Reinforcing the Infrastructure of Legal Research Through Court-Authored Metadata

Andrew J. Martineau

This article examines how the role of the court system in publishing legal information should be viewed in a digital, online environment. It then argues that courts should author detailed, standardized metadata for their written work product. This practice would result in immediate, identifiable improvements in free and low-cost case law databases, and may beneficially impact the next generation of AI-powered research tools as well.

Introduction .................................................................................................................................................................................. 6
Access to Legal Information in a Digital Age .................................................................................................................................................................. 6
Digital Legal Information: Theoretical and Practical Implications ..................................................................................................................... 7
Modeling the Drift Toward Digital: Winners and Losers ................................................................................................................................. 14
Improving Public Access to Case Law Through Structured Metadata ............................................................................................................. 20
Government as Data Provider ........................................................................................................................................................................... 21
Structured Metadata: The Big Picture ......................................................................................................................................................... 24
The Proof Is in the Patents: Comparing Google Patents with Google Scholar ..................................................................................................... 26
How Structured Metadata Can Reinforce the Infrastructure of Legal Research ................................................................................................ 28
How Might Court-Authored Metadata Affect Cutting Edge Research Tools? .................................................................................................... 36
Where Do Librarians Fit In? ............................................................................................................................................................................. 41
Conclusion ........................................................................................................................................................................................................ 44

** Instructional Services Librarian and Assistant Professor, University of Minnesota Law Library, Minneapolis, Minnesota. This is a revised, updated, and expanded version of Andrew Martineau, My Lawyer Has a First Name, It's G-O-O-G-L-E: Improving Online Access to Case Law Through Court-Provided Metadata (Culminating Experience Project for MLIS, University of Washington Information School, May 13, 2013). I would like to thank Penny Hazelton, whose comments and insights on early drafts of this paper significantly altered its ultimate trajectory. Likewise, I’d like to thank my classmates in the 2013 University of Washington Law Librarianship Program for letting me bounce ideas off of them (even though they had their own papers to write). The participants of the 2016 Boulder Conference on Legal Information deserve my thanks as well for helping me solidify and organize the ideas in this article. Thank you to the librarians at the University of Minnesota Law Library for their helpful feedback, and to Connie Lenz and Scott Dewey in particular for their detailed comments. Finally, I am especially grateful to Miranda Snyder who, after proofreading dozens of drafts of this article over the years, should be awarded a medal (and maybe an honorary J.D./M.L.I.S).
Introduction

¶1 Technology has dramatically altered the legal information landscape over the past half century. This extended, ongoing transition from print to digital has exposed conflicting interests among legal publishers, Internet-native companies, courts, the public, and law libraries. Many in the legal community now enjoy easy access to a much richer universe of information than they did in the print era. Others have not benefited equally from these technological innovations; in fact, as some of the basic tools of legal research cease to be available in print, their ability to access quality legal information may be in danger.

¶2 Law librarians have stood on the front line to confront, implement, and explain these technological advancements, and have guarded against their collateral harms. They regularly advocate for free public access to useful legal information, but this role becomes more difficult in times of lean budgets. Thus, law librarians should endorse policies that appropriately use the free market in furthering these goals.

¶3 This article argues that courts could help reinforce the “crumbling infrastructure of legal research” by including detailed, standardized metadata with their written work product. Part 1 describes how the creation, publication, and dissemination of legal information have evolved over the past few decades. Next, part 2 demonstrates how improved, court-authored metadata could help ensure that the public, solo practitioners, and small firms retain access to useful legal information as the legal publishing industry continues to evolve. Finally, part 3 outlines how law librarians might help the courts fulfill their roles as editors and publishers of legal information.

Access to Legal Information in a Digital Age

¶4 Many scholars have covered the long history of U.S. legal publishing, from a variety of angles and perspectives. This part begins by reviewing scholarship that

1. This is a reference to Robert C. Berring, The Heart of Legal Information: The Crumbling Infrastructure of Legal Research, in LEGAL INFORMATION AND THE DEVELOPMENT OF AMERICAN LAW 272 (Richard A. Danner & Frank G. Houdek eds., 2007) [hereinafter Berring, Crumbling Infrastructure].

theorizes the recent history of legal publishing, beginning when legal information first became available via remote access to databases like Lexis. Next, it describes selected scholarship on the transition of legal information from print to digital formats, framing a few of the most pertinent issues associated with the decline of print. Finally, this part provides a broad framework for modeling the general structure of the legal information marketplace, ultimately arguing that the shift from print to digital access has created new opportunities for government actors to participate in—and improve—the marketplace for legal information.

Digital Legal Information: Theoretical and Practical Implications

¶5 Robert Berring, Peter Martin, Thomas Bruce, and Ethan Katsh are among the scholars who helped conceptualize the evolution of legal information over the past three decades, a period during which the legal information system underwent a seismic shift due to the explosion of legal information available online. Two relevant themes emerge from this body of literature: first, that the shift to digital has fundamentally changed how researchers use and view legal authority (case law, especially); and second, that this shift has expanded the public’s access to the law in some ways, while limiting access in others. Recent scholarship by Susan Nevelow Mart demonstrates—with empirical evidence—another cause for concern: the subtle subjectivity but significant impact of the proprietary algorithms on which this research often relies.

¶6 In The Heart of Legal Information: The Crumbling Infrastructure of Legal Research and other works, Robert Berring describes how technological advances and resultant changes in the legal publishing marketplace have affected the concept of legal authority. For example, because publication of an individual court docu-
ment in an existing electronic database costs next to nothing (assuming that the infrastructure for hosting these digital files is already in place), researchers can now access sources of legal authority that they could not access (or access fully) in the print world. The most obvious example is unpublished cases, but researchers also enjoy wider access to administrative law, legislative history, and court dockets.

¶7 This all may seem like good news, but Berring argues that this phenomenon, combined with the ubiquity of full-text searching, may undermine traditional notions of precedent and authority. Rather than relying on the West Digest System to guide them to relevant, authoritative, and, usually, appellate case law, contemporary researchers tend to find cases using keyword searches. Legal researchers can thus circumvent finding aids like Shepard’s or the Digest System, and access the law directly. Because researchers can locate a case without learning its place in the greater structure of the law—or, at least, the structure imposed by an editor at West—the cases they find are divorced from context. Furthermore, the assumption that the law has an inherent structure at all becomes exposed as a myth—the law emanating from courts only appears to be innately, rationally structured because of the scaffolding provided by the Digest System. In effect, the sheer volume of available published and unpublished case law, floating untethered from context and much of it without real importance, dilutes the authority of the cases that really do matter according to traditional notions of precedent.

¶8 Attorneys in small firms, solo practitioners, and pro se litigants are especially at risk since they might lose access to crucial finding aids and secondary sources in print without gaining access to the digital tools necessary to thrive in this new information environment. Berring, however, sees the situation as transitional and temporary: before you know it, we might all be querying an artificial


8. See, e.g., Berring, Crumbling Infrastructure, supra note 1, at 287–90.
9. See, e.g., id. at 287–94.
10. See id. at 287–90.
11. See id. at 279; Robert C. Berring, Full-Text Databases and Legal Research: Backing into the Future, 1 High Tech. L.J. 27, 48 (1986) [hereinafter Berring, Full-Text Databases] (noting that this reliance on keyword searching can have negative consequences: because legal concepts can often be described in a variety of ways, formulating good searches is harder than legal research databases advertise to users).
12. See Robert C. Berring, Legal Research and Legal Concepts: Where Form Molds Substance, 75 Calif. L. Rev. 15, 26 (1987) [hereinafter Berring, Form Molds Substance] (“Now the researcher can search the entire corpus of law on a word-by-word basis, free from the constraint of a subject thesaurus. Custom-designed subject structures and searches based on entirely different groupings of subjects are possible once the intervening intelligence is removed.”).
13. See Berring, Full-Text Databases, supra note 11, at 54; see also Berring, Form Molds Substance, supra note 12, at 26.
14. Berring, Form Molds Substance, supra note 12, at 26–27. Ethan Katsh offered an alternative explanation for the “loss of faith in the metaphor” of law as a “seamless web”: even if law had an inherent structure when print was dominant, this structure could not survive in a digital world of “versatile and volatile” data points. See Ethan Katsh, Law in a Digital World: Computer Networks and Cyberspace, 38 Vill. L. Rev. 403, 405–06 (1993).
15. See Berring, Crumbling Infrastructure, supra note 1, at 289.
16. See Berring, Chaos, Cyberspace and Tradition, supra note 4, at 208.
intelligence-powered system in plain language for the answers to our legal woes. As long as the legal researchers of the future demand that these new systems provide authoritative, useful information, the market will catch up. In the meantime, we’ll be in for a bumpy ride.

¶9 Peter Martin describes the same phenomenon—the transition from print to electronic publishing—but focuses on the opportunities that the Internet creates for courts to directly disseminate information to the public, and the positive effect this could have on the development of the law in general. In Reconfiguring the Law Report and the Concept of Precedent for the Digital Age, Martin argues that access to a wider array of case law (such as unpublished and trial court opinions) would promote greater consistency among trial courts; increase the quality of judicial opinions and decisions; and encourage judges to adopt rules that are well reasoned, even if such rules are not necessarily “binding” authority. To fully achieve the promise of electronic publishing, however, courts would need to take a more active role in publishing their cases (in terms of using the right formats, encoding documents with high-quality metadata, and providing a means for authentication, among other things). In addition, these “slip opinions,” pulled directly from a court’s website, would be truly useful only if the state adopted vendor-neutral citation policies.

¶10 Thus, Martin tends to argue that we can avoid many of the problems identified in the shift to digital primary law if we implement the technology in a thoughtful, intentional way. Rather than a threat to the status quo, the transition to online legal information creates new opportunities for courts to assert control over their case law.

¶11 Thomas Bruce, like Martin, was instrumental in the creation of the Legal Information Institute (LII) at Cornell. In Public Legal Information: Focus and Future, Bruce describes the shifting roles of government bodies, private actors, and academic institutions as legal information moved to an online environment. Given the low cost of publishing on the web, Bruce argues that the government should act as the primary distributor of its data online. Academia, on the other

17. See Berring, Cognitive Authority, supra note 7, at 1706–07 (anticipating, however, that “black box” systems like these raise new concerns, such as the question of how to evaluate their effectiveness). In fact, it seems that these AI tools are now in the nascent stages of development; though, perhaps, the hype may not match reality—at least not yet. See generally Jamie J. Baker, 2018 A Legal Research Odyssey: Artificial Intelligence as Disruptor, 110 LAW LIBR. J. 5, 16, 2018 LAW LIBR. J. ¶ 40.

18. Berring, Cognitive Authority, supra note 7, at 1708.

19. See Berring, Crumbling Infrastructure, supra note 1, at 295 (“We are on the verge of a point where the system will stop working.”).


21. Id. at 37–38.

22. Id. at 30. See generally Peter W. Martin, Neutral Citation, Court Web Sites, and Access to Authoritative Case Law, 99 LAW LIBR. J. 329, 2007 LAW LIBR. J. 19. For a more recent update on this issue, see Michael Umberger, Checking Up on Court Citation Standards: How Neutral Citation Improves Public Access to Case Law, 31 LEGAL REFERENCE SERVS. Q. 312 (2012).

23. Martin, supra note 2, at 30. Although, Martin’s recent update on PACER shows that even in 2018, the federal judiciary has much room for improvement when it comes to digitally publishing case information. Peter W. Martin, District Court Opinions that Remain Hidden Despite a Long-standing Congressional Mandate of Transparency—The Result of Judicial Autonomy and Systemic Indifference, 110 LAW LIBR. J. 305, 2018 LAW LIBR. J. 14 [hereinafter Martin, District Court Opinions that Remain Hidden].


25. Id. at 26.
hand, should focus on the creation of standards for this data, to increase interoperability among different web-based systems. Further, to achieve “effective public access,” free online databases would need to offer better feedback to users.

¶12 Around the same time as the publication of Public Legal Information, Bruce authored Tears Shed over Peer Gynt’s Onion: Some Thoughts on the Constitution of Public Legal Information Providers. Here, he compared competing models for the legal information marketplace. For example, should government bodies self-publish their data, or should they outsource to private companies? Should a central authority control such a system, or should control be distributed among the creators of the data? In the end, Bruce concluded that the government should self-publish its work, initially through some central authority (e.g., to ensure that data standards were complied with), which would then shift to a decentralized model as the formats and standards became entrenched.

¶13 For Ethan Katsh, the transition from legal information in print to legal information in electronic format is just one aspect of the digital revolution that has swept through the legal community. Whether this ongoing transition is good or bad is less important to Katsh than the fact that it is inevitable and will continue to change how law is created, researched, and practiced. Like Berring, Katsh believes that traditional conceptions of legal precedent and authority were so intertwined with print as a medium of communication that the connection between the two was effectively invisible. Now that we have entered a world where digital has become dominant, however, the veil has been lifted. The massive volume of available case law will erode our concept of, and reliance on, precedent. The malleability and impermanence of digital law could undermine our faith in its authority. Boundaries between legal and nonlegal information, as well as between law practice and other sectors of the economy, will become porous, lessening lawyers’ value that traditionally flowed from privileged access to an exclusive, specialized set of

26. Id. at 26–28.
27. Id. at 31–34. John Joergensen, writing from his experience with the New Jersey Court Publishing Project, argued that users also expect legal information databases to search through many (or all) jurisdictions at once, as well as hyperlink to cited material. John P. Joergensen, Are Non-Profit Internet Publishers the Future of Legal Information?, 17 LEGAL REFERENCE SERVS. Q., nos. 1–2, 1999, at 33, 40 (“If the provision of [free] legal material on the Internet is to truly prosper and provide a real alternative to the large online services, the interactivity of sites needs to be improved so that the interactivity between the various sites that supply information at least approaches that of the major online services.”).
29. Id. § 4.
30. Id. § 7.
31. Id. § 8.
32. Id. § 9.
34. Id. at 268 (“The viability, effectiveness, and nature of law in the future depends on whether we understand the changes occurring to the law and are able to respond to them.”).
35. See id. at 35–40.
36. Id. at 44–46.
37. Id. at 89–94.
rule-oriented knowledge. Thus, lawyers will spend less time researching, analogizing, and synthesizing legal rules, and instead focus more on finding practical—often nonlitigious—solutions to client problems.

¶14 The shift to digital will impact ordinary citizens as well. Although nonlawyers have gained access to a vast universe of online legal material, much of it is incomprehensible to them. As Katsh describes the situation, the physical distance between the user and legal information has shrunk to zero, but the informational distance is still great:

Informational distance refers to how inaccessible a medium makes information. The medium may be difficult to use or the information may be presented in a difficult to understand format. Consequently, some information may be less accessible than other information not because it is far away or because it is conceptually complex but because of inherent qualities of the medium in which it is organized and stored.

¶15 Because online research platforms fail to give users adequate guidance, Katsh observed in 1995 that “[a] user logged on to most commercial databases today is physically close to relevant material but informationally distant from it.”

¶16 Much like Berring, however, Katsh believes that these issues seem troubling in large part because we are still transitioning from a print to a digital world, and that these digital problems will have digital solutions: “The barriers currently standing in the way of or lending confusion to accessing electronic materials, however, are, to a considerable extent, a consequence of poor software design that will gradually be remedied.”

¶17 In the works discussed above, Berring, Martin, Bruce, and Katsh focus on different but overlapping aspects of the shift from mainly print to mainly electronic sources of primary law. The transition they describe and anticipate in these articles is still underway. The Digest System has only fallen further out of favor, but many researchers do not have access to an adequate replacement, instead relying solely on basic keyword searching to find the law. Although the informational distance experienced by users of online databases has certainly diminished, online platforms still fail to give users adequate guidance and support—particularly in the low-cost or free segment of the market. We can now access a great amount of primary legal information in electronic format, but much of it is merely digitized versions of print documents, with the few features that take advantage of the digital format “tacked on” by legal publishers after the fact. As Katsh observes, “[w]e are still in an age in which

39. Id.
40. Katsh, supra note 14, at 450.
41. Cf. id. at 476. Part of this is because “word searching is not really very easy nor is it conceptually simple.” Id. at 475–76. For a librarian’s perspective on the problem of informational distance, see Haigh, supra note 5, at 253, which argues that in an online environment, the informational distance can be greater than that which was found in a traditional library, where at least patrons had reference librarians nearby and free access to a curated collection of secondary sources.
42. Katsh, supra note 14, at 478.
43. Katsh observed in 1989 that digital “[i]nformation need not be presented any longer in uniform and standardized form, since many of the constraints of print have been lifted.” Katsh, supra note 33, at 94. However, common practice has still yet to catch up with what the technology would allow. See generally Martin, supra note 22, at 345, ¶ 35 (“Most court Web sites remain locked onto the image of a decision as a printed document. The all-too-common approach at judicial sites is to present opinions in files designed to replicate the print slip opinions formerly distributed by the court.”). But cf. Berring, Crumbling Infrastructure, supra note 1, at 275 (with regard to commercial databases in the
...much of what emanates from computers strives to be similar to print.”

Although written in 1989, this observation still rings true today, particularly when it comes
to decisions published by courts. The problems stemming from the decline of print
have yet to be solved; in fact, these problems may have become so familiar as to
become somewhat invisible. Perhaps this explains why many of the insights offered
in the scholarship cited above feel as immediate and relevant as ever.

Meanwhile, Susan Nevelow Mart identifies another reason to be concerned
about the shift to digital legal research platforms: their lack of “algorithmic account-
ability.” Over the last several years, she has published a series of studies designed to
test the effectiveness, accuracy, and consistency of the major legal research plat-
forms. These studies have exposed great variation in the behavior of the current
crop of legal research databases, which is worrisome given their veneer as objective,
unbiased systems. For example, Mart uncovered significant discrepancies between
Lexis and Westlaw in how each platform generates and classifies headnotes, as well
as how individual headnotes link to other cases via each platform’s citator engine.
Later, Mart found a great deal of variation between the results returned by Lexis
Advance, Westlaw, Fastcase, Google Scholar, Ravel, and Casetext in response to rela-
tively simple case law searches, both in terms of relevancy rankings and—more
alarmingly—whether cases were included in the results at all.

On a practical level, these studies showed how the algorithms and hidden
metadata practices employed by research databases have concrete effects on
research outcomes. For comprehensive research, it may be necessary to use a
variety of search engines rather than just one. On a more conceptual level, these
year 2007, “new forms of the digital sources have emerged that do take advantage of the potentials
of the medium.”).

44. Katsh, supra note 33, at 261.
45. Databases like Ravel and Fastcase have now integrated features to visualize relationships
among court opinions; however, these tools are likely limited by the underlying data they draw from:
court opinions, the machine-readable Bluebook citations contained within, and very basic metadata.
46. See Mart, Algorithm, supra note 6; Susan Nevelow Mart, The Case for Curation: The Relevance
of Digest and Citator Results in Westlaw and Lexis, 32 LEGAL REFERENCE SERVS. Q. 13 (2013) [hereinafter
Mart, Curation]; Susan Nevelow Mart, The Relevance of Results Generated by Human Indexing
and Computer Algorithms: A Study of West’s Headnotes and Key Numbers and LexisNexis’s Headnotes
and Topics, 102 LAW LIBR. J. 221, 2010 LAW LIBR. J. 13 [hereinafter Mart, Human Indexing and Com-
puter Algorithms].
47. See Mart, Algorithm, supra note 6, at 394, ¶ 12 (“Technical bias is built into systems. We just
don’t see it because the systems we use are black boxes.”).
49. Id. at 244–49, ¶¶ 39–50.
50. In this study, Mart tested a collection of “medium-good starting searches” that would be
representative of what a typical attorney might initially use as search terms. Mart, Algorithm, supra
note 6, at 409, ¶ 36.
51. Id. at 390, ¶ 5 (“There is hardly any overlap in the cases that appear in the top ten results
returned by each database.”).
52. Id. at 420, ¶ 57 (“Even for returning results from searches in a specific case database, every
algorithm draws on a different set of sources and processes, whether those sources and processes are
classification systems, secondary sources, citation networks, internal case analyses, mined user search
history, or machine learning deployed in the unique environment each legal database provider offers.
These algorithmic variations in worldview lead to substantial variations in the unique and relevant
results each database provides.”).
53. Id. at 390, ¶ 6 (“From the law professor seeking to set up a corpus of cases to study, to
the trial lawyer seeking that one elusive case, to the legal research professor showing students the
articles posed difficult questions about the opaqueness and accountability of online research platforms. In print, the sources of legal research were basically transparent. Although we weren’t privy to the thoughts and reasoning of each West employee as they processed court decisions, the editorial process ultimately produced a classification scheme that was entirely perceptible and thus open to analysis and criticism. (Of course, most legal researchers were content with engaging the print research system on a surface level.) Early versions of keyword searching, an exercise in strict Boolean logic, were also fairly transparent. If you knew the coverage of the database, it should have been evident why any given case was present in, or absent from, the results list. Also, you would expect different databases with similar coverage to return similar results in a Boolean environment, though perhaps in a different order if ranked by relevancy.

However, when algorithms play a more active role in the search process and much of the underlying metadata is hidden from users, legal research tools become more of a “black box.” Search terms are fed in, results are churned out—but what happens in the middle is a bit of a mystery. Given this lack of “algorithmic accountability,” how can users evaluate the overall effectiveness of these databases and determine which tool is best suited for particular tasks? How can potential biases be spotted and either leveraged or counteracted? These questions will become only more pressing as algorithms, likely to be described in the marketing literature as “A.I. powered,” undertake more of the work previously left to the researcher.

limitations of algorithms, researchers who want full results need to mine multiple resources with multiple searches.”); see also Mart, Human Indexing and Computer Algorithms, supra note 46, at 244, ¶ 38 (“If comprehensive research is required, either both [Westlaw and Lexis] must be used, or the researcher must fill in the gap by making sure that enough secondary sources have been reviewed to assure a good complement of seed cases.”).

54. See, e.g., Mart, Algorithm, supra note 6, at 395–96, ¶ 14 (“Algorithmic accountability is the term for disclosing prioritization, classification, association, and filtering. What we need is a frank discussion with database providers about what it means to search in their databases.”).

55. Id. at 393, ¶ 11 (“Going beneath the surface of research systems, even in the predigital search environment, has never been the norm. There is a long history in legal research of researching with only a surface understanding of the underlying structure.”).

56. See id. at 391–92, ¶ 9.

57. “Hidden” metadata would include things like logged user behavior history, rather than user-facing metadata such as Topic and Key Number classification.


59. Id. at 389, ¶ 2 (“Legal researchers are not likely to be able to tell how the encoded biases and assumptions are affecting search results. Legal database providers have viewed their algorithms as trade secrets and so have been reluctant to discuss the algorithms.”). Though, as demonstrated in this study, carefully designed experiments can expose some of the biases inherent in a database’s algorithm. See id. (“In the absence of transparency from the database providers themselves, there may still be things that can be learned about system biases. This article sets out the results of a study designed to reveal how hidden biases and assumptions affect the results provided by some of the major legal database providers.”).

60. See id. at 420, ¶ 58 (“Black-boxing the research process is not helping educators or students achieve this goal. Algorithmic accountability will help researchers understand the best way to manipulate the input into the black box and be more certain of the strengths and weaknesses of the output.”).

61. See id.

62. Id. at 396, ¶ 15 (“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.”); cf. Berring, Chaos, Cyberspace and Tradition, supra note 4, at 209–10 (“The danger of the high-end products is that each step in the research process that is carried out automatically by the front end system, is a step taken away from the purview of the researcher. Each decision that is built into the system makes the human
¶21 In summary, the transition of legal information from print to electronic formats has been a hot topic in the law librarian community, and this was particularly true in the nineties and early aughts. The resulting books and articles sketched a rough, and often eerily prescient, outline of what might (and did) happen in the intervening decades. Many questions raised in this body of scholarship remain unanswered, while more recent scholarship has posed additional conceptual and practical concerns. For example, has the nature of legal authority changed, due to information overload and the reliance on decontextualized keyword searching? How can we protect the public’s access to useful and authoritative legal information, when sources crucial to the legal research process may cease to be available in print? How can researchers be confident in their research, when they do not have a good sense of how legal information databases work due to their “black box” quality? Outright answers or solutions are unlikely to be simple. However, by reimagining the role of the court system in the marketplace for legal information, we may at least mitigate these problems.

Modeling the Drift Toward Digital: Winners and Losers

¶22 At this point, it would be useful to model the essential elements of the legal information system and describe how the relationships among them have changed during the digital revolution. There are many ways to conceptualize the structure of the legal information system. Drawing inspiration from a model proposed by Richard A. Danner, an effective legal information marketplace requires (1) a “source” of the legal information; (2) an “editor” of this information, to make it digestible and useful; (3) a “publisher,” to fix this information in a stable format and to disseminate copies as needed; and (4) an “access point” for this information. Different parties or institutions can take on multiple roles or even share overlapping responsibilities for some of these tasks, depending on the circumstances.

who is doing the search one level further removed from the process. If each user of information was aware of these steps, if each user understood what was being done for her and could monitor results with a skeptical eye, the danger would not be so great. But the whole point of these systems is to work automatically.”). For a discussion of algorithmic accountability in the AI context, see Baker, supra note 17, at 22–25, ¶¶ 62–68.

63. See, e.g., Robert C. Berring, Collapse of the Structure of the Legal Research Universe: The Imperative of Digital Information, 69 Wash. L. Rev. 9, 16–17 (1994) [hereinafter Berring, Imperative of Digital Information] (“An information system is an ordering of any form of data in a way that makes it understandable and retrievable. Think of every information system as having two parts. The first part is the database of information, the second part is the organizing system.”); Bruce, supra note 28 (considering which actors should play which roles in the legal information system); Thomas R. Bruce, Legal Information, Open Models, and Current Practice, 30 Revue juriQuE THEMIS R.J.T. 181–87 (1996) (arguing that the government is a “wholesaler” of information, but that this “raw material . . . needs the added editorial and organizing value provided by academia, practitioners, and publishers if it is to be genuinely useful . . . .”); Katsh, supra note 14, at 454–55 (mapping the relationship between courts, libraries, lawyers, and the public to show how the model will shift in a digital environment); McGinnis & Wasick, supra note 2, at 998–1000 (“There are two basic ways for law to organize information: (1) a centralized, top-down approach [e.g., codified law] and (2) a distributed approach [e.g., common law.”).

64. Richard A. Danner, From the Editor: Big Things, 86 Law Libr. J. 185, 188 (1994) (asserting that the legal information environment comprises “(1) the creators of legal information, (2) its publishers and distributors, (3) law librarians, who acquire, organize and assist users in locating information, and (4) the users themselves.”).
In the print world, the source of primary legal information was the government, most commonly the legislature, administrative agencies, and the courts. Depending on the context, different parties—or multiple parties—could play the role of editor by writing annotations and headnotes, making classifications, and creating various tables, indices, and finding aids. Many state court reporters, for example, would add synopses and headnotes to decisions, though this became less common over the years. Some states would annotate their official codes. Often, however, the heavy-duty editorial work fell to profit-driven legal publishers. Similarly, the printing of legal information was sometimes done by the government, but publishing was commonly outsourced to companies like West and Lexis (or their predecessors). This made sense: publication in print is expensive and can be logistically complex. As far as access points go, these existed wherever the print volumes ultimately ended up, oftentimes libraries.

We are now at the tail end of an extended transition from the print to digital format. This has been a messy process, and stakeholders in the traditional print world have had to adapt and shift responsibilities. Of course, even in 2020, the government remains the initial source of legal information. Editorial work is still mostly done by the legal publishers (sometimes now algorithmically), though some states continue to publish case summaries in the form of synopses or headnotes.

The significant developments have occurred in the realms of publishing and access, both of which are increasingly accomplished digitally and online. Whereas the traditional roles of “publisher” and “access point” were distinct and

---

65. Id.
66. See Peter W. Martin, Abandoning Law Reports for Official Digital Case Law, 12 J. APP. PRAC. & PROCESS 25, 34 (2011) (“Today, far fewer than half the states have a judicial officer so denominated and there are no more than a baker’s dozen of jurisdictions (twelve states plus the United States Supreme Court) in which a public reporter of judicial decisions and staff perform the full range of functions traditionally associated with official case law publication.”); see also States with Court Provided Headnotes, Official Headnotes, and Syllabi, LexisNexis, http://lexisnexis.custhelp.com/app/answers/answer_view/a_id/1075428/~states-with-court-provided-headnotes%2C-official-headnotes%2C-and-syllabi [https://perma.cc/JPK8-VV8A] (note that while this chart includes good historical data, it does not seem to be regularly updated). Washington provides a good example of a state where the court reporter played an active role in editing cases, which included authoring headnotes. Tim Fuller, “The Most Accurate and Useful Law Books Possible,” Wash. Terr., Wash., Wn.2d, and Wn. App, Milestones of Official Case Reporting in Washington, https://www.courts.wa.gov/appellate_trial_courts/supreme/?fa=atc_supreme.milestones [https://perma.cc/BYT3-JT7H]. Today, Washington headnotes are written by Lexis editors under the supervision of the court reporter. Martin, supra, at 34 n.30.
67. Nevada, for example, still publishes its own annotated code. See Nev. LEGISLATIVE COUNSEL BUREAU, LEGISLATIVE COUNSEL BUREAU: AN OVERVIEW (2016), https://www.leg.state.nv.us/Division/Research/Publications/Misc/LCBOverview.pdf [https://perma.cc/HQ7C-8L77].
69. See, e.g., Mart, Human Indexing and Computer Algorithms, supra note 46, at 226, ¶ 8 (“West creates a direct correlation between a headnote (drafted by a human editor) and the related key number topic, relying primarily, but not exclusively, on human editing to assign headnotes to a point in a classification system. LexisNexis relies primarily, although not exclusively, on algorithms to assign a headnote (taken from the court’s language) to a topic in the classification scheme . . . .”).
70. For example, Kansas provides court-authored syllabi. Martin, supra note 2, at 12 (“Kansas . . . is also one of a small number of states in which summaries of the key points of law in an opinion (the syllabus or set of headnotes) are prepared by the court itself, rather than added by the reporter or a private contractor”); see also Cases and Opinions, KAN. JUD. BRANCH, http://www.kscourts.org/Cases-and-Opinions/default.asp [https://perma.cc/FF9V-Q6RY].
separate, now these roles are often blended.\textsuperscript{71} Many of the major legal publishers (Westlaw, Lexis, Bloomberg/BNA, and Wolters Kluwer, to name a few) “publish” their products in online databases that double as access points—at least, for the legal professionals who can afford access. Although the public often enjoys some access to expensive research databases through law libraries, these kinds of licensing arrangements should not be taken for granted.\textsuperscript{72}

\textsuperscript{71} The government also publishes and provides access online, usually without extensive editorial content or other enhancements.\textsuperscript{73} In the print world, the government was always the initial source of primary legal information, and sometimes acted as editor or publisher, too. The government’s expanding role as an access point for legal information, however, is somewhat new, or at least a departure from tradition. In the past, the government provided access points to legal information more or less exclusively through libraries; now, government entities such as courts, agencies, and legislatures can disseminate legal information to citizens directly over the Internet, without the library acting as intermediary.\textsuperscript{74}

\textsuperscript{72} Given our federal system of government, it should come as no surprise that the robustness and functionality of these databases vary greatly from state to state, branch to branch, and government level to government level. For the most part, though, state websites host appellate-level court opinions on the Internet,\textsuperscript{75} along with state codes, session laws, and other legislative or administrative material. Likewise, the GPO, at govinfo.gov, provides access to many federal district court, court of appeals, and Supreme Court decisions, along with the \textit{United States Code}, the \textit{Code of Federal Regulations}, legislative history material, presidential documents, and more.\textsuperscript{76}

\textsuperscript{71} In fact, some might argue that making a distinction between “publishing” and “providing access” breaks down entirely on the web. However, it’s easy to find examples where this distinction holds up, such as when one entity makes data and metadata available for reuse, and another entity harvests this data and displays it to the public. In \textit{Law in a Digital World}, Katsh is very critical of importing print metaphors into the digital sphere: these metaphors are usually somewhat imprecise and can impose limitations on how we think of and use new technologies. See, e.g., Katsh, \textit{supra} note 14, at 407–08.

\textsuperscript{72} See Berring, \textit{Chaos, Cyberspace and Tradition}, \textit{supra} note 4, at 207–08 (“The producer of electronic information does not need libraries. The heart and soul of electronic information has been direct marketing to the end user.”).


\textsuperscript{74} See Danner, \textit{supra} note 64, at 188–89.


\textsuperscript{76} As of now, govinfo.gov hosts circuit court, district court, and bankruptcy court decisions dating as far back as 2004, but much of this collection is incomplete. Coverage varies across different courts and jurisdictions. It is also possible to find opinions older than 2004, but this seems to be limited to opinions directly related to cases decided after the 2004 cutoff (for example, where a case was originally decided in 2002, but appealed or otherwise revisited in 2005). \textit{About United States
These efforts go a long way in ensuring that every citizen can access basic legal information. But even the best government legal information databases lack tools that would be considered fundamental in a traditional law library. On one hand, compared to its state-level counterparts, govinfo is likely the most robust government-run case law platform. In govinfo, advanced search fields include full-text, party name, court, nature of suit, case number, party role, citation, and others. Multiple fields can be searched simultaneously, making it possible, for example, to conduct a full-text search within specific jurisdictions. On the other hand, govinfo does not provide a citator, does not organize cases by legal issue, and includes only a fraction of the case law originating from the federal court system. In other words, although govinfo and similar government websites succeed in disseminating new cases to the public, they are not viable substitutes for, say, Lexis Advance or a law library.

In addition to the major legal publishers and various government entities, newer market entrants have begun filling the dual roles of “legal publisher” and “legal information access point.” These entities, such as Google Scholar, generally obtain legal information in the form of bulk data and then employ automated processes and algorithms to provide search tools. Court Listener, a nonprofit, provides access to millions of state and federal opinions and other court documents, sourced from volunteers, donations, and a system that scrapes data from court websites. The coverage of these databases is much closer to comprehensive than govinfo’s, case citations are often hyperlinked, and some even include basic citator functions. Still absent from these databases is a means to truly Shepardize a case, subject indexing, classification in any real detail or, of course, any human-generated editorial content.

---

COURTS OPINIONS, U.S. GOV’T PUBL. OFF., https://www.govinfo.gov/help/uscourts#about [https://perma.cc/HV8W-K5K8] (“United States Courts Opinions (USCOURTS) collection is a collaborative effort between the U.S. Government Publishing Office (GPO) and the Administrative Office of the United States Courts (AOUSC) to provide public access to opinions from selected United States appellate, district, and bankruptcy courts. The content of this collection dates back to April 2004, though searchable electronic holdings for some courts may be incomplete for this earlier time period.”).

77. Though it does include a field for “nature of the suit,” which is extremely broad (e.g., “patent”) compared to the specificity found in the Key Number System. At the state level, there are no court websites that include subject indexing for case law. SARAH GLASSMEYER, STATE LEGAL INFORMATION CENSUS: AN ANALYSIS OF PRIMARY STATE LEGAL INFORMATION 5 (2016).

78. In 2008, Ian Gallacher noted that “not only has the Internet not fulfilled its potential to make the law freely available to everyone, the legal community’s reliance on Internet-based legal information is helping the decline in book-based legal research, which in turn is helping to constrict open and free access to legal information.” Ian Gallacher, “Aux Armes, Citoyens!:“ Time for Law Schools to Lead the Movement for Free and Open Access to the Law, 40 U. Tol. L. Rev. 1, 21 (2008). The situation is somewhat improved now, but this observation still holds up over a decade later.

79. Although it is questionable whether some of these companies are true legal publishers, they are certainly access points for legal information.


81. COVERAGE, COURT LISTENER, https://www.courtlister.com/coverage/ [https://perma.cc/LCJ7-D2V7].

82. For instance, Google Scholar can display all of the cases citing to a particular case, as well as an estimate of the depth of discussion, but it will not really tell you how these later cases cited to the original case—for example, whether the treatments were negative, positive, or somewhere in between. Similarly, Court Listener lists citing references without state of the law information.
Still, these free Internet-based entities have shown that it is possible to be a functional “access point” for legal information, without having to worry too much about editing or really even “publishing” this information. As mentioned above, Google receives dumps of bulk data containing court opinions and then hosts this data on its servers—this is the extent of its role as publisher. In terms of editing, Google provides basic citator functions, hyperlinks, and field searching—but this is all done algorithmically and, it is likely, at little ongoing cost.

Similarly, a variety of lower-cost databases have entered the fray, among them Fastcase, Casemaker, Casetext, and Judicata, to name a few. Often, these newer market entrants employ innovative algorithms to replicate features of Westlaw and Lexis that traditionally required a considerable amount of human input. In some cases, these algorithms function independently from human editors; in other cases, the algorithms work in tandem with human editors, to increase efficiency and thus reduce costs for end users. As such, the functionality and reliability of the different components of these databases can vary.

Law libraries have adapted to new roles as well. Primarily access points in the past, they have increasingly taken on the role of publisher, spurred by the comparatively low cost of online publishing. For example, it is commonplace for academic law libraries to oversee the digital publication of law reviews and faculty scholarship. A few libraries have created (or played a role in creating) online databases of case law and other primary law documents. The most well-known example would be Cornell’s Legal Information Institute (LII), which hosts, among other things, the United States Code, Code of Federal Regulations, Supreme Court decisions, and a collection of original content including the Wex legal encyclopedia. The Rutgers Law Library partnered with New Jersey state and federal courts to publish case law through a database they built from the ground up. It captures a good amount of metadata, including citations, citing references, party names, decision date, and docket number. Recently, Harvard Law School Library partnered

---

83. Here, the distinction between "publication" and "providing an access point" is tricky to delineate; perhaps so many traditional "publishers" have also become online portals because the distinction between these two roles breaks down in an online environment. See supra note 71.

84. For example, Fastcase displays state of the law information, but this is handled by automatically scanning Bluebook citations for words like “rev’g.” Thus, a case won’t be flagged as overturned until another case comes along that explicitly recognizes (in a Bluebook-formatted citation) the fact that the original case has been overturned. Casemaker does provide complete state-of-the-law information, edited by humans. Greg Lambert, Casemaker Unique Among Legal Research Providers, Mich. Bar. J., Nov. 2010, at 54. Casetext uses a combination of algorithms and human editors to provide state of the law information. How Was Casetext Citator Built, CASETTEXT, https://help.casetext.com/casetext-citator/how-was-casetext-citator-built (last accessed May 17, 2019).

85. I say “primarily” because law librarians have a track record of editing and writing all kinds of research guides, bibliographies, books, law review articles, and so on.

86. These activities seem to combine different elements from three roles: editor, publisher, and access point.


88. Though, the quality of the metadata varies depending on the practices of the issuing court. See John P. Joergensen, The New Jersey Courts Publishing Project of the Rutgers-Camden Law Library, 94 LAW LIRN. J. 673, 677, 2002 LAW LIRN. J. 42, ¶ 21 (“One problem we have is that it is impossible to get 100% accuracy in automatically extracting metadata from documents prepared and formatted by various people. In the case of the U.S. district court decisions, the formatting and placement of useful information varies so greatly from document to document that we cannot do it at all. Since we do not have the time or personnel to go through these documents individually, we do not extract any metadata.
with Ravel (now owned by Lexis) to create a complete database of published case law, which required Harvard to scan a huge amount of historical case reporters.89

¶33 As outlined above, the contemporary crop of legal information databases break down into four rough types: expensive, one-stop shops run by major legal publishers; newer market entrants that use technological innovations to mimic the functionality of more traditional databases, but at a lower cost; a hodgepodge of government, library, and nonprofit websites, with varying degrees of coverage and functionality; and tools like Google Scholar that aspire to be free alternatives to Westlaw, Lexis, and Bloomberg but are missing crucial tools required for legal research. This marketplace-in-flux often duplicates efforts (e.g., instead of “just” Shepard’s, we now have Shepard’s, KeyCite, BCite, and Fastcase’s Bad Law Bot), and it seems reasonable to assume that many of these companies and organizations will either disappear or consolidate in the coming years.

¶34 Ownership consolidation in legal publishing is written about extensively, often from a critical point of view.90 But another kind of consolidation is underway in the legal publishing realm that blends previously distinct roles in the legal information ecosystem. Whether government entities or private companies, actors at every stage of the online legal information cycle are making some effort to become an access point for legal information, rather than just some combination of source, editor, or publisher of legal information.91

¶35 Whether the public has benefited from this trend is debatable, but certainly true is the increasingly large gap between the research tools available to a large law firm and those available to most pro se litigants. This is an especially frightening proposition if we envision a future where legal research is conducted exclusively online.92 Entities like Google Scholar host great volumes of legal documents, but they do not provide services like Shepard’s or KeyCite.93 Instead, these services rely from them. As a result, the only search options for this collection are the full-text search engine and by docket number.“).


90. See, e.g., Arewa, supra note 5, at 820–28; Berring, Chaos, Cyberspace and Tradition, supra note 4, at 198–99; Berring, Cognitive Authority, supra note 7, at 1698.

91. Cf. Katsh, supra note 14, at 431 (“An electronic network turns everyone into a publisher in a different and more meaningful way. The network provides the facilities for individuals to distribute their messages efficiently and cheaply, both widely and narrowly, to large groups as well as small.”).

92. This is just speculation, but I believe print, like vinyl, will never go away completely. However, it is possible to imagine a world where individual publications cease to be available in print (or where some law libraries cease to purchase them). If, for example, Shepard’s ceased to be available in print, then legal research would require an online component; this would also be the case if the West Digest System moved exclusively into the Westlaw database. Many, perhaps most, law libraries have cancelled or limited their subscriptions to many series of Shepard’s. Laura C. Dabney, Citators: Past, Present, and Future, 27 LEGAL REFERENCE SERVS. Q. 165, 181–82 (2008) (specifically noting the University of Washington’s Gallagher Law Library; anecdotally, 12 years later, it would seem that it is commonplace now for libraries to no longer receive Shepard’s in print).

93. Ravel has a citator algorithm, but it is available in only the premium version. And, in any case, its reliance on citation parentheticals in Bluebook format makes it an imperfect proxy for something like Shepard’s. See Negative Treatment, Ravel, https://ravellaw.zendesk.com/hc/en-us/articles/115002279974-Negative-Treatment [https://perma.cc/D3HN-Z2H8] (“Ravel’s treatment indicator is not a replacement for looking at a Lexis Shepard’s report and should not be relied upon exclusively.”).
on full-text searching combined with some basic forms of automated indexing (e.g., by identifying the year of the decision and the name of the court that issued the opinion). To better serve the needs of legal researchers, these free services must have the capability to tell users whether a case is still good law. Detailed subject indexing and classification, the ability to conduct faceted searching, and hyperlinked citations to statutes, regulations, and other sources of law would make these free services even stronger. Many free services already provide some of these functions, though in a limited way. As discussed below in part 2, with a little help from the courts, tools like Google Scholar might evolve into fully viable alternatives to Westlaw, Lexis, or Bloomberg Law for cash-strapped legal researchers.

In a world full of access points for legal information, the role of the law library might be less apparent to some, but libraries are as vital now as they always have been. Law libraries facilitate access in ways that other actors in the legal information marketplace cannot. They pay for subscriptions and public access terminals, manage and organize digital collections to ensure that patrons can access the sources necessary for legal research, and employ librarians who can show patrons how to use these databases. Furthermore, as discussed in part 3, law librarians could play a role in improving free or low-cost legal information databases by more actively participating in the case law reporting process.

**Improving Public Access to Case Law Through Structured Metadata**

One way to ensure that solo practitioners, small firms, and the public at large continue to enjoy access to the resources and tools necessary to conduct legal research is to help newer market entrants evolve into fully realized legal research prototypes. Giving the public the ability to keyword search their way through legal material without feedback or guidance might do as much harm as good. As observed by Katsh in 1993:

Even if companies provided legal databases cost-free to the general public, they might be relatively useless to lay people because the boolean scheme is difficult to master and there are considerable differences in the rules and conventions for searching different databases. . . . A user logged on to most commercial databases today is physically close to relevant material but informationally distant from it. This is even more true of the individual who is interested in finding information on sources linked to the Internet.

Katsh, supra note 14, at 449; see also Bruce, supra note 24, at 31–34. Even for lawyers, it can be difficult to capture legal concepts in a keyword search. See Glassmeyer, supra note 77, at 5 (“It is doubtful whether or not full text searching is sufficient or useful in accessing legal information. As it stands, no state provides an index to its case law.”); Daniel Dabney, *The Universe of Thinkable Thoughts: Literary Warrant and West’s Key Number System*, 99 LAW LIBR. J. 229, 237, 2007 LAW LIBR. J. 14, ¶ 35 (“At its heart, even the cleverest natural language search engines draw much of their power from the ability of the system to recognize individual words. To the extent that the ideas of interest to lawyers can be reliably associated with individual words, those systems excel. But to the extent that there is a gulf between the individual words and the ideas of interest to the searcher, free-text systems are limited.”). See generally Daniel P. Dabney, *The Curse of Thamus: An Analysis of Full-Text Legal Document Retrieval*, 78 LAW LIBR. J. 5 (1985).

See generally Arewa, supra note 5, at 838 (“One of the biggest potential barriers to new entrants in the legal information industry is the scope and depth of existing proprietary publishing business models. These databases have fostered a market in which consumers expect to have access and the ability to search a large range of potential legal information. Lexis and Westlaw have thus built legal databases with millions of documents; the creation of such databases and the functionality that such players bring to their databases has shaped many users’ expectations of what electronic legal databases should offer.”).
tools. Free private sector portals like Google Scholar currently lack critical features present in pay services. By committing to making court opinions freely available with detailed metadata, courts can use the free market to help mitigate these access issues.

¶38 Librarians intuitively understand why court metadata practices are important—in fact, we have been advocating about this issue since the birth of the Internet. The following sections explicate these traditional rationales, discuss why these traditional rationales are still valid (even in 2020), and identify how court-authored metadata may be increasingly vital given the continued decline of print and the rise of AI and litigation analytics in legal research.

Government as Data Provider

¶39 As long as the Constitution is in force, the government will continue to be the original source of U.S. primary law. As discussed above, the government also plays the roles of editor, publisher, and access point for legal information. In the online environment, and particularly with regard to case law, the government should focus its efforts on editing its legal content in a format conducive to reuse by third parties—that is, by supplying detailed, quality metadata—rather than focusing too much energy on the creation of web-based portals to access this information.

¶40 Some argue that the GPO (and the government in general) should not be in the business of providing permanent access to government information, due to concerns about adequate, sustained funding. Government spending, as a political issue, can be inconstant and unpredictable. Xiaohua Zhu, while not taking sides on this issue, notes the following additional concerns about government-sponsored case law databases:

[Critics point to] the difficulty of collecting historical case law, the expense of maintaining a large system, the lack of legislation, the often-changing information policies of govern-

---


97. See David Robinson et al., Government Data and the Invisible Hand, 11 Yale J.L. & Tech. 160, 160 (2009) ("It would be preferable for government to understand providing reusable data, rather than providing Web sites, as the core of its online publishing responsibility."); see also Richard Susskind, The End of Lawyers? 268 (2008) ("The purpose of this sharing is not simply to give citizens sight of more documents. Rather, in the spirit of Web 2.0, it is to make public information available as a raw material that citizens, entrepreneurs, charitable bodies and many others can fashion, re-organize, and supplement for their own purposes."); Bruce, supra note 24, at 26 ("It is becoming increasingly obvious that the issuers themselves—the courts, legislatures and agencies—are in the best position to maintain and publish their own collections of data.").

98. See James A. Jacobs, James R. Jacobs & Shinjoung Yeo, Government Information in the Digital Age: The Once and Future Federal Depository Library Program, 31 J. Acad. Librarianship 198, 203–04 (2005) (arguing that the GPO should focus on “helping agencies package and deliver their various information products,” rather than taking on “the role of permanent public preservation and access . . . the role traditionally fulfilled by FDLP libraries . . . .”).

99. See Berring, Cognitive Authority, supra note 7, at 1705 ("The government, at the state and federal level, has never been a reliable provider of information. The impetus to publish and maintain information, and to provide quality access to it, has always been the market."); see also Gregory M. Silverman, Rise of the Machines: Justice Information Systems and the Question of Public Access to Court Records Over the Internet, 79 Wash. L. Rev. 175, 184–85 (2004) (discussing funding for justice information systems at the state level).
ment agencies, and the different ideologies or value systems about governments’ roles in providing public access to the law. Many still hold the belief that the government should not compete with the private sector in the area of information dissemination but should leave the business opportunity to commercial information providers . . . . 100

¶41 Still, a strong case can be made for some kind of government involvement in publishing and disseminating government information. As Richard Susskind argues, “[e]nlightened public information policy would encourage or even require public bodies to take on the job of drawing attention to the laws, regulations, and rules that are so central to their daily work and ensuring they are made more accessible and digestible to the citizenry.”101 Similarly, the American Association of Law Libraries (AALL) has recognized the vital importance of the public sector retaining responsibility for the preservation of authentic government records.102 In the private sector, businesses come and go unpredictably—doubly so in periods of rapid technological transition—and when they go, they tend to take their information with them.103 Governments, on the other hand, tend to be much more stable and resilient.104 Thus, although the private sector is well suited for developing innovative searching and analytical tools for electronic government documents, it might be unwise to rely on private companies for archiving these documents in a permanent state.105

¶42 If the government has stored its data in a standardized format and in a way that can be harvested by third parties, there is reason to believe that the private sector can provide robust search tools should the government portal be inadequate.106 According to Robinson et al., building an interactive site that searches,
displays, and organizes information can be accomplished inexpensively, when the source data is freely available in a standardized format: “[w]eb hosting is cheap, software building blocks are often free and open source, and new sites can iterate their designs rapidly.” In other words, the creation of a Westlaw-style search interface is the (relatively) easy part; the difficult, expensive part is amassing the data to be searched, adding editorial content, and providing the necessary metadata to make it useful and easily searchable.

¶43 By focusing on the data itself, rather than finding tools and web portals, the government could minimize the risk of an expensive project quickly becoming obsolete. The Internet constantly changes: once popular destinations can become ghost towns overnight. Frankly, when one thinks of highly adaptable, flexible organizations, the government doesn’t jump to mind as an exemplar. An expensive, labor-intensive government information system could be replaced overnight by an innovative Internet startup. This isn’t a bad thing at all; in fact, this phenomenon is one of the strengths of a free market system. Government information is meant to be used, repackaged, and monetized by private parties, which is one reason why government documents at the federal level do not receive copyright protection.

¶44 Here, it may be useful to draw an analogy between legal information on the “information superhighway” and an automobile on a physical highway. Most people probably wouldn’t want the government to manufacture their automobile. A government-built car would be slow, expensive, uncomfortable, aesthetically bland, and not very much fun to drive (though, possibly, safe and generally functional). People do, however, tend to trust the government with constructing the infrastructure for their cars (i.e., highways, tunnels, etc.). In the world of online legal information, user interfaces, search engines, AI-powered analytical tools, and the like are.

accessibility of too many judicial opinions, Martin notes that “[t]he rapid development of sophisticated Internet search tools provides strong evidence that with the right combination of public sector involvement and private sector competition in the dissemination of legal information” this may never become much of a problem at all. See Martin, supra note 2, at 44. Examples exist of the private sector creating portals to government data when the government-provided portal was seen as inadequate. Google Patents, for example, is a much more user-friendly interface than the search platform hosted by the USPTO. A few years back, in response to perceived shortcomings of the Regulations.gov website, Cornell helped launch the “Regulation Room,” a website dedicated to providing access to federal regulations in a Web 2.0 context, complete with RSS feeds, blogs, and social networking features. Claire Cardie et al., Rulemaking 2.0, 65 U. Miami L. Rev. 395 (2011). Regulation Room has since gone offline, however.

107. Robinson et al., supra note 97, at 170.
108. For example, the social media sites Friendster and MySpace suffered severe declines in popularity due to the rise of Facebook. See generally Liat Clark, Researchers Conduct Autopsy of Social Network Friendster, WIRED (UK) (Feb. 27, 2013), http://www.wired.co.uk/news/archive/2013-02/27/autopsy-of-friendster [https://perma.cc/K9HL-N8FA] (noting that social media sites can be prone to sudden mass exoduses).
109. Robinson et al., supra note 97, at 173 (“The institutional workings of government make it systematically incapable of adapting and improving Web sites as fast as technology itself progresses.”). See generally Berring, Cognitive Authority, supra note 7, at 1705–06 (arguing that innovation in the legal information sphere will likely come from the private sector, as has been the case in the past); Robert C. Berring, On Not Throwing Out the Baby: Planning the Future of Legal Information, 83 Cal. L. Rev. 615, 616 (1995) [hereinafter Berring, Throwing Out the Baby] (“[M]arket forces, rather than governmental fiat, should dictate changes in the legal information system.”).
111. Cf. Katsh, supra note 14, at 440 (employing a similar metaphor, but focusing more on the physical structure of the Internet).
the cars we drive, and metadata is the infrastructure—the boring, heavy-duty stuff holding it all together.

¶45 For metadata to provide the necessary infrastructure for a full-fledged legal research tool (e.g., Westlaw), it must be consistently applied (you wouldn’t want your highway to have unpaved gaps), contain as much detail as possible (you’d want your highway to have all of the necessary signposts so you don’t miss your exit), adhere to the same metadata standard (you wouldn’t want to switch to the left-hand lane halfway through your trip), and use a common metadata language (you wouldn’t want your highway to suddenly turn into a monorail system). Happily, all of these components are feasible or extant (or in the midst of development). The metadata language is XML; one option for a standard is called LegalXML, which is still being perfected.112 These technologies are now commonplace in court e-filing systems, libraries, government repositories, and throughout the Internet in general. But, as discussed below, they have yet to fulfill their full potential with regard to online case law publication.

Structured Metadata: The Big Picture

¶46 In the term “eXtensible Markup Language” (XML), “markup” refers to “information embedded in the text of a document that is not intended for printing or display.”113 Rather, it is meant to be read by a machine. Through this “metadata,” XML can tell a computer what a document means, in a limited sense. The greater the detail included in the metadata, the more a computer can be programmed to “know” about a document.114

¶47 The key strength of XML is its standardized format, which is still flexible enough to apply across different information systems. Government repositories and e-filing systems commonly employ XML. If different courts used the same XML scheme, each court could use its own case management system, but the systems could still “speak” to each other by means of that shared XML language.115 Metadata applied in this way links to related data (in the same database or elsewhere), which gives it a potentially profound amount of power:116 when the terms “Semantic Web” or “linked data” are used, they are referring to this concept. In this way, government-wide adoption and implementation of a shared XML standard for legal materials would allow for better integration of court records with documents published by the legislative and executive branches, such as statutes, regulations, and administrative rulings.117

112. This article focuses on XML because it seems to be the most commonly used of these technologies, but any machine-readable markup language might work fine, as long as it was used consistently across different courts and institutions.
114. See id.
115. See id. at 185–86.
116. See Edward L. Rubin, Computer Languages as Networks and Power Structures: Governing the Development of XML, 53 SMU L. Rev. 1447, 1448 (2000) (“XML facilitates thought and allows knowledge to cumulate over space and time . . . it becomes more effective and more powerful the more widely it is used.”).
Gregory Silverman identifies three broad categories of XML metadata. First, there is procedural markup (also called presentational markup). This metadata tells a computer how to display a document. This includes instructions about font type, font size, margin size, and similar elements. Second, there is structural markup, also called descriptive markup. This metadata identifies the general type of data being tagged—for example, this would tell a computer the title of the document. Last, and most important for the purposes of this article, comes semantic markup. This markup tells a computer what pieces of data in a document mean, in a limited sense. Applications of semantic markup might include labeling legal rules or issues, or noting whether a court’s treatment of precedent is supportive or negative.

Metadata in XML format is added to a document through “tagging.” For example, the surnames of plaintiffs in court documents could be identified by a tag like “<PlaintiffLastName>.” Here, it doesn’t matter how the tag is characterized (e.g., “PlaintiffLastName” versus “LastName”) as long as the tags are applied consistently within the court and across different courts—that is, as long as each court uses the same metadata standard. A court document tagged in this manner would look something like this with the metadata visible:

```
<PlaintiffLastName> Doe </PlaintiffLastName>
```

“<PlaintiffLastName>” is the “start tag,” and “</PlaintiffLastName>” is the “end tag.” Tags can be given attributes (like a numerical value, e.g., <Date Received=“05.16.2018”>). And an “empty tag” contains no element. All tags can be nested, which allows for multiple tags to be applied to the same chunk of text as well as the tagging of tags themselves.

Metadata becomes even more powerful when it is tied to a metadata standard that is sufficiently detailed and complex, which requires that the relationships among metadata tags be specified and defined. For example, the ideal metadata standard for court opinions might define the hierarchical relationships among different causes of action and legal issues, much like the West Key Number System. In this way, XML standards allow free-floating bits of data to solidify into structured data, which can more effectively be organized, parsed, searched, arranged, and linked by search engines and analytical tools.

In its raw form, with the metadata tags visible, XML looks cumbersome. However, court personnel likely rarely see XML data in its raw format. With regard to simple metadata, such as party names and filing dates, much of this is generated
through e-filing systems automatically. More complex and granular metadata, such as identifying and classifying legal rules, would be trickier to apply and might require new tools and interfaces. Fortunately, a variety of computer applications (and computer application add-ons) have been developed to make this process more user-friendly. Anyone can download a plug-in enabling users to tag text in Word documents simply by highlighting it, though courts may opt to develop or purchase bespoke tools. Once a court settles on a metadata standard (e.g., “<holding>” to be applied to text containing the court’s holding), this process could involve someone, perhaps a judge, clerk, or other court personnel, highlighting text and applying tags from a preset (i.e., controlled language) list. Alternatively, court personnel might input parts of an opinion into a form that would store the data in XML format, with fields for headnotes, subject classifications, factual background, rules applied, outcomes, and so on.

The Proof Is in the Patents: Comparing Google Patents with Google Scholar

To illustrate the effects of robust government-authored metadata, the U.S. Patent and Trademark Office (USPTO) provides an instructive example. When the USPTO publishes a patent, it does so with a rich, detailed collection of metadata that includes the name of the inventor, name of assignee, citations, classification, information about the patent examiner and prosecution process, and the various document sections (summary, claims, description, etc.). All of this work is done up front, as patents are processed. The end result is an extremely powerful search tool on the USPTO website. Being a government product, however, the search engine is rigid, difficult to use, and 20 years old in appearance.

The USPTO does not hoard this data for itself, thankfully. Private sector entities like Google Patents have used this data to create similarly powerful but much more user-friendly patent search and retrieval tools. In addition, because Google Patents receives data from other national and international patent offices, it can search multiple jurisdictions simultaneously and draw connections between patents filed across different jurisdictions. Overall, its functionality is roughly equivalent to that of the basic patent search tool in Westlaw.

By comparison, Google Scholar’s case law search falls far short of the standard set by Westlaw, Lexis Advance, or Bloomberg Law. Most important, Google does not provide users with citator tools comparable to KeyCite or Shepard’s.
This means a user is not given an immediate indication of whether a case is still good law or whether the decision is the ultimate disposition of the case.\textsuperscript{134} This could be particularly dangerous to inexperienced lay users, who might not consider these issues unless the information were provided in a visible, easy-to-understand fashion.

\textsection{55} Google Scholar does have a rudimentary citator, which allows researchers to see where a case has been cited subsequently. Although Google estimates the depth in which the cited case has been discussed in subsequent cases, it does not indicate clearly and succinctly how subsequent cases treat the preceding case.\textsuperscript{135} Nevertheless, a determined and cash-strapped legal researcher could use Google Scholar to check whether a case is good law, if they had the legal training to read through each decision citing the case in question. This time-consuming task is probably completely impractical for a layperson, however.

\textsection{56} In addition to its lack of a fully functioning citator, Google Scholar fails to index or classify its case law by subject and, unsurprisingly, adds no editorial content.\textsuperscript{136} Finally, although Google Scholar hyperlinks to cases cited within an opinion, it does not do the same for legislation or administrative materials. This is likely because Google does not host materials from the other branches of government, but most of these sources could be hyperlinked if the will were there: most state and federal codes are freely available online.

\textsection{57} So why is Google Scholar’s case law search engine so weak compared to Google Patents? It would seem that the principal difference is the amount and quality of the metadata that these tools have to work with.\textsuperscript{137} The functionality of Google Scholar is limited by the data Google is given: unlike Westlaw, Google does not have an army of editors to summarize cases, index these cases by narrow legal topic according to a proprietary scheme, characterize citations, and the like.\textsuperscript{138}

\textbf{How Structured Metadata Can Reinforce the Infrastructure of Legal Research}

\textsection{58} By authoring detailed, high-quality metadata in every legal opinion, courts can help market entrants like Google Scholar become fully functioning legal research tools, thereby expanding meaningful access to legal information and fos-
tering a more vibrant and competitive legal information marketplace. As a beneficial byproduct, courts would gain greater control over how their decisions are found, interpreted, and applied, activities that legal publishers may have influenced too heavily in the past. To accomplish this goal, the public sector (courts and libraries, generally) will need to conduct some of the editorial work normally performed by legal publishers.

¶59 In a sense, by subscribing to legal research platforms like Westlaw and Lexis, courts are paying for the government’s work product, aggregated into a single database and accompanied by editorial content. This arrangement seems sensible in a print world: publishing is expensive, gathering and compiling court decisions from every jurisdiction is a logistical nightmare, and the value-added, labor-intensive components like Shepard’s and West’s Digest System are necessary to making case reporters useful to researchers.

¶60 In the digital world, courts purchase many of the same specialized tools; however, given the state of technology, it might make sense for them to create these tools themselves—or, better yet, to lay down the infrastructure necessary for their creation. The inclusion of detailed, structured metadata in electronically published court opinions is a relatively low-cost way to improve the quality of online research tools—whether free or otherwise. As Martin notes, “[t]aking this largely invisible step can have a positive effect on the usefulness of court websites and, at the same time, reduce the costs of redistribution through commercial systems.”

¶61 For example, the inclusion of metadata relating to how cases are cited and discussed within an opinion would allow for third parties to create programs that mimic Shepard’s or KeyCite, as well as more precise hyperlinking across court documents. KeyCite and Shepard’s rely on a combination of human editors and computer algorithms to function. In both systems, human editors provide some case analysis upfront, and the rest of the work is handled automatically. It is this

139. Cf. Berring, Crumbling Infrastructure, supra note 1, at 295 ("What happens when the legal system has to deal with the fact that it has lost control of the sources of the law itself?").
140. See id. ("Some basic work, work that lacks glamour and perhaps profit, needs to be done on the infrastructure of legal research").
141. Note that this is not a break from tradition. In the print era, courts often outsourced the publication of official reports to a legal publisher and then purchased the volumes back at a discount. See Martin, supra note 66, at 54. There are parallels here with academic publishing, where universities pay researchers, the resulting research is published in for-profit journals, and then the journals are sold back to university libraries.
142. Martin, supra note 2, at 38.
143. Id.
144. LegalCiteM seems to still be under development, as part of the LegalXML standard. Oasis Legal Citation Markup (LegalCiteM) TC, OASIS, https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=legalcitem [https://perma.cc/BE99-3S8M]. Many of these OASIS LegalXML initiatives are global in scope, which might be ambitious given the differences among legal systems, as well as the challenge of getting academics, programmers, and legal entities throughout the world to agree on standards. Some scholars have argued that the government should play a more active role in this process, given the potential power of XML standards. See Rubin, supra note 116, at 1449 ("[T]he political character of XML language creation requires that these organizations be regulated and that the more likely regulator is the government, not W3C. But . . . this regulation, like the regulation of the negotiated rulemaking process, should not specify substantive results, but simply specify the structure of the non-governmental decision making body that creates the language.").
145. Dabney, supra note 92, at 177 ("[T]oday both programs do case analysis editorially, and the rest of the process, including finding citations and headnote assignment (and, in KeyCite,
human component that essentially bars Google Scholar and its ilk from providing full-fledged citator tools.\textsuperscript{146} Computer algorithms require upfront costs to develop but are otherwise inexpensive; human editors require salaries and health insurance. If at the outset courts completed some of the work of Westlaw's human editors,\textsuperscript{147} Google would just need to develop an algorithm—a task Google's track record proves it does well. This would also have the effect of lowering Westlaw's cost of doing business\textsuperscript{148}—and given how expensive subscription services like this are, customers could benefit if Westlaw passed along its savings. More competition in the citator marketplace might be a good thing, as recent scholarship casts some doubt on the accuracy and consistency of KeyCite and Shepard's.\textsuperscript{149}

§62 Pay services like Westlaw and Lexis do much more than provide state-of-the-law information, however. They also provide hyperlinks across different sources of law, integrated secondary sources, and advanced search and filtering options. If courts tagged dates, party names, names of judges and attorneys, case holdings, docket numbers, and other basic information (much of which is already “tagged” as such by courts through e-filing systems),\textsuperscript{150} a system like Google

---

\textsuperscript{146} But see Fastcase's Bad Law Bot, which shows that an almost entirely automated citator can convey some (though possibly spotty) state-of-the-law information. Note that the Bad Law Bot carries with it the following disclaimer:

Keep in mind that Bad Law Bot determines negative case history by using algorithms, and that it is not intended to be a complete replacement for a full editorial citator or for reading all later-citing cases. . . . If a case has been overturned but no court opinion has cited to it yet, Bad Law Bot won't be able to find any citation signal information. Meet Our Newest Team Member, Bad Law Bot!, FASTCASE, https://www.fastcase.com/blog/badlawbot/ [https://perma.cc/6JVE-8FRB].

Also, such tools require adherence to Bluebook protocols, which isn't always done. For example, Korte v. Sebelius, 735 F.3d 654, 720 (7th Cir. 2013), discusses the overruled case Adkins v. Children's Hosp. of D.C., 261 U.S. 525 (1923), without employing Bluebook signals noting that the case is bad law. As a result, automated citators would miss the negative treatment of this case, unless it were cited according to Bluebook rules elsewhere (which happens to be the case here—Adkins has been cited hundreds of times).

147. This might not take that much extra effort, after the standards have been settled on: Bluebook citations almost qualify as machine-readable metadata as is, and they are treated as such by tools like the Fastcase Bad Law Bot. As described in the preceding note, relying solely on Bluebook citations does not allow these automatic citators to compete with KeyCite and Shepard's in terms of accuracy, currency, and comprehensiveness. For that, additional citation metadata would be required. Pablo D. Arredondo has demonstrated that Bluebook citations can be used in other ways as well, such as generating brief case summaries and improving relevancy rankings of search results. Pablo D. Arredondo, Harvesting and Utilizing Explanatory Parentheticals, 1 LEGAL INFO. REV. 31, 49–50 (2015–2016).

148. Robinson et al., supra note 97, at 171 (“When government provides reusable data, the practical costs of reuse, adaptation, and innovation by third parties are dramatically reduced.”).

149. See Paul Hellyer, Evaluating Shepard's, KeyCite, and BCite for Case Validation Accuracy, 110 LAW LIBR. J. 449, 465, 2018 LAW LIBR. J. 20, ¶ 50; Aaron S. Kirschenfeld, Yellow Flag Fever: Describing Negative Legal Precedent in Citators, 108 LAW LIBR. J. 77, 96, 2016 LAW LIBR. J. 4, ¶ 49; cf. Mart, Human Indexing and Computer Algorithms, supra note 46, at 244–49, ¶¶ 39–50 (showing a wide discrepancy between KeyCite and Shepard's when using these citators to find additional cases by topic).

150. See generally Martin, supra note 22, at 346, ¶¶ 37–38 (describing how the North Dakota Supreme Court and Oklahoma Supreme Court case law databases allow for retrieval by basic metadata fields like "opinion author," and that the databases seem to be tied to the docket-based e-filing system at large).
Scholar could mimic these functions, allowing for expanded field searching, faceted searching, and an expanded set of options for sorting and filtering search results.

¶63 Given the breadth of materials housed in each database, Westlaw and Lexis users can search across multiple jurisdictions, state or federal, and multiple sources of law simultaneously. Assuming that other branches of government continue to expand their use of standardized XML metadata, we could imagine free services allowing for searches across different content types and jurisdictions as well. By systematically tagging citations to primary legislative and administrative law, a system like Google Scholar could link from court documents to provisions in the United States Code or Code of Federal Regulations. This is crucial to the ability of a legal database to compete with the likes of Westlaw or Lexis; as Martin observes in his analysis of the Arkansas court website:

Those working in digital collections have come to expect that case and statutory citations in decisions will be linked to the provisions cited, that the statutory authority cited for a regulation will be equally accessible, and finally that statutory annotations will have this same functionality. The Arkansas case law archive, like those mounted by the courts in most other United States jurisdictions, cannot offer this degree of integration. As a consequence, even with an enhanced search engine and a deeper historical collection, this path-breaking public site will have a hard time competing with the commercial services that bring statutes and case law together.

¶64 Once courts become comfortable tagging “easy” parts of a case (i.e., the parts that are objectively identifiable, like citations or party names), they could begin experimenting with more complex and detailed tagging practices. This might include, for example, tagging specific elements of different causes of action or specific factors in multifactor tests. If this practice were widespread and standardized, it might be possible to design a program that could pull up the rules underlying a cause of action in specific jurisdictions. If courts issued headnotes, as is still done by some, and classified these notes according to a metadata standard

151. Though linking to secondary sources would be trickier; publishers are not inclined to give this content away for free. Perhaps links to official guidance material or nonprofit self-help sites would be feasible, or links to library catalogs. See generally Manriquez, supra note 117 (describing how linked data could increase the visibility of library resources on the Internet).
152. See generally Martin, supra note 2, at 38 (describing how metadata deficiencies limit the flexibility users have when interacting with a case law database).
153. For example, the Federal Election Commission offers bulk data in XML format, accessible to anyone with Internet access at no charge. See Downloadable Data Files, Fed. Elec-
tion Comm’n, http://www.fec.gov/finance/disclosure/ftp_download.shtml [https://perma.cc/C4WK-
154. Martin, supra note 66, at 87.
155. See Bruce, supra note 24, at 27 (“It is simply not that difficult to decide which parts of a judicial decision are useful meta-data.”).
156. See supra notes 66 and 70; see also Martin, supra note 2, at 39 (“With limited exceptions, the headnotes and issue summaries prepared for official print reports by public law reporters
mapped onto some open source taxonomy, free legal research services could provide users with the ability to browse legal materials by narrow topics. If standard subject tags were adopted across jurisdictions, this would enable browsing to occur across jurisdictions.

§65 Effectively, this would shift responsibility about how the law is classified and structured from the private to the public sector. And this might be an impor-

have not accompanied the decisions themselves onto the Internet or into commercial online collections.

157. See Martin, supra note 2, at 39 (advocating that courts should produce their own headnotes and synopses, host them along with the online version of the case, and allow them to be harvested and used by commercial entities). This practice could have especially important benefits for the quality of electronic legal research. Scholars have argued that researchers using search engines often choose search terms pertaining to the client's factual situation, rather than the legal rules, issues, and principles at play. See, e.g., Carol M. Bast & Ransford C. Pyle, Legal Research in the Computer Age: A Paradigm Shift?, 93 Law Libr. J. 285, 297–98, 2001 Law Libr. J. 13, ¶¶ 43–50; Barbara Bintiff, From Creativity to Computerese: Thinking Like a Lawyer in the Computer Age, 88 Law Libr. J. 338, 345–46 (1996); Richard Delgado & Jean Stefancic, Why Do We Tell the Same Stories? Law Reform, Critical Librarianship, and the Triple Helix Dilemma, 42 Stan. L. Rev. 207, 221 (1989) ("Computers may be excellent means of finding cases about cows that wander onto highways. They are less useful in finding cases that illustrate or discuss more complex or abstract concepts."); Stefan H. Krieger & Katrina Fischer Kuh, Accessing Law: An Empirical Study Exploring the Influence of Legal Research Medium, 16 Vand. J. Ent. & Tech. L. 757, 789 (2014) ("This study's findings suggest that electronic researchers can, in fact, be expected to emphasize fact terms as compared to legal concepts in their research and to rely more on primary sources and less on secondary sources than print researchers."); cf. Hanson, supra note 2, at 583–84, ¶¶ 54–55 (arguing that keyword searching can be effective for finding cases on narrow legal doctrines, but is less conducive to finding cases embodying high-level legal principles).

Searching by keywords pertaining to factual information alone can have negative consequences on research outcomes and, by extension, might impair the efficiency of the legal system as a whole. See generally Katrina Fischer Kuh, Electronically Manufactured Law, 22 Harv. J.L. & Tech. 223, 267–70 (2008) (arguing keyword searching has diminished reliance on editors, leading to greater diversity in how individual lawyers might "frame" a case and more "tilting at windmills," i.e., reliance on spurious legal theories). By embedding taxonomical information in opinions, free and low-cost databases might mitigate these problems by allowing users the option to find cases more easily by topic. Though it is debatable whether and to what extent users of electronic databases will use classification data even if it is made readily available, see Kuh, supra, at 245, we can speculate that research databases could leverage this metadata to provide expanded, less literal keyword search results. See generally Mart, Algorithm, supra note 6, at 392, ¶ 9 ("But it turns out that trying to make sense of information without underlying ontologies or classification systems can impede automation practices. Legal database providers may even make the human additives to their search explicit. LexisNexisboasts of the human indexing in Shepard's citations; Westlaw is proud of its human-generated Key Numbers; and Bloomberg BNA advertises that the human indexing in its BNA treatises significantly boosts search results."). For example, a search engine might link certain sets of fact-related terms to legal issues commonly involving these facts and then retrieve some cases that deal with the legal issue but do not include the actual search terms used. Some of this might be possible through citation analysis alone—but more explicit classification metadata could only help.

158. In many ways, this argument parallels—in kind of an inverted way—proposals advocating for greater use of public-private partnerships in providing basic government services. Usually, this means outsourcing services traditionally provided by the government to private companies. In the world of legal information, the private sector has historically been the dominant player in the market, and this article argues for greater participation by government actors. See generally Dominique Custos & John Reitz, Public-Private Partnerships, 58 Am. J. Comp. L. 555 (2010) (providing a history of public-private partnerships and arguing that these kinds of arrangements can result in insufficient oversight over private contractors providing public services); see also Joergensen, supra note 27, at 33 ("Traditional print is becoming increasingly cost bound and expensive, profit margins are shrinking, and consolidations are rampant. Because of this, and the relatively low startup costs involved in Internet publishing, the opportunity for non-profit and governmental institutions to gain control over their publications has never been greater.")
tant benefit because scholars such as Berring have argued persuasively that the Digest System was not merely descriptive; rather, it influenced how the law was interpreted and applied.159 One specific complaint lodged against the Digest System is that West has historically been slow to add categories for new species of legal theories, effectively hiding them from practitioners and the lay citizenry alike.160 If courts classified headnotes, they could add categories whenever they saw fit.161 In this way, courts would gain greater control over how their decisions are found, interpreted, and used; by extension, the judiciary would enjoy greater control over the evolution of the common law in substantive terms.162

159. See, e.g., Berring, Full-Text Databases, supra note 11, at 33 ("Lawyers began to think according to the West categories."); Berring, Imperative of Digital Information, supra note 63, at 23 ("How one organizes the law became the center of what the law could and did mean."). But see Peter C. Schanck, Taking Up Barkan’s Challenge: Looking at the Judicial Process and Legal Research, 82 LAW LIBR. J. 1, 17 (1990) ("[K]ey numbers, headnotes, indexes, and so forth have had little or no impact on either the content of our law or our understanding of the legal system."); Joseph A. Custer, The Universe of Thinkable Thoughts Versus the Facts of Empirical Research, 102 LAW LIBR. J. 251, 265, 2010 LAW LIBR. J. 14, ¶ 49 ("My research suggests that neither does the Key Number System influence the law nor does the law influence the Key Number System."). For a recent analysis of whether and how the West Digest influenced lawyers, see generally Danner, Influences of the Digest, supra note 2. By giving researchers the freedom to “self-index” virtually the entire body of case law through keyword searching, it might seem like this concern has been nullified. However, Richard Delgado and Jean Stefancic argue that keyword searching is a partial solution at best. Delgado & Stefancic, supra note 157, at 221 ("[C]omputerized research can 'freeze' the law by limiting the search to cases containing particular words or expressions. Research should encourage browsing and analogical reasoning. Paradoxically, computer-assisted research can discourage innovation and law reform."). Furthermore, some algorithms depend on traditional classification systems to provide relevant search results. For example, Westlaw’s algorithm is tied to the West Topic and Key Number System, though this link is not apparent to users. See 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, 610 (2014) ("A fundamental flaw in the assertion that online research transcends legal categories is that . . . legal categories are now critical components of legal database search algorithms: WestlawNext is the epicenter of such search algorithm innovation.").

160. Berring, Full-Text Databases, supra note 11, at 33–34 ("The editors were trained to ‘normalize’ judicial opinions that used strange language or strange analysis or otherwise appeared to be anomalous, to bring them back into the orthodox mainstream, to make them fit past cases and present expectations."); Delgado & Stefancic, supra note 157, at 215–16 ("[C]hange comes slowly; The topic ‘Labor’ received a heading in the 1950s, and until recently West classified ‘Workers’ Compensation’ under ‘Master and Servant’ law.").

161. If anything, the problem would be too many new categories of cases. But this is less of a problem than one might assume, given that (1) cases could be put in more than one spot, meaning that in addition to the more novel and unconventional category, a case could be given a tag that conformed to a standard, trans-jurisdictional taxonomy, perhaps created by some panel of experts; and (2) the provider of the user interface (e.g., Google Scholar) could always elect to ignore classification metadata that didn’t conform to the standard taxonomy, whatever that turned out to be.

162. At the very least, wider access to good classification schemes might improve the quality of arguments brought in front of a court. Cf. Kuh, Electronically Manufactured Law, supra note 157, at 263–64 ("[T]he digest and key systems provide a print researcher with a significant amount of information about a case before the researcher reviews the case text and, per cognitive psychology, the labels and categories imposed by the digest and key systems will have a strong influence on researcher understanding. . . . [W]e would expect electronic researchers to be less apt than print researchers to recognize faults in a case or theory that is at least superficially supportive of a research goal."). For additional background on the topic of how the shift to keyword searching has affected the research and analysis of lawyers, see, e.g., Berring, Form Molds Substance, supra note 12, at 21–27; and Berring, Thinkable Thoughts, supra note 7.
Above, we saw how metadata increases the findability of case law and other sources of law. Courts can also use standardized metadata to exert greater control over when opinions are kept hidden. According to Silverman, XML “permits information in court records to be shared with the public at the courthouse and over the Internet while respecting the legitimate privacy interests of litigants and others who come before our courts.”  

This is important because privacy concerns are cited as one reason courts should be cautious about accepting filings and publishing decisions via the Internet. XML would allow a court to tag elements of a document as public and accessible to anyone, or private and accessible to just authorized individuals. Such a system might be designed so that information marked as private would be automatically culled from the document as it is made available for harvesting to third parties like Google.

Furthermore, if courts took seriously their roles as editors and publishers of legal information, mistakes involving the accidental dissemination of sensitive information might be reduced. As Martin notes in a recent article, online case reporting is now often tied directly to case management systems. For example, whether a federal district court decision ends up in the case law databases on Westlaw, Lexis, or Google Scholar can depend on whether the judge (or court personnel) correctly labels the document in PACER as a “written opinion.” In some cases, failing to apply the correct label can keep potentially useful cases hidden; in other cases, such mistakes can result in the unnecessary dissemination of sensitive personal information, like medical records in routine Social Security disability appeals. To protect privacy, it might make sense to decouple case reporting from case management systems. Instead of merely selecting a document type from a drop-down menu to mark a case for publication, a court might be forced to undertake the extra step of uploading the document to a separate repository containing only written opinions meant for wide distribution. The case management system could reside behind a CAPTCHA (and thus be effectively hidden from Google's

---

164. Silverman notes how “[a]ccording to these doomsayers, inexpensive and convenient public access to court records over the Internet must be abjured if we are to preserve what remains of the collapsing catacombs of personal privacy . . . .” Id. However, Silverman argues that this fear is misplaced because courts could “use XML tags to mark up sensitive personal information and control access to it programmatically.” Id. at 211; see also Peter A. Winn, Online Court Records: Balancing Judicial Accountability and Privacy in an Age of Electronic Information, 79 Wash. L. Rev. 307, 327 (2004) (arguing for an approach that balances privacy against the public’s interest in accessing court documents online).
165. See Silverman, supra note 99, at 211; Turner, supra note 129, at 281 (“[D]ifferent users are permitted to view different fields out of the entire document. Because individual fields are recognized, for example, court software displaying an electronic brief could show court staff all data, while blocking out address information in the view accessible by the general public.”); cf. Julien Mailland, The Semantic Web and Information Flow: A Legal Framework, 11 N.C. J.L. & Tech. 269 (2010) (noting that these kinds of technologies can block access to information, which can be used for nefarious purposes by repressive regimes).
166. See Martin, District Court Opinions that Remain Hidden, supra note 23, at 323, ¶¶ 48–51.
167. Id.
168. Cf. id. at 329, ¶ 68.
169. Although this would add an extra step to the process of marking court documents as written opinions meant for wide distribution, this additional friction could have the benefit of forcing judges to more carefully consider their role as “publishers” in the legal information system.
indiscriminate gaze), whereas the case law repository would encourage just this kind of indexing and harvesting by search engines.

¶68 Relatedly, courts could use metadata tags to clearly state whether a decision should be considered precedential.¹⁷⁰ Tagging a document as “nonprecedential” would allow a search engine to give such a decision less weight or allow for appropriate filtering options—an especially important ability given the amount of unpublished case law making its way online.¹⁷¹

¶69 Good online publishing and metadata practices would also help ensure the accuracy and integrity of digital legal information. Martin notes that even the major legal databases have difficulty conforming digital texts to their official, print counterparts.¹⁷² Currently, “judges, reporters of decisions, and editors use delays inherent in the production of those reports to make post-release revisions . . . [resulting in] an indeterminate risk of version discrepancy.”¹⁷³ If a court were to publish cases in XML format and allow the metadata to be freely harvested by third-party websites, any changes made to the version stored on the court’s website could (sooner or later) be reflected on reliable third-party sites.¹⁷⁴ Then, when a finalized version of the opinion becomes available, an authenticated, official version in PDF format should be posted in conjunction with the XML data.¹⁷⁵ Thus, by seriously investing in the online publication of decisions and opinions, courts can ensure that researchers find the most up-to-date, authoritative version of a case when conducting their research online.

¶70 In an ideal universe, we could imagine every federal and state trial and appellate court marking up all of their written work product in minute granularity. But this might not be very realistic, especially in the short term. Fortunately, improving metadata practices isn’t an all-or-nothing proposition: courts can take it one step at a time, and every step would have real effects on the legal information marketplace. For example, given that appellate-level court opinions have the greatest legal significance, it might make sense to focus on these higher courts first. Further, even marking up the “easier” elements of court opinions would go a long

¹⁷⁰. Martin advocated for tagging each case with a note about its importance, but thought that every case should be citable and precedential. Martin, supra note 2, at 34 (“The digital environment allows appellate courts to tag those opinions they believe to involve routine application of settled law and for those conducting case research to focus initially on other opinions, without giving rise to all the problems that can flow from withholding opinions from general circulation on that ground or declaring those opinions non-precedential and uncitable.”).

¹⁷¹. See id. at 9 (“Vast numbers of ‘unpublished’ decisions of state and federal courts, decisions that have no volume and page numbers, are now collected and organized, linked and annotated in virtual law libraries.”); Berring, Crambling Infrastructure, supra note 1, at 287–90.

¹⁷². Martin, supra note 66, at 363, ¶ 82.

¹⁷³. Id.

¹⁷⁴. See Martin, supra note 22, at 363, ¶ 82.

way to improving the legal information environment. Much of this is captured by
court e-filing systems already, but going one step further and tagging citations with
treatment information would be a relatively small task that could improve services
like Google Scholar immensely.

¶71 The recommendations listed above would help U.S. courts avoid falling
behind foreign court systems when it comes to online case publication and access.
The European Union has adopted a number of policies in this area, including a
publisher-neutral system for case identification, a subject classification scheme
with multiple hierarchical levels and a decent amount of specificity, and an open
data policy that allows its content to be reused in downstream commerce.176 This
citation system (called the European Case Law Identifier) and the basic metadata
standard are being implemented by many national courts throughout Europe as
well.177 There is even a growing recognition that meaningful access to accurate legal
information is a basic human right.178 And, according to preliminary views of the
European Commission and the Hague Conference on Private International Law,
the availability of reusable, quality metadata might be considered a foundational
pillar of this right.179 Berring, in 1995, claimed that the United States had the best
legal information system in the world; we have a responsibility to ensure this
remains the case in the future.180

¶72 To sum up, the metadata practices of the judiciary can have concrete effects
on the cost and quality of legal information databases that provide online access to
case law. If courts used the same basic markup language for citations, free and low-
cost research databases could develop algorithmic citators with accurate state-of-
the-law information. The more detailed the metadata, the wider the array of search
and filtering options these databases could provide. If courts fully embraced this
editorial role, and began to write and classify their own headnotes and case sum-
maries, we could imagine free or low-cost case law databases allowing users to find
case law by topic, rather than just via keyword search. This presents an opportunity
for courts to take greater control over how their decisions are found, interpreted,
and applied.

176. See Marc van Opuijen, Gaining Momentum: How ECLI Improves Access to Case Law
in Europe, 5 J. Open Access L. 1, 3–4 (2017) (providing background on ECLI); Marc Van Opuijen et
al., Online Publication of Court Decisions in Europe, 17 LEGAL INFO. MGMT. 136, 141–43 (2017) (noting
that although the European Union adheres to open data practices that allow downstream commercial
and noncommercial use of court decisions, policies vary across member states). For an example of
how these policies allow for improved search capabilities, see InfoCURIUM, http://curia.europa.eu/juris
/recherche.jsf?language=en [https://perma.cc/7GBG-64ZT].

177. Opuijen, supra note 176, at 4–5 (providing a map showing implementation status
across Europe); see also European Case Law Identifier (ECLI), EUR. E-JUSTICE, https://e-justice.europa
facilitate easy access to—and citation of—national, foreign and European case law, the Council of
the European Union invited Member States and EU institutions to introduce the European Case Law
Identifier (ECLI) and a minimum set of uniform metadata for case law.”).

178. For background, see generally Mitee, supra note 175.

179. Id. at 1479 (citing The Hague Conference Guiding Principles to be Considered
in Developing a Future Instrument (2008), an annexure to Access to Foreign Law in Civil and
/upload/foreignlaw_concl_e.pdf [https://perma.cc/8TYJ-JPNA].

180. Berring, Throwing Out the Baby, supra note 109, at 618.
How Might Court-Authored Metadata Affect Cutting Edge Research Tools?

¶73 The preceding section describes some of the concrete benefits of robust, court-provided metadata. This section strays into the territory of speculation and prediction: how might court metadata practices impact the development of AI-powered research tools?

¶74 Artificial intelligence, we are often told, is already here. Although it is difficult to predict everything that private actors using AI might do with court-authored metadata, it is safe to say that a large disconnect exists between common definitions of “artificial intelligence” and marketers’ use of the term. What we usually think of as artificial intelligence is often labeled “Artificial General Intelligence,” or AGI. A computer system meets this stringent definition when it is able to read a plain text document (or process information delivered verbally) and “understand” what it means—human-generated metadata be damned. This technology might be a ways off—and that might be a good thing. When it lands, court metadata practices will be the least of our concerns since virtually every white-collar professional could be out of a job.

¶75 When marketers use the term “artificial intelligence,” they usually mean “Artificial Narrow Intelligence,” or ANI. Examples of ANI include tools that use machine-learning processes to accomplish narrow tasks that previously required human involvement (self-driving cars, playing chess or Jeopardy!); tools that illuminate or visualize trends and connections in large datasets (e.g., CARA or Ravel); or just about any software that incorporates voice recognition coupled with natural language processing (e.g., Siri or Google Assistant). When marketers describe these technologies as “AI powered,” what they probably mean is that the technology uses some form of machine learning or is able to create some illusion of AGI. So, when you ask Google Assistant, “Who was the president in 1922?” it is certainly impressive that it returns “Warren G. Harding” rather than just a link to a list of historical U.S. presidents. But the reason this magic works is because of a Wikipedia page containing structured metadata explicitly telling Google that (a) Warren G. Harding was a U.S. president, and (b) he was president in 1922. Thus, this answer is

---

181. For a deeper exploration of this topic, see Baker, supra note 17.
182. Martin, supra note 2, at 39 (“Beyond linked and searchable headnotes and case summaries are myriad possibilities.”).
183. See Tad Friend, Superior Intelligence: Do the Perils of A.I. Exceed Its Promise?, New Yorker, May 14, 2018, at 44. Classic fictional representations of this kind of AI would include HAL 9000 from 2001: A Space Odyssey or, as a more benign example, Data from Star Trek: The Next Generation.
184. Cf. Baker, supra note 17, at 20, ¶ 53 (“While lawyers can generally access the information they seek, computers do not yet have the ability to move beyond natural language processing to natural language understanding. It is impossible, then, for computers to truly perform effortless expert legal research.”).
185. At least, as Amara’s Law would suggest: “We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run.” Roy Amara, WIKIPEDIA, https://en.wikipedia.org/wiki/Roy_Amara [https://perma.cc/QM87-GQDP]; see also Baker, supra note 17, at 16, ¶ 40 (arguing that marketing departments overstate the capabilities of current-generation AI tools).
186. Or worse—see, e.g., Harlan Ellison, I Have No Mouth and I Must Scream, reprinted in American Fantastic Tales: Terror and the Uncanny from the 1940’s to Now 197 (2009).
187. See Friend, supra note 183, at 44 (providing a definition of “Artificial Narrow Intelligence”).
achieved more through decade-old Semantic Web technologies (structured, linked data and natural language search) than anything much resembling HAL 9000.

¶76 This isn’t to say that these kinds of ANI technologies aren’t very valuable and useful. In fact, they are almost uncanny when they work correctly. But they are not magic, and they often rely on human-generated, structured data to work.188 Even tools that employ machine learning techniques require humongous sets of unstructured and structured data with which they can train, learn, and improve. It stands to reason, then, that better court metadata practices will increase the accuracy of ANI products when it comes to answering legal questions, help them learn and improve more quickly, and lower their costs.189 Already, you can plug some legal questions into Google and return an “answer” rather than just a link to a website. Wouldn’t it be nice if that “answer” came from, say, a recent court opinion or government website rather than an online troll?190

¶77 Besides helping AI tools become better, cheaper, and more accessible, good metadata practices are important to the development of AI tools at a more basic level, in a way that impacts the integrity of the legal system. Decades ago, Berring made a good case that the West Digest System—which acted as a gatekeeper in the print world—not only reacted to the law but, because of the nature of the common law system, influenced the development of legal thought as well.191 After all, a case can’t be used as precedent if it can’t be found.192 Search engines and early ANI tools

188. See generally Jonathan Jenkins, What Can Information Technology Do for the Law?, 21 Harv. J.L. & Tech. 589, 603 (2008) (“Placing legal information—e.g., statutes, regulations, and judicial opinions—into the Semantic Web will enable search tools and decision support systems to operate on uniformly structured data, without relying on more uncertain methods for extracting information from plain text. Machine learning methods will be able to identify rules and patterns more accurately in such a data set.”): cf. Baker, supra note 17, at 12, ¶ 28 (describing how the effectiveness of medical diagnostic AI systems have been limited by medical data being “scattered across different computer systems in both structured and unstructured form . . . [making] it nearly impossible for one program to have a complete picture of the patient’s health record.”). Judicata, a startup that promises improved semantic legal search, relies on a combination of human editors and algorithms to build its database of structured case law metadata. See Ansel Halliburton, Judicata Raises $5.8M Second Round to Build Out Advanced Legal Research Systems; Keith Rabois Joins Board, TechCrunch (May 28, 2013), https://techcrunch.com/2013/05/28/judicata-raises-5-8m-second-round-to-build-out-advanced-legal-research-systems-keith-rabois-joins-board/ [https://perma.cc/5G49-K87N]; see also John O. McGinnis & Russell G. Pearce, The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Services, 82 Fordham L. Rev. 3041, 3049–50 (2014).

189. Gal & Rubinfeld, supra note 138, at 16 (“Indeed, research has shown that access to data can shape both the level and direction of innovative activity, thereby affecting both private as well as social welfare.”).

190. As a cautionary tale, see the exchange on Yahoo! Answers that became the “How is Babby Formed?” meme. Annalee Newitz, How is Babby Formed—The One Yahoo Meme That Perfectly Represents the Faltering Company, Ars Technica (Mar. 21, 2017), https://arstechnica.com/information-technology/2017/03/how-is-babby-formed-the-meme-that-will-define-yahoo-forever/ [https://perma.cc/3FFF-BK9W]. See also David Colarusso & Erika J. Rickard, Speaking the Same Language: Data Standards and Disruptive Technologies in the Administration of Justice, 50 Suffolk U. L. Rev. 387, 403 (2017) (“Litigants searching online for information about family law are far more likely to find a search engine-optimized site from an enterprising small law firm looking for new clients rather than either of the more neutral resources.”).

191. See Berring, Imperative of Digital Information, supra note 63, at 23.

192. See Berring, Chaos, Cyberspace and Tradition, supra note 4, at 193 (“Only if the decision could be found in a bound case reporter, which meant it could be found in the West system, could it be deemed real.”); cf. McGinnis & Wasick, supra note 2, at 1000 (“Inaccessible data points are useless to the decision maker.”).
are the gatekeepers now and could be susceptible to bias-reinforcing “feedback loops” similar to those theorized to have resulted from the Digest System.\textsuperscript{193} Looking toward the future, robust, court-authored metadata could help ensure that cases are found and used the way that courts intend them to be.\textsuperscript{194}

\textsuperscript{193} Without explicit guidance from the courts, AI systems would need to lean even more heavily on user behavior as a data source. ROSS learns, in part, by having subscribers “upvote” or “downvote” its outputs. Baker, supra note 17, at 15, \textsuperscript{¶} 36. This is fine as long as the tool is used only by a single constituency, like big law firms. However, if a judge or judicial clerk also relied on the tool, a bias-reinforcing feedback loop results, especially if we assume that the law firm user base would significantly outnumber the user base in the judiciary, and be weighted by the system accordingly. To illustrate step-by-step: (1) the AI tool would begin to reflect the biases of its largest user group, attorneys at big law firms; (2) the research results of a judicial clerk using the tool would reflect this bias; (3) the decision of the judge would be influenced by this biased research; (4) the decision would become part of the AI system’s dataset, further solidifying this bias. \textsuperscript{Cf.} Ronald E. Wheeler, Does WestlawNext Really Change Everything? The Implications of WestlawNext on Legal Research, 103 Law Libr. J. 359, 364–66, 2011 Law Libr. J. 23, \textsuperscript{¶}¶ 16–20 (noting that algorithms that mine user behavior data to provide relevant results will overlook the “legal oddities” that might be most interesting to academics or creative lawyers). See generally Baker, supra, at 24, \textsuperscript{¶} 65 (describing how the biases of programmers and users can become entrenched in algorithms).

\textsuperscript{194} See generally Berring, Thinkable Thoughts, supra note 7, at 317 (asserting that, despite reports that computer intelligence will soon rival human intelligence, he rather “would choose to hope that we humans will stay in control of our lives and of our legal system”).

\textsuperscript{195} See generally Berring, Cognitive Authority, supra note 7, at 1693 (“While the Topic and Key Number System was never deemed ‘authoritative,’ the power of the classification function that it performed was staggering. Generations of lawyers learned to conceptualize legal problems using the categories of the Topics and Key Numbers of the American Digest System.”); Berring, Full-Text Databases, supra note 11, at 32 (“The importance of the placement of the headnote into the Digest’s subject index cannot be overemphasized. This initial placement had a tremendous impact on any subsequent manipulations of the data.”).

\textsuperscript{196} See Berring, Cognitive Authority, supra note 7, at 1706–07 (“The major filter of information could become the search engine. It will require less and less from the researcher while doing more and more for her. In this scenario the researcher accords cognitive authority to the search system. She relies on the algorithm that drives the system to be accurate.”).

\textsuperscript{197} Even in this scenario, research databases would likely still be opaque with regard to their proprietary algorithms. In fact, when it comes to systems using advanced machine learning algorithms (like those modeled as “neural networks”), even the owners of the algorithm might not have a clear sense of the AI system’s decision-making process. But if these systems were trained, at least in part, using structured data supplied by the court system, we might have a better sense of what these systems were basing their decisions on. See generally Mart, Algorithm, supra note 6, at 399, \textsuperscript{¶} 18 (“Legal databases use similar primary law, but how it is readied for the algorithm differs: by the elements of metadata, relational, or object oriented database architecture, for example, or the categories of classification that are chosen.”).
¶79 In the past, we have viewed court opinions as quasi-literary creations with a narrative quality rather than a rigid, uniform set of facts, rules, and holdings. However, the future may require us to rethink this view. 198 Silverman argues that by “tagging all the information contained in a court document, it is possible to dispense with documents altogether—through dissolving them into structured information.” 199 In this way, structured metadata would enable a computer program, whether labeled as possessing AI or otherwise, to identify, isolate, manipulate, and display individual components of a court decision, 200 allowing a legal researcher to easily identify and access just the parts or aspects of legal opinions that are relevant to her purposes. This would bring online legal research sources closer to the “hypertext” envisioned by Katsh in 1995, particularly for the researcher without access to a premium legal database. 201

¶80 The thought of researchers finding—and relying on—decontextualized bits and pieces of court decisions to formulate legal advice surely would make many people nervous, but keep in mind that this isn’t far from how some attorneys conduct research online today. 202 Furthermore, in the aggregate, these bits of data could be boiled down into meaningful guidance by litigation analytics tools. 203 Again, courts should consider the audience for their decisions: the more machine-readable their written output, the greater chance that precedent will be found in the right contexts and have the anticipated effect on attorneys and subordinate courts. As new generations of attorneys enter the judiciary, a pivot to this kind of thinking about authoring court orders and opinions might occur naturally.

¶81 At a certain point, a legal information database can become so advanced as to encroach on the traditional role of the lawyer. We might even ask whether such a system is actually engaged in the practice of law. 204 Some have argued, in fact, that restrictions on the unauthorized practice of law are stifling innovation in the legal technology field. 205 As noted above, Google Scholar displays lists of citing cases, but

---

198. See generally Berring, Cognitive Authority, supra note 7, at 1704 (“The mummified and stylized prose of today’s judicial opinion will become a museum piece.”).
200. See Berring, Cognitive Authority, supra note 7, at 1704 (“In the future when, operating under a format-neutral regime, a court releases its opinion, that opinion is going to be manipulated, parsed, and repackaged by the legal information providers.”).
201. When we think of hypertext now, we think of a Web 1.0 world of hyperlinks; when Katsh discusses hypertext, he seems to be anticipating something more powerful, perhaps even the Semantic Web environment where information is linked but also identified in a standardized, machine readable format:

The use of hypertext in cyberspace requires a new image or conception of information—not of discrete volumes existing on shelves, or of discrete and numbered issues and editions, but of something more organic and dynamic, of bodies of information in which the links contribute to a work in which the whole is much more than the sum of its parts. Katsh, supra note 38, at 211.

202. See Molly Warner Lien, Technocentrism and the Soul of the Common Law Lawyer, 48 Am. U. L. Rev. 85, 88–89 (1998) (“[T]he methodology of researching in and working with electronic texts encourages work habits that prioritize speed and all too easily enable lawyers to find a kernel of phraseology that may support their often incorrect preconceived notions.”).
203. Currently, the litigation analytics tools available on Westlaw, Lexis, and Bloomberg Law hold a lot of promise, but their capabilities seem to be limited by the generality, inconsistencies, and gaps in the data provided by the state and federal case management systems on which these tools rely.
204. See Baker, supra note 17, at 27–28, ¶¶ 74–77.
205. Jenkins, supra note 188, at 605–06.
without editorial characterizations of these citations—for example, whether the citing case overturned or affirmed the prior case. Assuming that Google were willing to devote the time and energy to creating its own KeyCite-like service, would this constitute providing legal advice? This is a relatively benign example, but we can envision future AI tools that flaunt the divide between “research assistance” and “legal advice” in more troubling ways.\textsuperscript{206}

\textsuperscript{82} By providing their own characterizations of case treatment in explicit, machine-readable terms, courts can help organizations like Google steer clear of these muddy legal issues. If a Shepard’s-like system could be built off of the metadata included in court documents, then Google would be displaying this metadata in a particular way, arguably, rather than adding its own interpretation of the underlying data itself. A side effect of this practice is that it would require courts to be crystal clear about what they are up to in their legal opinions, which could have benefits as well as drawbacks.\textsuperscript{207}

\textsuperscript{83} Also worth considering is the prospect of automated (or AI-assisted) trial systems.\textsuperscript{208} The rigid and uniform classification of rules, standards, multifactor tests, and so on might be a side effect of requiring judges (or court personnel) to provide granular metadata on these points, and this data would only improve an ANI’s ability to decide a case (or, more realistically, provide guidance to the decider). Legal disputes may be difficult to automate now, given the law’s complexity, ambiguities, gaps, and inconsistencies, but through the process of tagging documents thoroughly, maybe some of those wrinkles could be exposed and smoothed out.\textsuperscript{209}

\textsuperscript{84} To summarize this part, XML is a powerful, flexible, and compatible way to embed metadata in electronically published documents. By including robust metadata in their opinions, and making this data available to be harvested, indexed, and linked by commercial and nonprofit entities, courts can help ensure that their

\textsuperscript{206} See generally McGinnis & Wasick, supra note 2, at 1018 (anticipating a future where users will be able to plug natural language questions into search engines and receive natural language answers).

\textsuperscript{207} See Berring, Cognitive Authority, supra note 7, at 1704 (“Think how much easier the law would be to understand if each opinion had to begin with an official judicially authored summary of the case. We could even ask judges to write these in a controlled vocabulary. We could ask them to tell us precisely how what the decision is intended to affect the law. Rather than major decisions being followed by fractious debate as to what the Court intended, we could simply ask the Court to tell us.”); cf. Martin, supra note 2, at 38 (speculating that requiring courts to provide detailed metadata might even have “a long range beneficial effect on the analytic structure of decisions”). But see Bruce, supra note 28, § 7.2.1 (“To be sure most markup schemes will be fairly general and most likely confined to fairly incontestable metadata like the name of the author of an opinion or the date of enactment of a statute. No matter how much law students might wish for it, it is not likely that we will ever see a judicial opinion containing tags like <PAYATTENTION> or <DICTUMclass=\"IMPORTANT\" duration=\"ETERNAL\">, even from a court wanting to lend weight to its own statements.”).

\textsuperscript{208} See generally Katsh, supra note 33, at 110 (“The specter of a computer that would render final judgment for the parties is more remote than the appearance of a computer that could answer particular kinds of questions, one that could help parties in a dispute to clarify what their argument is about and what kinds of solutions are possible, and one that could guide them through the problem-solving process.”).

\textsuperscript{209} See generally McGinnis & Wasick, supra note 2, at 996 (“Efficient legal search can become law itself”). Also, it is interesting that this parallels how the West Topic and Key Number System and the underlying body of court opinions congealed over time into a “seamless web” of law.
documents will be findable, useful, and machine-readable.\textsuperscript{210} Even if some of the recommendations above would be very difficult or even unrealistic to implement,\textsuperscript{211} simply tagging citations with basic metadata could create substantial benefits. It could be a boon, for example, to free sources of legal information, enabling cross-referencing, reverse-citation indexing, and features mimicking those of Shepard’s and KeyCite. If we have faith in the free market’s ability to spur innovation (especially in sectors of the economy with plenty of competition, as is increasingly the case in the market for free and low cost online legal research), this could be a cost-effective way to give the public access to high-quality, free legal research tools. The more structured data included with electronically published court documents, the better free sources of legal information can mimic the advanced features of Westlaw, LexisNexis, and Bloomberg Law. Peering into the future, we can speculate that good court metadata practices would improve the functionality, reliability, accountability, and cost of the next generation of AI tools.

Where Do Librarians Fit In?

\textsuperscript{¶}85 Librarians are especially well suited to helping courts improve how they publish their case law, both in a practical sense and in terms of advocacy.\textsuperscript{¶}86 First, they are experts in using and maintaining legal information systems—from a variety of perspectives given the diversity of material cataloged, patrons served, and reference questions faced. And they are accustomed to thinking about legal concepts categorically and relationally, which makes them particularly well suited for the task of creating legal taxonomies and ontologies (e.g., an open source Key Number System).\textsuperscript{¶}87

\textsuperscript{¶}87 The integration of a comprehensive legal ontology with large volumes of legal documents containing metadata in a standardized format could result in powerful legal research tools, with content linked across varied court systems, or even across different branches of government.\textsuperscript{213} Keele and Pearse note that a “shared

\begin{itemize}
\item[\textsuperscript{210}] See Martin, supra note 22, at 346, \textsuperscript{¶}37–38 (noting that the metadata included on the North Dakota Supreme Court and Oklahoma Supreme Court websites aid in finding court decisions); see also Martin, supra note 66, at 84–85 (describing the importance of allowing bulk data downloads in government case law databases).
\item[\textsuperscript{211}] Gallacher, supra note 78, at 38 (“Without using West’s copyrighted Key Number system as a guide, the archive’s administrators would have to develop an indexing protocol that is sufficiently detailed to satisfy lawyers working against a deadline and who have developed high expectations of accuracy and completeness from working with the West system. After establishing this protocol, every case in the archive would have to be analyzed and indexed, requiring a substantial investment of time and effort.”).
\item[\textsuperscript{212}] Cf. Keele & Pearse, supra note 130, at 399–400, \textsuperscript{¶}43 (describing the benefits of open source ontologies in the context of legal scholarship).
\item[\textsuperscript{213}] Cf. id. (describing the same phenomenon, but with a focus on open source legal scholarship); Bruce, supra note 28, \textsuperscript{¶}4 (“We imagine that a comprehensive public legal information regime will be an aggregation of different low- or no-cost providers acting under a variety of arrangements, principally self-publication and achieving interoperability through common standards and practices. If pressed, we would probably advocate the formation of something like a W3C for law, a consortium that could develop and promulgate interoperability standards, but we would not imagine that it would take on responsibility for comprehensive publication at any level, including service as a comprehensive portal.”). For more detailed background on this topic, see, e.g., Kevin D. Ashley, The Case-Based Reasoning Approach: Ontologies for Analogical Legal Argument, in Approaches to Legal Ontologies 100 (G. Sartor et al. eds., 2011).
\end{itemize}
taxonomy/ontology emanating from the legal academy could also be ‘mapped’ to taxonomies/ontologies developed for more practical or public uses such as projects in the open law movement or internal governmental use.” In this way, search engines could see connections among not only primary law documents but secondary sources emanating from legal academia as well.

In this way, search engines could see connections among not only primary law documents but secondary sources emanating from legal academia as well. Second, law librarians have a track record of advocating for the public interest in the face of legal publishing monopolies. They have pushed for vendor-neutral citation, for example, which would play a necessary role in removing barriers of entry for new legal information databases. Other sources of resistance might include the courts, whose commitment to open access has been mixed. Ian Gallacher notes that some courts might be resistant to open access in part “because open access would likely encourage more, and more complicated, pro se filings.” Martin describes restrictive contracts between some courts and West or Lexis (sometimes, in exchange for discounts and other benefits) as another obstacle to open access.

Finally, as also pointed out by Martin, state governments “quite commonly assert copyright in all law report editorial additions.” Although the copyrightability of state-authored editorial enhancements is questionable in some circumstances, states should explicitly disclaim these copyrights to encourage reuse by free and low-cost legal information providers.

Persuading courts to spend time and money on tagging legal documents will certainly be an uphill battle. During the debates on vendor-neutral citation, Martin described the resistance of court staff to simply adding paragraph numbers to their decisions. However, one way forward might be to focus attention on tangible benefits: if there existed freely available alternatives to Lexis or Westlaw, then courts wouldn’t need to subscribe to these expensive services. The freed

---

214. Keele & Pearse, supra note 130, at 400, ¶ 43 (footnote omitted).
215. See id.
216. See generally Martin, supra note 66 (describing how some state courts have taken their online case reporting responsibilities very seriously, while others have not).
217. Gallacher, supra note 78, at 22.
218. Martin, supra note 2, at 28–30; cf. Leslie A. Street & David R. Hansen, Who Owns the Law: Why We Must Restore Public Ownership of Legal Publishing, 26 J. Intell. Prop. 236 (2019) (describing “clickwrap” agreements on commercial websites that forbid bulk downloads and harvesting by third parties, even where the commercial website contracted with the state to be the official online host of the state’s legal information). See generally Bruce, supra note 24, at 21 (“The danger I perceive lies in the idea that even the unrealized potential for such a cutting-out of the private sector would be seen as sufficiently threatening to justify strict—and artificial—limits on the level of value that could be offered by issuers themselves.”).
221. See generally Glassmeyer, supra note 77 (showing that it is commonplace for states to include copyright claims on websites hosting legal information); Katie Fortney, Ending Copyright Claims in State Primary Legal Materials: Toward an Open Source Legal System, 102 Law Libr. J. 59, 2010 Law Libr. J. 3.
222. Martin, supra note 22, at 355 n.127. Including paragraph numbers in court opinions, and requiring pinpoint citations to specific paragraphs instead of page numbers, could help improve legal analytic tools that rely on citation analysis. After all, Bluebook citations are much like machine-readable metadata, and the more specific the information included within each citation, the more power they would have. For an in-depth analysis of the different ways that Bluebook citations can be used to improve research databases, see Arredondo, supra note 147.
resources could help pay for a full-time metadata professional who could focus on writing headnotes and taking care of the more complex metadata tasks (e.g., characterizing citations). This would require a leap of faith—an “if you build it, they will come” sort of attitude. A tough sell, but maybe not an impossible one.

¶90 Third, law librarians have the skills to digitally catalog and mark up the gigantic backlog of historical cases, a task that would require either the cooperation of a vendor or a major scanning project. Law librarians are experts in processing metadata for print and digital cataloging efforts; and their institutions are well familiar with digitizing historical scholarship libraries. Yes, tagging all of these old documents would be a herculean task, but perhaps algorithms could do this work in a “good enough” way or at least assist in the process. And although this article argues that courts ought to take the reins when it comes to future work product, historical legal documents might be good candidates for crowdsourced tagging (with librarians taking the lead, of course).

¶91 Finally, librarians—particularly in academia—could help facilitate Thomas Bruce’s suggestion that an open source case law repository be initially centralized in a single authority, much in the way that the GPO publishes federal case law. That way, the central authority could (1) ensure that courts comply with data standards, and (2) offer technical assistance (and possibly technical infrastructure) to help get these projects off the ground, especially in places short on personnel and funding. Eventually, when courts became comfortable with the standards and technology, they could decouple from the central repository. As more courts became independent publishers, the system would transition to a distributed, decentralized model.

¶92 Academia would be a good initial home for a central repository of newly published case law. Academic law libraries commonly employ specialists familiar with digital repositories and metadata standards; furthermore, while it is commonplace to pay vendors to host institutional repositories, some law libraries do host repositories based on open source software, like DSpace. Once a repository is created and standards settled on (no small task), court staff or court librarians could

223. See Gallacher, supra note 78, at 27. Recently, the Harvard Case Law Project proved this to be possible; it completed this task in a partnership with Ravel. Dulin & Krible, supra note 89. Somewhat ironically, Ravel has since been acquired by Lexis; it seems, however, that Harvard wisely anticipated this eventuality in its agreements with Ravel. See Adam Ziegler, Continued Support for the Caselaw Access Project, Er Seq. (June 8, 2017), http://etseq.law.harvard.edu/category/caselaw-access-project/ [https://perma.cc/8KQV-LDXN].

224. See Gallacher, supra note 78, at 38–39 (“The prospect of developing such a protocol, indexing all existing case law according to it, and applying the protocol to all future court opinions is, perhaps, too large a task to consider, even for an army of law students. Therefore, for an open-access site to be able to offer indexed research, it seems inevitable that some form of automated indexing process is necessary.”).

225. See generally Timothy K. Armstrong, Crowdsourcing and Open Access: Collaborative Techniques for Disseminating Legal Materials and Scholarship, 26 SANTA CLARA COMPUTER & HIGH TECH. L.J. 591 (2010) (arguing that crowdsourcing could play a crucial role in the open access movement and lead to free databases that are viable alternatives to Lexis or Westlaw); see also Gallacher, supra note 78, at 47–49.


227. Id.

228. Id.

229. See id. § 8.5 (“In order to have a distributed model that supports common interfaces and capabilities we need to formulate workable standards. This is challenging to say the least, and the
be given direct access to their pocket of the repository, allowing them to upload content and metadata. Again, the goal would be to allow court personnel and librarians a space to become comfortable with metadata standards and digital repositories, in the hope that one day they would strike out on their own.

¶93 Librarians should heed the dangers of inaction. What happens when Shepard’s, KeyCite, or BCite become unaffordable, when these are the only true state-of-the-law tools on the market? If a library can afford only one computer terminal with limited Westlaw access, what happens if the books go away? What if the “cognitive authority” of court opinions is further undermined by free, but ad-supported, case law databases dependent on “crowdsourced” descriptive and editorial content, where popularity acts as a proxy for authoritativeness—legal information as click bait?230 As Berring warns:

If the market picks the slickest, easiest-to-operate system with the glitziest front end, the integrity of the legal information system is in peril. As legal information commingles with other forms of information, there could be a significant debasing of legal information. With no informed, critical intelligence making choices, the marketeers will be in control. The thought of Rupert Murdoch controlling legal information makes my blood run cold. That would be a grim future indeed.231

Conclusion

¶94 A quick scan of Law Library Journal tables of contents over the past decades reveals that the ascendency of the digital format has brought with it a whole slew of challenges for law librarians, both practical and theoretical.

¶95 On a practical level, public access to quality legal information is in danger of declining—ironic given that the public enjoys remote access to a vast amount of primary legal authority over the Internet. Free websites lack tools that are crucial to the research process, such as fully functioning citators, editorial content, and authoritative secondary sources. For now, the public can still access these materials in law libraries, but it should give us pause that libraries commonly provide some of the basic tools solely through online research databases. Research databases are often leased, not owned, and access can easily be restricted to a class of privileged patrons. In fact, we may be entering an age of worsening information inequality, with the best AI-powered tools affordable only to elite law firms.232 Given our adversarial justice system, this raises real access-to-justice issues.

230. See Berring, Thinkable Thoughts, supra note 7, at 316 (“Popularity on the Internet is usually the result of skilled marketing. The best advertisers become the most authoritative sources. It scares me.”).

231. Berring, Cognitive Authority, supra note 7, at 1707.

232. See Berring, Chaos, Cyberspace and Tradition, supra note 4, at 204–06 (noting that large firms can afford the cutting-edge tools that others cannot, while also conceding that “[t]here have always been information elites. Those with resources could always get better data, better service, indeed, better librarians. There has never been real parity of access to information.”).
However, robust, court-authored metadata would go a long way toward improving free and low-cost research databases—in effect narrowing the sizeable quality gap between the legal information available to a pro se litigant versus a large law firm. Metadata that characterizes the nature of citations would help low-cost and free databases mimic the functions of KeyCite and Shepard’s. Headnotes would aid researchers both in finding and understanding the case law in these databases. Marking up legal rules and holdings would allow low-cost and free research tools to display or highlight important elements of a case. Links to other resources (e.g., statutes and regulations) would help nudge inexperienced researchers to other relevant sources of law.

In addition to these concrete concerns, the decline of print raises a number of questions more conceptual in nature. For example, is it realistic to expect that legal research can be conducted effectively and efficiently by keyword searching alone, without the use of a classification system? Can we expect AI tools to accurately grasp the law, absent a framework of authoritative, structured metadata? How can we hold the accuracy, objectivity, and fairness of these tools accountable when their underlying metadata and algorithms are proprietary? Could these changes in technology and research methodology chisel away at the structure of the law itself?

In *Legal Research and the World of Thinkable Thoughts*, Berring makes a plea for a new Blackstone to impose order and structure on the current mess that is online legal information. Rather than a single Blackstone, ours might be a bit more diffuse and decentralized: structured, standardized metadata written by individual courts and graphed onto the Semantic Web. In the future, instead of inserting legal documents into a classification scheme, the necessary organizational infrastructure might be embedded within the documents as they are generated. As librarians, we have a role to play, both in responding to the growing advancements in information technology and in preserving our profession’s values as the changes swirl all around us.

---

233. See supra notes 94 and 157.


235. See generally Berring, *Crumbling Infrastructure*, supra note 1.