One-Year Report to the Center for the Public Domain: Open US Code

This is our second report to the Center for the Public Domain on the work supported by the Center's grant to the Legal Information Institute at Cornell – a grant that is enabling us to make quantum improvements to the core resource in our open US law collection, the US Code. As we estimated in the initial project budget and confirmed in our six-month report, this overall effort will in the end represent roughly a year and a half of programming. We are pleased to report that we are proceeding on schedule, and are by now extremely close to releasing the first fruits of this project to the public. Consequently, this might be characterized as our final report on the first phase of the US Code work. Consistent with our mid-project report, we still anticipate completion of the full system funded by the Center's grant within a few months.

Not only has the project progressed as we had hoped, it has opened up exciting opportunities that lead us to see its ultimate benefits as extending beyond the value of the product as a public resource and high profile example of open access to law. Until now our Institute's open law initiatives have largely influenced public bodies through example. The Supreme Court's official web site was prompted, in part, by the example of ours, as was the official site of the New York Court of Appeals. But while we offered technical assistance to the staffs of both courts, the offers were not accepted. The sophistication of this data project and the associated expertise show promise of opening up more direct exchange with those responsible for electronic dissemination of US legislation. This report outlines those developments and sketches our thoughts about how we might build on them. In short, as we look beyond this project we think we see important ways we can, with further Center support, leverage both its example and the know-how it has generated.

Timing and timelines.

At the time of our first report six months ago, we had just hired our first programmer with project funds, and had completed only the first of several steps in the process (see below).

This report finds us ten days away from the release of a finished, static HTML version of the Code and two weeks away from a presentation of our work for technical staff at the US House of Representatives. At the latter meeting, we expect to show our work to individuals working on XML-based legislation initiatives at the Senate, House of Representatives, National Archives and Records Administration, and the Library of Congress. We are approximately six weeks away from release of our "rigid" XML DTD, and three months from completion of a "point in time" version.

While it would have been nice to wait a few weeks and report that these important near-term successes had been achieved, we felt it wiser to report on schedule. We will follow up with a detailed report of the HTML release and House of Representatives meeting should the Center wish.

A brief recap of the technical process

Our report of six months ago sketched briefly the steps involved in reaching the point-in-time version that is our goal. It is worth repeating them, by way of orientation. The overall process involves:

1) Creating scripts that process the "raw ASCII" text of the Code, as we receive it from the House of Representatives, into "loose" XML – that is, XML that conforms to a DTD that flexibly marks up structure and some notable text features, but little else;
2) Assessing the accuracy of step 1 using a battery of small analytic and metrical programs, as well as using these tools to further reveal the deep structure of the Code. More is said about the importance of

this step below. This is both a deeper and more important step than we had initially believed, and is the basis for significant improvements in the quality of the product we are creating.
3) Using XSLT, create a static HTML version of the Code for public release.
4) Using information discovered in steps 1 and 2, design a "rigid" DTD that describes a full-featured, archival XML version. At the same time, design a SIM database schema that encompasses the DTD
5) Using XSLT, create the rigid XML version from the "loose" version of step 1.
6) Load the XML version into SIM and create suitable applets for serving.

It should be noted that perhaps 85% of the effort involved in the project in is the first three steps; we have essentially completed the first four. Document-analysis projects tend to be heavy in up-front effort; understanding the text and its structure is usually the hardest part, and this is certainly the case with the US Code. Too, the Code is uniquely problematic. A number of factors combine to make it a difficult challenge for electronic publishers:

- Sheer bulk and complexity. The raw ASCII text of title 42 alone runs approximately 45 Mb; the complete ASCII text consumes about 8 times that; and our current HTML version runs to approximately 102,000 HTML pages.
- Deceptive structure. While the division of the code into Titles (at the highest level of structure) and sections (at the lowest) is simple and straightforward, intermediate structural levels may be differently named and ordered from one title to the next.
- Inconsistent numbering and naming practices. Section numbers (e.g.) vary widely in format and composition, and may or may not contain numbers, dashes, alphabetic characters, and so on.
- Simple human errors and inconsistencies spread over a large collection. In perhaps the most notorious example, Title 17 contained two equally valid (and quite different) Section 512s for a year. Less dramatic examples include incorrect cross-references (about .4% of the total) and inconsistent subsection labeling and structuring practices.

These factors act in combination to create an environment in which (first) it is difficult if not impossible to create software that will correctly handle inconsistencies in the input, and (second) difficult to assess how well the software has performed. The size and complexity of the output is such that proofreading is impractical and errors are likely to be deeply buried.

As a result, we have devoted considerable time and effort to constructing small analytic scripts that help in probing the deep structure of the Code's text and in assessing how well our software has done in detecting and processing important features. Analytic information provided by these scripts has been invaluable in perfecting our conversion software. It also represents something of an advance over other efforts in the field; our analytic tools frequently provide us with answers sought by technical people at the House of Representatives (see below).

While it has consumed considerable "up-front" time, this careful approach has resulted in a version of the Code that is far more accurate in presentation and content than any we have published previously and (we believe) one that is superior to any other available on the Internet whether from an open or private source. In the HTML version that is to be released in two weeks, we provide the following unique features:

- fully-formatted tables, detected with 100% accuracy;
- fully-labelled, correctly indented, addressable (externally-linkable) substructure in each section of the Code, with 99.4% accuracy;
- linking of all explicit internal crossreferences, with 99.2% accuracy
- accurate formatting of notes sections
- superior typography and layout, controllable via XSLT and CSS stylesheets

Each of these in itself represents a considerable advance over what the LII has done in the past and goes far beyond what is available from any other public source. We estimate that the new version will attract some one million page views per week.

Going forward, the care taken in constructing the "loose" XML and HTML versions should serve us well. Because they are stable and well-constructed, the creation of the rigid XML version and its loading into a database should be simple. We anticipate successful completion by the end of February 2002, almost exactly one year after the project hired its first staff member. We will report in full at that time. Indeed,

should the Center board or staff be interested we would be pleased to demonstrate or present the work funded by this grant.

House of Representatives

About four months ago, we were contacted by technical staff at the US House of Representatives who are engaged in various XML-based legislation projects there. They had come across our early work on the "loose XML" version on the Web, were curious about our work, and eager to see what we had done. We arranged a meeting and demonstration for mid-October, which unfortunately coincided with the height of the Washington anthrax scare. The meeting was postponed to November 29th, 2001. At that time we will show our work to a joint technical committee from the House, Senate, Library of Congress, and National Archives and Records Administration.

In the meantime, informal technical exchange has sprung up between our project team and the organizers of the Washington meeting. It is clear from our interaction with the House staffers that there is a great need to share information, programming techniques, and best practices surrounding the publication of legislation. There appears to be a growing recognition that complementarity exists between what we do and what the House does. On the one hand, they have better access to people with substantive legislative expertise; on the other, we seem to have more experience and sophistication technically. The upcoming demonstration/workshop will provide an opportunity to discuss forming a community of practice around public-domain publication of legislation, something that we hope will proceed with the Legal Information Institute in a leadership role.

Further thoughts and efforts

In the absence of context and experience, it is hard to know how much weight to place on relationships like the one developing with the House staff. It is worth noting that, to our knowledge, exchange of the sort we have begun is unprecedented in the US. With few exceptions, government legal-publishing operations have obtained technical expertise through outsourcing or by simply abdicating responsibility to the private sector. The level of serious communication catalyzed by this project suggests some intriguing possibilities for partnership, collaboration, and further effort.

Courts, legislatures, agencies, and other creators of primary law have for some time been under pressure to make their work product available to the public via the Internet. For the most part, this has been a case of what one observer has called "technology by purchasing agent" – that is, each provider has sought a different set of off the shelf solutions. The result is a hodgepodge of systems that are all to some extent less transparent and useful to the public than they might be. We have already remarked that the government lacks experience with "open code" in the legal sense – they are used to seeking publishing expertise in the highly proprietized private sector. All this suggests opportunities for transformative change in the direction of a digital commons.

In the past, the LII has chosen to lead by example in a way that is largely implicit, much as we are with the US Code. Our collections stand as examples to others, and to a degree we have made technical information and expertise available that might be helpful to others in constructing similar things. But – in part because we are by nature doers rather than promoters -- we have been slow to issue white papers or detailed documentation that would help others to understand our techniques and to make use of them themselves. We could do much more in this vein, and it seems that it would fill a need that is being experienced by the staffs of public law-making bodies at the local, state, and federal levels.

We have begun to envision a place on the Web where implementors of public law dissemination systems could look for and exchange expertise, assistance, recommendations, and techniques. Such a site would consist in part of community-based knowledge-capture apparatus like archived discussion of techniques, and in part of publications intended to communicate and advance technical knowledge in the field. Our

aim, as always, would be to be practical rather than academic – the publications we have in mind would be closer in spirit to a FAQ or HowTo than a formal academic journal.

When this Center-funded US Code project is finished, we should like to document its methods. We now think about doing so in a way that would invite similar contributions from others and thus launch a forum for exchange. An important aim of this potential follow-on project would be simply to lower the threshold for government bodies taking responsibility for public access to their own output by identifying and explaining relevant nonproprietary, "open code" techniques.

We also contemplate continuing to develop the tools and know how made possible by this project by applying them to portions of the Legal Information Institute's on-line collection. Logical extensions include the Code of Federal Regulations or state legislation (New York, for example). We now think about making the choice, at least in part, based on material's potential for generating interest and involvement by those in the public sector.

No doubt it is premature to be speaking of a next grant proposal, but we do hope that as this first project nears completion the Center will share our enthusiasm for building on it in the ways suggested above. At this point, we merely want to open the subject and ask that the Center let us know when in terms of the likely completion of this work and its own grant-making schedule a follow-up proposal would be timely.