stripped of external words and possible extensions and rootExpanders were added—removing the “of” and “for,” for example, in the “statute of limitations for personal injury” search—and second, when the results were ranked.

All computer searches are ranked, but not all are ranked the same. In the commercial databases, Boolean search results are typically ranked by order of court importance—highest court first—and then by date. So if you search a database containing all federal cases, your results will be ranked as follows: Supreme Court (most recent case first, least recent last), Courts of Appeals (most recent case first, least recent case last), District Court (most recent case first, least recent case last). You can see a listing of all results, so you can opt to start reviewing your results with, say, some Circuit Court cases rather than looking through all the Supreme Court cases. But all cases that match your search will be listed.

Searches using Natural Language protocols, though, are usually ranked according to how closely the results match your search. So a District Court case might come first and a Supreme Court case might come last if the District Court case uses language more closely related to your search than does the Supreme Court case. And the number of returned results is limited; not all results that match your search are available for review.

These rankings, and the limitations placed on the number of results generated by a Natural Language search, arguably went some way to addressing the relevance/precision problem—one could tell fairly quickly how successful the search had been and if it was not returning the type of cases the researcher expected to see, a new search was readily available. And by ranking the results in terms of how closely they matched the inputted search, the program was attempting to return precise results, without drowning the researcher in information.
But many researchers were unhappy with Natural Language searching. They felt that it limited their ability to recover relevant documents, and they felt the loss of control inherent in allowing an algorithmic disambiguation of a complex legal search. Then, Google showed up, changing everything.

5. The Google Generation

There are many ways to describe your generation. The most accepted way is to refer to you as “millennials,” as opposed to the older “generation X,” or “baby boomer” cohorts. This isn’t the place to discuss the merits of grouping individuals together and ascribing generational traits to you all, regardless of your individual backgrounds. Suffice it to say that for some, at least, you are all more properly thought of as the “Google generation.”

Google is, without question, a remarkable phenomenon. As a search engine, it appears to capture more quickly and more accurately the thought behind our searches and it returns results that are better suited to our search, often, than could fairly have been expected based on the search we entered. For those who had grown used to less effective search engines, it once again seemed like magic.

It isn't magic, of course, just very skillful programming combined with a new way of looking for information. This isn't the time for a detailed description of the Google search engine. For our present purposes, the general knowledge idea of a web crawler, combined with a ranking process that ranks the results in the context of what other people searching for similar information have done with it, is more than sufficient. For one thing, it should tell us that while the Google approach works extraordinarily well, when looking for general information, it presents some challenges for legal researchers.

Ranking, in general, is something that should cause legal researchers concern, because it means that the information we receive in response to our query is being organized according to how someone else thinks we should receive it and, since the process is automated, that decision is not conscious, but is rather the result of some general decisions about how information should be organized made by a programmer whose interest and knowledge about legal information is unknown to us. That doesn't mean that the decisions are necessarily bad ones, or that we'll disagree with them. But it does mean that information is being presented to us in an order not of our choosing. And with legal information in particular, the decisions other people have made about the materials responsive to our searches might be helpful, but they also might not.

It's this uncertainty that is problematic to legal researchers: the results might be definitive, and they might be missing crucial information; we might have uncovered all we need to know, and there might be a lot we still need to learn. We can risk it and finish our research quickly, but that likely isn't the best strategy in the long term. This

---

3. If the way to describe generational differences is through the technology that defines a cohort's era, then my baby boomer generation might be better thought of as the "television" generation and generation X might be thought of as the "personal computer" generation. Or not.

4. In fact, one useful definition of who makes up the "Google generation" is "anyone who hasn't found Google magical when using it." If you're so used to Google's results that there's no surprise when it does what it does, then you're probably a member of that generation.
problem is inherent in forced-ranking research; we’ve learned to so trust Google’s speed and accuracy in day-to-day research that we might believe that the first few result panels we review contain most of the relevant information, and believe that most of the other panels will contain duplicative or less relevant information. It’s not necessarily true, though, as you will have learned if you’ve drilled deeply through the search results for something you’ve looked for on Google; often highly relevant information will be revealed well down in the search results. The information is down there because not many people have looked at it or because of some other artifact of Google’s ranking process, forcing it deeper into the substratum.

But just because other people haven’t looked at it doesn’t mean that it’s not helpful to our analysis. In fact, in legal analysis, sometimes it’s the thought that others haven’t had that helps us to overcome a conceptual block and craft a persuasive argument. We’ll come back to the danger of forced-ranking searching when we talk about specific databases; for now, though, please remember that there are issues relating to this type of research of which we should be aware before engaging in it.

6. Other Problems Associated with Computer-Assisted Research

There are four other problems associated with computer-assisted research we should touch on briefly here. These problems are obvious, and will come as no surprise to anyone who’s used a database to conduct research. They are, nonetheless, problems, and we should bear them in mind when we conduct computer-assisted legal research.

A. Coverage

The first problem is coverage. Obviously, we can’t find something in a database if it’s not there to be found. Most of the time, this isn’t really a problem for the large commercial databases, especially LexisNexis and Westlaw. They both have deep coverage for all primary databases—cases, statutes, regulations, and so on. They might have slightly less profound coverage for some of their secondary and tertiary databases—law reviews, for example, and newspapers and other reference materials. But these are small concerns; for most of the research you’ll be doing in law school and as junior lawyers, these databases are more than deep enough.

That’s not necessarily true of other legal information sources. Some have complete coverage for some databases but only partial coverage for others, some have partial coverage for all databases—the only thing that’s consistent is the inconsistency between various databases and services. And this isn’t necessarily a problem; sometimes partial coverage is good enough and will get you the results you need. But sometimes you will need to do some historical research; when a statute or regulation has been changed since the event precipitating litigation, for example, you likely will need the relevant language at the time of the incident, not the present day. It’s important to know what data set you’re working with, in order to be sure that the results you’re generating are sufficiently deep for the specific research you’re doing.
It would be a simple task to decide upon your favorite source for legal information and note down the coverage for that source. Then you would always have that information ready to hand and could check it if a coverage question came up. Unfortunately, as we all know, the information stored in databases is dynamic and ever-changing. It’s not likely that a legal information database would take away coverage, and much more likely that it would add coverage to its already-existing information. But it’s not impossible that once-available data might no longer be available, especially on sites that are affected by changes in political administrations (information on federal administration agency websites, for example, often changes when a new president is elected), and even if a site adds coverage, that changes the nature of the results you might generate. The safest practice, then, is to check the coverage for every database you use, at the time you use it. You probably don’t have to do this for LexisNexis or Westlaw, but it’s advisable to do it for every other site you use.

B. Accuracy

The second problem associated with computer-assisted research is obvious to anyone who has typed a word, only to see a different result on the screen. Put simply, the computer will search for exactly what you asked, even if you misspelled a word or a name. That’s something we all understand, and we’ve all had experience of typing the wrong thing into a search engine and having it come back with results we didn’t understand. But for lawyers, typos can have important consequences when researching.

Imagine, for instance, you’re a junior lawyer working for a law firm, and you type “statute of limitations” instead of “statute.” It’s a mistake that happens all the time. But if you do that when you search, you might get no results, or, at least, no meaningful ones. Most likely, you will realize your mistake quickly and retype the search. But depending on the pricing structure your firm uses, that search might have incurred some costs for your client. If the client doesn’t want to pay for it, your firm will have to pay for that research itself. The firm won’t be too happy with you if you force it to pay for too many mistakes you’ve made.

C. Specificity

The third problem is more subtle. If you search for my name—Ian Gallacher—on Google, for instance, you’ll find that I’m either a law professor, a pastry chef, a jeweler, an actor in a movie, or a designer of scenic backgrounds for computer games. But if you search for the more common spelling of my name—Ian Gallagher—you’ll find that I sing for an Irish show band based in New Jersey, I’m a member of the Northern Ireland Policing Board, I’m a fictional character in a British soap opera (that one feels

5. The large legal databases now have context-sensitive spell-checkers that are designed to fix this type of mistake and to search instead for what the database assumes you wanted to type. You’ve doubtless seen the same sort of thing on Google. These spell-checkers can minimize the accuracy problem if you make a mistake the system recognizes, but nothing can eliminate it completely. And, of course, if you want to search for a term the computer wants to change, you’d have to wrestle with the computer so that it actually runs the search you want it to run.
oddly accurate sometimes), or that I am, or have been, the strength and conditioning coach for the Vancouver Giants (certainly not true).

The obvious point here is that spelling matters, and in a matter like a name search, one wrong letter can mean the difference between a correct result and an incorrect one. Even with the correct spelling you'd have to do some work to decide which of the various "Ian Gallacher" possibilities best fits me, but if you spelled my name incorrectly you might never know you were looking at the wrong series of people.

There's nothing surprising about this, and almost nothing to be done about it either. Both LexisNexis and Westlaw have some checks in place to look for words that might have been misspelled, but even they can't make allowances for less-than-common misspellings. We just have to be as careful as possible when using computers that we don't set ourselves on the wrong path, dooming our research to failure.

D. Language

The final issue that often comes up with computer-assisted research is, in a way, related to the typo problem. Because computers can only search for those things you ask them to search for, you must correctly predict how a court, or a legislature, or a rule-making body has discussed a topic if you're using a word-based search to find information. If you search for a statute or regulation using a table of contents, this problem is eliminated, which is one reason that's such a good strategy to use when searching those databases. But if you're looking for a case, you have to predict what words or terms the court will use, or you will not find all relevant information.

If you look for "spousal abuse," for example, you won't find "domestic violence." The two concepts might be related, and the two terms might be used interchangeably, but the words themselves are completely different and computers can only search for the words you ask them to search for. This puts a tremendous burden on the legal researcher to predict the language courts use when describing legal issues and to search for all possible variations. It's possible to do, but it's still an important limitation to remember.

It's also worth remembering that this is a limitation that doesn't apply to pre-indexed legal research. The indexer has read the case before indexing it, and has made the adjustment that is easy for us as humans to make and impossible for computers to make. A case that discusses spousal abuse and a case that discusses domestic violence will likely receive the same coding, allowing the researcher to find both with equal ease.

This is why one of the safest approaches to legal research is to combine pre-indexed and self-indexed techniques. Searching for the same information in two different ways might seem redundant but this is one area where redundancy makes sense. Why that is, and how you might do it, is a topic for the next chapter.
A Brief Discussion of Algorithms and Classification

¶8 At the simplest level, an algorithm is “a set of step by step instructions, to be carried out quite mechanically, so as to achieve some desired result.”14 The Pythagorean theorem is an algorithm, and so is the set of instructions that Netflix uses to recommend movies. Although algorithms have always had a role in modern life,15 it is the role that algorithms play in selecting what legal information we see that is critical for legal researchers.16 As we increasingly rely on algorithms for the assessment of information, algorithms dominate in mediating our information environment.17 If researchers are not aware that the information they seek may be missing from a database, or that the results that might be helpful may not be privileged in the result set, or that the list of documents suggested may have been generated by a legal worldview that opposes the path the researcher is trying to forge, a research session may terminate with no helpful results when helpful results actually exist.18 So we have to have sets of questions like these to ask each algorithm: How is information included or excluded from a system? How does the resource use predictive algorithms to anticipate use? How is relevance evaluated? Does the “black box”19 of the algorithm’s work lend a seeming objectivity to the results? How does use of the system change result patterns?20 For attorneys, learning to navigate black boxes is part of the ethical duty to do competent research: knowing something about why you received the results that you did is a critical skill. For legal research professors, teaching this skill may involve passing on some understanding of how the systems we use today evolved.

¶9 Online legal information systems did not arise as completely new structures. The initial transition to any new technology is frequently fairly literal. Think of the “horseless” carriage or the first bicycles. The first legal information that made the

15. As just one example from industry, Gant charts were simple instructions for scheduling that have been in use since the mid-1890s. Starting in 1958, those instructions were computerized with algorithmic instructions, using the Naval Ordnance Research Calculator, the most powerful computer in existence at the time. Jeffrey W. Herrmann, A History of Production Scheduling, in Handbook of Production Scheduling 1, 11–12 (Jeffrey W. Herrmann ed., 2006).
17. Id. at 167–68. The study of algorithms as mediators of all public information is a rich field on its own, but one that is beyond the scope of this article, which will limit its focus to algorithms that mediate legal information systems.
18. See Gregory J. Downey, Making Media Work: Time, Space, Identity, and Labor in the Analysis of Information and Communication Infrastructures, in Media Technologies, supra note 16, at 141; see also Nicholas F. Stump, Following New Lights: Critical Legal Research Strategies as a Spark for Law Reform in Appalachia, 23 Am. U. J. Gender Soc. Pol’y & L. 573, 639 (2015) (stating that for the law review articles promoted to the researcher as “Context & Analysis,” the researcher has no way of knowing the criteria used by the publisher in picking those articles, and the work that went into making those decisions has a definite influence on the course of the research).
19. On one level, a black box is any “technical object that operates as it should. When this occurs, the complex sociotechnical relationships that constitute it are rendered invisible, or black-boxed.” Darryl Cressman, A Brief Overview of Actor-Network Theory: Punctualization, Heterogeneous Engineering & Translation 1, 6 (Ctr. for Policy Research on Sci. & Tech. No. 09-01, 2009), http://summit.sfu.ca/item/13593.
transition online was the full text of cases, made searchable with Boolean logic.\textsuperscript{21} Headnotes, case summaries, statutes, news, business information, and finally law reviews were subsequently added to the systems.\textsuperscript{22} The freedom to search full text without the constraints of classification systems was supposed to unmoor the law from its structure.\textsuperscript{23} But it turns out that trying to make sense of information without underlying ontologies or classification systems can impede automation practices.\textsuperscript{24} Legal database providers may even make the human additives to their search explicit. LexisNexis boasts of the human indexing in Shepard’s citations;\textsuperscript{25} Westlaw is proud of its human-generated Key Numbers;\textsuperscript{26} and Bloomberg BNA advertises that the human indexing in its BNA treatises significantly boosts search results.\textsuperscript{27}

\textsuperscript{¶10} The complexity of the source material may require classification to aid relevant search results. Even the current “Google-like” legal databases provide extensive prefiltering, postfiltering, and word wheel\textsuperscript{28} options for granular classification by source, authority, jurisdiction, and content type, and by value-added indexing by humans.\textsuperscript{29} Some of the changes in the levels of prefiltering and word wheel options have been in response to user demand; lawyers seem to need clasi-
sification in the law.30 Lawyers, after all, are human, and we are all hardwired to impose structure on the world.31

¶11 The transition to online searching has increased the complexity of the search task; as more information becomes available, more research is required.32 As is frequently the case, automation has made a task more complex for the humans involved.33 In exchange for instant access, the user has to master increasingly complex tasks to recover information effectively. The human reasoning, classification schemes, design decisions, and other work that went into the creation of the systems the researcher is using are mostly hidden.34 Going beneath the surface of research systems, even in the predigital search environment, has never been the norm.35 There is a long history in legal research of researching with only a surface understanding of the underlying structure. Speaking of lawyers at the time of transition to online searching, Bob Berring has noted that most were unaware of the details of the classification systems imposed by the Key Number system,36 and in the early days of online searching, most users were unaware of the structure underlying the system.37 This is almost certainly still true. But that is not to say that some basic understanding of the forces at work would not be helpful to researchers in the

30. Maggie Nelson, The Argonauts 53 (2015) (positing an "Aristotelian, perhaps evolutionary need to put everything into categories"). Recent studies on the human mind illustrate the deep-seated desire to classify and categorize, and, in response, lawyers push online systems to recreate the systems and categories. Daniel J. Levitin, The Organized Mind: Thinking Straight in the Age of Information Overload 25 (2014) ("The formation of categories in humans is guided by a cognitive principle of wanting to encode as much information as possible with the least possible effort. Categorization systems optimize the ease of conception and the importance of being able to communicate about those systems."). Even in the evolution of online databases, where the first databases were just the stripped-out text of cases, the momentum has always been toward more structure and classification in the online systems. F. Allan Hanson, From Key Numbers to Keywords: How Automation Has Transformed the Law, 94 Law Libr. J. 563, 569–72, 2002 Law Libr. J. 36, ¶¶ 19–25; see also William G. Harrington, A Brief History of Computer-Assisted Legal Research, 77 Law Libr. J. 543 (1984–1985).

31. Levitin, supra note 30, at 32. Lawyers may be singular in their need for control and order. See, e.g., Margaret Hagan, Do Lawyers Want Bad Visual Design?, Open Law Lab (June 28, 2016), http://www.openlawlab.com/2016/06/28/do-lawyers-want-bad-visual-design/ [https://perma.cc/7FWQ-AFQ5] ("Lawyers want maximum overload of information in response to queries they do; They want it listed out in detail, with lots of information packed onto the page; They don't want white space, they want text covering as much of the screen as can fit. They want lots and lots of controls, all kinds of filters and sorting mechanisms.").

32. Berring, supra note 21, at 1683–90 (tracing the differences between the forms and content of an 1891 Supreme Court case and a 1996 Supreme Court case).


35. Berring, supra note 21, at 1694.

36. See id.; Robert Berring, Chaos, Cyberspace and Tradition: Legal Information Transmogrified, 12 Berkeley Tech. L.J. 189, 210–11 (1997). In the pre–online research world, many lawyers were aware that the West classification system missed a lot. See Daniel P. Dabney, The Curse of Thamus: An Analysis of Full-Text Legal Document Retrieval, 78 Law Libr. J. 5, 14 (1986) ("This short review of ideas in indexing shows that the indexing process is prone to many sorts of errors and uncertainties. Manual indexing is only as good as the ability of the indexer to anticipate questions to which the indexed document might be found relevant. It is limited by the quality of its thesaurus. It is necessarily precoordinated and is thus also limited in its depth. Finally, like any human enterprise, it is not always done as well as it might be.").

37. Berring, supra note 21, at 1697.
brave new world of information overload and satisficing that we now live in. And that is what we need to teach legal researchers.

¶12 Lawyers are not alone, of course. Most people do not think about the underlying structures of the technologies they use. But some inquiry into the forces at work in the legal research environment, at this moment when so much of the work is truly invisible, is certainly called for. It is time to examine technical bias in legal computer systems. Technical bias is built into systems. We just don’t see it because the systems we use are black boxes. The following attributes contribute to the biases that programmers embed in the black box:

- prioritization (“emphasiz[ing] . . . certain things at the expense of others,” like relevance ranking);
- classification (putting an “entity [in a] constituent . . . class.” Data training may import human biases);

38. Satisficing is a time-honored information-seeking activity; it means to settle for what is “most readily available with little or no regard for costs and benefits,” when you want quick answers and may not have time to optimize the search. Brian C. O’Connor et al., Hunting and Gathering on the Information Savanna: Conversations on Modeling Human Search Abilities 131 (2003).

39. The classifications, design decisions, and choices made every day by information scientists in our technological environment frequently embody “moral and aesthetic choices” that impact our own decisions and thoughts. Bowker & Star, supra note 34, at 3–4. These hidden choices can have very deep effects, as the recent “great recession” has shown. The great recession was in part a failure of algorithmic oversight. See Morton Glantz & Robert Kissell, Multi-Asset Risk Modeling: Techniques for a Global Economy in an Electronic and Algorithmic Trading Era 437–39 (2014).


41. Friedman & Nissenbaum, supra note 5, at 330. Technical bias is one of the three biases that computer systems can display; the other two are emergent and preexisting. Emergent biases are those that arise in the actual use of the database; these biases can relate to new societal knowledge, differing expertise, differing values, or a mismatch between the user and the system design. Id. at 335. The preexisting bias in legal databases is, at a minimum, the structure of the law itself, and the content and classification systems that have been imposed on the law by legal vendors. Id. at 333; see also Hanson, supra note 30, at 569–72, ¶¶ 19–25. This was a gradual change, as the first computer retrieval systems included only the simple text of cases, with no structure or classification; the evolution to structure and classification was gradual. Harrington, supra note 30, at 543; see also Berring, supra note 21, at 1693, 1696. For an excellent history of the evolution of ideas about the effect of classification schemes on legal thinking, see Richard A. Danner, Legal Information and the Development of American Law: Writings on the Form and Structure of the Published Law, 99 LAW LIBR. J. 193, 2007 LAW LIBR. J. 13. The classification systems may also hide assumptions about the nature of the law that mask paths to justice. See Stump, supra note 18, at 573; see also Hildebrandt, supra note 40, at 148–49 (discussing the effect computational “nonreading” of texts—the pattern recognition that is the language of machine learning—may have on legal interpretation).

42. Cressman, supra note 19, at 6.
 association (“marks relationships between entities”); and
filtering, which “includes or excludes information according to various rules or
criteria.”

§13 An interesting example of how assumptions or biases inform results is from
a study on coding algorithms to enforce that exemplary seemingly simple rule of
law, the speed limit.44 What seems a relatively straightforward problem becomes
dense with assumptions when one thinks about how to implement the law: do you
enforce the letter of the law or the intent of the law; is every second you exceed the
speed limit a separate violation; do weather or road conditions matter; how often
should a driver be given a ticket; does context matter?45 There were three groups of
coders using actual driving data taken from a vehicle’s computer: the first group was
asked to implement the letter of the law; the second group was asked to implement
the intent of the law; the third group was asked to follow a detailed written specifi-
cation.46 The differences in the results of each coding scheme are stark: the number
of tickets issued by the algorithms varied from zero to 661, for the same driving
pattern.47 All of the groups made assumptions independent of their instructions; for
example, there were significant differences in how the first two groups coded toler-
ances for exceeding the speed limit, and all of the groups assumed, without instruc-
tion, perfect driving conditions.48 As the authors of the study note, transparency
about coding assumptions may be the only solution to success in implementing
automated legal compliance or enforcement in a fair and open manner.49

§14 “Algorithmic accountability” is the term for disclosing prioritization, clas-
cification, association, and filtering.50 What we need is a frank discussion with
database providers about what it means to search in their databases. Trade secrets
should not prevent algorithmic accountability. Some database providers do provide
search tips that can help their users understand what happens between input and
output;51 legal database providers also publish basic search information for their

43. Nicholas Diakopoulos, Algorithmic Accountability: Journalistic Investigation of Computational
44. Shay et al., supra note 8, at 1.
45. See id. at 20 for a chart summarizing the differences in results for three groups of coders who
were given three different assignments for coding violations.
46. Id. at 4–5.
47. Id. at 20.
48. Id. at 7–8, 14–15. The variations and subtle assumptions are quite varied, and the Shay article
discusses only a few.
49. Id. at 30–31.
51. For example, this information is from a help protocol from Summon, a discovery-layer search
product, transmitted in an e-mail to the author because the information is behind the password-
protected administrative module. E-mail from Joan Policastri, Collection Servs. & Research Librarian,
Univ. of Colo. Law Sch., to author (Jan. 8, 2016, 12:05 PM MST) (on file with author). It would be
more helpful if the information were readily available to users:

**Boolean search and Summon relevancy algorithm:** Boolean queries get processed by the
same relevancy algorithm as any other query. This means relevancy enhancements that come from
the application of stemming, character normalization, etcetera will apply in Boolean queries as
well.

Applying the relevancy algorithm to Boolean queries is particularly helpful in Boolean searches
using multiple search terms. For example: **paint drying time (glass OR wood).**
The more we understand about the input into the black box, even without knowing the code for the algorithm, the more we can see how the system operates in practice. At the moment, we can really only see the output of the systems, and that is what this study investigates.

The need to know about the input, the paths that mark the way to the results, only increases as the amount of work being done by the algorithms increases. A case in point is the use of IBM’s artificial intelligence program, Watson, by the medical community. Watson is IBM’s supercomputer, which uses artificial intelligence and machine learning to leverage large amounts of data. Watson is better than humans at reading through documents and is starting to be used as a “quick-witted digital assistant” in oncology clinics, but with a caveat: doctors use it in conjunction with Watson Paths, a visual tool that allows a doctor to see the underlying evidence and inference paths Watson took in making a recommendation. “It’s not sufficient to give a black-box answer,” said Eric Brown, IBM’s director of Watson technologies. As decision makers, doctors want knowledge, not technological determinism. Legal researchers need to demand the same kind of transparency.

What Legal Database Providers Say About the Search Experience

Some information on the search experience is available on each legal database. In terms of the basic types of search, researchers tend to refer to Boolean searching—meaning that the researcher uses terms and connectors, such as “and,” “or,” and “not,” to construct a search—and natural language searching—meaning

---

Id. Legal professionals who were expert searchers had compiled their own special tips and tricks for getting the best results, but those tips and tricks were for pre-2010 Boolean searching. New tips and tricks are necessary for keyword searching in the current set of search algorithms, but for the most part, these tips and tricks have not been revealed.

52. See infra ¶ 16.
56. See Julie E. Cohen, Network Stories, LAW & CONTEMP. PROBS., Spring 2007, at 91, 92, (“What makes the network good can only be defined by generating richly detailed ethnographies of the experiences the network enables and the activities it supports, and articulating a normative theory to explain what is good, and worth preserving, about those experiences and activities.”).
57. George Boole was a nineteenth-century mathematician, and his work on the analogies between algebraic symbols and symbols that represent logical forms and syllogisms resulted in the application of what is known as Boolean logic to searching. George Boole, Britannica.Com, http://