Testimony of
Thomas R. Bruce

Co-Founder and Director of the Legal Information Institute
A Research and Publication Activity of the Cornell Law School

before the
Committee on House Administration
Subcommittee on Oversight
United States House of Representatives

on
Modernizing Information Delivery in the House

June 16, 2011

Comments of the Cornell Legal Information Institute
before the

Committee on House Administration
Subcommittee on Oversight

June 16, 2011

Chairman Gingrey, Ranking Member Lofgren, and members of the Committee, thank you for inviting me to appear before you today.

My name is Tom Bruce, and I am the co-founder and Director of the Legal Information Institute, a research, engineering, and publishing activity of the Cornell Law School. In 1992, we were the first to make primary legal information available on the Web, and, in 2000, the first to create a version of the United States Code in XML. In collaboration with USGPO and the Office of the Federal Register, we have developed an innovative version of the Code of Federal Regulations, currently available in beta test. We have been engaged by the Library of Congress to develop functional data models for legislative data that are very different from those that currently underpin the familiar THOMAS and LIS systems.

The results of our work have been undramatic but pervasive. For example, we caused legal section and paragraph symbols to be incorporated into the basic symbol set for HTML in 1993 [i], and have worked on standard practices for Internet delivery of legal information ever since[ii]. We have consulted on legislative and judicial publishing and administration systems in 14 different countries on 4 continents, sometimes for government and sometimes for independent legal publishers, both noncommercial and commercial, including Thomson Reuters West Group and Lexis-Nexis. Recently, as part of work that we are doing in developing countries, we have become concerned with the effects of legal information policy on trade and on the climate affecting businesses large and small. Those effects are equally visible at home in the United States.

In our role as an Internet provider of primary legal information, we served more than 14 million unique individuals with over 71 million page views last year. Roughly 22 per cent of our referred traffic is sent to us by government web sites, notably the IRS. For the last few years, the IRS has widely distributed our version of Title 26 of the US Code for use by its tax-assistance programs. We are linked to by half a million web sites [iii].

Speaker Boehner and Majority Leader Cantor have already voiced support for new electronic data standards at the House, including (especially) the creation of documents in open, machine-readable formats [iv]. Today, I would like to say a little about the implications of that strategy and sketch the shape and size of its benefits. I will also urge you to consider some specific ways to make it happen. The manner of its implementation will strongly affect its usefulness to the Congress, and to the American people.

Goals

The use of open standards to create interoperable, accessible legislative information will benefit Congress and the American people in four ways:

1) Making the work of Congress easier

There are many inside government who could more skillfully identify ways to make the use of electronic documents within the House better and easier. Nevertheless, an outsider like myself can identify some compelling ways in which an XML-based, lifecycle-oriented legislative information system would make the work of the Congress faster and more efficient. Some of the tools I discuss here are already in use within offices and programs responsible for discrete portions of the legislative process, but to my knowledge none has been implemented over the entire lifecycle of legislation.

- *Smart word processing for legislation*
  XML, accompanied by data models that reflect legislative process, provides the foundation for a series of process aids and improvements that might best be described as "smart word processing for legislation". Typically, these are rich environments that provide functions and features helpful in legislative drafting. Many such improvements involve pulling data from a well-architected legislative data environment outside the document itself, such as automated incorporation of language from related or referenced statutes, automated construction of hyperlinked cross-references, and so on. [v]

- *Document management and status tracking*
  Many inside and outside government are interested in knowing what the law is, and in keeping track of the status of pending legislation. Independent, transparency-oriented operations like govtrack.us [vi] have done a commendable job of creating status-tracking applications by scraping data published via THOMAS and other Congressional web sites. While govtrack.us is a good job, it is not a perfect one -- nor can it be without bulk access to significantly better data created and published in bulk at the direction of the Congress. For example, the availability of timely legislative status information would enable a cascade of current-awareness services developed for many different niche markets, much as weather data from NOAA has been differentiated into a series of different weather forecasting products for different audiences.

- *Summarization and "dashboard views"*
  Smart tools are needed to provide overview and summary of Congressional actions. The Congressional Record Daily Digest currently serves this purpose, in print and in two online versions that have different capabilities. Like many non-digital products, the printed version is necessarily a compromise between depth and overload. It is isolated from the data sources it summarizes. The online version in GPO Access contains no links to the text of legislation under discussion; the version offered in THOMAS does, but neither links to information about the other people, places, and things it mentions [vii]. At the same time, it may not be concise enough for a truly high-level summary. Outsiders

have developed applications that very quickly summarize the mood or actions of Congress for a particular time period[viii]. As yet, these are toys, but they do show that there is a need to bring considerable ingenuity to the problem of accurate and timely summary of Congressional events and documents. Clearly, too, those summaries need to provide pathways to full and complete information. With time and better data, the same ingenuity that produced these prototypes could produce a wealth of helpful products.

2) Further reducing the cost of Congressional work

Those ideas realize savings by easing small, frequently-performed tasks within particular stages of the legislative process. A look at the whole legislative document lifecycle may reveal further efficiencies. The cost of moving bills and resolutions from stage to stage within the legislative lifecycle can be high [ix]. An XML-based system architected with the entire legislative lifecycle in mind would substantially reduce those costs, eliminating the need for repetitive reprinting and re-proofing at each stage of the process. There may well be other process savings that can be realized through careful consolidation and rethinking of the document management process as an integrated process taking place across the full lifecycle of a bill or resolution.

It is tempting, in this context, to try to maximize return on investment through rigid enforcement of centralized approaches and apparatus. Such an approach was tried, to a degree unsuccessfully, in the Federal e-rulemaking initiatives of the late 1990s [x]. No matter the source or force of standardization efforts, internal constituencies can and will remain intransigent in the face of centralization if they believe that it increases burdens and not benefits. The best approaches to centralization may, in fact, resemble the South Beach Diet: not the most effective diet science can imagine, but the most effective in practice if only because it is one that people will follow. With that in mind, the should be to maximize effective return on investment by creating standards and practices that respect careful analysis of use cases important to stakeholders, rather than mandating theoretical efficiencies that prove unsustainable. The result is likely to be a highly-connected federation of activities, linked by common standards and protocols, operating under the oversight of different administrative entities.

3) Making the work of Congress easier to find and understand

People use information retrieval systems by taking something they know -- a term or phrase -- and using it to find something they don't. Outsiders often have no idea where to begin. They don't know the particular terms of art used in legislation, and they understand little about how the process is organized and documented. A major design goal for government information systems should be to lower the threshold for information discovery as much as possible. That requires improvement in the systems offered to the public by Congress itself, and will be further realized through independent innovation and a vigorous market for products and services based on legislative data, including free-to-air offerings by parties outside government. The first goal would be served by a series of discrete improvements in THOMAS or by the construction of successor systems, and the second by the offer of legislative data in bulk, in XML.

Usually, we talk about this kind of informational threshold-lowering in terms of "transparency". That is often a code phrase for "public accountability". Transparency and accountability are excellent, important goals, as Speaker Boehner and Majority Leader Cantor have remarked [xi]. But "transparency" has another meaning: opening legislative data to a range of vital, concrete

information-seeking activities used for personal and professional purposes. Among those, the predictive value of legislative information for business planning looms large. For example, data about the legislative activity that creates and surrounds the tax code is as much a predictor for the business climate as the weather data provided by NOAA is for the climate itself -- and there is an equally broad interest in using its predictive value to plan strategies and activities [xii]. In that way, the primary legislative data provided by Congress meets a huge public need whose fulfillment stimulates and shapes business activity at all levels. That in turn creates a marketplace for information products and services where editorial and technical innovation can be rewarded.

4)Enabling technical communities inside and outside government to carry those aims further

There are a lot of products and services waiting to be created from legislative data. At this writing there are just under 19,000 different items in Amazon's catalog whose name uses the phrase "income tax". Most of them are printed books. In the pre-digital world, primary legal information provided the raw material for editorially-innovative products and services that repackage and explain legislative data for a huge range of audiences. Many represent particular professions, industries, or classes of private individuals. In the world of modern software applications, much less of this has happened -- yet. A search of Apple's app store for tax products shows 33 iPhone apps and about half that many for the iPad. Clearly, there are a lot of products and services waiting to be created.

A few have been. My own organization has, for more than a decade, created "mashups" of Federal data that help in legal research, primarily applications that facilitate movement across disparate collections of judicial opinions, statutes, and regulations, or provide current-awareness services. More recently, independent developers have built services like govtrack.us, which shows the current status of proposed Federal legislation, and created iPhone apps that offer primary materials like the US Code and the CFR. There is much, much more that can be done.

To see just how much, we should put aside popular, romantic visions of caffeinated high-tech hipsters building apps for mobile phones, and look instead at something solidly old-school and middle-class: TurboTax. TurboTax, and other tax-preparation aids like it, show what a mature software product built atop Federal law can do. Because it is well-designed and helpful, 20.7 million copies of TurboTax were purchased last tax season. The use of its Web-based version grew by 18%. It is a wildly successful product. TurboTax is also valuable to government. It serves as a funnel into IRS e-filing programs, which have allowed the IRS to close half of its tax service centers and realize other operational savings. How much of that does TurboTax account for? It is difficult to say with any accuracy, but an informed guess would be around 15%, given its market share, the number of taxpayers filing electronically, and what is known about user behavior [xiii]. Through follow-on effects, TurboTax saves a great deal of money for the government.

That is a dramatic success, generated by the impact of a series of complex statutory requirements on a mass market. It has been facilitated by active collaboration between government and private industry in establishing standards and data flows[xiv]. The result is an old-school "killer app". Those are rare.

But as the success of "app stores" for mobile platforms indicates, a marketplace of low-priced, narrowly-purposed applications can easily grow to match the most massive market for one-size-fits-all consumer software products. The availability of timely legislative data, delivered in XML designed for openness and interoperability, will form the basis for such a market in specialized, professional applications -- a market that will reward government with savings and efficiency as well as rewarding the innovators who create these new products.

What is needed to make this happen?

Cleaning and opening up the data

The data provided under any modernization initiative needs to be:

- *Compliant with open standards*
  Legislative data needs to be created and presented in open, interoperable, machine-readable formats with documented schemas and metadata models. In modern practice, XML is the preferred format for this. Page-description formats like PDF fail the test of machine-readability, as well as being far more difficult to work with.

- *Clean*
  Misformatted data is expensive to repair. When misformatting or data corruption occurs at the head of a value chain, the liability for repair is transmitted to every consumer of the data, resulting in duplicative, expensive effort [xv]. For that reason, government needs to ensure the quality of the data it issues, and to do so without introducing undue delay in transmission.

- *Consistent over time*
  Often, the success of a computer text-processing application depends on being able to detect and match patterns in the data itself. For instance, automatic conversion of cross-references into Web links relies on matching certain patterns of words and numbers that make up citations; extracting the names of parties from the header of a judicial opinion requires foreknowledge of the way that the text is arranged. Software built for such purposes inevitably makes assumptions about what it will encounter, and breaks when those assumptions are invalidated by changes in the format or arrangement of text [xvi]. For that reason, consistency and coherence in the format and arrangement of data greatly reduce the difficulty of writing and maintaining useful applications over time.

- *Timely*
  People need to know the current state of the law, but that is not all. Properly-built systems that make current law available can evolve, over time, into systems that provide legislative information extending into the future as well as into the past [xvii]. Such a point-in-time system -- one that makes it possible to know what the state of the law was at a particular time in the past, or what it will be at some point in the future when pending laws come into effect -- would be a very valuable tool.

  Right now, if you are outside government, it is very difficult even to work out what the current state of the law is. At this writing, the LII's US Code updating feature shows that

988 changes have been made to USC Title 26 since the last electronic release of a full Title update by the Office of the Law Revision Counsel[xviii]. That is in part because changes to the tax code are frequent, and in part because the public update-release practices of the Office of the Law Revision Counsel can combine with accidents of the calendar to leave the most recent official release of a given Title as much as 18 months out of sync with current legislation.

Most users with a need for timely information thus rely on more-or-less speculative codifications done by commercial publishers such as Lexis and Westlaw. To say that they are speculative is perhaps an exaggeration. Because most amending text refers directly to current legislation and relatively little is ever completely new, it is possible to guess very accurately how codification of particular provisions will be done. But it is, nevertheless, a guess -- one that is less likely to be accurate in new areas of the law or in places (eg. Title 6) where the Code reflects recent changes in the organization of government.

- *Clear as to provenance and authority*
  Government data should be authoritative and authentic. But -- as the above section on timeliness makes clear -- there are intervals when we need to know the text of a law, whether it is completely settled or fully in force or not. Thus, data about what the law says needs to be accompanied by data about where the text has come from and how authoritative it might be. That is well within reach of current practice in metadata modeling.[xix]

  The current debates about "authenticity" largely fail to account for this need for information about things not yet in full force, or in an indeterminate state. Many incorrectly bind the idea of "authenticity" to the use of specific document formats or encodings. In reality, it is possible to use a number of techniques to verify the status and accuracy of a particular piece of legal text. While the resemblance of page-description formats like PDF to printed text may comfort those who equate accuracy and authority with the fixity of print, there are many other ways to ensure that the text we are viewing is an accurate representation of the text issued by an official body. At least some of those techniques interfere far less with the useful qualities of digital text than PDF encoding does, and XML excels at facilitating processing and reuse.

- *Available in bulk*
  Bulk availability of legislative data is necessary for three reasons. Most collections of legal text are fairly useless unless they're comprehensive. Processing legal data is easier and more efficient in larger packages. Finally, significant numbers of applications are reduced in value (or flatly impossible to create) if the whole of a corpus is not available for concurrent processing. Certain kinds of finding aids that summarize information from across an entire corpus, such as a subject index, are good examples. Hard-won experience at the LII tells us that this is also true of automated quality-control and repair apparatus, which often relies on a survey of an entire corpus to detect and repair anomalies or markup problems in some portion of it [xx].

- *Available through well-documented access methods*
  Consistency and clarity are virtues not only for the data, but also for the means by which it is exposed to outside use. Well-documented application program interfaces (APIs) [xxi], document schemas, file-naming practices, metadata registries, identifier regimes[xxii], and access to the expertise of government specialists via blogs and other documentation of principles and best practices are essential to practical use of the data by outside parties. In this respect, Google's documentation of its APIs and its openness to building-out by outside developers are exemplary [xxiii]. Government should do these things as well.

Reaching these goals by implementing standards and creating partnerships

The House should encode, manage and promulgate its in-process and finalized legislative work products in ways that meet the above five goals for the data itself. Reaching that state, in turn, requires that it solve two problems.

The first is the creation of an appropriate, functional model or models for legislative data and metadata, embracing the entire legislative lifecycle in a considered and comprehensive way. The models should be specified as XML application profiles, and account for document structure and for relevant metadata expressed in RDF. That effort can usefully draw on several similar undertakings underway inside and outside Congress [xxiv]. It needs to be aimed at both the modernization of systems and workflows inside the House, and at the free provision of high-quality, open, interoperable bulk data to outside innovators and markets. The specifications for that project might best be created by an advisory group drawn from government, the technology and legal-publishing sectors, and the legal information science and engineering community.

The second need is for an appropriate framework in which to foster public-private partnerships designed to make use of such data. Remarkable things are possible when data is carefully leveraged to promote both efficiencies and services in an environment of collaboration between inside and outside stakeholders. Collaborative projects like the IRS e-filing system make the most sense when they are aimed at particular constituencies affected by defined categories of legislation. That implies that the best results will be achieved by chartering multiple small projects based on public-private partnerships. Development of a suitable framework for chartering such projects will be critical. The framework might itself be developed by a public-private collaboration similar to ETAAC at the IRS[xxv].

What about print?

The fate of printed versions under such a regime is uncertain. Some who wish to retain them will point out that there are many in the United States who do not have access to digital information via the Internet. That group of have-nots comprises about 23% of the population, and is heavily skewed toward the elderly and toward households with incomes under $30,000. [xxvi]

First, it is worth pointing out that digital files in XML can be readily expressed as print. The reverse is not true. It is possible to imagine a system in which print-on-demand facilities can make available as many copies as are needed, where they are needed, when they are needed. That would be better than what we have; the number of freely-distributed printed copies

mandated under the present system is simply too small to provide any kind of effective public access. There are still many valid reasons to distribute and archive printed copies[xxvii], but public access is probably not one.

There will be universal informational needs that it is in the public interest to meet comprehensively. Some are already being addressed through intermediaries who, in effect, relay information from the Internet to Internet-disadvantaged (or unaware) populations; others can be. The remainder, it seems to me, are best done in a targeted way. The IRS tax-assistance programs, which are coincidentally aimed at the same populations that are least well-served by the Internet, provide an example. And that suggests that the mechanism for identifying, prioritizing, and creating programs that meet specific needs of Internet-disadvantaged groups might well be the same as that needed to develop sensible data-publishing programs in the first place: targeted public-private collaborations of the sort I described earlier.

Conclusion

Creating clean, interoperable legislative data for bulk distribution to innovators and developers inside and outside government will significantly improve the efficiency and lower the cost of internal operations of the House. It will create new markets for legal information, and result in products and services that will benefit millions of Americans. It will have enormous predictive and practical value for American businesses of every size and shape. That will happen most quickly and efficiently if the effort is kicked off by a process of standards development, accompanied by the administrative innovation needed to effectively develop public-private collaborations around the use of legislative data.

Thank you for the opportunity to testify today. I look forward to your questions.

---

[i] See http://www.intercom.co.cr/www-archives/1993-q2/0194.html , note from Tim Berners-Lee memorializing the request.

[ii] For example, the URN:LEX standard for unique document identifiers. See http://tools.ietf.org/html/draft-spinosa-urn-lex-01

[iii] These statistics are taken from Google Analytics and Google Webmaster Tools for the www.law.cornell.edu site, from June 1 of 2010 to May 31 of 2011. They undercount by roughly 10 percent, as they do not include accesses to the Wex legal encyclopedia we provide at topics.law.cornell.edu.

[iv] Letter to the Honorable Karen Haas from Speaker John A Boehner and Majority Leader Eric Cantor (April 29, 2011), available at http://scr.bi/inig4d.

[v] A number of useful "wish lists" written by legislative drafters can be found on the Web, including one from Ed Hicks of Justice Canada (at http://www.opc.gov.au/calc/docs/Article_Hicks_UltimateLegislationSystem_2009.pdf) . XML.house.gov provides a list of such features already incorporated into House drafting systems (http://xml.house.gov/drafting.htm).

[vi] Govtrack.us is an independently developed system for tracking the status of federal legislation, and for searching the legislative corpus in innovative ways. It was developed by Joshua Tauberer, and can be found at http://www.govtrack.us/

[vii] While straightforward hyperlinking to other documentary sources is well understood, the connection of legislative data to real-world entities that are not documents on the Web (eg. for purposes of name-authority control) are more the province of newer Semantic Web technologies. Such an approach informs our current work for the Library of Congress.

[viii] See, eg., John Wonderlich's writeup of these apps at http://www.theopenhouseproject.com/2008/06/19/capitol-words/

[ix] During its design phase, I was told by an insider that the XML legislation system contemplated by Justice Canada had reduction of these inter-stage transfer costs as an explicit design goal, and that it was expected that those savings would cover the cost of the system. Unfortunately, I've been unable to find a post-mortem report assessing this claim.

[x] See generally "*Achieving the Potential: The Future of Federal e-Rulemaking, A Report to Congress and the President*", a report of the ABA Committee on the Status and Future of e-Rulemaking. Available online at http://ceri.law.cornell.edu/erm-comm.php .

[xi] Letter to the Honorable Karen Haas from Speaker John A Boehner and Majority Leader Eric Cantor (April 29, 2011), available at http://scr.bi/inig4d.

[xii] See generally Bruce, "*Some thoughts on the Constitution of Public Legal Information Providers*", originally published 2004 in the Journal of Information Law and Technology, available online at http://www.law.cornell.edu/working-papers/open/bruce/warwick.html . More recently, Robinson et al have addressed government web sites in their very influential paper "*Government Data and the Invisible Hand*", available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1138083

[xiii] The main reason for uncertainty is that some users of tax software continue to file by mail. That number is known to be declining, but an exact figure is hard to come by. A look at the 2010 report of the Electronic Tax Administration Advisory Committee (available at http://www.irs.gov/pub/irs-pdf/p3415.pdf) holds a few clues. Roughly 77% of taxpayers now e-file individual tax returns. TurboTax sales would equal about 19% of that total, but that figure should be discounted by whatever percentage of TurboTax users file manually. If a quarter of TurboTax users still file manually, then TurboTax accounts for about 15% of all e-filers.

[xiv] See, generally, the ETAAC report at http://www.irs.gov/pub/irs-pdf/p3415.pdf.

[xv] See, for example, Frug, "*Ground-up law: Source quality, access, and the CFR*", at http://www.hklii.hk/conference/paper/2B3.pdf

[xvi] For example, there have been 3 unsuccessful attempts made to create an external federated search apparatus for the United States Courts of Appeal -- two by us, and one by the now-defunct AltLaw site at Columbia Law School. All three were frustrated by shifting, ongoing inconsistencies in the labeling and organization of data by the 13 Circuit Courts, which among them use at least 7 different systems for file-naming alone. A successful attempt by Justia.com requires extensive manual maintenance by programming staff on an average of once every two weeks.

[xvii] One example of such a system, built in Australia, is described here: http://www.austlii.edu.au/austlii/research/2008/pit/ . Similar systems exist in Canada and Papua New Guinea among other places.

[xviii] The feature is created by mashing up data created by parsing the current Classification Tables published by the Office of the Law Revision Counsel and combining it with data taken from THOMAS. Parsing the Classification Tables is itself a task that would be made much easier by making them available in XML. They provide a very good example of something whose design is nicely optimized for human consumption in print, but can only laboriously be made machine-readable (the Parallel Table of Authorities and rules is another such). Too, one might question why there is no resource available to the public that fills the same need with respect to the US Code that the e-CFR does for the Code of Federal Regulations.

[xix] See, eg., Hillmann, Dushay, and Phipps, "*Improving Metadata Quality: Augmentation and Recombination*" [2004] at http://dcpapers.dublincore.org/ojs/pubs/article/viewArticle/770 for ideas about how this might be done, and why.

[xx] A typical example of such an approach would be the use of authority files to validate legal citations. The general idea is to survey the entire corpus to collect a list of referenceable documents, from which it is possible to assemble a canonical file of valid possible citations. Citations within the corpus can then be compared to the canonical file to determine validity. We use similar techniques to assemble a database of valid US Code section numbers, since these cannot be calculated according to any rational algorithm.

[xxi] See Wikipedia's explanation of APIs at http://en.wikipedia.org/wiki/Api . In general, APIs specify methods by which external programs may access data or methods implemented in software running independently.

[xxii] See http://tools.ietf.org/html/draft-spinosa-urn-lex-01

[xxiii] Somewhat self-referentially, a Google search on the terms "google API documentation" turns up a substantial number of useful hits.

[xxiv] A partial list of examples would include the work at xml.house.gov, the LII's work on legislative metadata modeling for the Library of Congress, and some of the work that has gone into FDSYS at the Government Printing Office, as well as exemplary efforts with legislation in the UK.

[xxv] ETAAC is described at http://www.irs.gov/efile/article/0,,id=136216,00.html . A look at the linked biographies of ETAAC members provides some idea of the scope of involvement by diverse industries, and a look at the ETAAC annual reports paints a picture of robust and focused collaboration.

[xxvi] These figures are drawn from the latest demographic data available from the Pew Trust Internet and American Life Project, available online at http://www.pewinternet.org/Static-Pages/Trend-Data/Whos-Online.aspx

[xxvii] For a concise summary of useful ideas on this point, see the Ithaka S+R study of the Federal Depository Library Program, at http://www.ithaka.org/ithaka-s-r/research/documents-for-a-digital-democracy