A WWW manifesto

The upcoming CALI/LEAP conference and the LII Hypertext Authoring Workshop which will follow quickly thereafter provide good opportunities for discussion of why and how we might derive some "community standards" for legal-information publishing on the Net. We've thrown together this document in the interest of getting such a discussion started.

No doubt a lot of it will be old hat for those of you who have been exposed to these arguments on TEKNOIDS and in other venues over the past eighteen months. We need to remember that there are significant new providers of legal information emerging on the Net, and that these people have not had the dubious benefit of our collective wisdom, and that faculty within law schools have not really been exposed to it at all. For that reason we decided to construct a broader document which would address the standards and coordination issues, including along the way an introduction outlining the reasons why anyone should be concerned with this in the first place.

Why publish on the Internet?

Several law schools, both in the United States and abroad, have seen the virtues of mounting electronic information on the Internet. Many more have plans to do so. It also seems likely at this point that the traditional online services will be offering legal information via the Net, as will a galaxy of new private- sector legal information providers, professional associations, and law firms. Many of these efforts are in the proof-of-concept stage, but all will be offering significant collections of information in the not-too- distant future. Some, like Cornell's, Indiana's, and six or seven others which will come on-line in the next few months, have gone or will go well beyond that. The international legal community, in particular, has already started to gravitate to online publishing, and several servers have appeared in Germany, Norway, and New Zealand.

The reasons why a law school would wish to do this sort of electronic publishing are varied but easily understood, inasmuch as they're a superset of the reasons why law schools have traditionally published journals and law teachers have written for them. The reasons to publish electronically on the Internet include all of the reasons why one might want to publish in any medium including print. Beyond that, the Internet offers many advantages over print as a publication vehicle. Reasons for preferring electronic publishing in general over print have been so extensively presented and discussed that there is probably no need to repeat them here, except to briefly mention that electronic systems in general provide advantages of searchability, malleability, and instantaneous delivery and updating which are familiar to most if not all legal people. Beyond that, we should add that publication on the Internet is inexpensive, and that it reaches a much wider and diverse audience than traditional legal print materials. The combination of these two factors makes it economically feasible to mount materials of interest to groups of scholars, students, and others who are "beneath notice" under traditional print publishing schemes because the interest group has been perceived as too small to constitute a viable market. For the time being, too, the mounting of materials on the Internet offers a fair degree of visibility to the law school doing so, something which is noticed by potential students, alumni, and others in a way which warms the hearts of senior administrators, admissions officers, and others. Finally, there are whole classes of information other

than scholarly materials which law schools may wish to place on the Net, ranging from information about admissions policies and maps of the campus to library pathfinders, course schedules, and other things which either serve the local population or attract others to join it.

For most law schools, the delivery system of choice has been the cluster of technologies known as the WorldWideWeb (hereafter the "Web"). Some academic providers have adopted a "go-slow" posture with respect to the Web; these few individuals cite lack of an adequate hardware base on which to run Web client software (and occasionally other administrative and technical problems with their local infrastructure) as a reason for caution or for steadfast adherence to older, less capable technologies such as Gopher or simple Telnet sessions. While there are a lot of low-end machines out there, we believe that this is a problem which can be worked around as matters stand, and which will disappear totally in the future, in part because of improved law-school infrastructure and in part because of improved Web technology. The commercial world is certainly placing its bets on the Web, as are the majority of other academic disciplines.

A full description of the Web technology is outside the scope of this document. Two aspects of the Web are crucially important to our ideas here, and should be borne in mind throughout. First, it is a distributed hypertext system which permits document authors to link documents and other multimedia-type objects such as graphics, sound, and video together across machines on the Net. There are, of course, many implications to this, but perhaps the most important is that large bodies of related text can be built from smaller collections mounted at different sites belonging to different organizations for different purposes. For example, 50 law schools might independently mount materials related to a particular area of law as it is treated in each of 50 states, and then link the entire collection together through a single table of contents. A teacher might draw on materials mounted by law schools, commercial firms, news services, and law firms in structuring a collection of text for a particular class; she might also draw on the work of other teachers who have done similar things in the same area. The second important aspect of Web technology is that it incorporates a system of typographic markup which can also be used as the basis for directed searching. In this first respect it differs from WESTLAW and LEXIS, which currently offer no typographic capabilities, and in the second (search capability) it behaves, or could behave, in a way similar to the "field searches" offered by those two services.

Why standardize?

Standardization and coordination of effort are meant to address problems which have already shown themselves, and to anticipate problems which are likely to arise, as use of this technology by "law people" proliferates. There are problems (like concocting schemes for link naming and navigation) which arise from the Web technology specifically, and there are issues which would complicate any effort, regardless of the technology used, in which multiple authors and institutions are or will be involved. These two categories we consider separately as problems of standardization, which are concerned with the style in which we use the technology of choice, and issues of coordination, which are concerned with the way in which we collectively construct a body of substantive material. Overlapping these two areas is a set of concerns involving information quality, which we would wish to address in a standardized way not so much as a way of establishing quality control practices, but as a way of offering

standardized mechanisms by which users may determine the quality and authority of the information we offer as it applies to their own concerns and interests.

Needless to say many of the reasons for standardization flow from the things one proposes to standardize. One can't discuss them independently, and so the "What to standardize" section below provides the real rationale for many of the things we are proposing. But decisions about what to standardize also need justification (as will be seen shortly) and should not be made lightly, especially insofar as they impose extra burdens on authors. However, once we agree that certain scenarios for the usage of our material are likely, it becomes necessary to think about standardization. For purposes of this document we have thought about three problems in particular: the problem of resource location, the problem of quickly structuring teaching materials out of existing electronic text, and the problem of conveying information about accuracy, provenance, and the authority of a particular document to a user who may fall almost anywhere on a spectrum of legal expertise.

The resource location problem is well known to anyone who has spent more than fifteen seconds cruising the Net. Things are hard to find. To make them easier to find we need catalogs which may or may not be automatically constructed, but which in any case will be easier to construct provided certain levels of standardization of structure and nomenclature exist among the things being cataloged.

The second problem is, in fact, a bundle of problems such as those which might face a law teacher wishing to quickly structure a set of linked hypertext materials assembled from multiple sites, with the aim of supporting a particular course or even a particular pedagogical unit within a course. In short, we try to anticipate the problems which would face the teacher preparing materials on Sunday night for a class on Monday morning. How might she go about writing a hypertext outline for the class, linking materials from different law school sites to it as needed? How would she know what's available and what it's called?

The third problem is purely a product of the Internet environment as we have experienced it thus far. It is common knowledge that there have been serious problems with the quality of information offered by volunteers who are primarily interested in the workings of delivery systems rather than in what is being delivered (the infamous "periodic table" incident comes to mind here). A corollary challenge lies in the fact that, even if one does create quality documents, it is often necessary to provide some indication of exactly how reliable the document is, and of what context should be considered when interpreting it, for a highly diverse audience which may be more or less familiar with the law in general and with US law in particular. For example, one might put a multilateral treaty on the Net, and then be faced with questions such as whether or not the draft is authoritative, whether a particular reader's country is bound by it or not, whether amendments not available online have superseded its provisions, etc.

Before going further we should point out that there are some disadvantages to standardization which have primarily to do with the process used to achieve it. In general we need to keep three things in mind. First, the history of the Internet (particularly when compared to the OSI standards process) has shown time and time again that debating standards without reference implementations is a time-consuming and ultimately useless

process. Second, standardization should not become a constraint on thoughtful experiment. We should concentrate instead on building successive supersets of standards, firming them up as there is real evidence that one or another approach deserves to become the common method. (Needless to say we will need reference implementations to make valid comparisons.) Finally, and perhaps somewhat cynically, we should remember that there will be those who prefer debating standards to actually doing anything of use to anyone, and those who would impose standards which try so hard to account for all cases that they become (if they can ever be agreed on, and documented) so burdensome that authors will ignore them. We should try to keep the process a reasonable one which maintains roots in reality.

## What to standardize

### Extensions to HTML markup

The ability to apply a very flexible typographic scheme to legal text of course implies a need to have similar texts look similar to one another and to traditional print forms, assuming that one wishes to avoid confusing the reader. It also suggests that law schools have something at stake in the discussion of those standards; as a community, we have need for typographic entities which many other disciplines don't use and would not consider a part of any base standard. Many of us remember the tricks we had to play -- and not so long ago -- to persuade WordPerfect to insert a section symbol into a document, or to produce small-cap fonts for citation purposes. Of course, once commercial software developers recognized that the legal market was a significant one, these features began to be incorporated into their products. A somewhat similar process surrounds the development of software for the Net. We have actually had an influence on the Web markup standard which is out of proportion to our numbers, at least to the extent that Cornell has been actively involved in Web client development and to the extent that the "clerk of the works" on the typographic standards document (the HTML+ specification) has been disproportionately sympathetic to the needs we have articulated. But there has as yet been no effort amongst "law people" to agree on a uniform (and, one hopes, lightweight) set of extensions to the standard Web markup which would be of use to the community as a whole. One is needed, especially if we expect to lobby for it effectively in the Net standards process.

### A core set of searchable fields

The preceding paragraph is directed primarily at typographic markup; the desire to add field searching capability presents a slightly different problem. First, we can take it as given that almost anything we do will be ignored by the rest of the Web community, as the field names we might use would for the most part be so discipline-specific as to be ignorable by the majority of authors, and because Web client software will (unless designed otherwise) ignore any markup used for this purpose without ill effects. The need here is primarily for agreement within the legal hypertext community for a system of fieldnames which will be supported in different kinds of materials and which users can reasonably assume will be present in collections they wish to search. It is a little dismaying to note that the online providers, who of course have considerable experience with this, have never fully standardized these even across collections within one service, let alone with respect to each other. Nonetheless, it could be done up to a point, and that level of standardization would be sufficient for most.

Note, too, that this issue is closely coupled with the idea of

cross-site search capability. So long as search engines have existed in close proximity to the materials being searched (ie. were put together by the same information provider), the critical need has been for documentation which explains to the user how a particular search engine works with respect to a certain body of material. Examples of this include (at the trivial level) the explanatory help text put next to Cornell's Directory of Legal Academia and (at the non-trivial level) the help texts used by WESTLAW and LEXIS respectively. Once engines are usable across sites (Indiana's search engine plowing through Cornell's material, and so forth) the need for standardized search fields and engine interfaces becomes much greater. Standardization of query protocols (a la Z39.50) and of engines (use of common search software at all sites, or provision of standard-format indexes for pickup) do not, of course, address this problem ; the need for standardization lies in the data itself insofar as the tagging is a part of the data for this purpose.

Document granularity

Document "granularity" is a term which takes in a set of decisions made by authors about how text should be divided and marked for linking. There are several factors which bear on these decisions:

* Speed of delivery. We can anticipate that for some time many users of our systems will, at least some of the time, be limited to retrieval speeds which can be supported via modem (eg. SLIP and PPP technologies). This places a somewhat fuzzy upper limit on the size of document parts, one which is tied to the longest retrieval time which we believe that most users will find acceptable. It also suggests that we might discuss, along with the rest of the Web community, useful schemes for link weighting and prefetching as ways to work around slow line speeds.

* The structure of the document itself. Statutes and regulations have a structure which in and of itself strongly suggests the ways in which one might divide them into smaller sections. Others, such as judicial opinions, do not. For those which have "natural" subdivisions we need to agree on the level at which actual file divisions should take place and at which link destinations should be marked (eg. section, subsection, paragraph and so on). For those where there are no "natural" divisions, we need to agree on a general approach. For example, with judicial opinions one might agree on a system of abstracts or headnotes, finding a way to isolate the holding or holdings, and so on. It is entirely possible that this problem cannot be easily solved in a standard way even if the method is a highly abstract one, and we will want to be careful to identify a point of diminishing returns in the discussion. This suggests that a first step would be to identify a list of general document classes for which a standardized approach is possible and desirable.

* Logical unit of text retrieved in a search. Most if not all search engines commonly in use retrieve text at the file level; the searcher receives, in effect, a list of files which contain the text being searched for. Experience (in particular with Gopher) indicates that if a given document has been divided into chunks (files) which are too large, search mechanisms become useless. The files retrieved by a search are simply too big in themselves to be useful. In the extreme case, a search run against (say) the Copyright Act might return one result: a file containing the entire Copyright Act.

* Future use of the material. Much of our effort over the next year or two will most likely involve the mounting of "core texts"

which can be extensively leveraged by others. Authors need to be aware that their materials may be referenced and used by others, and structure appropriately. In general, this will probably mean dividing them (or at least providing link targets) at a level somewhat finer than is needed for immediate purposes, no matter that "extra" work is involved at the outset.

Link naming and referencing schemes

While we expect that some scheme for exporting catalogs of available links (possibly for general use and cataloging purposes, possibly for "pickup" by an Archie-like mechanism or other agent software) is going to be suggested and become prevalent within the next year -- perhaps especially because we anticipate that this will happen -- we think that we should agree on relatively standard schemes for naming target links. There are three reasons for this. First, a level of standardization helps authors who wish to make links which reach inside documents provided by others; they can easily guess or construct the name of the target link without necessarily resorting to a catalog or to inspection. Second, a level of standardization decreases the amount of work needed insofar as authors need not reinvent a system for link naming each time a new document or collection is approached. Finally, such a scheme can potentially act as the basis for a system of citation.

It seems likely that such a scheme would actually comprehend several subschemes, with each of the subschemes strongly related to document class (as described above in the section on granularity).

A Web bluebook?

Ultimately, the work that we do in discussing stylistic standards and in trying reference implementations ought to be aimed at the construction of a sort of style book for electronic legal text on the Net. This, we are afraid, implies the existence of a person or group of people who would document our decisions as we go, and from time to time summarize those standards which are evolving from practice without benefit of discussion. It would be our hope that someone would volunteer to fill this role as soon as possible, with the aim of having an evolving style guide available continuously, and a settled one completed before another year goes by. Some of this will of course depend on other processes, particularly the parallel effort to standardize HTML+, but there is no reason we should not construct something to which newcomers could be pointed as the process continues.

Why coordination?

"Coordination" as we have used it here is a code word which takes in two concepts: avoiding duplication of effort, and agreeing upon a logical evolution of the totality of legal materials on the Net which will most effectively promote the general usefulness of that material and, for the present, make it an attractive environment for potential authors to consider.

Avoiding duplication of effort is, of course, crucial. Law schools are not publishing houses and they are not computing centers. Only one or two have any staff whose time is fully dedicated to this sort of activity; some have no staff whose time is exclusively dedicated to any kind of computer support. Time for these activities is at a premium, and we do not wish to squander it.

The second notion proceeds from the idea that authors in print

enjoy (if that's the word) a huge body of material on which they can, and do, build by reference through footnotes, citation, and other mechanisms. We, of course, have hypertext linking -- but in some cases there is not much to link to. We need to give some consideration to the order in which we collectively put things on the Net, and try to provide at the earliest possible opportunity those materials which teacher/authors will most want to build on.

Needless to say, along with this goes the idea that we have to construct mechanisms for letting people know that these things exist, either actively through some form of notification or passively through catalogs, topical indexes, and cross-Web search tools.

The downside of coordination efforts is that they can serve to deter legitimate, alternate approaches to similar or identical bodies of material. Moreover, they can be even more time consuming and burdensome than stylistic standards if given the chance. For example, the "registration" of a project or document as described below -- simple notice to the community that one group has or will be working in a particular substantive area -- should not be assumed to be a "hands off" notice. While (particularly in the case of core materials) another group might want to think hard before working with the same substance already dealt with by another group's effort, they should feel free to do so if they feel they can add value. Nor should we become preoccupied with constructing a single taxonomy of materials as a framework for future efforts. There is unlikely to be agreement about the proper taxonomy to use, difficulty in finding the proper "pigeonhole" for some contemplated project, and so on. The current thinking at Cornell is that it is possible to apply a minimum of three perfectly legitimate high-level taxonomies to any topical index of legal material on the Net, and the upper limit is certainly much greater. In short, coordination of effort requires notification of others that work is underway or completed. It does not require that everyone adhere to some master plan, nor should it. There needs to be room for experiment and differing perspectives in this area, and there is no one body which anyone will now or would in the future agree on as having any authority in this area.

What to coordinate

Project notification and document registry

Given that a primary concern is avoiding duplication of effort, it's obvious that some clear channel needs to exist through which groups might notify each other of planned or completed work, and that these notifications need to be archived in searchable form somewhere so that others might see them. This was the original plan behind LAWSRC-L. LAWSRC-L has not been effective for two reasons: it never caught up with the legal resources which were already available at the time that it started, and providers have not been diligent about its use for announcing new projects. It is fairly clear that even if most providers suddenly become diligent about using LAWSRC-L, or indeed another notification system such as a USENET News group, some will not be. It would therefore seem useful to have an archivist or archivists for the notification channel whose responsibility it is to take up the slack for providers who forget to send notifications. This will not, of course, deal with people who do not send notification of work underway, since there is no way to read their minds. But it would help deal with the backlog and with those who neglect or do not know about the notification channel.

Such a registry should, of course, be divided into "works in

progress" and "works completed". What is not clear is how far beyond that we wish to go. Someone will no doubt suggest that we add a topical keywording scheme along with some sort of search capability which would allow someone to determine if work is going on or is completed in a particular area. Such a keywording scheme clearly needs a controlled vocabulary, and efforts to construct one by consensus will run afoul of the same problems discussed above with respect to a "universal taxonomy". It is probably best, then, to simply leave the construction of the keywording vocabulary at the discretion of the archivist who runs the archive. A workable scheme is under construction currently at Cornell, based on work done by Milles, Martin, and Bruce on standardized topical outlines; the results should be available by the end of the summer.

A first step

Another crucial element in the process of notifying all servers of any changes in location or creation of information resources is the establishment of a means of communication among WWW server maintainers. A mailing list called "legal-webmasters@fatty.law.cornell.edu" has been set up on an experimental basis to provide this channel of communication. It has several purposes, loosely defined as:

1) sharing of information about installation, specific servers, etc.
2) sharing of information about scripts and scripts themselves
3) coordination of effort, so that everyone knows about everyone else's resources in a timely and easily applicable fashion (I.e., posting URLs, HTML to be included in files, etc.)
4) posting of changes to servers that might affect others
5) discussion of new and/or needed elements for servers
6) posting of notice of new additions to servers
7) discussion of ways in which WWW servers can be better utilized

Legal-webmasters will be archived in the LII gopher: gopher://gopher.law.cornell.edu/11/listservs

A full text index?

Strong arguments can be made for a full text index of all materials on all legal servers. It is not at all clear at this point how such a thing would scale should the body of material grow at the rate we anticipate, or how it could be mounted in any one place without overstraining hardware and software. Nor is it clear that it could be effectively updated as institutions make changes and revisions to individual texts. Development of a distributed full-text indexing system that can overcome these difficulties and also utilize the search capabilities of structured hypertext (HTML) is currently under development. It will probably be at least a year before many of these problems are worked out, however.

A full-text index of abstracts or synopses might be more workable (essentially a "file card" plus synopsis and, perhaps even more significantly, some "contextualizing" information useful to non-legal audiences, non-US lawyers, and so on). In this case it would become the responsibility of the group originating the material to provide the "file card". One might think of such a scheme as a mini-OCLC plus synopses. Something similar to this has been contemplated by the LII, but the entire issue needs further discussion. It may be, too, that construction of such an

object should be deferred until a workable URN scheme appears, but since that's not likely to happen very soon one must consider the retroactive work which would be piling up in the meantime. At present, search tools exist which can be enhanced and developed for use in electronic legal publishing. The key to the success of these tools is, again, coordination of effort among the various legal information providers.

The entire issue needs further discussion.

Software tools

Just as it is desirable to avoid duplication of effort in the construction of materials it is desirable to avoid duplication of effort in building software tools to handle and prepare them. FSR and Perl scripts, CGI scripts, and word-processing macros all have broad application. Again, the best course would be for one person at one institution to take responsibility for keeping current and coordinated on these matters, with the site perhaps offering Web and/or FTP retrieval of tools. As a notification channel for these purposes, the legal-webmasters listserv seems to be quite adequate for the present, though that will probably change as players outside law schools assume more importance.

What's on first?

Ultimately, electronic publication of legal material on the Net needs active participation by legal scholars and lawyers if it is to succeed. Those busy individuals will find it a much more desirable medium for authoring if they can apply their perspectives -- as represented by outlines and essays -- to core material such as statutes and regulations. In this context we might well substitute the phrase "link via hypertext" for the word "apply", and of course this assumes that those core materials are available to be linked to. While we realize (and have discussed above) the disadvantages of trying to prescribe what groups should work on, we do think it is possible to achieve some consensus about what should be done first. The thinking at Cornell is that materials which would appear in print as "statutory supplements" across broad stretches of the law-school curriculum are good candidates. This not the only possible perspective, and for various reasons (including the desire to showcase the expertise of particular faculty members) others might wish to proceed differently. Whatever the case, it seems we should construct an "it would be nice if..." list of starting materials soon.

Information-quality issues

Context

Certainly any conscientious information provider should endeavor to provide the most up-to-date version of a particular document which they practically can. More important, we need to arrive at a standard method by which readers may be told which version of a document they are reading, where it came from, what copyright restrictions may apply to it, who to send notice of errata, and so on. Legal information providers need also to bear in mind that they are dealing with an audience which is international and may be more or less expert in the law. The system in use at Cornell, in which "Context", "Structure", and "Copyright" documents are provided alongside (and linked to) all large collections of material is one we believe useful and deserving of examination by others.

Citation

Given the current state of affairs any mention of the citation issue is likely to provoke fireworks. Whatever the concerns of existing commercial publishers may be, it is clear that we need to work toward two goals. First, we need to provide a means of citing documents which have no existence anywhere but on the Net. The URL/URN scheme is a workable step in this direction, but it may lack sufficient granularity. To some extent this problem can be addressed within our community by the provision of "hanging" link targets (see above), but we may not expect that others outside that community will pay much attention, and more thought and discussion are needed.

Second, there is a need for a media-neutral citation scheme of the sort currently in use in the Lousiana courts and in the Sixth Circuit -- hopefully, one which would be compatible with those systems. We need to derive such a system and put it in place. The considerations surrounding the construction of such a system are complex and would take too long to elaborate fully here. Nonetheless, we view it as essential that such a system be constructed and be put in place as quickly as possible, hopefully within the year.

Maintenance

Conscientious maintenance of documents can't be enforced, but it is a matter of sufficient concern to all of us that we believe guidelines should be set out. Out-of-date, superseded information is worthless. On the other hand, maintenance of the information is problematic given the size of collections and the variability of funding and interest once the information is initially mounted. In general, the problem may be settled by the market -- people won't link to bad information. It may be that this "friends don't let friends go out of date" approach will be enough, but we ought to consider whether other means are needed and viable.

Where do we go from here?

Obviously this document will undergo extension and revision as the discussion continues; probably a good starting point for any discussion would be to think about what's been left out of this preliminary attempt. Insofar as this document represents an agenda for future efforts, we think it should be discussed and revised at CALI/LEAP. Have we covered everything which needs to be standardized or coordinated? Have we tried to standardize too much?

We would like to give the conclusions reached at CALI/LEAP a "reality check" based on the experience of authoring teams attending the Cornell workshop in late June. With those responses available to us, we ought to be able to construct a more comprehensive and accurate document. From there, we would suggest the construction of a listserv list or newsgroup specifically to deal with the construction of a bluebook.

Needless to say, your comments at any stage are welcome.

Tom Bruce (trb2@cornell.edu)
Peter Martin (martin@law.mail.cornell.edu)
Will Sadler (will@polecat.law.indiana.edu)

May 1994.