Uniform Tools for Legal Referencing
[citation needed]

Frank G. Bennett

8 October 2012

Law via the Internet 2012
Cornell University

[excerpt from Bennett, CITATIONS, OUT OF THE BOX (2012)]

Chapter 1

Introduction

> *I can see the dancin'*
> *The silhouettes on the shade*
> *I hear the music, all the lovers on parade*
> *Open up, I wanna come in again*
> *I thought you were my friend*
>
> —Yvonne Baker and the Sensations[1]

About this Book

This book is a coming-of-age celebration for Multilingual Zotero (MLZ), a suite of software that offers an alternative to the time-honoured practice of hand-crafting citations in legal and multilingual publishing. The long-term aim of the project is to improve the quality of our research lives by allowing us, as a community, to spend less time assembling documents and more time thinking about what should go into them.

The concept is simple, and as this Introduction explains in some detail, it is not particularly new. Reference management software to assist in organizing research materials and formatting citations is widely available. Yet while there are many products in circulation, none has yet offered robust support for legal or multilingual research. For a comparative lawyer with terrible handwriting and a mechanic's bent for computer programming, the temptation to meddle has proven too strong to resist.

There is a reason legal and multilingual support has been lacking in reference management tools: implementing these features is really very hard. Building the `citeproc-js` citation formatter, the MLZ prototype and the companion suite of legal styles described in this volume has taken me the better part of four years. That effort has been informed and motivated by supervision of international students in the faculty where I hold my appointment. The Nagoya University law programs are a microcosm of "globalisation", representing ten or more language domains in any given academic term. American lawyers are wont to protest, perhaps too much, at the burden of the "uniform" American legal citation system;[2]

[1] THE SENSATIONS, LET ME IN (Chess Records 1961), *available at* http://www.youtube.com/watch?v=ef1znNdZA1k.

[2] *Compare* James D. Gordon, *How Not to Succeed in Law School*, 100 THE YALE LAW JOURNAL

but researchers in the wider world face the harder task of navigating sources in multiple languages, from a mixture of jurisdictions. The arcana of citation conventions in any one country pale against the demands of comparative research.

If these tasks can be simplified, the method of doing so can and should be shared. For this reason, both the MLZ software and this text are freely available. If you have already purchased a copy of this book (paper or ebook) there is no cause for disappointment or alarm; sales of this volume help to assure the continued health of the MLZ project. If you are reading the freely distributed PDF version of this text and find that MLZ is useful in your work, please consider making that purchase. The version for sale has an elegant cover, and if we should one day meet, I will be happy to sign your copy in my terrible handwriting.

To introduce the software itself, a full MLZ installation is made up of three plugins for the Firefox browser:[3]

MLZ Client: This provides the same core facilities as Zotero proper, but with (unintrusive) extensions to the user interface that allow alternative versions of many item fields to be added. The MLZ client also allows the selective inclusion of alternative field content in generated citations.

Abbreviation Filter: This support plugin allows abbreviations to be applied to citation elements in a variety of ways on a per-style basis.

Word Processor Plugins: The MLZ system uses the same word processor plugins as official Zotero, providing on-the-fly citation support for LibreOffice Writer, Word for Windows, and Word for Mac.

MLZ is closely related to official Zotero at the code level, but please note that the two are separate projects. MLZ should always be referred to by that name when seeking support, and not by the name "Zotero". There are important differences between the two systems under the surface, and this clarity will help avoid confusion and yield a quicker, more accurate response to queries.

The remainder of the Introduction explains where MLZ comes from, why it has arrived so late on the scene, and how it fits into its surroundings as an open source, third-party product. Impatient readers may wish to skip forward to the Getting Started chapter, which covers the essential steps for installing and running the software.

1679, 1692 (1991) ("The worst part of legal writing is having to learn the legal citation system. This is set forth in literally thousands of subrules in a book whose name nobody can remember, but which everybody calls the Bluebook, mostly because it's blue...."), *with* C.M. Bast & S. Harrell, *Has the Bluebook Met Its Match -- The ALWD Citation Manual*, 92 LAW LIBR. J. 337, ¶ 6 (2000) ("[K]nowledge of correct legal citation distinguishes those who have legal education from those who do not.").

[3] Available via `http://citationstylist.org/tools`

The next section of this Introduction, *Hard Cases*, reviews some of the more demanding requirements of legal and multilingual authoring, with brief notes on how MLZ handles each. The section can used as a self-test questionnaire: if the citation examples it contains look at all familiar, you are within the target audience for this book; otherwise you may wish to look at other reference manager offerings.

This book does not presume to be the final word on reference management or metadata standards. The software it introduces is a work in progress, which I hope demonstrates the power of consistent metadata practices in a working system of this kind. If you find the system useful, item data accumulated in the course of writing projects will have lasting value, as it can be migrated to future versions, or to other reference manager platforms, as support technology improves apace.

That is the view going forward. We now rewind a bit to take a look at the current state of play in legal and multilingual reference management.

Hard Cases

To be concrete, there are five particularly challenging use cases that a reference manager with multilingual and legal support must address. These are reviewed here, with a note of how each is handled by MLZ. In the discussion below, "CSL" refers generally to the Citation Style Language that drives the formatting magic of MLZ and other modern reference managers.

Multilingual: extra details

When citing resources outside the primary language of the document, adjustments to the content may be needed to make the reference accessible to the target audience. The most common case is transliteration. In a publication aimed at an English-speaking audience, for example, a reprint of the Japanese novel 坊っちゃん (bō, chan) might be cited as follows:

> Natsume Sōseki, *Botchan* (Modernized edition, Shinchōsha 2003)

When citing across language boundaries, a transliterated title may not be sufficient. In this case it may be desireable to add a translation, set off with distinctive punctuation:

> Natsume Sōseki, *Botchan* [The Little Master] (Modernized edition, Shinchōsha 2003)

Author names from some language domains may be difficult to distinguish in their romanised form. A recent trend is to permit inclusion of author names in their original script as a supplementary parenthetical, for clarity:

> Natsume Sōseki (夏目漱石), *Botchan* [The Little Master] (Modernized edition, Shinchōsha 2003)

Some publishers ask that transliterated titles be forced to italics in citation forms that would otherwise use plain roman type:

> '*Yūsen shakuchi ken o minaoshi e: hisaichi taishō, haishi o kentō*' [Reconsideration of preferential lease rights: abolition in disaster relief zones under review], Asahi Shinbun (1 August 2012)

but

> Matthew L Wald, 'Court Weighs an Order on Yucca Mountain', New York Times (3 August 2012)

In the preferred citation form for theses of our own faculty, author names and titles are given in the original script, followed respectively by a transliteration in parentheses or a translation in square brackets:

> 夏目漱石 (Natsume Sōseki), 坊っちゃん [The Little Master] (Modernized edition, Shinchōsha 2003)

In some academic environments the hints might be reversed:

> Natsume Sōseki (夏目漱石), *The Little Master* [坊っちゃん] (Modernized edition, Shinchōsha 2003)

There are two problems to be solved: supplementary details (transliterations, translations) must be attached to item fields; and those details must be incorporated into citations in a controlled way. MLZ supports alternative field values in any language or script. Most multilingual formatting requirements can be addressed through a language preference panel used to meld multilingual data into finished citations. This permits use of any of the existing 600+ CSL styles in a multilingual context. See page 28 for details.

Multilingual: style by language

Citation style conventions vary across language domains. For example, "title case" is a property of English-language citations only:

> Edmund Curll, *A Complete Key to The Tale of a Tub; with Some Account of the Authors, the Occasion and Design of Writing It, and Mr. Wotton's Remarks Examin'd* (London, 1710)

but

> René Macé and Giovanni Bocace (trs), *Les trois justaucorps, conte bleu, tiré de l'anglois du Révérend Mr Jonathan Swif [sic]* (Dublin, 1721)

More demanding adjustments may be needed when publishing for a polyglot readership, or where the local citation format has quite specialised requirements. In such environments it is common to cite foreign materials in a style appropriate to their own language domain:

> Swift, Jonathan (著)、深町弘三(訳)「桶物語」190p 岩波文庫 1953.[4]

but

> Swift, Jonathan, *A Tale of a Tub* (London, J. Nutt 1704).

In MLZ, an extension to the CSL formatting language allows entirely separate formats to be applied on a per-language basis, based on the standard language code set in the **Language** field of the item (in these examples, `en` for English, `fr` for French, and `ja` for Japanese).

Legal: style by jurisdiction

Even where document and citations are in the same language, citations to primary legal materials require distinctive formatting for particular jurisdictions.

> *Palsgraf v Long Island Railway* 248 NY 339 (1928)

but

> *Swiss Bank Corp v Air Canada* (1987) [1988] 1 FC 71 (CA)

Similarly, special formatting may also be required for documents issued by certain frequently cited international organizations:

> Universal Declaration of Human Rights, GA Res 217(III)A UN Doc A/RES/217(III)

Such specialised adjustments to citation form are unavoidable in the law, and like language discrimination they require a hint in the item data. In MLZ, a `jurisdiction` variable set on legal item types from a controlled list of value provides this hint.

This is one of several extended fields used by MLZ and its companion version of the CSL language. The extended values are candidates for inclusion in official Zotero; but pending review and acceptance by the main project, they are managed in MLZ via menus accessible with a right-click over the **Extra** field. The pull-down list used to set the `jurisdiction` value on legal item types is shown in the illustration above.

[4]This example is hand-crafted, as MLZ does not yet offer a Japanese citation style. The other examples in this section were generated using the MLZ OSCOLA style.

Legal: meaningful fragments

In a reference manager, the basic unit of content is the *item*. In most cases, it is obvious what constitutes an item. It is simply the thing that we cite:

(1) DAVID LODGE, SMALL WORLD 23 (Penguin Books 1995).

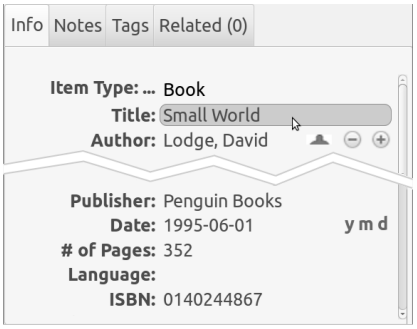
(2) *Id.* at 97.

The first reference is to page 23 of *Small World*. The second is to page 97 of the same novel. The citation details are replaced with *id.*, as appropriate for an immediate back-reference in the American Law style followed in these two examples.

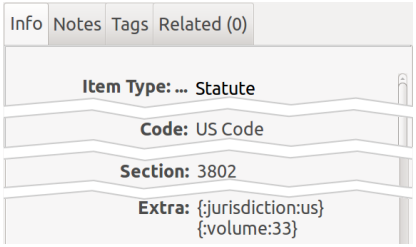
The American Law style applies the same logic to statutes, using *id.* for immediate back-references to the same statute or code:

(3) 33 USC § 3841(a).

(4) *Id.* § 3802(b).



This all seems quite straightforward, until we consider how these resources are stored in the reference manager database. Examples (1) and (2) above are generated from the single database item shown to the right. The citation formatter sees the data for the two as "the same" because it carries the same (hidden) ID number associated with the database entry. For ordinary resources, "the item" in the database is thus the same as "the item" understood by the citation formatter.



Now consider the second pair of references. The U.S. Code is a monolithic codification of all U.S. federal statutory law, with thousands of individual provisions. Each is a potential target of intense scrutiny by litigants and commentators, and a potential subject of analysis in its own right. Storing "U.S. Code" as a single database item would be sufficient to produce nicely formatted citations, but it would result in a system that is useless as a tool for legal research.

To allow fine-grained note-taking, a reference manager for law should permit individual provisions of a statute to be stored as primary reference items. This is possible in MLZ, by setting the pinpoint information in the **Section** field, as shown in the illustration. See page 40 below for details.

Legal: parallel citations

Court judgements in certain jurisdictions may be published in multiple reports. Some citation systems require parallel citation in this event, with a reference to major services where the text of the judgement can be found:

> Hanna v. Plumer, 380 US 460, 461, 85 S. Ct. 1136, 1137 (1965).[5]

In MLZ, case reports are stored as individual items, and it is up to the author to arrange them in the appropriate order; but as the example above shows, the citation formatter will take over from there, producing a "collapsed" parallel citation that begins with the title common to the items in the set, and ending with any common trailing matter (in this case, the date).

The MLZ citation engine will also handle parallel citations to statutory law and treaties:

> White Slave Traffic (Mann) Act, ch. 395, 36 Stat. 825, 826 (1910).
>
> National Environmental Policy Act of 1969, § 102, 42 USC § 4332 (1969).
>
> Department of Transportation Act, Pub. L. No. 89-670, § 9, 80 Stat. 931, 931 (1966).
>
> Treaty of Friendship, Commerce and Navigation, United States-Japan, art. X, 4 U.S.T. 2063, 75 U.N.T.S. 135.

Here as well the items must be cited in the appropriate order, but the formatter will then make the necessary adjustments to produce a correct parallel citation.

Software and Science

Reference management software—smart, personal electronic libraries populated with items that know how to cite themselves correctly—is familiar technology that holds promise of streamlining legal and multilingual research and writing. Existing tools arise from three cohorts of development. The first can be found in document processing software with automated bibliography support that emerged from computer science faculties in the 1970s.[6] This cohort is today represented primarily by the open-licensed LaTeX/BibTeX document system. A second cohort arose in the 1980s, when bibliographic software for use with WYSIWYG word processors and personal computers first appeared. The leading

[5] In this example, the numbers 461 and 1137 are pinpoint page references, indicating the exact page on which the statement supporting the author's argument appears.

[6] *See e.g.* BRIAN K. REID, SCRIBE: INTRODUCTORY USER'S MANUAL (Computer Science Department, Carnegie-Mellon University 1978); LESLIE LAMPORT, LaTeX: A DOCUMENT PREPARATION SYSTEM (Addison Wesley 1985) (documenting the companion BibTeX bibliography management system by Oren Patashnik).

example today of this latter cohort is the proprietary EndNote® product.[7] The utility of these tools to researchers is perhaps best shown by the extent of their use, with LaTeX being very nearly ubiquitous in maths-intensive disciplines, and EndNote® claiming a user base in the millions.[8]

A common characteristic in projects of these early cohorts is that they tended to be initiated by practicing researchers, who were themselves intimately familiar with the publishing bottleneck posed by referencing requirements, and undertook development with a view to streamlining the research process.

The licensing terms of the leaders in the two cohorts differ, a fact perhaps best explained by skill levels in their respective target communities. LaTeX was developed within and directed at a scientific community where computing skills are plentiful. EndNote® (like other members of the second cohort) is aimed primarily at an audience familiar with word processors, but not necessarily with computer programming. That difference may explain a great deal: while opening up the source code of a project to users able to extend and improve it may be beneficial, where that is not the case, the revenue stream from proprietary distribution may offer a more certain path to sustainability. Both models obviously work; which works best may depend to some extent on the makeup of the target audience.

Neither of the early cohorts has produced a general solution for legal or multilingual referencing. In part, this is the result of timing. The law has highly specialised citation requirements, and expertise in technology support for law was initially concentrated in commercial projects dedicated to the needs of individual jurisdictions.[9] The basic standards necessary for robust multilingual support, on the other hand, simply did not exist in 1980.[10] Multilingual capabilities have been added to the early-cohort products over time,[11] but their multi-language functionality remains awkward and incomplete.

---

[7] *See e.g.* Ruth E. Wachtel, *Personal Bibliographic Databases*, 235 SCI. 1093 (New Series, 1987) (reviewing five offerings: Reference Manager, Scholar's Bibliofile, Ref-11, Pro-Cite, and SciMate); *Oral History of Ernest Beutler*, LEGENDS IN HEMATOLOGY (American Society of Hematology Nov. 6, 1990), http://www.hematology.org/Publications/Legends/Beutler/1599.aspx (placing the first release of Reference Manager in mid-1984); ABOUT NILES & ASSOCIATES, INC. (Nov. 12, 1996), http://web.archive.org/web/19961112110744/http://www.niles.com/home/Company.htm (indicating that the first version of the EndNote® program was written in 1985); *also* Personal email from Victor Rosenberg, (no subject) (Jun. 14, 2012) (indicating that ProCite was developed at University of Michigan, licensed to a commercial firm in 1982, and first marketed in July 1983).

[8] *EndNote*, THOMSON REUTERS: PRODUCTS A-Z, http://thomsonreuters.com/products_services/science/science_products/a-z/endnote/.

[9] See discussion *infra* at pages 10 to 19.

[10] The design of Unicode was proposed in 1988, and the Unicode Consortium that serves as caretaker of the standard was launched in 1991. Joseph D. Becker, Unicode 88 (1988); *Chronology of Unicode Version 1.0*, THE UNICODE CONSORTIUM, http://www.unicode.org/history/versionone.html.

[11] *See* Donald E. Knuth, *The New Versions of TeX and Metafont*, 10 TUGBOAT 325 (1989) (introducing TeX 3.0, with support for 256 character sets); Oren Patashnik, *BibTeX Yesterday, Today, and Tomorrow*, 24 TUGBOAT 25 (Proceedings of the 2003 Annual Meeting, 2003) (listing multilingual support as a goal for the next phase of BibTeX development).

MLZ is a contribution to the "third wave" in reference management, a fresh cohort fueled by changes the surrounding environment. Programming skills are far more broadly distributed than previously. Powerful high-level languages and toolsets allow more work to be done with less code. Collaborative software development tools are more advanced and more accessible, and multilingual text processing is now simpler and more standardised.[12] These developments are behind the emergence of Zotero, as well as projects such as ColWiz, Mendeley, Papers and Qiqqa. In today's environment, detailed feedback and code contributions by a proportion of users (not all users, but some) is more likely and more productive, which helps explain why important elements of the "third wave" are distributed and maintained as free *and open* software.[13]

Openness and user participation are enabling, in three respects. The first point concerns development incentives. Closed-source development works best when aimed at existing audiences. In reference management, this means automating well-established and consistent citation practices used in specific academic fields. Where the market and the product are clearly defined, the gains (in time saved and money earned) are relatively immediate. But as illustrated in the previous section, the technical hurdles for legal and multilingual support are high. Recovery of development costs is less certain in this context, because potential users are dispersed across national, jurisdictional and disciplinary boundaries. From the standpoint of proprietary development and marketing, building out a finished system with such capabilities in the hope that it will soon pay for itself would be a risky proposition at best.

An open source model does not make these issues vanish, but user involvement in development does permit an incremental approach, in which a basic framework can be extended by degrees through contributions by interested user communities.

The second point concerns the needs of users themselves. A reference manager is a personal library, and as such its content must remain accessible across the career of the researcher. To achieve the full benefits of a uniform referencing platform, it must similarly be available for sharing with potential collaborators. Exclusive licensing makes access less certain, which impedes dissemination. Proprietary products are not without a role in this space, but broad appeal across the full spectrum of potential users requires a free and open platform at the core, permanently accessible, and sustainable over the long term.

A third factor, relevant to international dissemination of bibliographic technology, is security policy. Potential stakeholders in uniform reference management include government bodies that are understandably sensitive about national reliance on software systems controlled by players beyond the reach of regula-

[12] *See supra* note 10.

[13] The `citeproc-js` citation formatting engine written by the author is distributed under alternative free software licenses (AGPL and CPAL), and runs in the core of Mendeley, Qiqqa, Zotero, and other projects in the third wave. Styles in the CSL style repository on which `citeproc-js` depends are distributed under a Creative Commons license. Both Zotero itself and the MLZ system introduced by this book are distributed under an AGPL license.

tory discipline.[14] With that thought in mind, let us now turn to consider the current state of legal information technology in the United States.

Law and Order

Legal resources have particularly demanding citation requirements, and to some extent this is unavoidable. It is important that both legal citation styles and legal research software attend to the special needs of the field, while taking scrupulous care to prevent complexity from running out of control.

By their nature, *legislation* and *administrative rules* are subject to revision and recompilation. A target text may be an "original act" creating entirely new law, an "amending act" specifying only the changes to be made to existing law, or a "consolidated act" in which the changes are merged into a finished revised text. Finally, current acts (original or consolidated) may be reorganized for inclusion in a "codification scheme" such as the US Code. Citations in legal argument must identify the exact text relied upon; and because rulemaking processes vary across jurisdictions, citation methods must do so as well. Hence some degree of complexity, not to say arbitrariness, is inevitable.

On the other hand, *judicial judgments* seem at first blush to be less daunting. A legal judgment is an immutable document, issued by a decision-making body at a single point in time. In terms of citation, such material appears to differ little from journal articles and the like: given a canonical document, all that would seem to be required is a uniform scheme for describing it. However, the significance of a legal judgment is heavily dependent on context. A decision may be subject to appeal, and while an appellate judgment is also a single immutable text, the overall procedural history of a given legal case is essential to understanding the significance of the individual judgments contributing to the final result. In addition, the rule or interpretation expressed in a final judgment may be modified or overturned by entirely separate judgments in other cases at a later point in time. The need to track all of this detail is more a matter of reference management than citation practice, but it is critical to the legal research endeavor, and must be born in mind.

The wrinkles described above are inherent features of the law itself. Further complications arise from the ways in which legal text is published. Once issued by the court, a legal judgment may be disseminated through multiple channels, some of which may be more readily accessible to some readers. Accordingly, many legal styles in the US (and certain other jurisdictions) require that parallel references be given when a judgment is available through multiple reporters or services. Such citations typically have special formatting requirements, omitting some elements of each cite in the series of parallels to save space and improve

[14] *Information Warfare: Running for Linux*, STRATEGY PAGE (Jan. 9, 2011), http://www.strategypage.com/htmw/htiw/20110109.aspx (Government departments in Russia and China pressed to adopt domestic or open source operatings systems and office suites out of security concerns).

readability. For example:[15]

> Harvard Crimson, Inc. v. President and Fellows of Harvard Coll. 445 Mass. 745, 840 N.E.2d 518 (2006).

Legal citation styles are thus complex things, in part because of the difficulty of the underlying material and in part due to idiosyncrasies of the publishing chain (factors aggravated, as we have seen, by variation across the world's legal systems).

Compleat Rules of Citation

The leading US legal style began life as a short pamphlet entitled "A Uniform System of Citation: Abbreviations and Form of Citation", written by the outgoing editor of the Harvard Law Review to support the work of classmates in the following year. The document opened with a disclaimer that would remain substantially the same through nine subsequent editions over the ensuing forty years:[16]

> "This pamphlet does not pretend to include a complete list of abbreviations or all the necessary data as to form. It aims to deal with the more common abbreviations and forms to which one has occasion to refer."

From this modest beginning, the Uniform System grew in scope, size and influence, and is today an established national fixture of the legal research and writing process. The "pamphlet" became a "booklet" in 1934, having grown to some 48 pages. Two years later, it was printed with a copyright notice for the first time, with the names of three other leading law reviews listed as joint proprietors. At the first national conference of law review editors, held in 1949, the style was the sole candidate put forward as a national form of citation. Beginning with the eleventh edition (published in 1967), the tone of the guide began to take on a more assertive quality, dropping the declaration of incompleteness and offering the following guidance:

> "The editors are unable to recommend that the Third Edition *Merriam-Webster New International Dictionary* replace the Second Edition as a general authority for definition and italicization. The new edition fails to distinguish those foreign words which should be italicized in English writing, and is in general *insufficiently prescriptive*." (latter italics supplied by the current author)

[15] In civil law jurisdictions where extra-judicial commentary plays an important role in legal interpretation, case notes and the like may be appended to a case reference in a similar shorthand fashion.

[16] A UNIFORM SYSTEM OF CITATION (1926).

In the twelfth edition (1976) mimeographed supplements formerly supplied on request were folded into the main publication. From that point forward, the guide has progressively expanded in bulk, to become the "511-page tome"[17] that it is today.

As a counterpoint to the progress of these events, a minor literature of satirical review has grown up around the guide in recent decades. The leading voice is of course that of Judge Richard Posner, whose distaste for the complexity of the guide is well known:

> "The particular faults of the Bluebook ... place it in the mainstream of American legal thought. ... The vacuity and tendentiousness of so much legal reasoning are concealed by the awesome scrupulousness with which a set of intricate rules governing the form of citations is observed." (Posner, 1986)[18]

Similarly pointed sentiments had been expressed by others before:

> "[The Bluebook's detailed rules on citation] increase the speed at which the legal enterprise slows down." (Strasser, 1977)[19]

and have been expressed by others since:

> "For those who think too intensely about law—including anyone who ever edited or wrote a law review article—the Bluebook serves as a morality play too dull to endure but too conspicuous to ignore." (Chen, 1991)[20]

There have been two attempts to launch competing guides.[21] Neither has caught fire in a big way: despite positive reviews of the new entrants in some quarters, the leading style has not been dislodged as the basic standard of citation in legal publishing and, with some variation, the nation's courts.[22]

The inertia of the unpopular incumbent has been explained in terms of customer adherence to a leading standard under so-called "network effects".[23] As that theory has it, where the unit value of a product increases with expansion of its customer base, existing customers will become progressively more reluctant

[17] The phrase is from Richard Posner. Richard A. Posner, *The Bluebook Blues*, 120 YALE L.J. 850, 852 (2011).

[18] Richard A. Posner, *Goodbye to the Bluebook*, 53 U. CHI. L. REV. 1343, 1343–44 (1986).

[19] Alan Strasser, *Technical Due Process: ?*, HARV. C.R.-C.L L. REV. 507 (1977).

[20] Jim C. Chen, *Something Old, Something New, Something Borrowed, Something Blue*, 58 THE UNIVERSITY OF CHICAGO LAW REVIEW 1527, 1528 (1991).

[21] *See* Posner, *Goodbye to the Bluebook*, *supra* note 18 (announcing the Chicago Manual of Legal Citation, or "Maroonbook"); Bast & Harrell, *supra* note 2 (arguing the case for a manual published by the Association of Legal Writing Directors).

[22] Adoptions of the ALWD Citation Manual can be viewed online at `http://www.alwd.org/publications/adoptions.html`.

[23] *See* Bast & Harrell, *supra* note 2.

to shift to the incompatible products of a competitor, even when those products are of superior quality.[24] Under certain conditions, such markets may be "path-dependent".[25]

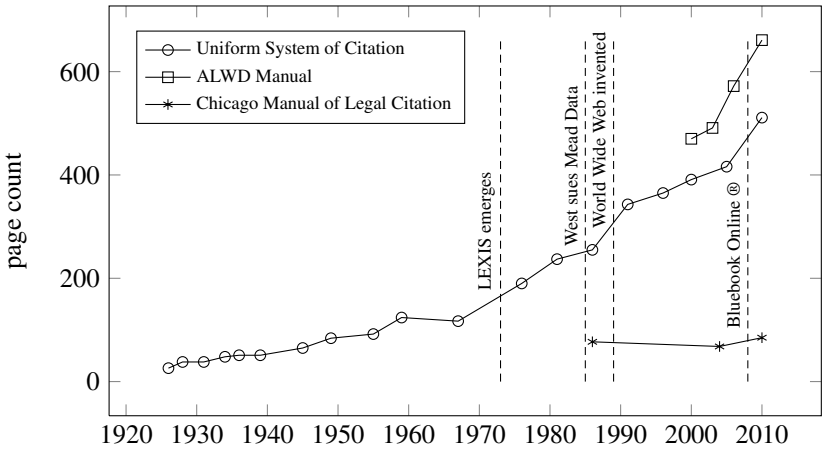


Figure 1.1: US legal style guides and information technology

Unfortunately, the path in this case is pretty clearly headed over a cliff. Figure 1.1 shows the relative size of the leading style and its would-be competitors as a function of time. As reflected in the page counts, the Chicago guide (the "Maroonbook") has attempted to reverse the leading style's trend toward ever more detailed prescriptive rules.[26] The ALWD guide, on the other hand, self-consciously strives for "compatibility", diverging only in marginal simplification and smoother exposition of the leading style's precepts.[27] By all accounts the ALWD guide has achieved the greater uptake of the two.[28] This fits nicely with the expectations of network effects theory,[29] but it is hardly cause for celebration, since the primary complaint directed at the leading style is its sheer bulk.

From a technologist's perspective, the most striking feature of this thread of discourse is how consistently it misses a sweet irony: the self-defeating spiral of unmanageably detailed rules, this inefficiency trap, is a byproduct of the rise of information technology.

---

[24] Joseph Farrell & Garth Saloner, *Standardization, Compatibility, and Innovation*, 16 THE RAND JOURNAL OF ECONOMICS 70 (1985).

[25] *Id.*

[26] Posner, *Goodbye to the Bluebook*, *supra* note 18.

[27] *See* Bast & Harrell, *supra* note 2; Christine Hurt, *Network Effects and Legal Citation*, 87 IOWA L. REV. 1257 (2002).

[28] Hurt, *supra* note 27.

[29] *See id.*

The public launch of the LEXIS service by Mead Data Central in 1973, flagged in Figure 1.1, marked the beginning of electronic search and retrieval systems as a tool for legal research. West Publishing Company was at that time (and remains today) the dominant legal publisher in the US, boasting comprehensive coverage of state and federal case law through a range of reporters. West was prompted to respond to the LEXIS challenge, introducing its own electronic service in 1975. In 1985, LEXIS announced a plan for "star pagination"—markup in electronic text pages flagging page boundaries and page numbers in West case reports, which would eliminate the need for its subscribers to refer to (or subscribe to) the West reports themselves. West sued LEXIS over copyright infringement, and the parties ultimately settled. The terms of the agreement remained confidential until 1998, when West was compelled to disclose them in litigation with a third party over the same issue.

If any doubts existed over the position of electronic research systems in the law, they were dispelled by the serious ambitions behind *Mead Data Central v. West*. Today, the US legal system is heavily reliant on electronic text retrieval and associated systems. These are expensive services, but they have real value because of their scope, speed and accuracy, extending the lawyer's access to the raw stuff of legal research.

Market incentives have been crucial to the development of more efficient legal research platforms, but market incentives are a two-edged sword. The other blade comes into view when we consider the role played by metadata in these platforms, and in third-party reference managers like MLZ.

A Market for Consistency

In modern electronic library systems, documents can be retrieved in one of three ways. The most familiar to readers will be the URL, a specially formatted string first defined by Tim Berners-Lee while working at the CERN research institute in 1989.[30] As everyone reading this is *well* aware, a URL looks something like this:

```
http://digitalcommons.law.yale.edu/fss_papers/498/
```

URLs can be very long and awkward to manage. They are also tied to one and only one copy of the document, access to which might be blocked (for example) by a paywall. To provide a more robust means of tracking down documents, unique identifiers have been developed that can be assigned to a published work independent of its location. Typical examples are PMID[31] (for articles in the medical field), DOI [32] (for articles generally) and ISBN[33] (for books). Such

[30] Tim Berners-Lee, Information Management: A Proposal (CERN 1989).

[31] *See* National Center for Biotechnology Information, *Search Field Descriptions and Tags*, PUBMED HELP (2005), http://www.ncbi.nlm.nih.gov/books/NBK3830/.

[32] *See* International DOI Foundation, *Overviews and Standards*, THE DOI SYSTEM, http://www.doi.org/about_the_doi.html.

[33] *See* INTERNATIONAL ISBN AGENCY FAQS, http://www.isbn-international.org/faqs.

identifiers form part of the *metadata* describing the work. On the World Wide Web, identifiers can be submitted to a special website (a "resolver") to obtain the specific URLs that lead to actual copies of the target resource. A DOI, to take one example, might look something like this:

```
10.1111/j.1747-4469.1976.tb00951.x
```

By dropping the text of this identifier into a resolver (or, indeed, into a search engine) we can obtain a set of links to the article in the preceding example.

Unique identifiers are a relatively recent innovation, and not all articles and books have them. Even if they do, it might not be known to the researcher—for example, the DOI of the article referenced above is not mentioned in the text of the article itself. In such cases, *structured metadata* that describes the resource can be used to find it, in effect by automating the process of looking the work up in a library card catalog or the like.

Structured metadata can take many forms, but a couple of examples will suffice to illustrate the concept. A description of the same article in the BibTEX format might look like this:

```
@article{langbein_market_1976,
  title = {Market Funds and {Trust-Investment} Law},
  volume = {1},
  issn = {0361-9486},
  url = {http://www.jstor.org/stable/827950},
  number = {1},
  journal = {American Bar Foundation Research Journal},
  author = {Langbein, John H. and Posner, Richard A.},
  month = jan,
  year = {1976},
  pages = {1--34}
}
```

In the RIS format, a description of the same article would look something like this:

```
TY  - JOUR
ID  - 8052
T1  - Market Funds and Trust-Investment Law
JF  - American Bar Foundation Research Journal
A1  - Langbein,John H.
A1  - Posner,Richard A.
VL  - 1
IS  - 1
PY  - 1976/01/01/
SP  - 1
EP  - 34
SN  - 0361-9486
UR  - http://www.jstor.org/stable/827950
ER  -
```

Technologists can grow quite agitated about the details of metadata formats, but for our purposes the key point is simply that structured metadata *is structured*, so that computers can easily parse out the content and do interesting and useful things with it. Using either of the structured descriptions above, a computer can easily retrieve a list of other works by the authors, reconstruct the table of contents of the journal issue in which their article appears, or search for other articles in which it is cited. Rich and plentiful metadata is the lifeblood of modern information systems. It makes them smart, responsive, and unintrusive.[34]

While citations and metadata share the same general purpose, they have distinct roles, and the difference is easy to miss. In a review of the nineteenth edition of the Bluebook, Richard Posner contrasts the leading style guide with a simpler, less prescriptive stylesheet used by clerks in his chambers. In a (sympathetic) column published in response, Stephen and Jonathan Darrow question the wisdom of diverging too sharply from the conventions of the leading style:

> *Although Westlaw properly processed most of Posner's Bluebook-defying citation forms, it choked on some seemingly reasonable abbreviations that we postulated. For example, abbreviating the word "Technology" as "Tech." in "13 Albany Law Journal of Science and Tech. 751" resulted in a Westlaw error message.*

A researcher in a field other than law might well ask, "Why on earth does this matter?" After all, citations exist for the convenience of people. For the convenience of computers, we have structured metadata and unique identifiers. The problem (and the reason the point made by Darrow and Darrow is valid as far as it goes) is that we *don't have* structured metadata or unique identifiers for US legal materials. If we are most lawyers, we probably don't even know what they are.

Despite the extraordinarily demanding citation requirements of the US jurisdiction, American lawyers are seldom exposed to metadata in a structured form. The article referenced in the examples above is available in the popular Westlaw legal research service (after we once overcome the "Westlaw error message"), but in contrast to aggregator services in other fields,[35] Westlaw provides only the text of the article, with no structured metadata. The same is true throughout the service, and throughout the archives of every commercial provider of primary legal text in the US market. No DOIs or other unique identifiers. No structured metadata.

As stated in their literature, Westlaw, Lexis, and other aggregators of US case law aim to provide comprehensive research support.[36] Such services depend

[34] Posner, *The Bluebook Blues*, *supra* note 17, at 853.

[35] A welcome oasis in the American legal metadata desert is HeinOnline, a leading legal publisher and aggregator of law review content.

[36] *See, e.g.* Thomson Reuters, *Research Fundamentals: Getting Started with Online Research*, WESTLAW 1 (Jun. 2010), http://lscontent.westlaw.com/images/content/GettingStarted10.pdf ("The Westlaw legal research service is comprehensive, easy to use, and up-to-date. It will help you perform accurate and effective legal research.").

internally on the sort of unique identifiers and fine-grained metadata of the kind described above, but that is an internal matter; at the customer level, the only identifier shared in common between Westlaw, Lexis and other services is the one established in 1926: [37] the properly formatted citation, in the leading style, as it would appear on the printed page.

The escalating page counts shown in Figure 1.1 are the end result of forcing citations (intended for humans) to serve as machine-readable metadata: achieving uniformity by dint of an instruction manual is possible, but it requires a very *long* instruction manual.

Consistency alone is not quite enough to make citations serve (approximately) as document identifiers. Some mechanism must enable machines to identify and interpret human-readable citations within a given document, so that citations can be resolved to proper identifiers and, ultimately, addresses. For better or for worse, such a mechanism does exist, in the form of *regular expression* pattern-matching, a common feature of all major scripting and programming languages. Regular expression code looks something like this:

```
/L=\|(?<volume>\d+)?\s?U\.\s?S\.\s?(?<page>\d+)
\s*?\|>(?<anchored>\d+)/
```

If the purpose of the code in this example seems obscure, that itself would be the point. Regular expressions are powerful, but also opaque, complex, prone to error and difficult to debug. In the words of Jamie Zawinski, lead developer for the Netscape browser:[38]

> *Some people, when confronted with a problem, think 'I know, I'll use regular expressions.'*
>
> *Now they have two problems.*

Widespread reliance on human readable citations in contexts that call for proper structured metadata is poor design. Systems that rely heavily on such code can be expected to break from time to time, and legal information systems in the metadata-starved US jurisdiction tend to do just that.[39]

Metadata starvation in the law is not complete, and the benefits are readily apparent where it does exist. HeinOnline[40] serves the article cited in the above example with structured metadata (it does not have a DOI as far as I am aware). The metadata format used there (COinS) is somewhat frightening to the untrained eye so I will not reproduce it here; but the full details are provided. Visiting the article's HeinOnline page in Zotero, we can do some useful things:

[37] *See* Posner, *The Bluebook Blues*, *supra* note 17, at 857; *citing* James W. Paulsen, *An Uninformed System of Citation*, 105 HARV. L. REV. 1780, 1782–85 (1992).

[38] See `http://regex.info/blog/2006-09-15/247`; `http://kagan.mactane.org/blog/2011/08/16/the-problem-with-jamie-zawinski-and-regular-expressions/comment-page-1/#comment-572`

[39] *See e.g.* `http://www.examiner.com/civil-rights-in-portland/justia-gate`

[40] William S. Hein, *List of Libraries*, HEINONLINE, http://home.heinonline.org/content/list-of-libraries/.

- Add an item for the article to our Zotero research database (with a single click).
- Find copies of the article supplied by other vendors (with a single click).
- Attach a copy of the article to the Zotero item just created (ditto).
- Insert a citation for the article into a document (again with a single click).

Note that none of these operations involve laboriously typing out citation details. That was done once by the maintainers of HeinOnline when the article was published, and there is no need to do it again. Systems based on structured metadata just work.

In a normal market, the leading style would have plenty of competition. The Maroon Book, the ALWD guide, the McGill guide or the OSCOLA guide, not to mention Judge Posner's brief stylesheet, are all perfectly worthy alternatives. There is far greater variety in citation style in other disciplines. Because they are rich in metadata, other disciplines have the freedom to choose how they want their cites to appear. In the law, because we have no publicly accessible metadata in the world's largest jurisdiction, we do not have that freedom.

Electronic library systems require unique, uniform identifiers for machine-driven referencing. Citations written according to the leading style guide currently serve this role, in the US jurisdiction and in many others. The uniformity of written, human-readable citations is important, as adoption of variant styles by court systems and law reviews across the country could easily cause the cross-referencing on which our fragile information infrastructure depends to completely fall apart.

Legal citations thus play a role similar to that of URLs, as well-defined "addresses" that identify referenced resources in an electronic repository. Ordinarily, the specifications for such things are managed as a public resource by the firms or persons that benefit from them. For example, the standards document that defines the format of URLs carries the following note concerning republication:[41]

> *Distribution of this memo is unlimited.*

Standards development is neither easy nor cheap, but the standards underpinning the Internet can be distributed freely because the major players see interoperability as good for business. They support a collaborative standards process out of self-interest.

The leading style guide is an important common asset in legal publishing, but the two monolithic research services have little incentive to promote interoperability. That being the case, the leading style has remained the responsibility of a small group of law students, and their work has come to be supported primarily by sales of the guide. In contrast to the RFC protocols that drive the

[41] Roy Fielding et al., Hypertext Transfer Protocol—HTTP/1.1 (The Internet Society 1999).

Internet, the online version of the leading US legal style is published with the following restriction:[42]

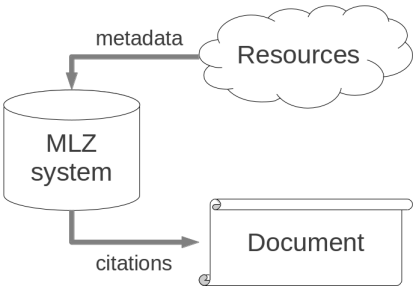
> *Except as expressly provided by this Agreement, any use of the Site and its content is strictly prohibited without our written consent.*

The uncommonly strong demand for uniformity in the US jurisdiction has led to a pay-per-view model for funding legal style maintenance and development. This approach rewards expansion of the guide itself,[43] discourages moves toward proper automation, and appears to have reached a point of diminishing returns as far as users are concerned.

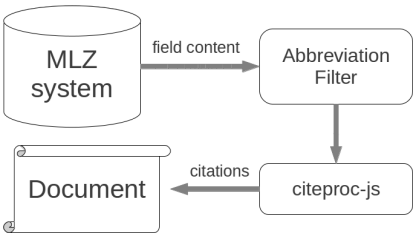
Moving Forward

Citations are complex things, and a piece of software will not, by itself, move us past the current impasse. Data must be organised and arranged consistently to achieve good results across styles, and across systems. In MLZ, citations are generated by combining input from two sources: the MLZ database; and the abbreviation lists associated with the target style.

The MLZ database holds the raw metadata from which citations are generated, in the workflow illustrated to the right. Metadata can be input manually, or it can be captured automatically with one click when an appropriate "site translator" is available (a small snippet of code that extracts metadata from pages on a particular website). In either case, the elements of metadata (court name, title, editor, etc.) should be assigned to MLZ fields using the *Item Examples* appendix as a guide. The MLZ styles will produce correctly formatted citations for items described in the appendix, without awkward adjustments to content.



In addition to citation forms themselves, the conventions for abbreviation vary between styles. In MLZ, abbreviations are applied by the *Abbreviation Filter*, a small Firefox plugin that stands between the raw content of the MLZ database and the formatting engine, as shown in the illustration to the right. Abbreviation lists are style-specific, and for legal styles, in



[42] Harvard Law Review Ass'n, *Terms of Use*, THE BLUEBOOK (Feb. 15, 2008), https://www.legalbluebook.com/public/TermsOfUse.aspx.

[43] *See* Posner, *The Bluebook Blues*, *supra* note 17, at 852.

particular, the logic coded into the lists is necessary to proper operation of the style for legal content.

The remaining chapters and Appendices of this book concern specifics of the MLZ system. Like any complex software, the project is a moving target, and the current version as you read this may differ in some details from the description provided here. Significant changes introduced after the time of writing will be documented (together with errata) at the following URL:

`http://citationstylist.org/errata`

The following chapters, *Getting Started* and *Under the Bonnet*, offer basic information on installing, operating and extending MLZ. The instructions are not comprehensive; links to relevant resources on the Web are provided for users who wish to dig deeper than the outline view supplied here.

The Appendices provide guidance on the data entry patterns expected by the MLZ styles, as well as notes on the syntax of abbreviation entries, and on special features of the MLZ extended version of the CSL citation formatting language. This is the heart of the system, and the portion of this book likely to have the most persistent utility to readers.