Designing Software to Shape Open Government Policy

Harlan Ming-Tun Yu

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Computer Science
Advisor: Edward W. Felten

September 2012

Abstract

Modern information technologies have transformed the meaning and promise of "open government." The term originally stood for the ideas of government transparency and public accountability. But with the rise of the Internet, "open government" has grown to encompass a wide range of civic goals—greater public participation and increased government efficiency, among others—newly enhanced by the potential of digital technologies. Software now plays a key mediating role between governments and citizens, and the design of software can both inform and shape the effectiveness of open government policies.

In this dissertation, we explore government's role as an information provider in the digital age. Rather than struggling, as it currently does, to keep up with the rapid pace of technological change, we contend that government should focus on enabling others to innovate, by publishing its data in bulk, machine-readable formats. This approach allows citizens to easily adapt government data for any desirable purpose using the latest technological tools, rather than relying on a single government-provided interface.

Despite its benefits, government may refuse to publish adaptable data for a variety of reasons. Such is the case with the U.S. Courts, who maintain a harmful paywall policy that limits access to electronic court records. We describe our pursuit to change the Courts' policies through the development of the RECAP browser extension, which we built to liberate records from the Courts' online access system. We analyze how RECAP's core design features contributed to its widespread adoption, and impacted the policy discourse. But even where government data are readily available, they may still be difficult to comprehend. We study how the U.S. Congress' age-old legislative process hinders the development of automated software with immense efficiency and transparency benefits. We outline the steps that Congress would need to take to modernize its process and embrace these improvements.

Finally, we discuss how recent "open government" policies have blurred the distinction between political and technological openness. We propose a clearer framing that separates the politics of public accountability from the technologies of open data, which we hope will make both ideals easier to achieve.

Acknowledgements

I would like to thank my advisor, Ed Felten, for relentlessly supporting my academic pursuits and allowing me to explore my intellectual interests. Over the years, he has always been (and, I am convinced, will always be) one step ahead of me no matter what I am doing, and has continually guided me in the right direction. He has given me opportunities for a lifetime.

At the Center for Information Technology Policy, I had the great fortune to have been surrounded by an incredibly talented and boisterous group of fellow researchers: Joe Calandrino, Will Clarkson, Ian Davey, Ari Feldman, Shirley Gaw, Josh Goldstein, Alex Halderman, Josh Kroll, Tim Lee, David Robinson and Steve Schultze. We have learned and grown together, both in research and in life, and I am lucky to consider them some of my closest friends. I especially wish to remember my dear friend, Bill Zeller, for his loyalty, his braveness, and his boundless curiosity and witty charm. I miss him deeply.

The wonderful camaraderie of the greater CITP community contributed to many interesting and fruitful research discussions. I am particularly thankful for my friendships with Deven Desai, Jens Grossklags, Joe Hall, Nadia Heninger, Nick Jones, Jen King, Ronaldo Lemos, Tom Lowenthal, Sajid Mehmood, Steven Roosa, and Wendy Seltzer. During my summers, I learned how to make myself useful in the real world, under the valuable guidance of Andrew McLaughlin, Beth Noveck, Dave Roberts, Derek Slater, Jason Waddle, and David Wagner.

I thank my committee—Ed, Beth, Jen Rexford, Mike Freedman, and Matt Salganik—for their thoughtful feedback and direction throughout the dissertation process. I'd also like to recognize helpful contributions to my research by Dhruv Kapadia, Carl Malamud, and the good folks at the Electronic Frontier Foundation and the Internet Archive. Thanks to Laura Cummings-Abdo and Melissa Lawson for making logistics a breeze.

I spent many fond years—from the very beginning of Graduate School—together with Tony Capra, Forrester Cole, Alex Golovinskiy, Janek Klawe, Ted Laird, and Haakon Ringberg. They are brothers to me, and more than anyone else at Princeton, they have positively shaped my views of the world. For that, I am forever indebted to them.

And most importantly, I would like to thank my family. Every success of mine has only been possible because of my parents' loving support and sacrifice. Whenever I think about where they and my grandparents started, and where we are now, I have no words to express how truly thankful I am for all of the opportunities they have given me. I also wouldn't be where I am today without my brother, Byron, who has always set the right example and paved the way for me. Last, but certainly not least, I thank Waiching Wong for believing in me and sharing her world with me.

For my parents.

Contents

Abstract . . . iii
Acknowledgements . . . iv
List of Figures . . . viii

1 Introduction **1**
1.1 Designing Software to Shape Policy . . . 2
1.2 Contributions and Roadmap . . . 3

2 Government Data and the Invisible Hand **5**
2.1 The Federal Internet Presence Before Data.gov . . . 6
2.2 Innovating for Civic Engagement . . . 12
2.2.1 Government Provides Data . . . 12
2.2.2 Private Parties Present Data to Citizens . . . 13
2.3 Practical Policy Considerations . . . 15
2.4 Alternatives and Counterarguments . . . 18
2.5 Conclusion . . . 20

3 RECAP: Turning PACER Around **23**
3.1 PACER: Public Access to Court Electronic Records . . . 25
3.2 Liberating Court Records . . . 31
3.3 The Design of RECAP . . . 35
3.3.1 Technical Design Overview . . . 36
3.3.2 Key Design Lessons . . . 39
3.4 Policy Challenges . . . 47
3.4.1 The Decline of Practical Obscurity . . . 47
3.4.2 Judicial Appropriations . . . 49
3.5 Conclusion . . . 50

4 Debugging the United States Code **51**
4.1 The U.S. Code and Positive Law . . . 56
4.2 Bugs in the U.S. Code . . . 61
4.3 Designing a Structured U.S. Code . . . 66
4.3.1 Benefits of a Structured Approach . . . 69
4.3.2 A Proof-of-Concept . . . 69
4.4 Practical Barriers to Implementation . . . 74
4.5 Conclusion . . . 76

5 The New Ambiguity of "Open Government" 77
5.1 Conceptual Origins 80
5.1.1 Conceptual Origins of Open Government 81
5.1.2 Conceptual Origins of Open Data 84
5.2 "Open Government" Meets "Open Data" 87
5.2.1 Early Roots of the Convergence 87
5.2.2 "Open Government" Becomes a Label for Both Technological Innovation and Political Accountability 90
5.2.3 Assessing the Merger 99
5.3 Our Proposal for a Clearer Framing 103
5.4 Conclusion 104

6 Conclusion 106
6.1 Future Work 107
6.2 Final Remarks 108

Selected Bibliography 109

List of Figures

3.1 High-level architecture of the RECAP system. 36

5.1 Conceptual framework separating the technologies of open data (vertical) from the politics of open government (horizontal). 80
5.2 Conceptual framework filled with several examples. 105

Chapter 1

Introduction

A popular government, without popular information, or the means of acquiring it, is but a prologue to a farce or a tragedy; or, perhaps both.

James Madison, 1822[1]

Modern information technologies have transformed the meaning and promise of "open government." When the term was first used in the 1950s, it was synonymous with the idea of "freedom of information" and government disclosure of politically sensitive information. Early advocates of the idea won the right to request and obtain paper copies of certain government records in the landmark Freedom of Information Act in 1966. Advocates around that time also campaigned for the right to oversee the government's decision-making process through "open" agency meetings. The open government policies of that era reflected the technologies available at the time—paper records and in-person deliberations.

But as information technologies have improved, citizens have raised their expectations about what open government entails. As in other areas of society, the commercial Internet has made communications between government and its citizens far cheaper, faster and more convenient than ever before. Citizens now expect an open government to publish all of its public data online and provide opportunities for the public to participate in key policy deliberations over the Internet. Governments, too, have increased their expectations: They see opportunities to use the Internet to tap into the "wisdom of the crowd" and collaborate with interested citizens to make policymaking more effective. The open government mantle is also used today to promote innovation and economic growth, and to drive initiatives to better inform consumer choices in the private marketplace.

While the policy goals attached to open government are diverse, the concept's allure is widespread and trending. In the past few years, governments around the world, and at all levels, have begun making broad open government promises to deliver

[1] Letter from James Madison to W.T. Barry (Aug. 4, 1822), *reprinted in* The Writings of James Madison (Gaillard Hunt ed.).

various tailored combinations of these goals. The hallmark of these new initiatives is their emphasis on modern technologies, and in particular, on the Internet and structured government data. Indeed, open government inherently pertains to the flow of information between government and its citizens, and technology can act as a critical amplifier, or suppressor, of information.

Governments may be well-meaning in their open government promises and their desire to adopt new technologies, but they often fall short of expectations. The quality of information flows depends heavily on government information policies, the adoption of new technologies, and internal staff culture to embrace new notions of information exchange. Many longstanding government information policies were hatched before the current generation of Internet technologies, using outdated assumptions about the capacity to process and communicate information. Remarkable advances in digital technologies over the past decade have radically changed these assumptions, and given the rapid pace of innovation, it is no surprise that many governments find it challenging to keep up: Information policies quickly fall out of date, recently purchased information systems are soon deemed legacy, and staff expertise about technology slowly lags behind the times. And at a time of widespread fiscal pressure, governments also find it difficult to adapt to the growing array of physical devices and social media that citizens are increasingly using to communicate.

1.1 Designing Software to Shape Policy

This dissertation explores how the design of software can shape—and ultimately improve—open government policies. We first suggest that governments in the Internet age should focus on publishing its data in bulk, structured formats, rather than building complicated interfaces for each passing technology. How difficult it is for government to publish structured data depends on the software that government uses to initially collect or create the data. When government does collect and publish data in a reusable way, government enables third-party stakeholders—like advocates, academics, journalists and others—to powerfully adapt its data in any way they see fit using the latest technologies, and to add value in unexpected ways. Third parties can use government data to experiment in parallel, in order to discover what innovations work best in different and changing technological environments.

However, governments may be unwilling to publish structured datasets for various reasons, even if the data are already published in other public mediums, like on a website or in print. In these cases, the traditional approach is to lobby the institution for changes in its publishing strategy. Alternatively, it may sometimes be possible to design and use software to create structured data from the outside, however painstaking the process. For example, software can be used to scan printed government tomes for easier online viewing, or new software can be developed to scrape certain information off government websites. As one case study, this dissertation examines the policies set forth by the U.S. Courts regarding the publication of federal court documents. We discuss how specially designed software can incrementally reassemble the collection of expensive-to-obtain electronic records, and make the collection far more accessible

to the public in an open online repository. But even in cases where the government does provide data in an easy-to-use format, the data may not be released in time for stakeholders to meaningfully respond, or may be too complicated for lay citizens to easily understand. In a second case study, we look at how laws are created by the U.S. Congress, and explain how certain changes to the legislative drafting and codification processes could lead to substantial efficiency and transparency gains. We demonstrate using software prototypes the potential benefits, and practical difficulties, of implementing such an approach.

In today's world, open government is inextricably linked to software, whether the goal is increasing political accountability, making service delivery more efficient, or promoting other public interests. Both within and outside of government, well-designed civic software needs to encapsulate a deep understanding of the human process that it intends to affect, in order steer the government toward specific open government goals. The design and use of software can inform policymakers about the range of policy choices that are feasible, and can demonstrate how governments could make more effective use of modern information technologies.

1.2 Contributions and Roadmap

The remainder of this dissertation is organized as follows. Chapter 2 proposes a counterintuitive policy idea: To better provide government information to citizens, governments should reduce their role building user-facing websites, and focus its publishing efforts on the release of bulk, machine-processable datasets. We argue that private parties are better-suited than governments to develop user-facing civic interfaces, and that governments should publish reusable data that simply enables outside innovation. While our discussion centers on the U.S. federal executive branch, the policy proposal generalizes to any governmental entity. This work originally appeared as an article in the Yale Journal of Law and Technology in 2009, and is joint work with David G. Robinson, William P. Zeller and Edward W. Felten.[2] Our original article has been widely cited in the open government literature, and its core idea has been implemented extensively by governments around the world.

Chapter 3 presents the RECAP system, whose goal is to make federal court records more publicly available, and to positively influence the U.S. Courts' public access policies. We built the RECAP system to liberate federal court records from the Courts' paywalled online access service called PACER. The system uses a Firefox browser extension to crowdsource the purchase of federal court records and combine them in RECAP's public repository. The repository now contains more than 2.7 million federal court records from 640,000 cases, and the extension is used by thousands of PACER users. We discuss the key design decisions that contributed to RECAP's

---

[2] David Robinson, Harlan Yu, William P. Zeller & Edward W. Felten, *Government Data and the Invisible Hand*, 11 Yale J. L. & Tech. 160 (2009). An early draft of the article first appeared online in May 2008, *available at* http://ssrn.com/abstract=1138083.

widespread adoption and use, as well as the policy barriers that impede successful changes in the Courts' policies. The RECAP project is joint work with Stephen Schultze, Timothy B. Lee and Edward W. Felten. Portions of this chapter were originally published in ACM XRDS in 2012, co-authored with Stephen Schultze.[3]

Chapter 4 examines the transformative potential of digital technologies in the U.S. Congress. We delve into the many peculiarities of the legislative process, which eventually outputs the massive—and buggy—consolidation of federal laws called the U.S. Code. We analyze the U.S. Code from the perspective of software development, and find that it is ultimately too unstructured for robust civic technologies. We reimagine the activities of Congress as a more structured process, and lay out a future where Congressional activities are far more efficient and transparent than they are today. While this future may be a long way off, we discuss how Congress would need to change its process to make itself amenable to a radically modern system.

Chapter 5 clarifies the current discourse about the purpose of "open government," which has increasingly blurred the technologies of open data with the politics of public accountability. We trace the term to the 1950s, when it referred only to politically sensitive government disclosures, and we explore its recent convergence with the Internet-era "open data" movement. We argue that the prefix "open" is deeply ambiguous, especially in the context of "open government data," and we propose new terminology to clearly distinguish between the technological means and the political ends of open government. This work originally appeared as an article in the UCLA Law Review Discourse in 2012.[4]

Chapter 6 concludes and discusses future work for computer scientists in the ripe—and increasingly familiar—field of open government policy.

---

[3] Harlan Yu & Stephen Schultze, *Using Software to Liberate U.S. Case Law*, 18 ACM XRDS 12 (2011), *available at* http://doi.acm.org/10.1145/2043236.2043244.

[4] Harlan Yu & David G. Robinson, *The New Ambiguity of "Open Government"*, 59 UCLA L. Rev. Disc. 178 (2012).

Chapter 2

Government Data and the Invisible Hand

In a word, let every sluice of knowledge be opened and set a-flowing.

John Adams, 1765[1]

If the federal government really wants to embrace the potential of Internet-enabled transparency, it should follow a counter-intuitive but ultimately compelling strategy: *Reduce* the federal role in presenting important government information to citizens. Today, government bodies consider their own websites to be a higher priority than technical infrastructures that open up their data for others to use. We argue that this understanding is a mistake. It would be preferable for government to understand providing reusable data, rather than providing websites, as the core of its online publishing responsibility. This core policy argument provides a backdrop for the technical work in this dissertation.

This Chapter examines the wide gap between the exciting uses of Internet technologies by private parties, on the one hand, and the government's lagging technical infrastructure on the other. Citizens today use an ever-changing variety of different devices to access information online, from traditional desktop environments, to an array of new platforms for mobile phones, tablets and electronic book readers. As the private sector continues to create more, increasingly powerful devices and platforms, governments are struggling to keep pace with the rapidly evolving nature of Internet technologies. Federal government webmasters are hindered by a minefield of compliance rules, while the aggregate cost of developing—and then maintaining—software for each new platform quickly becomes unsustainable, particularly in today's austere fiscal landscape.

In order for public data to benefit from the same innovation and dynamism that characterize private parties' use of the Internet, the federal government must reimagine its role as an information provider. Rather than struggling, as it currently does,

[1] John Adams, *Dissertation on the Canon and the Feudal Law* (1765), *reprinted in* The Works of John Adams, Second President of the United States 463 (1865) (Charles Francis Adams ed.).

to design websites that meet each end-user need, it should focus on creating a simple, reliable and publicly accessible infrastructure that *exposes* the underlying data. Private actors, either nonprofit or commercial, are better suited to deliver government information to citizens and can constantly create and reshape the tools individuals use to find and leverage public data. The best way to ensure that the government allows private parties to compete on equal terms in the provision of government data is to require that federal websites themselves use the same open systems for accessing the underlying data as they make available to the public at large.

Our approach follows the engineering principle of separating data from interaction, which is commonly used in constructing websites.[2] Government must provide data, but we argue that websites that provide interactive access for the public can best be built by private parties. This approach is especially important given recent advances in interaction, which go far beyond merely offering data for viewing, to providing services such as advanced search, automated content analysis, cross-indexing with other data sources, and data visualization tools. These tools are promising but it is far from obvious how best to combine them to maximize the public value of government data. Given this uncertainty, the best policy is not to hope government will choose the one best way, but to rely on private parties in a vibrant marketplace of engineering ideas to discover what works.

2.1 The Federal Internet Presence Before Data.gov

The Internet's transformative political potential has been clear to astute nontechnical observers since at least the mid-1990s, but progress toward that transformation has been sporadic at best. In January of 1995, when the Republicans regained a Congressional majority, they launched THOMAS, a website that details every bill in Congress.[3] But by 2004, the site was so out of date that seven Senators cosponsored a resolution to urge the Library of Congress to modernize it.[4]

The Federal Communications Commission—the agency most closely involved in overseeing digital communications—had a website whose basic structure had remained

---

[2] Most sophisticated websites use separate software programs for data and interaction, for example storing data in a database such as MySQL, while interacting with the user via a web server such as Apache. Many government websites already use such a separation internally. Government sites that currently separate these functions are already partway to the goal we espouse.

[3] *About THOMAS*, Libr. Congress, http://thomas.loc.gov/home/abt_thom.html (last visited June 30, 2012).

[4] S. Res. 360, 108th Cong. (2004) ("A resolution expressing the sense of the Senate that legislative information shall be publicly available through the Internet.").

unchanged for a decade, before 2011.[5] Regular users of the system reported that in order to obtain useful information, they had to already know the docket number for the proceeding in which they are interested.[6] Materials could be searched by a few criteria such as the date of submission or name of the submitting attorney, but the site did not allow users to search the actual content of comments and filings even when these filings have been submitted to the agency in a computer-searchable file format.[7] Even Google, which was severely handicapped by its lack of access to the agency's internal databases, did a significantly better job of identifying relevant information.[8]

Federal webmasters are eager to embrace the Internet's full potential, and in some cases, they have been remarkably successful in the context of their challenging environment. Compared to technologists in the private sector, federal webmasters face a daunting array of additional challenges and requirements. An online compliance checklist for designers of federal websites identifies no fewer than twenty-two different regulatory regimes with which all public federal websites must comply.[9] Ranging from privacy and usability to FOIA compliance to the demands of the Paperwork Reduction Act and, separately, the Government Paperwork Elimination Act, each of these requirements alone is, considered on its own, a thoughtfully justified federal mandate. Each one reflects the considered judgment of our political process, informed by the

---

[5] *Compare Federal Communications Commission (FCC) Home Page*, INTERNET ARCHIVE (Sept. 17, 2001), http://web.archive.org/web/20010917033924/http://www.fcc.gov/, *with Federal Communications Commission (FCC) Home Page*, INTERNET ARCHIVE (Mar. 17, 2011), http://web.archive.org/web/20110317192731/http://www.fcc.gov/. The FCC launched a major redesign of its website in 2011, adding significant improvements to its usability and functionality. *See* Alex Howard, *FCC.gov reboots as an open government platform*, O'REILLY RADAR (Apr. 5, 2011), http://radar.oreilly.com/2011/04/fcc-website-reboot-open-source-cloud.html.

[6] *See* Jerry Brito, *FCC.gov: The Docket that Doesn't Exist*, TECH. LIBERATION FRONT (Nov. 1, 2007), http://techliberation.com/2007/11/01/fccgov-the-docket-that-doesnt-exist/; *see also* Cynthia Brumfield, *The FCC is the Worst Communicator in Washington*, IP DEMOCRACY (Sept. 5, 2007, 9:17 AM), http://www.ipdemocracy.com/archives/002640the_fcc_is_the_worst_communicator_in_washington.php.

[7] Jerry Brito, *Hack, Mash & Peer: Crowdsourcing Government Transparency*, 9 COLUM. SCI. & TECH. L. REV. 119, 123-25 (2007), *available at* http://www.stlr.org/html/volume9/brito.pdf.

[8] Jerry Brito, *FCC.gov: Searching in Vain*, TECH. LIBERATION FRONT (Oct. 29, 2007), http://techliberation.com/2007/10/29/fccgov-searching-in-vain.

[9] The checklist also includes an additional thirteen "best practices" for building a government website. *Requirements and Best Practices Checklist*, FEDERAL WEB MANAGERS COUNCIL, http://www.howto.gov/web-content/requirements-and-best-practices/checklists/long (last visited June 30, 2012).

understanding of information technology that was available when it was written. But the cumulative effect of these requirements, taken together, is to place federal web designers in a compliance minefield that makes it hard for them to avoid breaking the rules—while diverting energy from innovation into compliance.[10] The stultifying compliance climate is an undesirable side effect, not a choice Americans endorsed through our political process.[11] Indeed, there is no guarantee that these requirements interact in such a way as to make total compliance with all of them possible, even in principle.[12]

These problems attend any individual federal website; a second layer of challenges can emerge when the federal government seeks to impose coordination or consistency across the remarkably broad range of rulemaking processes and data. This happened with Regulations.gov, a government-wide docket publishing system created in response to the E-Government Act of 2002 and launched in 2003. It is used today by the vast majority of federal agencies[13]—in fact, the policy of the Office of Management and Budget (OMB) not only requires its use but also precludes the agencies from using "ancillary and duplicative" docketing and rulemaking systems of their own design.[14] This exclusivity rule, combined with the difficult interagency politics involved in honing system features, have led to a bare-bones approach that leaves out

---

[10] In contrast, private developers are not bound by the same compliance rules. However, they could voluntarily determine that certain government practices—like "Section 508" requirements to make content accessible to individuals with disabilities—also make sense in the private setting.

[11] For example, several different requirements that were developed independently of one another require certain content to be included on homepages. Overall, these rules prevent certain kinds of simple, intuitive interfaces that might in fact be desirable. Our proposal, by reducing the importance of homepages, helps resolve this issue. By making all data available and allowing non-governmental actors to structure interactions around their own aims, information technology professionals can avoid the problem of being mandated to clutter their homepages with boilerplate disclosures.

[12] And compliance is, in any case, a difficult practical challenge. One survey found that only 21% of federal agencies post on the Web all four types of FOIA data required under the 1996 Electronic Freedom of Information Act Amendments. *See* Kristin Adair et al., *File Not Found: Ten Years After E-FOIA, Most Federal Agencies Are Delinquent*, 2007 Nat'l Security Archive 7, *available at* http://www.gwu.edu/~nsarchiv/NSAEBB/NSAEBB216/index.htm.

[13] *About Us—Partner Agencies*, Regulations.gov, http://www.regulations.gov/\#!aboutPartners (last visited June 30, 2012).

[14] *See* Office of Mgmt. & Budget, Executive Office of the President, Expanding E-Government: Partnering for a Results Oriented Government 4 (2004), *available at* http://www.whitehouse.gov/sites/default/files/omb/assets/omb/budintegration/expanding_egov12-2004.pdf.

the agency-tailored functionality found in many of the systems it replaced. Concerns about cost-sharing have also led the system to omit even features whose usefulness and desirability is a matter of broad consensus.[15]

Regulations.gov was launched with a limited search engine and no browsing capability, so that only those who already knew the terms of art used to categorize rulemaking documents were able to use it effectively.[16] Five years later, a re-launched version of the site offered up its limited inventory of computer-readable data directly to the public (in this case, using a single RSS feed) which allowed any interested person or group to create an alternative, enhanced version of the website.[17] This has permitted the creation of OpenRegs.com, which competes with Regulations.gov by offering "an easy-to-navigate regulatory portal" with "features not available anywhere else," like a more sensible set of RSS feeds, one for each individual agency.[18]

However, because the engine behind Regulations.gov gathers and integrates only very basic information about the many documents it displays—such as a title, unique identifier, and author name—the decision to share this information with the public can offer only limited benefits. Most of the information relevant to the rulemaking process remains locked away in computer files that are images of printed documents, which cannot be easily reused. A 2008 report sponsored by the American Bar Association concluded that Regulations.gov "continues to reflect an 'insider' perspective"[19] and lacks a comprehensive, full-text search engine over all regulatory data.[20] The same report also emphasized that individual executive branch entities such as the Environmental Protection Agency and the Department of Transportation have been forced to close down their own more advanced systems, which offered deeper insight into docket materials, in order to comply with the prohibition on redundancy.[21] A congressional panel was similarly critical, finding that "[m]any aspects of this initiative are fundamentally flawed, contradict underlying program statutory requirements

---

[15] Our discussion of Regulations.gov draws heavily on a report by the ABA-chartered Committee on the Status and Future of e-Rulemaking. *See* Cynthia Farina et al., *Achieving the Potential: The Future of Federal e-Rulemaking*, Sec. Admin. L. & Reg. Prac. Am. Bar Ass'n 1 (2008), *available at* http://ceri.law.cornell.edu/erm-comm.php.

[16] *Feds Open Portal for Online Comments on Regulations*, 9 CDT Pol'y Post 3 (Jan. 23, 2003), *available at* http://www.policyarchive.org/handle/10207/bitstreams/2159.pdf.

[17] *See* Heather West, *Regulations.gov Unleashes Wealth of Information for Users*, Center for Democracy & Tech. Blog (Jan. 15, 2008), https://www.cdt.org/blogs/heather-west/regulationsgov-unleashes-wealth-information-users.

[18] *About*, OpenRegs.com, http://openregs.com/about (last visited June 30, 2012).

[19] Farina et al., *supra* note 15, at 20.

[20] Farina et al., *supra* note 15, at 30.

[21] Farina et al., *supra* note 15.

and have stifled innovation by forcing conformity to an arbitrary government standard."[22]

There are a number of potential ways to improve Regulations.gov. These include changing the funding model so that government users will not face higher costs if they encourage their stakeholders to make more extensive use of the system and streamlining the decision making process for new features. If the ban on ancillary agency systems were also relaxed, the focus on structured, machine-readable data that we suggest here could be used to explore new functionality while still continuing to contribute documents to the existing Regulations.gov infrastructure.[23]

The tradeoff between standardization and experimentation, and the concerns about incomplete or inaccurate data in centralized government repositories such as Regulations.gov, are inherently difficult problems. USASpending.gov, created by legislation co-sponsored by Barack Obama and Tom Coburn in 2006,[24] presents another example: There, the desire to increase data quality by adopting a uniform method of identifying the recipients of federal funds has led to proposed amendments to the original legislation, aimed at improving data accuracy and standardization across agencies.[25] It is encouraging to see legislators take note of these intricate but significant details.

As long as government has a special role in the presentation and formatting of raw government data, certain desirable limits on what the government can do become undesirable limits on how the data can be presented or handled. The interagency group that sets guidelines for federal webmasters, for example, tells webmasters to manually check the status of every outbound link destination on their websites at least once each quarter.[26] And First Amendment considerations could complicate, if not outright prevent, some efforts to moderate online fora related to government documents. Considerations like these tend to make wikis, discussion boards, group annotation, and other important possibilities impracticable for government websites themselves.

Meanwhile, private actors have demonstrated a remarkably strong desire and ability to make government data more available and useful for citizens—often by going to great lengths to reassemble data that government bodies already possess but are not sharing in a machine-readable form. GovTrack.us integrates information about

---

[22] H.R. Rep. No. 109-153, at 138 (2006).

[23] *See* Farina et al., *supra* note 15 (detailing specific steps toward a better Regulations.gov).

[24] Federal Funding Accountability and Transparency Act, Pub. L. No. 109-282, 120 Stat. 1186 (2006).

[25] Strengthening Transparency and Accountability in Federal Spending Act, S. 3077, 110th Cong. (2008).

[26] *Establish a Linking Policy*, Web Managers Advisory Council, http://www.howto.gov/web-content/manage/categorize/links/linking-policy (last visited June 30, 2012).

bill text, floor speeches and votes for both houses of Congress by painstakingly reprocessing tens of thousands of webpages.[27] It was created by a then-graduate student in linguistics in his spare time.[28] Carl Malamud, an independent activist, painstakingly took the SEC's data online,[29] while the RECAP project (as described later in §3) is now attempting to open up judicial records, which are currently housed behind a government paywall.[30]

In some cases and to some degree, government bodies have responded to these efforts by increasing the transparency of their data. Key congressional leaders have expressed support for making their votes more easily available,[31] and the SEC has adopted a structured format called XBRL that increases the transparency of its own data.[32] In 2004, the OMB even asked that government units "to the extent practicable and necessary to achieve intended purposes, provide all data in an open, industry standard format permitting users to aggregate, disaggregate, or otherwise manipulate and analyze the data to meet their needs."[33] We argue below for a stronger impetus to provide open data: not "to the extent . . . necessary to achieve intended purposes," but as the main intended purpose of an agency's online publishing.

The federal government's current steps toward reusable data are valuable and admirable. But these efforts are still seen and prioritized as afterthoughts to the finished sites. As long as government bodies prioritize their own websites over infrastructures that will open up their data, the pace of change will be retarded.

---

[27] GovTrack.us: Tracking the U.S. Congress, http://www.govtrack.us (last visited June 30, 2012).

[28] Joshua Tauberer, *Case Study: GovTrack.us*, *in* Open Government: Collaboration, Transparency, and Participation in Practice, 201, 212 (Daniel Lathrop & Laurel Ruma eds., 2010).

[29] Posting of Taxpayer Assets, tap+essential.org, to listserver+essential.org, SEC's EDGAR on Net, What Happened and Why (Nov. 30, 1993, 10:36:34 EST), *available at* http://w2.eff.org/Activism/edgar_grant.announce.

[30] Previous efforts to open up judicial records were led by Malamud. *See* John Markoff, *A Quest to Get More Court Rulings Online, and Free*, N.Y. Times, Aug. 20, 2007, https://www.nytimes.com/2007/08/20/technology/20westlaw.html.

[31] *Open House Project Calls for New Era of Access*, OMB Watch (May 15, 2007), http://www.ombwatch.org/node/3287.

[32] *See generally XBRL*, U.S. Sec. & Exch. Comm'n, http://xbrl.sec.gov (last visited June 30, 2012).

[33] Clay Johnson III, Exec. Office Of The President, Memorandum No. M-05-04, Policies For Federal Agency Public Websites 4 (2004), *available at* http://www.whitehouse.gov/omb/memoranda/fy2005/m05-04.pdf.

2.2 Innovating for Civic Engagement

Our goal is to reach a state where government provides all of its public data online[34] and there is vigorous third party activity to help citizens interact and add value to that data. Government need not—and should not—designate or choose particular parties to provide interaction. Instead, government should make data available to anyone who wants it, and allow innovative private developers to compete for their audiences.

2.2.1 Government Provides Data

Government should provide data in the form that best enables robust and diverse third party use. Data should be available, for free, over the Internet in open, structured, machine-readable formats to anyone who wants to use it. Using "structured formats" such as XML makes it easy for any third party service to gather and parse this data at minimal cost.[35] Internet delivery using standard protocols such as HTTP provides immediate real-time access to this data to developers. Each piece of government data, such as a document in XML format, should be uniquely addressable on the Internet in a known, permanent location.[36] This permanent address allows both third party services, as well as ordinary citizens, to link back to the primary unmodified data source as provided by the government.[37] All public data, in the highest detail available, should be provided in this format in a timely manner. As new resources are made available, government should provide data feeds, using open protocols such as

---

[34] Freedom of Information Act, 5 U.S.C. §552 (2002), *as amended by* Electronic Freedom of Information Act, Pub. L. No. 104-231, 110 Stat. 3048 (1996).

[35] To the extent that nontrivial decisions must be made about which formats to use, which XML schemas to use, and so on, government can convene public meetings or discussions to guide these decisions. In these discussions, government should defer to the reasonable consensus view of private site developers about which formats and practices will best enable development of innovative sites.

[36] Using the usual terms of art, the architectural design for data delivery must be RESTful. REST (short for Representational State Transfer) defines a set of principles that strives for increased scalability, generality, and data independence. The REST model adopts a stateless and layered client-server architecture with a uniform interface among resources. *See* Roy Thomas Fielding, Architectural Styles and the Design of Network-based Software Architectures (2000) (unpublished Ph.D. dissertation, University of California, Irvine), *available at* http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm.

[37] Concerns about data integrity—for example, possible modification by an intermediate service—can be addressed by using digital signatures. The originating Department or Agency can sign each primary source in such a way that data is verifiable and modification by an intermediary can be detected by the data recipient.

RSS, to notify the public about the additions. These principles are consistent with the Open Government Working Group's list of eight desirable properties for government data.[38]

In an environment with structured data, the politics of what to put on a home page are avoided, or made less important, because the home page itself matters less. And technical staff in government, whose hard work makes the provision of underlying data possible, will have the satisfaction of seeing their data used widely—rather than lamenting interfaces that can sometimes end up hiding valuable information from citizens.

2.2.2 Private Parties Present Data to Citizens

The biggest advantage of third party data processing is to encourage the emergence of more advanced features, beyond simple delivery of data. Examples of such features include

- *Advanced search*: The best search facilities go beyond simple text matching to support features such as multidimensional searches, searches based on complex and/or logical queries, and searches for ranges of dates or other values. They may account for synonyms or other equivalences among data items, or suggest ways to refine or improve the search query, as some of the leading web search services already do.

- *RSS feeds*: RSS, which stands for "Really Simple Syndication," is a simple technology for notifying users of events and changes, such as the creation of a new item or an agency action. The best systems could adapt the government's own feeds (or other offerings) of raw data to offer more specialized RSS feeds for individual data items, for new items in a particular topic or department, for replies to a certain comment, and so on. Users can subscribe to any desired feeds, using RSS reader software, and those feeds will be delivered automatically to the user. The set of feeds that can be offered is limited only by users' taste for tailored notification services.

- *Links to information sources*: Government data, especially data about government actions and processes, often triggers news coverage and active discussion online. An information service can accompany government data with links to, or excerpts from, these outside sources to give readers context into the data and reactions to it.

---

[38] The group identified that government data must be complete, primary, timely, accessible, able to be processed by machines, non-discriminatory, non-proprietary and license-free. *See 8 Principles of Open Government Data*, Open Government Working Group, http://www.opengovdata.org/home/8principles (last visited June 30, 2012).

- *Mashups with other data sources*: To put an agency's data in context, a site might combine that data with other agencies' data or with outside sources. For example, MAPlight.org combines the voting records of members of Congress with information about campaign donations to those members.[39] Similarly, the nonprofit group Pro Publica offers a map showing the locations of financial institutions that have received funds from the Treasury Department's Troubled Asset Relief Program (TARP).[40]

- *Social media, discussion fora and wikis*: A site that provides data is a natural location for discussion and user-generated information about that data; this offers one-stop shopping for sophisticated users and helps novices put data in context. Such services often require a human moderator to erase off-topic and spam messages and to enforce civility. The First Amendment may make it difficult for government to perform this moderation function, but private sites face no such problem, and competition among sites can deter biased moderation. Private sites can also more easily integrate unfiltered social media content, from existing platform APIs like Facebook[41] and Twitter.[42]

- *Visualization*: Often, large data sets are best understood by using sophisticated visualization tools to find patterns in the data. Sites might offer users carefully selected images to convey these patterns, or they might let the user control the visualization tool to choose exactly which data to display and how.[43] Visualization is an active field of research and no one method is obviously best; presumably sites would experiment with different approaches.

- *Automated content and topic analysis*: Machine-learning algorithms can often analyze a body of data and infer rules for classifying and grouping data items.[44]

---

[39] MAPLIGHT.ORG, http://www.maplight.org (last visited June 30, 2012).

[40] *Map: Show Me the TARP Money*, PRO PUBLICA, http://www.propublica.org/special/bailout-map (last visited June 30, 2012).

[41] *Facebook Developers*, FACEBOOK, https://developers.facebook.com (last visited June 30, 2012).

[42] *Twitter Developers*, TWITTER, https://dev.twitter.com (last visited June 30, 2012).

[43] "Many Eyes," for example, makes it simple for non-experts to dynamically visualize any custom dataset in a variety of different styles. *See* MANY EYES, http://www-958.ibm.com/software/data/cognos/manyeyes/ (last visited June 30, 2012).

[44] For example, software developed by Blei and Lafferty computed a topic model and classification of the contents of the journal Science since 1880. *See* David M. Blei & John D. Lafferty, *A Correlated Topic Model of Science*, 1 ANNALS APPLIED STAT. 17 (2007).

By automating the classification of data, such models can aid search and foster analysis of trends.

- *Collaborative filtering and crowdsourced analysis*: Another approach to filtering and classification is to leverage users' activities. By asking each user to classify a small amount of data, or by inferring information from users' activities on the site (such as which items a user clicks), a site might be able to classify or organize a large data set without requiring much work from any one user.

Exactly which of these features to use in which case, and how to combine advanced features with data presentation, is an open question. Private parties might not get it right the first time, but we believe they will explore more approaches and will recover more rapidly than government will from the inevitable missteps. This collective learning process, along with the improvement it creates, is the key advantage of our approach. Nobody knows what is best, so we should let people try different offerings and see which ones win out.

For those desiring to build interactive sites, the barriers to entry are remarkably low once government data is conveniently available. Web hosting is cheap, software building blocks are often free and open source,[45] and new sites can iterate their designs rapidly. Successes thus far, including the GovTrack.us site that Joshua Tauberer built in his spare time,[46] show that significant resources are not required to enter this space. If our policy recommendations are followed, the cost of entry will be even lower.

2.3 Practical Policy Considerations

Our proposal is simple: The federal government should specify that its primary objective as an online publisher is to provide data that is easy for others to reuse, rather than to help citizens use the data in one particular way or another.

The policy route to realizing this principle is to require that federal government websites retrieve their published data using the same infrastructure that they have made available to the public. Such a rule incentivizes government bodies to keep this infrastructure in good working order, and ensures that private parties will have no less an opportunity to use public data than the government itself does. The rule prevents the situation, sadly typical of government websites today, in which governmental interest in presenting data in a particular fashion distracts from, and thereby impedes, the provision of data to users for their own purposes.

Private actors have repeatedly demonstrated that they are willing and able to build useful new tools and services on top of government data, even if—as in the

---

[45] For example, the "LAMP stack," consisting of the Linux operating system, the Apache web server, the MySQL database software, and the PHP scripting language, are available for free and widely used.

[46] *See* Tauberer, *supra* note 28.

case of Joshua Tauberer's Govtrack.us[47], Carl Malamud's SEC initiative[48], or court records projects like RECAP[49]—they have to do a great deal of work to reverse engineer and recover the structured information that government bodies possess, but have not published. In each case, the painstaking reverse engineering of government data allowed private parties to do valuable things with the data, which in turn created the political will for the government bodies (the SEC and Congress, in these cases) to move toward publishing more data in open formats.

When government provides reusable data, the practical costs of reuse, adaptation, and innovation by third parties are dramatically reduced. It is reasonable to expect that the low costs of entry will lead to a flourishing of third party sites extending and enhancing government data in a range of areas—rulemaking, procurement, and registered intellectual property, for example.

This approach could be implemented incrementally, as a pilot group of federal entities shift their online focus from finished websites to the infrastructure that allows new sites to be created. If the creation of infrastructure causes superior third party alternatives to emerge—as we believe it typically will—then the government entity can cut costs by limiting its own web presence to functions such as branded marketing and messaging, while allowing third parties to handle core data interaction. If, on the other hand, third party alternatives to the government site do not satisfactorily emerge—as may happen in some cases—then the public site can be maintained at taxpayer expense. The overall picture is that the government's IT costs will decline in those areas where private actors have the greatest interest in helping to leverage the underlying data, while the government's IT costs will increase in those areas where, for whatever reason, there is no private actor in the world to step forward and create a compelling website based on the data. We expect that the former cases will easily outnumber the latter.

One key question for any effort in this area is the extent of flexibility in existing regimes. A number of recent laws have explicitly addressed the issue of putting government information on "websites." The E-Government Act of 2002, for example, asks each agency to put its contributions to the Federal Register, as well as various other information, on a public website.[50] This opens up a question of construal: Does an Internet location that contains machine-readable XML—which can be displayed

---

[47] Govtrack.us: Tracking the U.S. Congress, *supra* note 27.

[48] *Next-Generation EDGAR system - Better Data. Stronger Markets.*, U.S. Sec. & Exch. Comm'n, http://www.sec.gov/edgar/searchedgar/webusers.htm (last visited June 30, 2012).

[49] *See infra* §3; *see also* Markoff, *supra* note 30.

[50] Pub. L. No. 107-347, 116 Stat. 2902 (2002).

directly in a web browser and deciphered by humans but is designed to be used as input into a presentation system or engine—count as a "website"?[51]

If not, these statutory requirements may require government bodies to continue maintaining their own sites. It could be argued that XML pages are not webpages because they cannot be conveniently understood without suitable software to parse them and create a human-facing display. But this objection actually applies equally and in the same way to traditional webpages themselves: The plain text of each page contains not only the data destined for human consumption, but also information designed to direct the computer's handling or display of the underlying data, and it is via parsing and presentation by a browser program that users view such data.

One virtue of structured data, however, is that software to display it is easy to create. The federal government could easily create a general "government information browser" which would display any item of government information in a simple, plain, and universally accessible format. Eventually, and perhaps rapidly, standard web browsers might provide such a feature, thereby making continued government provision of data browsing software unnecessary. Extremely simple websites that enable a structured data browser to display any and all government information may satisfy the letter of existing law, while the thriving marketplace of third party solutions realizes its spirit better than its drafters imagined.

We are focused in this Chapter on the government's role as a publisher of data, but it also bears mention that governmental bodies might well benefit from a similar approach to *collecting* data—user feedback, regulatory comments, and other official paperwork. This could involve private parties in the work of gathering citizen input, potentially broadening both the population from which input is gathered and the range of ways in which citizens are able to involve themselves in governmental processes. But it would raise a number of questions, such as the need to make sure that third party sites do not alter the data they gather before it reaches the government. Alternatively, the government could mandate that certain data be reported in specific, well-defined formats, which facilititates downstream analysis.[52] Another example is the White House's Smart Disclosure initiative, which aims to facilitate "the timely release of complex information and data in standardized, machine readable formats in ways that enable consumers to make informed decisions."[53] These issues deserve further exploration but are beyond the scope of this work.

---

[51] Requirements that data be put "on the Internet" suffer no such ambiguities—providing the data in structured, machine-readable form on the Internet is sufficient to meet such a requirement.

[52] For example, the SEC has mandated that companies report financial information in XBRL format. *Office of Interactive Disclosure: History*, U.S. Sec. & Exch. Comm'n, http://www.sec.gov/spotlight/xbrl/oid-history.shtml (last visited June 30, 2012).

[53] Cass R. Sunstein, Exec. Office of the President, Memorandum for the Heads of Executive Departments and Agencies, Disclosure and Simplification as Regulatory Tools 2

2.4 Alternatives and Counterarguments

We argue that when providing data on the Internet, the federal government's core objective should be to build open infrastructures that enable citizens to make their own uses of the data. If, having achieved that objective, government takes the further step of developing finished sites that rely on the data, so much the better. Our proposal would reverse the current policy, which is to regard government websites themselves as the primary vehicle for the distribution of public data, and open infrastructures for sharing the data as a laudable but secondary objective.

The status quo has its virtues. As long as government websites themselves are the top priority, there is no risk that a lack of interest by private parties will limit citizens' access to government data. Instead, the government creates a system that every citizen can use (if not from home, then from a library or other public facility) without the need to understand the inner workings of technology. It might be argued that government ought to take a proprietary interest in getting its data all the way to individual citizens, and that relying on private parties for help would be a failure of responsibility. There is also a certain economy to the current situation: Under the current system, the costs of developing an open infrastructure for third party access are typically incurred in response to specific interest by citizens in accessing particular data—for example, Carl Malamud's campaign to move SEC data online.[54] These costs could be quite high if, for example, text extracted from a large number of scanned documents needs to be manually cleaned and re-formatted. If there is limited perceived interest in a particular dataset, it may not be worth the taxpayer cost of performing an expensive conversion. In those cases, perhaps it may make overall economic sense to simply allow the interested individual to scrape the data, however imperfect and time-consuming the process.

But, as described above, the status quo also has marked drawbacks. The institutional workings of government make it systematically incapable of adapting and improving websites as fast as technology itself progresses. No one site can meet as many different needs as well as a range of privately provided options can. And the idea that government's single site for accessing data will be a well-designed one is, as noted in §2.1, optimistic at best. Moreover, the government already relies heavily on private parties for facilitating aspects of core civic activities—traveling to Washington, calling one's representatives on the phone, or even going to the library to retrieve a paper public record all require the surrounding infrastructure within which the federal government itself is situated.

Another strategy—always popular in single-issue contexts—would be trying to "have our cake and eat it too" by fully funding *both* elaborate government websites and open data infrastructures. We have no quarrel with increasing the overall pool of resources available for federal web development, but not only is this unrealistic in

---

(2011), *available at* http://www.whitehouse.gov/sites/default/files/omb/inforeg/for-agencies/informing-consumers-through-smart-disclosure.pdf.

[54] Posting of Taxpayer Assets, *supra* note 29.

today's fiscal climate, we do not think that any amount of resources would resolve the issue fully. At some point in each federal IT unit, there is apt to be someone who has combined responsibility for the full range of outward-facing Internet activities, whether these include an open infrastructure, a polished website, or both. Such people will inevitably focus their thoughts and direct their resources to particular projects. When open infrastructures drive websites, the infrastructure and site each rely on what the other is doing; it is extremely difficult to innovate on both levels at once.

Some people might want government to present data because they want access to the "genuine" data, unmediated by any private party. As long as there is vigorous competition between third party sites, however, we expect most citizens will be able to find a site provider they trust. We expect many political parties, activist groups, and large news organizations to offer, or endorse, sites that provide at least bare-bones presentation of government data. A citizen who trusts one of these providers or endorsers, based on its reputation, will usually be satisfied. But if a private party intentionally modifies data, as to mislead information-seeking citizens, such misdeeds would likely receive strong public criticism by the press and other stakeholders. To the extent that citizens want direct access to government data, they can access the raw data feeds directly. Private sites can offer this access, via the "permalinks" (permanent URLs) which our policy proposal requires government-provided data items to have. If even this is not enough, we expect at least some government agencies to offer simple websites that offer straightforward presentation of data.

To the extent that government processes define standardized documents, these should be part of the raw data provided by the government, and should have a permanent URL. To give one example, U.S. patents should continue to be available, in standardized formats such as PDF, at permanent URLs. In addition, the Patent and Trademark Office should make the raw text of patents available in a machine-readable form that allows structured access to, for example, the text of individual patent claims.

Where it is necessary for a citizen to convince a third party that a unit of government data is authentic, this can be accomplished by using digital signatures.[55] A government data provider can provide a digital signature alongside each data item. A third party site that presents the data can offer a copy of the signature along with the data, allowing the user to verify the authenticity of the data item by verifying the digital signature without needing to visit the government site directly. As an alternate online solution, the government could provide authentic data over a secure HTTP connection, which would allow the user to audit the accuracy of a third party

---

[55] Digital signatures are cryptographic structures created by one party (the "signer") that can be verified by any other party (the "verifier") such that the verifier is assured that the signature could only have been created by the signer (or someone who stole the signer's secret key), and that the document to which the signature applies has not been altered since it was signed. *See, e.g.*, Nat'l Inst. Standards & Tech., U.S. Dep't. of Commerce, FIPS PUB No. 186-2, Digital Signature Standard (DSS) (2000), *available at* http://csrc.nist.gov/publications/fips/archive/fips186-2/fips186-2-change1.pdf.

site, by comparing randomly selected portions of data with the authentic government versions.[56]

2.5 Conclusion

In this Chapter, we have proposed an approach to online government data that leverages both the American tradition of entrepreneurial self-reliance and the remarkable low-cost flexibility of contemporary digital technology. The idea, though it can be implemented in a comfortably incremental fashion, is ultimately transformative. It leads toward an ecosystem of grassroots, unplanned solutions to online civic needs.

Since the 2008 release of *Government Data and the Invisible Hand*, from which this Chapter is derived, governments around the world, and at all levels, have begun to adopt the guiding principles behind our approach. In the U.S., federal websites still play a prominent role in disseminating crucial information to citizens, but a parallel "open government data" movement has rapidly gained institutional momentum.[57] Soon after coming into office, the Obama administration, through its Open Government Initiative,[58] promoted these ideas by creating the federal Data.gov

---

[56] With both digital signatures and secure HTTP, authenticity only assures the user that a particular government unit is the source of the data, and that the data haven't been modified in transit. The idea does not suggest that the contents of the data are correct, true or up-to-date. For instance, a dataset that is digitally signed by the government may inadvertently contain erroneous information.

[57] We suggested in §2.3 that the policy route to more reusable data should be "to require that federal government websites retrieve their published data using the same infrastructure that they have made available to the public." Up until now, this policy suggestion has turned out to be too strong: We have instead seen governments create separate open data publishing platforms, which are more or less independent from their existing websites. This is much easier to achieve technologically—it does not depend on restructuring the underlying architecture of legacy web content management systems. It is also more politically manageable, since the contents of datasets can carry a significant amount of political baggage. Opening up data often means sacrificing some amount of control and power, which individual civil servants and governmental units are naturally hesitant to relinquish. Creating a separate open data repository allows government to move forward with the "low hanging fruit," while postponing the publication of datasets that are politically- or bureaucratically-sensitive. The tradeoff, however, is that many datasets of strong public interest may never be published without a stricter rule, like the one we suggested.

[58] *Open Government Initiative*, WHITEHOUSE.GOV, http://www.whitehouse.gov/open (last visited June 30, 2012).

repository to serve as a catalog of raw agency information.[59] New policies outlined in the Open Government Directive[60] then required the executive agencies to publish datasets to Data.gov, of which there are thousands today.[61] Because of the Initiative, important publications like the Federal Register—the "daily newspaper of the Federal government"[62]—were published in machine-readable XML for the first time.[63] The Register's release stimulated a flurry of innovative activity that reimagined how the Register could be used.[64] Many of these new features—first demonstrated by private entities—were later *absorbed* by the government into its own official Federal Register offering.[65]

---

[59] Peter Orszag, *Democratizing Data*, White House Blog (May 21, 2009, 1:53 PM), http://www.whitehouse.gov/blog/Democratizing-Data.

[60] Peter R. Orszag, Exec. Office of the President, Memorandum No. M-10-06, Open Government Directive 1 (2009), *available at* http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf.

[61] *See Federal Agency Participation*, Data.gov, http://www.data.gov/metric (last visited June 30, 2012).

[62] *About the Federal Register*, National Archives, http://www.archives.gov/federal-register/the-federal-register/about.html (last visited June 30, 2012).

[63] Ray Mosley, *Federal Register 2.0: Opening a Window onto the Inner Workings of Government*, White House Open Gov Blog (Oct. 5 2009, 9:14 AM), http://www.whitehouse.gov/blog/Federal-Register-20-Opening-a-Window-onto-the-Inner-Workings-of-Government.

[64] *Id.*

[65] *See* David Ferriero, *Federal Register 2.0*, White House Open Gov Blog (July 26, 2010, 12:53 PM), http://www.whitehouse.gov/blog/2010/07/26/federal-register-20 ("In March 2010, the Office of the Federal Register approached the trio to repurpose, refine, and expand on the GovPulse.us application to bring the Federal Register to a wider audience. Federal Register 2.0 is the product of this innovative partnership and was developed using the principles of open government."); *see also* Michael White, *What FedThread Has Sewn*, Office Fed. Register Blog (July 28, 2011), https://www.federalregister.gov/blog/2011/07/what-fedthread-has-sewn ("The FedThread experiment worked so well, it helped convince us at OFR and GPO that we too could seize the free-roaming XML bull by the horns to build Federal Register 2.0, along with the GovPulse.us founders. The FedThread project gave us confidence as we built FR 2.o, and subsequent consultation with CITP helped us launch what may be the first-ever 100% open source, cloud-hosted U.S. Government web application.").

Following the White House's lead, local governments—like the cities of San Francisco[66] and Chicago[67]—have adopted similar open data catalogs and initiatives, primarily geared toward improving municipal services. Foreign governments have also taken notice: More than 55 countries have signed on to the nascent Open Government Partnership (OGP),[68] where the signatories must commit to "pro-actively provid[ing] high-value information, including raw data, in a timely manner, in formats that the public can easily locate, understand and use, and in formats that facilitate reuse."[69] While open data formats—as a technical matter—can make published information far more useful, the question of *what* information government should publish is an inherently political question—and one that technology, by itself, cannot solve. We explore this tension, between the technologies of open data and politics of open government, in §5.

The ideas proposed in this Chapter are simple yet powerful. Open data can amplify a wide range of policy goals, from rooting out corruption, to increasing citizen participation, and spurring innovation in private industries. The policy trends around open data are still in their early stages, and governments are still exploring their potential benefits. But while the prescription is new, many of the underlying principles that motivate the open data movement—like government transparency and accountability—are not. In the next two Chapters, we explore how thoughtfully-designed software can help push government entities towards more open data, and in turn, increased transparency and citizen understanding.

---

[66] San Francisco Data, https://data.sfgov.org (last visited June 30, 2012).

[67] City of Chicago Data Portal, https://data.cityofchicago.org (last visited June 30, 2012).

[68] *See* Maria Otero, *How the Open Government Partnership Can Reshape the World*, Guardian Prof'l—Open Gov't Brasilia 2012 (May 11, 2012, 3:30 AM), http://www.guardian.co.uk/public-leaders-network/blog/2012/may/11/open-government-partnership-reshape-world ("55 countries have committed to taking steps towards openness through OGP.").

[69] *Open Government Declaration*, Open Gov't Partnership 1 (Sept. 2011), http://www.opengovpartnership.org/sites/www.opengovpartnership.org/files/page_files/OGP_Declaration.pdf.

Chapter 3

RECAP: Turning PACER Around

Ignorantia juris non excusat.

"Ignorance of the law will not excuse." This principle—that one shall not be held innocent for breaking a law, by claiming to have been unaware of it—is deeply rooted in the legal doctrine of the United States.[1] It rests on the basic assumption that citizens have the opportunity to learn about the laws that govern them. In a common law system, like the one in the United States, this means that citizens need to have access not only to the statutes and regulations, but also to the interpretation of laws that happens during the judicial process.

Throughout its history, the U.S. Courts have maintained a tradition of transparency. According to the Courts, "[w]ith certain very limited exceptions, each step of the federal judicial process is open to the public."[2] In the physical sense, federal courthouses are generally open to the public. Courtrooms are architected with public galleries that allow any interested person to observe the Courts' proceedings. More importantly, the Courts maintain a detailed record of documents for every case, which is made available to the public. The record preserves the legal arguments and reasoning of the courts over time.

Public access to court records is fundamental to democratic process.[3] Access promotes fairness, by allowing the public to check that the laws are being applied in

---

[1] *See generally* Edwin R. Keedy, *Ignorance and Mistake in the Criminal Law*, 22 Harvard L. Rev. 75 (Dec. 1908) (tracing the origins of the doctrine in the United States back to both Roman and English law).

[2] Administrative Office of the U.S. Courts, Understanding the Federal Courts 6 (2003), http://www.uscourts.gov/uscourts/EducationalResources/images/UFC03.pdf.

[3] *See generally* Peter Winn, *On-Line Access to Court Records* 5-13 (unpublished manuscript, presented at Privacy Law Scholars Conference, June 13, 2008), *available at* http://docs.law.gwu.edu/facweb/dsolove/PLSC/

a consistent way.[4] It also helps to establish stable precedents—the legal principle of *stare decisis*—which is core to our common law system. Access also provides stability to the legal system, because without transparency, it is difficult for the courts to be "perceived as legitimate by the community."[5] Indeed, the Supreme Court has recognized that the public has "a general right to inspect and copy public records and documents, including judicial records and documents."[6]

As technology has evolved, the meaning and extent of public access to the courts has also changed. The earliest court records, called "plea rolls," were kept on sheepskin by the English central courts in the 12th century.[7] They recorded the outcome of proceedings but did not contain the reasoning of the court; nor were the rolls distributed to the public.[8] Handwritten pleadings were introduced in the 16th century, which helped court reporters compile legal arguments and case outcomes into "named reports," like Plowden's Reports.[9] Around that same time, the invention of printing revolutionized legal information and education. Printing enabled *identical* copies of court records to be spread widely, and it helped to better preserve these texts over time.[10] Case reporting in the U.S. followed the traditions in England, with information disseminated primarily through privately-prepared reports, for a profit.[11] Starting in the 1870s, the West Publishing Company dominated hard-bound law publishing for nearly a century, until the arrival of LEXIS's computerized service in 1973, which ended West's monopoly.[12] Electronic access to the law has been dominated by these two services since.

But as relied upon as LexisNexis (the successor to LEXIS) and Westlaw (West's flagship legal offering) are, they do not offer the *entire collection* of primary court

---

PLSC-Papers/Winn-Peter.pdf (summarizing U.S. case law on access to public records); *see also* Peter W. Martin, *Online Access to Court Records – From Documents to Data, Particulars to Patterns*, 53 Vill. L. Rev. 855, 856, 860 (2008) (reviewing the many purposes justifying citizen access to legal proceedings).

4 James Grimmelmann, *Copyright, Technology, and Access to the Law: An Opinionated Primer* (June 19, 2008), http://james.grimmelmann.net/essays/CopyrightTechnologyAccess.

5 Peter A. Winn, *Judicial Information Management in an Electronic Age: Old Standards, New Challenges*, 3 Fed. Cts. L. Rev. 135, 136 (2009).

6 Nixon v. Warner Communications, Inc., 435 U.S. 589, 597 (1978).

7 George R. Grossman, Legal Research: Historical Foundation of the Electronic Age 5 (1994).

8 *Id.*

9 *Id.* at 16-17.

10 *Id.* at 322.

11 The reports of Kirby, Dallas, Wheaton, Peters and Cranch set the early stage in American case reporting. *See generally id.* at 29-53.

12 *Id.* at 84.

records.[13] This collection has been maintained by the Courts themselves. Before the Internet, primary records were available only through hardcopy inspection at the courthouse where the case occured. This required interested parties to travel physically to the courthouse to obtain paper records. The process of accessing court records therefore was very expensive, since records of interest could be spread at various courthouses around the nation. This skewed access towards those with the most resources, who could either physically travel to courthouses or afford expensive legal information services. For those with limited resources, legal research would be—and still often is—prohibitively expensive.

As in other areas of modern society, the Internet presents the Courts with an extraordinary opportunity: They can now publish large amounts of primary legal information instantaeously to anybody who wants it, and at very low cost. Electronic distribution of records can significantly lower the barriers to public access, and can level the playing field for those who previously could not easily obtain court documents. It can also spark innovations in the legal information industry, as described in §2, by bringing the power of digital networked technologies to bear on legal research tools and services. Such innovations can help the public better understand what happens in the Courts, improve confidence in the Courts' processes, and increase equality in representation in our justice system.

But regrettably, the Courts have not yet embraced the full range of benefits that digital technologies and the Internet have to offer. While most court records today are either born digital or converted into digital formats, providing open public access to these records over the Internet has been a slow work-in-progress.

3.1 PACER: Public Access to Court Electronic Records

The U.S. Courts were among the first in government to recognize the enormous potential of information technology. The Judicial Conference of the U.S. Courts established its program for Electronic Public Access in 1988.[14] At the core of this program is a remote public access system called PACER, which stands for Public Access to Court

---

[13] Both services only purchase and re-distribute a subset of important court records, specifically, court opinions. They typically don't offer all the records in any given case.

[14] Judicial Conference of the United States, Report of the Proceedings of the Judicial Conference of the United States 83 (Sept. 14, 1988) ("On recommendation of the Committee, the Judicial Conference authorized an experimental program of electronic access for the public to court information in one or more district, bankruptcy, or appellate courts . . .")

Electronic Records.[15] When it was first developed, PACER used a dial-up bulletin board system—the common network technology at the time. Each court managed its own PACER infrastructure, which included server computers accessible through local phone numbers and a backend database of case information for that court.

The amount of case information available through dial-up PACER was limited. Users could retrieve lists of cases—searchable by name—and summary records for each case including frequently updated docket sheets.[16] However, it was not possible at the time to retrieve the actual text of briefs, written opinions, and other filed documents. The service charged an access fee of $0.60 per minute of connect time.[17]

Despite these shortcomings, the establishment of the PACER system was a considerable step forward for public access to judicial records. For the first time, court practitioners and the public had around-the-clock access to updated case information, which obviated the need to travel to the courthouse to obtain many basic records. In 1998, the Courts developed a new web-based version of PACER that is still currently in use, that replaced the increasingly obsolete dial-up technology.[18] As computing technologies rapidly decreased in price, the Courts could provide much more information online—most importantly, digital copies of all documents produced in each case—through a modern web interface.

Today, PACER puts online an extensive collection of raw documents from federal district, bankruptcy and appellate court proceedings, with some records dating back to the 1950s.[19] By the Courts' count, there are 500 million documents in the database covering 41 million cases.[20] PACER has without a doubt expanded public access to

---

[15] PACER is the public access complement to a system called CM/ECF, or Case Management/Electronic Case Filing, which is used by attorneys and court officials to electronically file and maintain case records.

[16] *See generally Public Access to Court Electronic Records—User Manual*, PACER Service Center (Mar. 1, 1998), http://www.pacer.gov/documents/pacer.txt.

[17] *Id.*

[18] Judicial Conference of the United States, Report of the Proceedings of the Judicial Conference of the United States 64 (Sept. 15, 1998) ("With the introduction of Internet technology to the judiciary's current public access program, the Committee on Court Administration and Case Management recommended that a new Internet PACER fee be established to maintain the current public access revenue while introducing new technologies to expand public accessibility to PACER information.").

[19] Some districts provide records from as far back as 1950, while others provide less coverage. For older records, dockets may be available, but the scans of documents may not be. *See PACER Case Locator—Court Information*, Admin. Off. U.S. Cts., https://pcl.uscourts.gov/courts.

[20] *See Public Access to Court Electronic Records, What Information is Available on PACER?*, Admin. Off. U.S. Cts., http://www.pacer.gov (stating "PACER currently hosts 500 million case file documents.") (last visited June 19, 2012);

federal court information, but the Courts have yet to harness digital technologies in a way that realizes the full potential for equal and open access that they present.

How PACER Limits Public Access

The first hurdle to using PACER is the need to register for a PACER account. Registration requires entering credit card information, which allows the Courts to charge users for accessing records. While most people do have credit cards, this initial requirement can already pose a high hurdle for public access, even if no charges are immediately incurred. Some citizens, with a one-off interest in a certain case, may be unwilling to enter credit card information just to obtain a few public documents. Or they may not want to run the risk of accidentally running up lots of charges, or may hesitate because of general uncertainties about security and privacy when providing their billing information online.[21] The registration requirement turns many potential PACER users away at the door.

Once the user registers and is logged into the PACER system, finding relevant information can be quite a challenge. The system is composed of more than 200 separate PACER software installations—one installation for each of the federal district, bankruptcy, and appellate court jurisdictions. Each installation is built on its own hardware stack, and each contains its own individual silo of case information for that court. Built primarily on technologies that are now a decade old, the PACER interface provides a complicated search interface that relies heavily on legal jargon, making it difficult—especially for non-lawyers—to find individual documents.[22] To look for a specific case, users need to use a separate online tool, called the PACER Case Locator,[23] to perform a nationwide search across all of the PACER silos.

---

see also Administrative Office of the U.S. Courts, Next Generation CM/ECF: Additional Functional Requirements Group Final Report 1 (Feb. 27, 2012), *available at* http://www.uscourts.gov/uscourts/FederalCourts/Publications/ASFRG-Final-Report.pdf ("[PACER] contains 41 million cases. . .").

[21] *See*, *e.g.*, Joseph Turow, et al., *Open to Exploitation: American Shoppers Online and Offline*, Departmental Papers, Annenberg Pub. Pol'y Center U. Pa. (2005) (finding high levels of misunderstanding and uncertainty among American online shoppers about how their personal information is used).

[22] *See* Martin, *supra* note 3, at 9 ("The federal courts did not establish computer-based case management systems or subsequent electronic filing and document management systems in order to provide the public with better access to court records. Those systems were created because they offered major gains for judges and court administrators. Remote access to them was also of immediate and direct benefit to lawyers . . .").

[23] *PACER Case Locator*, Admin. Off. U.S. Cts., http://www.pacer.gov/pcl.html (last visited June 19, 2012). The Courts launched the PACER Case Locator in April 2010.

But the biggest problem with PACER by far is its pay-for-access model. The Courts charge PACER users a fee of ten cents per page to access its records.[24] This means that, when looking for a case, searches will cost ten cents for every 4320 bytes of results—one "page" of information according to PACER's policy.[25] Once the case is found, obtaining a docket that lists all of its documents could—for lengthy cases—cost another dollar or two. To download a specific document in the case, say a 20-page PDF brief, the user would be charged another $2.00. While each individual charge may seem small, the cost incurred by using PACER for any substantial purpose racks up very quickly.

Even at many of our nation's top law schools, access to the primary legal documents in PACER is limited for fear that their libraries' PACER bills will spiral out of control.[26] Academics who want to study large quantities of court documents are effectively shut out. Also affected are journalists, nonprofit groups, *pro se* litigants, and other interested citizens, whose limited budgets make paying for PACER access an unfair burden. Even the Department of Justice paid $4 million in fees in 2009 to access these public records.[27] From the Court's own statistics, nearly half of all PACER users are attorneys who practice in the federal courts, which indicates that

---

[24] The per-page fee has risen in recent years, even as the cost of information technologies has trended downward. The PACER access fee increased from seven to eight cents per page in 2005, and again from eight to ten cents per page in 2012. *See New Fees in 2005*, U.S. Cts.—Third Branch News (Feb. 2005), http://www.uscourts.gov/news/TheThirdBranch/05-02-01/New_Fees_in_2005.aspx; *see also PACER Fee Increase To Take Effect April 1*, U.S. Cts.—Third Branch News (Sept. 30, 2011), http://www.uscourts.gov/news/newsView/11-09-30/PACER_Fee_Increase_To_Take_Effect_April_1.aspx.

[25] *PACER - Frequently Asked Questions*, Admin. Off. U.S. Cts. http://www.pacer.gov/psc/faq.html (addressing the question "How do you determine what a "page" is for billing purposes?") (last visited June 19, 2012).

[26] At Yale Law School, for example, the law library tells students and faculty that the library "may be able to cover some of these [PACER] costs, however, [they] must meet with [a librarian] to discuss [their] research and obtain prior authorization before [they] incur the costs." E-mail from John Nann, Assoc. Librarian for Reference and Instructional Services, Goldman Law Library, Yale Law School, to Yale Law School Community (Apr. 10, 2012, 1:37 PM EST) (on file with author). *See also* Erika V. Wayne, *PACER Spending Survey*, Legal Research Plus Blog (Aug. 28, 2009), http://legalresearchplus.com/2009/08/28/pacer-spending-survey.

[27] *See* Letter from U.S. Department of Justice, to Carl Malamud, Public.Resource.Org, in response to a Freedom of Information Act Request Relating to PACER Fees, at 15 (Nov. 25, 2009), *available at* http://bulk.resource.org/courts.gov/foia/gov.doj_20091125_from.pdf.

PACER may not be adequately serving the general public.[28] PACER's walled-garden approach also means that the major search engines are unable to crawl any of its contents. The pay-for-access model bears much of the blame.

Furthermore, the Courts do not provide any consistent machine-readable way to index or track cases, even though PACER gathers all of this information electronically and stores it in relational databases. Anyone seeking to comprehensively analyze case materials in a given area faces an uphill battle of reconstructing the original record.[29]

PACER is the sole mechanism from which the public can obtain electronic court records directly from the Courts. Ideally, the Courts would freely publish all of their records online, in bulk machine-readable formats, as suggested in §2, to allow any private party to index and re-host all of the documents, or to build new innovative services on top of the data. But while this would be relatively cheap for the Courts to do, they haven't done so for a number of reasons. One reason is privacy which is discussed in §3.4.1. Another reason is to protect the PACER revenue stream, which is used to fund the public access system and other court IT costs.

Congress first authorized the Courts to charge user fees for electronic public access in 1991.[30] A decade later, the E-Government Act of 2002 clarified that the

---

[28] Harlan Yu, *Assessing PACER's Access Barriers*, Freedom to Tinker Blog (Aug. 17, 2010), https://freedom-to-tinker.com/blog/harlanyu/assessing-pacers-access-barriers. Statistics offered by Michel Ishakian, from the Administrative Office of the U.S. Courts, also support this claim. The Honorable Ronald Leighton, et al., *Panel Three: Implementation—What Methods, If Any, Can Be Employed To Promote the Existing Rules' Attempts to Protect Private Identifier Information From Internet Access?*, 79 Fordham L. Rev. 45, 47 (2011) ("PACER has several categories of users. They are fairly discrete. Fully 75% are from the legal sector or are litigants, 10% are commercial users, approximately 5% are background investigators, which we have sorted out from commercial in stitutions, 2% belong to the media, and 2% represent academia.").

[29] Various academic and private groups attempt to compile case records in specific interest areas. These groups typically purchase records from PACER, and combine them with information from other sources, while adding other useful research features. *See, e.g.*, Lex Machina, Inc., https://lexmachina.com (last visited June 19, 2012) (providing for-profit access to their Intellectual Property Litigation Clearinghouse); The Civil Rights Ligitation Clearinghouse, http://www.clearinghouse.net (last visited June 19, 2012) (providing civil rights case materials, hosted by University of Michigan Law School); Legal Threats Database, http://www.citmedialaw.org/database (last visited June 19, 2012) (providing media law case materials, hosted by the Citizen Media Law Project).

[30] Departments of Commerce, Justice, and State, the Judiciary, and Related Agencies Appropriations Act, 1991, Pub. L. No. 101-515, §404, 104 Stat. 2101 (1990) ("The Judicial Conference shall prescribe reasonable fees . . . for collection by the courts . . . for access to information available through automatic data processing equipment.").

Courts were still allowed to charge fees for PACER access, but that the fees must be prescribed "only to the extent necessary" to provide the service.[31] With this clarification, Congress sought to authorize a fee structure "in which [case] information is freely available to the greatest extent possible."[32]

However, the Courts' current fee structure collects significantly more funds from users than the actual cost of running the PACER system. Stephen Schultze examined the recent Courts' budget documents and found that the Courts claim PACER expenses of roughly $25 million per year.[33] But in 2010, PACER users paid about $90 million in fees to access the system.[34] The extra revenue is used to purchase a range of unrelated technology enhancements for the Courts, like flat screen monitors in courtrooms and state-of-the-art AV systems.[35] The Courts' adoption of these modern technologies should be encouraged, but out of democratic principle and statutory limits, these improvements shouldn't come at the direct expense of open access to court records.

The reported figure of $25 million per year to run the PACER system is a reflection of a system that's built extremely inefficiently, at least in light of today's "cloud" technologies. Its overall architecture mirrors the traditionally decentralized nature of the federal court system.[36] The PACER software is developed centrally by the Administrative Office of the U.S. Courts (the AO). However, rather than running the PACER system centrally, the software is distributed to each of the over 200 district, bankruptcy and appellate courts. The software is then installed locally by each court on its own separate hardware. Each court has the ability to tailor the software to suit local preferences. By and large, jurisdictions only make cosmetic changes to the PACER interface (like changing the background color of the website) and leave the functionality of the system largely in tact as it is originally developed.

Previously, the cost of making a few hundred terabytes of data available on the Web was not low, but in an era of abundant and ever-cheaper cloud storage and

---

[31] E-Government Act of 2002, Pub. L. No. 107-347, §205(e), 116 Stat 2899 (2002).

[32] S. Rep. No. 107-174, at 23 (2002).

[33] Stephen Schultze, *What Does it Cost to Provide Electronic Public Access to Court Records?*, Managing Miracles Blog (May 29, 2010, 6:26 PM), http://managingmiracles.blogspot.com/2010/05/what-is-electronic-public-access-to.html.

[34] *Id.*

[35] *Id.*

[36] *See History of the Federal Judiciary*, Federal Judicial Center, http://www.fjc.gov/history/home.nsf/page/talking_ej_tp.html (last visited June 19, 2012) (". . . a system of federal trial courts, organized within state borders, reflected the legal traditions of each judicial district and facilitated citizen access to federal justice. The decentralized federal judiciary ensured that individual federal courts had a strong local orientation, while at the same time it united a geographically dispersed nation within a consistent system of federal law.").

hosting services, the Courts can now reduce their PACER operating expenses nearly tenfold by taking advantage of modern technologies.[37] The Courts' computing infrastructure may have made sense two decades ago, but today it only perpetuates higher than necessary costs and barriers to citizen access.

3.2 Liberating Court Records

Because all of the documents in PACER are public records, there appear to be no legal restrictions on how a document can be reused, once it is legitimately paid for.[38] For instance, it would be legal to e-mail the document to a colleague, or paste its contents into a blog post. Likewise, one would be free to share the document widely, by uploading it to a public repository tailored specially for court records.

---

[37] A back-of-the-envelope calculation suggests that it would be *easily* feasible to host the PACER database on Amazon S3 for less than $2 million per year. The entire PACER database has 500 million documents, and let's generously assume that each document is 1 MB in size. For storage, the Courts need 500 TB of space, which costs approximately $600,000 per year on S3. For bandwidth, the courts could serve more than 1 PB of data each month from S3—approximately ten times the current rate of public access—for less than $1 million per year. *See Amazon S3 Pricing*, Amazon Web Services, https://aws.amazon.com/s3/pricing/ (last visited June 25, 2012).

[38] Works prepared by the U.S. government, like a court opinion or a case docket sheet, are not eligible for copyright, 17 U.S.C. §105 ("Copyright protection under this title is not available for any work of the United States Government . . ."). However, other works in PACER, such as a brief filed by a private attorney are as usual protected by copyright. The question of whether the redistribution of such copyright works is legal has never been tested directly in court. Legal scholars suggest that sharing, in the case of RECAP, would be legal based on a fair use or implied license argument. *See*, *e.g.*, Rajiv Batra, *RECAP Attempts to "Turn PACER Around,"* Columbia Science and Technology Law Review Blog (Dec. 5, 2009), http://www.stlr.org/2009/12/recap-attempts-to-turn-pacer-around. In addition, the Courts assert that "The information gathered from the PACER system is a matter of public record and may be reproduced without permission. However, the PACER user assumes all responsibility for consequences that arise from use of the data." *PACER - Frequently Asked Questions*, Admin. Off. U.S. Cts., http://www.pacer.gov/psc/faq.html (addressing the question "What are the acceptable uses of the data obtained from the PACER system?") (last visited June 19, 2012).

Previous Efforts

Building a public repository of court records is what open government advocate Carl Malamud of Public.Resource.Org[39] tried to do in early 2008. He asked the public to "recycle your PACER documents" by uploading previously purchased files through an online interface on his website.[40] He would then re-publish those uploaded documents online in an organized way, allowing anyone to access them. His recycling program excited law librarians and other openness advocates,[41] but his initial effort did not gain significant traction for a few reasons.

First, recycling a PACER document was a manual process. It required users to expend extra effort—to visit Malamud's website and upload PACER documents they had previously downloaded to their local filesystems. Second, uploads did not include standard metadata about which case the document was from, nor the document's number on the case docket, both of which are necessary to index the document. This meant that Malamud had to look at each uploaded document manually in order to index and re-publish it. In short, the recycling system did not scale.

A separate endeavor that did scale, however, managed to bolster Malamud's repository, later in 2008. Unfortunately, the effort was short-lived. The U.S. Courts, in conjunction with the Government Printing Office (GPO), had launched a pilot project to provide computer terminals with free PACER access at sixteen federal depository libraries around the nation.[42] Malamud suggested that a "Thumb Drive Corps" visit depository libraries and copy PACER documents on to USB drives.[43] On that suggestion, an enterprising activist named Aaron Swartz visited his local depository library and managed to automatically download more than 19 million pages of PACER doc-

---

[39] Public.Resource.Org is a nonprofit organization whose mission is "to make the law available to all citizens." PUBLIC.RESOURCE.ORG, http://public.resource.org (last visited June 19, 2012).

[40] Recycle Your PACER Documents, INTERNET ARCHIVE (Feb. 6, 2008), http://web.archive.org/web/20100928233803/http://pacer.resource.org/recycling.html.

[41] Erika Wayne, *Recycle Your Pacer Documents*, FREE GOVERNMENT INFORMATION BLOG (May 1, 2008, 12:01 AM), http://freegovinfo.info/node/1815.

[42] *Free Access to Court Records Offered at 16 Libraries*, U.S. CTS.—THIRD BRANCH NEWS (Dec. 2007), http://www.uscourts.gov/News/TheThirdBranch/07-12-01/Free_Access_to_Court_Records_Offered_at_16_Libraries.aspx.

[43] John Schwartz and Robert Mackey, *Steal These Federal Records—Okay, Not Literally*, N.Y. TIMES BLOGS—THE LEDE (Feb. 13, 2009, 3:34 PM), http://thelede.blogs.nytimes.com/2009/02/13/steal-these-federal-records-okay-not-literally. *See also 16 Frequently Asked Questions about Recycling Your PACER Documents*, PUBLIC.RESOURCE.ORG 5 (Apr. 3, 2008), http://www.scribd.com/doc/2436299/Frequently-Asked-Questions-About-PACER.

uments, including case metadata.[44] When the Courts and the GPO realized what had happened, they shut down the entire pilot project, claiming that "the security of the Pacer service was compromised."[45]

Earlier efforts also made headway into liberating case records, but with more limited scope. Hyperlaw—one of the earliest such efforts—scraped federal appellate opinions from the Courts' dial-up systems starting in 1993, and released these opinions on CD-ROM.[46] AltLaw.org launched a modern version of that effort in 2007, by writing scrapers for each appellate court website, compiling more than 170,000 decisions.[47] In 2009, Google Scholar launched its Legal Opinions and Journals search engine[48] whose database exceeded AltLaw's collection.[49] Google accomplished this by licensing records from an unnamed legal information vendor and combining them with other available online sources.[50] One of these online sources, called Justia, buys what they consider to be important court records from PACER, and re-publishes them for free on their website.[51] Each of these projects primarily focus on collecting

---

[44] John Schwartz, *An Effort to Upgrade a Court Archive System to Free and Easy*, N. Y. Times, Feb. 12, 2009, *available at* https://www.nytimes.com/2009/02/13/us/13records.html.

[45] *Id.*

[46] *History of Hyperlaw*, Hyperlaw (Sept. 4, 2007), http://www.hyperlaw.com/history.html.

[47] *Columbia Law School Launches AltLaw.org*, Columbia Law School, (Aug. 23, 2007), http://www.law.columbia.edu/media_inquiries/news_events/2007/august07/altlaw_launch.

[48] *See* Anurag Acharya, *Finding the Laws that Govern Us*, Google Official Blog (Nov. 17, 2009, 12:05 PM), http://googleblog.blogspot.com/2009/11/finding-laws-that-govern-us.html.

[49] *See* Joe Hodnicki, *AltLaw Shuts Down But Remains an Open Access Success Story*, Law Librarian Blog (May 14, 2010), http://lawprofessors.typepad.com/law_librarian_blog/2010/05/altlaw-shuts-down-but-remains-an-open-access-success-story.html (". . . the AltLaw team announced that it would be shutting down its website and search service explaining that "[e]verything we have done or planned to do with AltLaw, Google has [sic] does better. . .."").

[50] Mark Giangrande, *Google SLOJ Details Emerge on Law Librarian Blog Talk Radio*, Law Librarian Blog (Dec. 8, 2009), http://lawprofessors.typepad.com/law_librarian_blog/2009/12/google-sloj-details-emerge-on-law-librarian-blog-talk-radio.html ("We did learn some new information about legal opinions in Scholar. One is that the case law database is licensed from a major legal information vendor, who Acharya could not name.").

[51] *U.S. District Courts and Bankruptcy Courts*, Justia, http://www.justia.com/courts/federal-courts (last visited June 19, 2012).

opinions; they don't attempt to collect *all* of the primary documents in each case, which include pleadings, motions, orders, and other associated case matierals.

While opinions may be the most important type of record, access *only* to opinions prevents the public from fully understanding *how* courts reach their decisions. A written opinion typically summarizes the basic facts and legal arguments made in a case, in order to justify the decision. But how would one know whether the judge overlooked a significant, potentially case-changing fact? Or whether a plaintiff tried to employ a certain legal strategy? Answering these kinds of questions requires the ability to inspect the entire primary record—all of the legal filings, presented evidence, and procedural details that reveal how a case develops, from the initial complaint to the final judgement. Moreover, without access to the entire corpus of records, journalists and researchers are unable to perform many types of data-driven studies that could uncover unexpected, but insightful, trends into how the court system functions. It is also extremely difficult for innovators to build better and cheaper online legal research services that are comprehensive enough for lawyers to use on a regular basis. The only way to obtain all of these records today is through PACER. But because of the paywall policy, these opportunites are currently squandered.

Introducing RECAP

Building on these past efforts, we created an extension for the Mozilla Firefox web browser called RECAP.[52] The goal of RECAP is to "turn PACER around."[53] by crowdsourcing the purchase of court records in PACER: Once a document is purchased by one user, it is liberated from the PACER paywall and placed in a public online repository. The repository is then shared with the world.

RECAP provides the user with two main benefits. First, the extension helps the user contribute to the public good, by automatically uploading purchased PACER records to the public repository. Second, using RECAP saves users money: When the user views a case docket, the extension determines whether any of the documents listed is available for free from the public repository. If so, the extension injects a RECAP link, next to the paid link, into the docket HTML for those available documents.

When RECAP was released in August 2009, the Courts reacted somewhat impulsively. They issued e-mail warnings to PACER users that RECAP might be dangerous to use because it is "open source" software.[54] Many courts posted similar warning

---

[52] The RECAP team consists of Timothy B. Lee, Stephen J. Schultze and myself. Prof. Ed Felten advises the project.

[53] RECAP: Turning PACER Around, https://www.recapthelaw.org (last visited June 19, 2012).

[54] Paul Alan Levy, *Federal Court Using Scare Tactics to Block Sharing of Public Records*, Public Citizen Consumer Law & Policy Blog (Aug. 21, 2009, 6:30 PM), http://pubcit.typepad.com/clpblog/2009/08/federal-court-using-scare-tactics-to-block-sharing-of-public-records.html.

messages on their PACER login pages, some of which still exist to this day.[55] Ultimately, the Courts had no legal recourse against a tool that simply helped citizens exercise their right to share public records.

Since release, RECAP has been installed by thousands of PACER users. The public repository contains more than 2.7 million documents from over 640,000 federal cases. Purchasing these documents from scratch from PACER would cost more than $1.5 million. And while our collection still pales in comparison to the 500 million documents purportedly in the PACER system, it contains many of the documents of highest public interest.

3.3 The Design of RECAP

The major challenge in designing RECAP is not technical. Rather, the software design was born primarily from an understanding of the PACER policy landscape and the way that PACER users currently interact with the system. Like traditional system design, designing RECAP was guided by a number of goals and constraints.

The primary goal of RECAP is to liberate as many useful court records as possible, and to do so in a legal, scalable way. Doing so required many PACER users to install our extension, which meant building a tool that PACER's primary user demographic—lawyers[56]—could easily use. It also meant understanding the legal research process and the incentives that would drive a potential user to download and

---

[55] *See, e.g.*, Eastern District of New York, CM/ECF Filer or PACER Login, https://ecf.nyed.uscourts.gov/cgi-bin/login.pl (last visited June 19, 2012). The warning reads: "Notice for CM/ECF Filers: The Eastern District of New York would like to make CM/ECF filers aware of security concerns relating to a software application, or "extension," called RECAP, which was designed by a group from Princeton University to enable the sharing of court documents on the Internet. . . . Please be aware that RECAP is "open-source" software, which can be freely obtained by anyone with Internet access and modified for benign or malicious purposes... Accordingly, CM/ECF filers are reminded to be diligent about their computer security practices to ensure that documents are not inadvertently shared or compromised. This District Court and the Administrative Office of the U.S. Courts will continue to analyze the implications of RECAP and related-software and advise you of any ongoing or further concerns."

[56] Approximately half of the 1 million registered PACER users also use CM/ECF (that is, they are attorneys who practice and file documents in the federal courts). *See A Look at Electronic Public Access in the Federal Courts*, U.S. Cts.—Third Branch News (Aug. 2010), http://www.uscourts.gov/news/TheThirdBranch/10-08-01/A_Look_at_Electronic_Public_Access_in_the_Federal_Courts.aspx; *see also Preliminary Findings: Satisfaction High Among PACER Users*, U.S. Cts.—Third Branch News (May 2010), http://www.uscourts.gov/news/TheThirdBranch/10-05-24/Preliminary_Findings_Satisfaction_High_Among_PACER_Users.aspx.

install our extension. At the same time, the design was constrained by a number of policy, legal and ethical considerations, which will be detailed in this section.

Our design of RECAP has enabled it to gain widespread adoption and distinguish itself from previous efforts to liberate case law in the United States. This section describes the overall architecture of the RECAP system, followed by six key design lessons drawn from our experience.

3.3.1 Technical Design Overview

The RECAP system consists of three primary components: (1) the RECAP Firefox extension, which is installed by individual PACER users, (2) the RECAP extension server, which processes uploads from the extension, and (3) the Internet Archive[57] public repository, which serves as the system's storage endpoint. Figure 3.1 shows a schematic of the high-level architecture of the RECAP system, along with the primary interactions between components. Each component of the system will be described in turn.

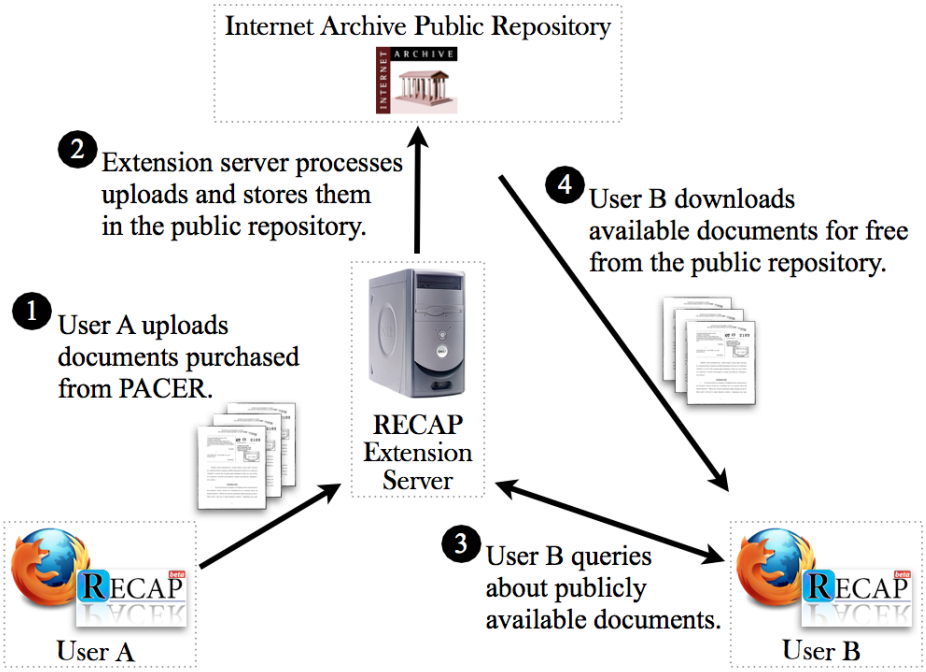

Figure 3.1: High-level architecture of the RECAP system.

[57] *See* Internet Archive, *infra* note 65.

The RECAP Firefox Extension

Creating the RECAP extension required a significant amount of reverse-engineering of the PACER software.[58] As previously described, the RECAP extension performs two primary functions: to help the user automatically contribute purchased records to the public repository, and to notify users when records are already available for free.

The RECAP extension is active only when the user is browsing a PACER site, and also logged into PACER.[59] Once active, the extension monitors the browser's traffic for records of interest—case dockets and individual documents[60]—based on known URL patterns. Case dockets in PACER are provided in HTML format, while individual documents are PDFs. When the user buys either type of record from PACER, the extension will upload a copy of it to the RECAP extension server, in the background, using an asynchronous request. In that request, the extension will include the necessary metadata for the record, so the server will know how to index it.[61]

Furthermore, when the extension notices the purchase and display of a case docket listing, the extension will parse the listing for links to individual documents. It will then make an asynchronous query the server to check whether any of those documents already exist in the public repository. For the publicly available documents, the extension will inject a small RECAP icon into the docket HTML, just after the page loads, adjacent to the for-a-fee document link. Clicking on a RECAP icon will pull up

---

[58] RECAP only operates on federal district and bankruptcy court PACER sites. A considerably different version of the PACER software is used by the federal appellate courts, which is not currently supported by RECAP.

[59] Each PACER site follows a common domain naming scheme. For example, the District of New Jersey PACER site is located at http://ecf.njd.uscourts.gov, while the Northern District of California is found at http://ecf.cand.uscourts.gov. The domain always starts with "ecf"—which stands for "electronic case filing"—and is followed by the court name abbreviation. *See Individual Court PACER Sites*, Admin. Off. U.S. Cts., http://pacer.psc.uscourts.gov/cgi-bin/links.pl (last visited June 19, 2012).

[60] Each court case has a *docket*, which is a chronological listing of documents that are associated with a case. At the top of the docket is a header, which describes the name and number of the case, important dates, and other summary information about the case. Each entry in the docket listing describes and links to an individual *document*. Some documents have supporting records—like an evidence exhibit—that we call a *subdocument*. Many lengthy cases will have tens or even hundreds of docket entries.

[61] Fortunately, it is possible to piece together the necessary metadata for each record type, using elements in the current and referer URLs. To index individual documents, RECAP also needs to have previously ingested the case docket on which the document is listed.

a full screen interstitial, from which the user can download the free document directly from the public repository. In a similar way, the extension will inject a RECAP link to freely available dockets, on the PACER page for docket searches.

Other aspects of the extension's design, and the purpose of their inclusion, will be detailed later in §3.3.2.

The RECAP Extension Server

The extension server is the intermediary between the browser extension and the Internet Archive (IA) storage endpoint. It acts as a proxy for uploads, performs intermediate processing of documents, hosts the central database of document metadata, and responds to queries from the extension about available documents.

The extension server is necessary for a few reasons. As explained below, the IA provides a very simple storage API that mirrors the capabilities of Amazon's Simple Storage Service (S3).[62] The API requires an authorized key in order to write to the RECAP storage collection. Rather than exposing our API key in the extension code, or requiring each extension to register its own API key, we store a single key on our server that is used for all extension uploads. In addition, the extension server runs scripts on uploaded documents and stores document metadata in a SQL database—neither capability is currently offered by the IA.

When the server receives an uploaded PACER case docket, it parses the HTML and extracts all the useful pieces of information from it. The header of the docket includes case-level information, like the case's name and number, its beginning and ending dates, and the names of the judge, parties and attorneys-of-record. Following the header is a table of documents, sequentially numbered and chronologically ordered. The table may or may not contain information about all of the document in the case, because PACER allows users to purchase arbitrary ranges of the docket listing. Since some cases may contain hundreds of documents, PACER allows cost-conscious users to purchase only a subset of the docket, *e.g.*, only documents 1-10 or only documents filed between a range of dates. For rows that are displayed, each row describes a single document, with a document number, date of filing, and short description of its contents (e.g., "Motion to Dismiss"). This document-level metadata is inserted into the extension server's document database. Together with the case-level metadata, all of the extracted information is then reformatted into a standard, structured XML docket, which follows our ad hoc standard (as described in §3.3.2).[63] The docket is then merged with existing docket data from the repository and uploaded to the IA.

---

[62] Amazon Simple Storage Service (Amazon S3), https://aws.amazon.com/s3 (last visited June 19, 2012).

[63] Because each individual court can upgrade and modify its version of the stock PACER software, PACER sites sometimes exhibit subtle differences, especially in the docket's HTML output. The RECAP docket parser adjusts for all of these differences that we're aware of, and outputs a normalized XML docket.

When handling individual document PDF uploads, the server first runs a best-effort script on the PDF to screen it for Social Security numbers.[64] If no sensitive information is detected, the server marks the document as available in the database and uploads it to the public repository. Finally, the server responds to document availability queries from the extension, which is a straightforward lookup in the document database.

The Internet Archive

We partnered with the Internet Archive[65] (IA) to serve as the storage endpoint for RECAP uploads. We chose to do so both as a cost-saving measure—they generously donated as much storage and bandwidth as we needed for the project—and also to ensure the longevity of the public repository. The repository's collection on the IA is called "usfederalcourts."[66]

The IA exposes a "S3-like" API for managing the contents of collections.[67] Each collection is separated into "items," or "buckets" in S3 lingo. Each item in our collection represents an individual court case, and within each item resides the XML docket and PDF documents for that case. All writes from extension installations are proxied through the RECAP server, using a single access key. Trusted third-party bulk uploaders, like Justia, have separate access keys to write directly to our IA collection.

We tweaked the IA's default collection settings such that it would be appropriate for storing court records. The IA typically performs optical character recognition (OCR) on PDF documents that are uploaded to its system, but we disable that feature for our collection. We also disabled search engine crawling for our collection outside from basic case-level metadata information. Both of these tweaks are for privacy reasons that will be discussed below.

3.3.2 Key Design Lessons

This section covers six key design lessons that contributed to RECAP's ability to increase the amount of available government data in a distributed and collaborative

---

[64] The script reads the text layer of the PDF if it exists, and tries to detect Social Security numbers using a simple regular expression. Documents that test positive are held back from the public repository and quarantined on the extension server for manual review.

[65] The Internet Archive is a nonprofit organization that "was founded to build an Internet library" and has the tagline "Universal access to all knowledge." Internet Archive, http://archive.org (last visited June 19, 2012).

[66] *RECAP US Federal Court Documents*, Internet Archive, http://archive.org/details/usfederalcourts (last visited June 19, 2012).

[67] *The Internet Archive's S3 like server API*, Internet Archive, http://archive.org/help/abouts3.txt (last visited June 19, 2012).

fashion. Many of these design elements will be useful in the development of other civic technologies, especially those with a policy purpose.

Distributing the Work

The first important design decision is in the system's high-level architecture: RECAP harnesses the existing work of thousands of current PACER users. The system effectively combines the PACER purchases of many users together, making the whole more useful than the constituent parts.

At 10 cents per page, a centralized solution would not have been a cost-effective solution: Buying all the records through automated scraping would literally cost hundreds of millions of dollars. Using divide and conquer, however, no extra money needs to be spent buying records for the public good. As long as the system can stitch together purchases in a usable way, the repository can grow constructively over time.

An advantage of this design is that that RECAP repository receives the documents of highest interest to its users. Interest and access to court documents is almost certainly a power law distribution—some high-profile cases will be accessed very frequently, while the long tail of routine cases will rarely be of any public interest. The significance is that while the current RECAP repository only has an estimated 0.5% of all PACER documents (approximately 2.7 million out of 500 million), the repository is far more useful than the percentage might suggest.

The model that RECAP uses to crowdsource public effort—with the goal of making government information more available—could be applied in other similar situations. For example, in Canada, a company called Geolytica was able to compile a comprehensive database of postal codes—information that was held closely by Canada Post. Geolytica accomplished this by crowdsourcing the lookups of Canadian street addresses. But unlike RECAP's situation, Canada Post has asserted copyright over the postal code data, and litigation is ongoing.[68] Other innovative initiatives, like Peer to Patent,[69] similarly crowdsource public effort for a policy purpose, but focus on efficiently gathering outside expertise (input) rather than publishing government information (output).

Establishing Ad Hoc Standards

In order to effectively merge the work of thousands of RECAP users, we needed to establish an ad hoc XML standard to combine case dockets. Merging dockets is necessary because PACER allows users to purchase dockets in a piecemeal fashion, as

---

[68] Michael Geist, *Canada Post Files Copyright Lawsuit Over Crowdsourced Postal Code Database* (Apr. 13, 2012), http://www.michaelgeist.ca/content/view/6415/125.

[69] Beth Simone Noveck, *"Peer to Patent": Collective Intelligence, Open Review, and Patent Reform*, 20 Harv. J. L. & Tech. 123 (2006); *see also* Peer to Patent, http://www.peertopatent.org (last visited June 19, 2012).

previously described. For example, one cost-conscious user, looking for the complaint in the case, may only purchase docket entries 1-10 from the docket. A second user is looking for other documents in the same case, purchasing entries 11-25. By converting each of these uploaded HTML dockets into normalized XML dockets, these extension server can more easily merge these two dockets together. In addition, by defining the ad hoc standard, we were able to allow RECAP's trusted bulk uploaders to contribute directly to the public repository, as concurrent uploads were coming in from our extension users, which significantly boosted the size of our collection.

The XML schema for dockets is relatively straightforward. It mirrors the semantic structure of the HTML case docket: a case header followed by a list of documents from the case. The Internet Archive stores RECAP's "master XML copy" of dockets in each case. Each time a case docket is uploaded, the extension server will first download the "master XML copy" from that case from the IA, merge in the contents of the uploaded docket, and upload the combined docket as the new "master XML copy." Because the IA RECAP collection has multiple authorized uploaders, we defined a "case locking" protocol to prevent race conditions when two uploaders attempt to update the same docket at the same time. This protocol helped to keep the public repository consistent with our document database, and allowed multiple users to upload into our repository at the same time.

Providing Meaningful Benefits

RECAP depends on user contributions, so the design needed to incorporate features that would benefit its primary users. This meant understanding how lawyers use PACER, and how RECAP can improve their PACER experience. Meaningful benefits would incentivize PACER users to install the RECAP extension, and more users would mean faster growth of the public repository.

We seeded the public repository with existing collections of documents made RECAP useful from the start.[70] Even on launch day, some cases were already well-populated with freely available documents, saving users money from the start. But in addition, we built new features into RECAP that would "fix" many of the common frustrations that PACER users typically encounter.

One frustration is with how PACER fails to set useful filenames for downloaded records. For example, when users download a PDF document from PACER (without RECAP), the site does not set a useful Content-disposition filename in the HTTP header. Rather, most major web browsers will simply default to the name of the PACER script that serves the file: show_temp.pl.[71] Since the file type looks like a Perl

---

[70] We used the tarballs published by Public.Resource.Org, which contain records from their recycling program, the "Thumb Drive Corps," and other public collections. *See* Directory listing for PACER tarballs, PUBLIC.RESOURCE.ORG, http://bulk.resource.org/courts.gov/pacer (last visited June 19, 2012).

[71] *See, e.g.*, Finis Price, *PACER Problems with Firefox*, TECHNOESQ BLOG (Dec. 3, 2008), http://www.technoesq.com/technology/2008/12/03/pacer-problems-with-firefox.

script, browsers won't handle the file using its the default PDF file type handler. Once the file is saved, users need to rename downloads to something more relevant (and probably ending in ".pdf") manually after the fact. Otherwise, previously downloaded PDFs become more difficult to find later. RECAP fixes this problem by offering users the option of more descriptive filenames. The extension rewrites the filename header in one of two ways: "lawyer style" (e.g., N.D.Cal._3-08-cv-03251_46_0.pdf, which mimics legal citations with the name of the court, the case number, and the document number) or "Internet Archive style" (e.g., gov.uscourts.cand.204881.46.0.pdf, which is how documents are named in the public repository, using internal PACER case numbering). "Internet Archive style" naming is on by default.

Another quirk of some versions of the PACER software is that it would set the "Cache-control" HTTP header to "no-cache," which instructs the browser to refrain from caching the loaded contents.[72] Slyly, PACER users would then be "double charged" for dockets, every time the user hit the browser's back and forward buttons, for example, to return to an already-purchased docket sheet. This significantly increases the cost of PACER use, often unbeknownst to the user. To help users save even more money, RECAP clears the "Cache-control" HTTP header and sets other cache-related headers in a way that's beneficial to the user.

Finally, RECAP tweaks the PACER defaults in minor ways to improve the user's overall experience. For example, PACER offers users a checkbox option to "Include headers when displaying PDF documents," which adds useful case metadata to the top of every page of downloaded PDFs. Rather than having users manually check the option every time (and sometimes forgetting to do so), RECAP users can set a preference to automatically check the box.

Many of these options are technically simple to implement, and they significantly improve the overall PACER experience, especially for lawyers who deal with large numbers of PACER records.

Minimizing User Effort

Fourth, it was essential in the design of RECAP that it didn't get in the way of the user's usual PACER workflow. In other words, RECAP had to be backwards-compatible with how lawyers currently perform their legal research. All of the copying and sharing functionality happens automatically in the background, so RECAP users need to do little more than install the plug-in to contribute their purchases to the public repository. Especially for practicing lawyers who are billing by the hour, it would not serve the best interests of their client, if they needed to spend extra time to contribute to the public repository. RECAP needed to be effort-free.

Using RECAP shouldn't even slow down a lawyers' work—the extension couldn't add delays to PACER's existing user interface, which meant making the client as simple as possible and doing the bulk of the processing on our central server. It takes a few steps for a document to get to its final storage location, so the central server is responsible for batching these requests and uploading them to the Internet Archive.

---

[72] PACER versions 4.0 and greater no longer exhibit this behavior.

Minimizing user effort should be a key design goal of any software designed to increase data transparency. Transparency is typically not the end goal for the user, but rather a desirable side-effect of the work product or process. If the transparency features of software get in the way of actual goals, the user will be frustrated at best, and will skirt the transparency features at worst. In the case of RECAP, the goal is to locate the right court documents, as quickly and as cheaply as possible. Minimizing effort to upload documents is especially important for RECAP, because the system relies on network effects to succeed. Without many active users, the RECAP system would not receive many uploads, and the repository would be significantly less useful for everyone. Usability is always important—whether the software liberates court records, adds structure to government documents, or provides other transparency benefits—but it was a particularly acute design goal in the case of the RECAP extension.

Providing Elements of Trust

Since RECAP uploads documents transparently in the background, we anticipated that our users—primarily lawyers—would be concerned about how exactly the extension functions, and whether their own privacy would be compromised. To assuage some of these concerns, we built various user interface features into the software which would expose to the user what the extension is doing.

Since RECAP makes real-time modifications to PACER pages, we wanted to minimize confusion about which page elements were provided by PACER, and which elements were added by the extension. Whenever a user clicks on a "little R" RECAP icon to download a free document, we use a full-window interstitial—graying out the current PACER page—to provide visual separation. The interstitial prominently displays the RECAP logo, the date RECAP received the cached copy, and a brief disclaimer about document authenticity.[73]

When uploading copies of records asynchronously in the background, the extension notifies the user with a momentary pop-up notification about what has just been uploaded. Some lawyers were concerned that RECAP would upload files when the user was using CM/ECF, the parallel online system that lawyers use to file documents in a case. They were wary that using RECAP would expose documents to the repository before they became public on PACER. The visual notifications made it easier for our users to understand when records were being uploaded.

Further, installing the extension places a permanent "little R" RECAP icon in the browser's chrome, which changes color depending on whether RECAP is active on the current page—blue for "active" and gray for "inactive." The icon lets users confirm visually that the icon's color is gray for CM/ECF pages and other non-PACER pages. A notification also pops up whenever RECAP's "active" state changes.

---

[73] The interstitial's disclaimer states: "RECAP is not affiliated with the US Courts. The documents it makes available are voluntarily uploaded by PACER users. RECAP cannot guarantee the authenticity of documents because the courts themselves have not implemented a document signing and authentication system."

By clicking on the permanent RECAP icon, the user can choose to temporarily disable RECAP entirely when using PACER. This feature responds to specific concerns from careful lawyers about sharing documents in cases they are involved in, because of potential ethical issues with adverse publicity. By offering this option, RECAP users are less likely to uninstall the extension entirely, just to deal with these infrequent but important concerns. However, when RECAP is temporarily disabled, the extension also will not query the server for freely available documents, which incentivizes users to re-enable RECAP whenever possible.

User privacy is another significant consideration. At the bottom of purchased case dockets, PACER appends a "transaction receipt"—in the form of an HTML table—that details the purchase. The receipt also contains, among other things, the cost of the download, the user's PACER login name and the "client code" (which is typically used by lawyers to track and pass on incurred PACER fees to their appropriate clients). We took care, however, to ensure that these private details contained in the receipt did not end up in the public repository. If we did not take this precaution, RECAP users would expose their PACER usage to the public, and could theoretically tip off opposing lawyers about which cases were being researched. While the extension does upload transaction receipts together with records (as to not complicate the extension's implementation with client-side HTML parsing), we promise our users, in the RECAP privacy policy, that personally-identifiable information about the user is immediately discarded.[74] As a measure of extra caution and added comfort for RECAP users, we also promise in our policy to purge our web server logs after 14 days.[75]

Some lawyers also expressed concerns about the legal risks of using RECAP. Specifically, PACER users are required to consent to an "Acknowledgement of Policies and Procedures" when initially signing up for a PACER account.[76] The policy states:

> "Any attempt to collect data from PACER in a manner which avoids billing is strictly prohibited and may result in criminal prosecution or

---

[74] *Privacy Policy*, RECAP Firefox Extension (Oct. 6, 2010), https://www.recapthelaw.org/privacy ("Some of these [uploaded] pages may contain personally-identifiable information such as a PACER username. Personally-identifiable information is immediately discarded by the Extension Server and is never saved on the Extension Server or transmitted to IA.").

[75] *Id.* ("The RECAP Extension Server keeps logs of all queries and uploads it receives. The Extension Server logs the client IP address, URL, and time of access. The Extension Server also separately logs information on the documents uploaded, including upload time, court name, case number and what transpired. . . . CITP will purge the above mentioned logs after 14 days.").

[76] *PACER On-Line Registration*, Admin. Off. U.S. Cts., https://www.pacer.gov/psco/cgi-bin/regform.pl (last visited June 19, 2012) (exhibiting a required form field to "acknowledge you have read and understand the Policies and Procedures listed above.").

> civil action. PACER privileges will be terminated if, in the judgment of judiciary personnel, they are being misused. Misuse includes, but is not limited to, using an automated process to repeatedly access those portions of the PACER application that do not assess a fee (i.e. calendar events report or case header information) for purposes of collecting case information."[77]

Prior to launch, we were uncertain whether the Administrative Office would interpret the provision broadly to prohibit the use of RECAP. For many lawyers who practice in the federal courts, losing their PACER privileges would be disastrous for their practice. So, an another added measure of caution, we designed RECAP such that it would be more difficult for the AO to immediately detect whether or not a PACER user was using our extension: All PACER page modifications happen locally, immediately after HTTP responses are received by the browser.[78] Fortunately this became a moot point as the AO stated soon after launch that they "have no problem with counsel using RECAP."[79]

Finally, two other measures added to user trust and increased adoption of the RECAP extension. First, we applied an open source license to the RECAP extension source code, and made it available for public inspection.[80] Second, we publicized our partnerships with other well-known entities in the field, like the Internet Archive, Public.Resource.Org and Justia.[81] These partnerships added to the legitimacy and momentum of our effort.

Handling Public Concerns

One significant challenge in opening up federal court records is the concern over the privacy of litigants and other individuals involved in cases. Many of life's dramas play themselves out in our public courtrooms. Consider cases concerning divorce, domestic

---

[77] *Acknowledgement of Policies and Procedures*, Admin. Off. U.S. Cts. 1 (May 1, 2012), https://www.pacer.gov/documents/pacer_policy.pdf.

[78] The one minor exception to this is with our feature to automatically "check" the option to display PDF headers. While this option is transmitted to the PACER users when the form is submitted, it simply mimics what the user could already do manually.

[79] Paul Alan Levy, *Official Word from US Courts – Feel Free to Use RECAP With Our Blessing*, Public Citizen Consumer Law & Policy Blog (Aug. 25, 2009, 6:05 PM), http://pubcit.typepad.com/clpblog/2009/08/official-word-from-us-courts-feel-free-to-use-recap-with-our-blessing.html.

[80] "CITP/RECAP" Code Repository, GitHub, https://github.com/citp/recap (last visited June 19, 2012).

[81] *Turning PACER Around*, RECAP Blog (Aug. 14, 2009), https://www.recapthelaw.org/2009/08/14/welcome.

abuse and bankruptcy. Many intimate private details are revealed in these proceedings and are part of public case records. There is growing concern that sensitive information in court records will become widely available online.

We rely on the public to report privacy issues in individual documents. As RECAP republishes more and more records online, privacy complaints—whether legitimate or not—have continued to roll in. When we receive a complaint, we manually review the document in question, and if its contents violate the Courts' privacy rules[82] in our estimation, we take the document out of the repository. To minimize the number of complaints and potential privacy harm, our central server tries to take a proactive role in screening out documents that contain Social Security numbers, as they are uploaded. However, there are many types of sensitive data—like names of minor children—which are currently difficult for even advanced machine learning algorithms to detect with sufficiently high probability.

Moreover, our informal policy forbids search engines from crawling individual documents. We make no effort to apply optical character recognition (OCR) on scanned PDFs to make their contents searchable. We do, however, allow search engines to crawl the contents of case dockets. Dockets themselves will rarely contain information worthy of redaction, even if documents in that case do. This policy improves public access, relative to the Courts' policies, by making it easier for people to find cases using popular search engines. At the same time, we exercise an abundance of caution, by maintaining a relatively higher level of obscurity for individual documents. The tradeoffs that we currently make are constrained by the Courts' current practices, which will be discussed below in §3.4.1.

Even with these restrictions in place, we still receive a few privacy complaints each week from the public. Typically, these complaints come from people who have searched their own name online, and found links in the results to past court cases in which they have been a part. On the one hand, information contained in court records can be highly embarrasing, and individuals should be able to overcome past transgressions and move on with their lives. On the other hand, public court records are just that—public—and in many situations, the community can legitimately benefit by learning about even the most uncomfortable details in court proceedings. It is debatable in each case whether the disclosure of a document, on balance, would benefit or harm society.

While the balance between open access and privacy is not clear-cut, neither is currently served well by PACER and the Courts' current policies.[83] While the hope is that the Courts will formulate a better policy on individual privacy going forward, the archived backfile of court records continues to be littered with sensitive information.

---

[82] *See* Fed. R. Civ. P. 5.2 *and* Fed. R. Crim. P. 49.1.

[83] *See* Winn, *supra* note 5, at 162 ("In sum, the existing federal electronic filing system does not appear to have been designed with the competing goals of facilitating access and protecting privacy in mind, and there remains considerable room for improvement in both of these two respects.").

This means that those who run public repositories of court records, like RECAP, need to handle privacy complaints on an ongoing basis.

Summary

The design of the RECAP system is based heavily on the intricacies of how the judicial process works, and how lawyers, litigants, court staff and the public interact within this system. Understanding the system—the goals, concerns and incentives of the individual players—is essential in designing usable software that makes a significant open government policy impact. But the goal of RECAP is not to run a mirror of the PACER database. The goal is to open up federal court records and induce the Courts to change their paywall policy, which would obviate the need to use RECAP at all.

It's not difficult to imagine how an open repository of court records might benefit the public. With PACER's many limitations, there are ample opportunities to use modern web technologies to build more intuitive and innovative online interfaces for records access. As a concrete example, a group of Princeton undergraduates spent a semester building a new front-end—much more usable than PACER's own—on top of our public repository, calling it the RECAP Archive.[84] Increased access would also spur new academic research that examines how the Courts function and exposes hidden judicial trends that impact society. For instance, political scientists have already used the RECAP dataset to perform empirical studies on the content of civil complaints: The researchers used spectral clustering analysis to classify complaints and build a taxonomy of civil litigation strategies.[85]

But while improved public access would lead to significant public benefits, it's unlikely that the Courts will voluntarily drop the PACER paywall anytime soon.

3.4 Policy Challenges

3.4.1 The Decline of Practical Obscurity

Before electronic records, the Courts relied on the notion of "practical obscurity" to protect sensitive information. That is, because these documents were only available by physically traveling to the courthouse to obtain the paper copy, the sensitive data contained therein—while public—were in practice obscure enough that very few people, if anyone, would ever see them.

---

[84] RECAP Archive, http://archive.recapthelaw.org (last visited June 19, 2012). The site was developed by Jen King, Sajid Mehmood, Daniel Roberts, and Brett Lullo in Spring 2010.

[85] Christina L. Boyd, et al., *Building a Taxonomy of Litigation: Clusters of Causes of Action in Federal Complaints* (Apr. 2012), *available at* http://ssrn.com/abstract=2045733.

Only in 2007 did the Courts define formal procedural rules that required attorneys to redact certain sensitive information from court filings.[86] But the policy did not apply retroactively, so many older documents available electronically still contain substantial amounts of sensitive information, including Social Security numbers.

A preliminary audit conducted by Carl Malamud found that more than 1,500 documents in PACER, out of a sample of 2.7 million documents, contain unredacted Social Security numbers and other sensitive information.[87] Research by Timothy Lee studied the rate of "failed redactions" in PACER, where authors simply drew a black box over the sensitive information in the PDF, leaving the sensitive information in the underlying file. He estimated that tens of thousands of files with failed redactions exist in PACER today.[88]

PACER's paywall attempts to extend practical obscurity, at least temporarily, to the digital realm, since it prevents the documents behind it from being indexed by major search engines. But over time, these documents will ultimately make their way into wider distribution, whether through RECAP or other means. Large data brokers already regularly mine PACER for personal data. The resulting decline in practical obscurity will ultimately force the Courts to deal more directly with the privacy problem.

This may mean that documents will need to be more heavily redacted before they are filed publicly, or in some cases, full documents will need to be sealed entirely from public view.[89] Properly aligning personal privacy with open access is a tricky proposition, but what's clear is that the Courts ought to make more explicit determinations about which data are sensitive and which are not, rather than relying simply on the hope that certain records won't often be accessed.[90] The risk is that the Courts will be overly conservative in their determinations—suppressing far more

---

[86] *Privacy Policy for Electronic Case Files*, Admin. Off. U.S. Cts., http://www.uscourts.gov/rulesandpolicies/JudiciaryPrivacyPolicy.aspx (last visited June 19, 2012).

[87] Letter from Carl Malamud, Public.Resource.Org, to the Honorable Lee H. Rosenthal, Chair, Committee on Rules of Practice and Procedure, Judicial Conference of the United States (Oct. 24, 2008), *available at* https://public.resource.org/scribd/7512583.pdf.

[88] Timothy B. Lee, *Studying the Frequency of Redaction Failures in PACER*, Freedom to Tinker Blog (May 25, 2011), https://freedom-to-tinker.com/blog/tblee/studying-frequency-redaction-failures-pacer.

[89] Other proposals have suggested, *e.g.*, a differential access approach, where technology could help grant selective access to certain pieces of sensitive information, depending on contextual factors such as the role of the requester. *See* Amanda Conley, et al., *Sustaining Privacy and Open Justice in the Transition to Online Court Records: A Multidisciplinary Inquiry*, 71 Md. L. Rev. 772, 844, 845 (2012).

[90] *Contra* Hartzog and Stutzman have argued that obscurity is a critical component of online privacy and could be used as a protective remedy. Woodrow Hartzog

information than necessary, at the expense of public understanding and oversight. However, given the status quo, the Courts will need to develop more robust policies about privacy protections, while establishing transparency (potentially *ex post*) into how determinations are made, and offering public recourse when determinations are questioned.

Somewhat counter-intuitively, more openness can help lead to more privacy if litigants and their counsel become aware of what information is contained in public records and more proactively choose what to include and when to petition for redaction or sealing.[91] A more accessible corpus also provides opportunities for researchers to devise new methods to protect personal privacy while enhancing the accessibility of the law.

3.4.2 Judicial Appropriations

The more immediate policy challenge is with judicial appropriations and the Court's budget. The Courts rely on Congress for appropriations, and until Congress is willing to replace the revenues generated from PACER's user fee with taxpayer funding, it is unlikely that the Courts will unilaterally make their electronic records openly available. The reliance on appropriations, and inadequate budgets, is a frequent barrier to judicial improvements[92]—and the policy conundrum here is no different. While members of Congress have written stern letters to the Courts inquiring about the efficacy of PACER fees,[93] the relevant House appropriations subcommittee, to date, has not seriously considered taxpayer funding for PACER in the current budgetary environment.

& Frederic D. Stutzman, *The Case for Online Obscurity*, 101 Cal. L. Rev. (forthcoming 2013).

[91] *See, e.g.*, Grayson Barber & Frank L. Corrado, *How Transparency Protects Privacy in Government Records* at 29-37 (May 23, 2011), http://ssrn.com/abstract=1850786.

[92] Peter W. Martin, *Rewiring Old Architecture: Why U.S. Courts Have Been So Slow and Uneven in Their Take-up of Digital Technology*, Cornell Law Faculty Working Papers No. 84, at 9 (2011), http://scholarship.law.cornell.edu/clsops_papers/84.

[93] *See, e.g.*, Letter from Joseph I. Lieberman, Chairman of the United States Senate Committee on Homeland Security and Governmental Affairs, to the Honorable Lee. H. Rosenthal, Chair of the Committee on Rules of Practice and Procedure, Judicial Conference of the United States (Feb. 27, 2009), *available at* www.hsgac.senate.gov/imo/media/doc/022709courttransparency.pdf.

3.5 Conclusion

The wide adoption of the RECAP software demonstrates to the Courts that the public has a strong interest in better access policies. By building a public repository, RECAP shows the Courts that providing access need not be as expensive or inaccessible as the technologies they're currently using. By publishing millions of court records, RECAP forces the Courts to deal head-on with their privacy problems, by accelerating the death of practical obscurity and obviating a common excuse to maintain their paywall policy. The paywall—while digital—extends the status quo of the analog world, and serves neither goal of open access nor personal privacy.

In the meantime, RECAP provides significantly increased access to the most frequently-requested court documents. It also enables others—developers, researchers, and other potential users of court data—to demonstrate the many downstream benefits that are possible, if only all of the records were freely available. The introduction of RECAP has also reinvigorated an open government focus on the judicial branch, in a landscape where significantly more attention is typically paid to the executive and legislative branches.

The U.S. Courts are traditionally a slow-moving institution, but they have much to gain by hastening their pace to keep up with digital technologies. There are ample reasons for the Courts to make better use of modern technologies and computer science expertise—from automated redaction of sensitive information, to preserving the authenticity of digital documents, to making their IT infrastructure far more efficient.

But today, the PACER paywall is still a significant barrier to public participation in the U.S. justice system. Online access to court records, through innovative third-party services, has the potential to serve as the spectators' gallery of the 21st century. Without free and open access to the underlying data, it is nearly impossible for developers to build new, useful services for citizens. Digital technologies present our Courts with a key opportunity to advance our Founders' vision of forming a more perfect Union, with equal justice under the law. Opening up free access to all electronic court records would mark a significant step in our collective journey.

Chapter 4

Debugging the United States Code

All our work, our whole life is a matter of semantics, because words are the tools with which we work, the material out of which laws are made, out of which the Constitution was written. Everything depends on our understanding of them.

Felix Frankfurter, Associate Justice, U.S. Supreme Court, 1964[1]

Access to government information forms the foundation of many open government policies. These policies tend to focus on releasing previously undisclosed information to the public,[2] or mandating that already-published information be made available in more accessible formats and mediums.[3] But increased access does not automatically mean that government processes are easier for citizens to understand. In many cases,

---

[1] Felix Frankfurter, *quoted in* Reader's Digest, June 1964, *reprinted in* James B. Simpson's Contemporary Quotations 66 (1998).

[2] For example, President Obama deems the Freedom of Information Act as a central part of open government. *See* Presidential Document, Memorandum of January 21, 2009, Freedom of Information Act, 74 Fed. Reg. 4683 (Jan. 26, 2009), *available at* http://www.whitehouse.gov/the_press_office/Freedom_of_Information_Act (declaring that "[in] our democracy, the Freedom of Information Act (FOIA) . . . is the most prominent expression of a profound national commitment to ensuring an open Government."). Similarly, the Open Government Partnership uses "access to information" as one of its four minimum eligibility criteria for country membership. *See* Eligibility, Open Gov't Partnership, http://www.opengovpartnership.org/eligibility (last visited July 6, 2012) (stating that "[an] access to information law that guarantees the public's right to information and access to government data is essential to the spirit and practice of open government.").

[3] Recent open government initiatives have included extensive open data components. *See infra* §5.

it is not enough to merely make information more available: If citizens are unable to interpret the information and extract useful knowledge—which they can use to make better decisions, participate meaningfully in the public sphere, and hold government officials accountable—democracy suffers.[4] The reality is that many government processes are extremely complex, and it often requires experts to translate the process into terms that lay citizens can understand.[5] Open government initiatives should aim to lower costs, not only to obtaining critical government information, but also to understanding how the process works.

In this Chapter, we examine the United States Congress and the understandability of the federal legislative process. In particular, we study two related subprocesses—drafting and codification—that are essential to the creation of the formal written outputs of Congress. Drafting clear bills requires "great skill, knowledge, and experience."[6] Clear bills produce laws that citizens can more easily follow, and reduces the number of disputes over what the laws mean. After each bill is passed, the process of law codification incorporates the bill into a subject matter compilation, which makes the entire collection of laws more readable. But flaws in these processes—based on decades-old technological assumptions—create inefficiencies and inaccuracies in the broader lawmaking process. In short, the process today is too imprecise, and it hampers the development and use of new technologies that could make Congress itself more efficient, and make it less opaque to the general public.

Historically, Congress has been "the most transparent national government institution."[7] Unlike the executive agencies and the courts, Congress allows much of its legislative activities to be recorded on video, broadcast live on television and streamed online.[8] A significant amount of legislative data is also available online: The Library

---

[4] *See* Josiah Ober, Democracy and Knowledge: Innovation and Learning in Classical Athens 2 (2008) (finding that "[the] history of Athenian popular government shows that making good use of dispersed knowledge is the original source of democracy's strength."); *see also id.* at 218 (stating that the instruments of participatory democracies "must manifest two general properties if it is to work effectively to lower transaction costs: it must be open, and it must be fair. By open, [Ober] mean[s] that the instrument is accessible in respect to entry (as opposed to restricting entry according to extraneous criteria) and clear in respect to interpretation (as opposed, for example, to being interpretable only by insiders "in the know").").

[5] *Id.* at 219 ("The complexity of modern rules, and the technical legal language in which they are cast tend to raise transaction costs.").

[6] John V. Sullivan, How Our Laws Are Made, H.R. Doc. No. 110-49, at 5 (2007), *available at* http://thomas.loc.gov/home/lawsmade.toc.html.

[7] Walter J. Oleszek, Cong. Research Serv., R42108, Congressional Lawmaking: A Perspective on Secrecy and Transparency 1 (2011), *available at* http://www.fas.org/sgp/crs/secrecy/R42108.pdf.

[8] The non-profit station C-SPAN has been broadcasting from the House floor since 1979, and from the Senate floor since 1986. *About the Congressional Chronicle*, C-

of Congress runs a website called THOMAS, which publishes the full text and the current status of bills, the history of recorded votes, committee documents, and a variety of other legislative information.[9] These data have been usefully reassembled by numerous third parties, most notably GovTrack and its progeny,[10] to make the activities in Congress more publicly discernable.

But despite how transparent Congress tries to be, our laws are born through "an exceedingly complex and evolving legislative process—much of it unique to each House of Congress."[11] The Constitution grants each House the ability to "determine the Rules of its Proceedings."[12] The House and the Senate have each established its own set of rules and practices, which have evolved and compounded over the past two centuries. To illustrate, the "official manual of House rules is more than a thousand pages long and is supplemented by more than 25 volumes of precedents, with more volumes to be published in coming years."[13] Tracking the activities of the House

---

SPAN Video Library, http://www.c-spanvideo.org/videoLibrary/aboutCC.php (last visited July 6, 2012). Since 2010, C-SPAN has made its entire video archive available online for free streaming. Brian Stelter, *C-Span Puts Full Archives on the Web*, N.Y. Times, Mar. 15, 2010, https://www.nytimes.com/2010/03/16/arts/television/16cspan.html. The House of Representatives also now runs its own live streaming video service. *House of Representatives Live Video*, H.R. Office of the Clerk, http://houselive.gov (last visited July 6, 2012).

[9] *About Thomas*, Libr. Congress, http://thomas.loc.gov/home/abt_thom.html (last visited July 6, 2012). While THOMAS publishes an extensive amount of data, advocacy groups have recently called significant improvements for THOMAS, including bulk data access. *See* Daniel Schuman, *Improve Public Access to Legislative Information*, Sunlight Foundation Blog (Apr. 10, 2012, 10:00 AM), http://sunlightfoundation.com/blog/2012/04/10/improve-public-access-to-legislative-information.

[10] GovTrack.us: Tracking the U.S. Congress, http://www.govtrack.us (last visited July 6, 2012); *see also Other Websites Reusing GovTrack Data*, GovTrack.us, http://www.govtrack.us/developers/downstream (last visited July 6, 2012).

[11] Sullivan, *supra* note 6, at V.

[12] U.S. Const. art. I, 5, cl. 2. *See also* Sullivan, *supra* note 6, at 3 ("The Constitution authorizes each House to determine the rules of its proceedings. Pursuant to that authority, the House of Representatives adopts its rules anew each Congress, ordinarily on the opening day of the first session. The Senate considers itself a continuing body and operates under continuous standing rules that it amends from time to time.").

[13] Christopher M. Davis, Cong. Research Serv., 95-563, The Legislative Process on the House Floor: An Introduction 1 (2010), *available at* http://www.dtic.mil/dtic/tr/fulltext/u2/a470219.pdf.

may not require understanding the entire rules manual, but it is fair to say that the legislative process is far more complicated than what most citizens are taught, and more intricate than most citizens should reasonably be expected to understand.

Congress produces a voluminous amount of information. For instance, in the 109th Congress, which was the two-year session from 2005 to 2006, members of Congress introduced more than 10,700 bills for consideration.[14] They debated these bills in more than 4,000 committee hearings and meetings, and proposed thousands of legislative amendments to modify these bills during debate.[15] Nearly 1,900 votes were taken on by the full House and Senate, which led to the eventual enactment of 482 bills into law.[16] These new laws generated 7,323 pages of statutory text.[17]

Proposed bills often attempt to amend previously enacted laws. For efficiency reasons, rather than restate an entire amended law to show a few desired changes, bills will succinctly describe the modifications in descriptive terms. Each modification typically takes the form of a command, such as "strike X and insert Y" at a specific citation in existing law.[18] Understanding the effect of each modification, by "reading" the bill, is a manual and tedious process: One must first locate the existing laws being modified, and then apply the proposed changes by hand, one at a time. For a lengthy bill that modifies a large number of existing laws, parsing its cumulative effect can be extremely time consuming. Moreover, understanding individual bills is getting increasingly difficult: In recent years, "Congress is passing fewer bills . . . but the ones it passes are much longer. Omnibus bills, sometimes thousands of pages in length, have gone from rarity to commonplace."[19] As a bill gets longer, the cost of parsing it goes up, making it less likely that regular citizens will be able to understand it, and giving well-funded entities—like lobbyists—a built-in advantage. In the long run, lawyers and judges are left to navigate and debate complicated legislative histories, which raises the cost of participating in the judicial process.

Congress has long recognized the usefulness of "comparative prints" as aids for understanding the effects of a proposed bill.[20] Since 1929 the House has had a rule

---

[14] Norman J. Ornstein, et al., Vital Statistics on Congress 2008, at 124-5 (2008).

[15] *Id.*

[16] *Id.*

[17] *Id.*

[18] Legislative drafting styles vary widely depending on who the author is. More often than not, the professional drafting offices in the House and the Senate play a role in crafting bills, and to the extent practicable, they follow established style guides.

[19] Ornstein, *supra* note 14, at 19.

[20] Richard S. Beth, Cong. Research Serv., RS20617, How Bills Amend Statutes 1 (2008), *available at* http://assets.opencrs.com/rpts/RS20617_20080624.pdf ("This comparative print can be of great aid in ascertaining the intended effect of amendatory legislation.").

known as the "Ramseyer Rule" that requires committee reports to show the "redlined" version of amended or repealed statutes.[21] The Senate has a similar rule called the "Cordon Rule."[22] Committee reports summarize the purpose and scope of bills, and are written when bills are reported out of committee to the full House or Senate.[23]

Despite the value of comparative prints, the Ramseyer and Cordon rules have very little impact on the understanding of legislative proposals at time they are being considered. The rules only require the creation of prints at the conclusion of the committee process, during which bills have undergone much debate and revision. The prints are also created only once, for the formal committee report, rather than dynamically as amendments to the bill are proposed and applied. The static nature of comparative prints reflects the fact that creating them can be a painstaking endeavor, especially for large bills. In fact, to lessen the burden in some cases, the rules to create prints may be waived by unanimous consent or special rule in the House,[24] or "to expedite the business of the Senate."[25] Worse for the public, the availability of committee reports is spotty. They are not published on any regular schedule and are sometimes filed many months after a bill has been reported out of committee.[26]

---

[21] Karen L. Haas, Rules of the House of Representatives, One Hundred Twelfth Cong. 26 (2011), *available at* http://clerk.house.gov/legislative/house-rules.pdf (establishing in Rule VIII(3)(e)(1) the following requirement: "Whenever a committee reports a bill or joint resolution proposing to repeal or amend a statute or part thereof, it shall include in its report or in an accompanying document—(A) the text of a statute or part thereof that is proposed to be repealed; and (B) a comparative print of any part of the bill or joint resolution proposing to amend the statute and of the statute or part thereof proposed to be amended, showing by appropriate typographical devices the omissions and insertions proposed.").

[22] *See* Senate Rule 26.12. Matthew McGowan, Senate Manual, One Hundred Twelfth Cong. 50 (2011), *available at* http://www.gpo.gov/fdsys/pkg/SMAN-112/pdf/SMAN-112.pdf.

[23] Sullivan, *supra* note 6, at 15-16.

[24] *See* Deschler's Precedents of the United States House of Representatives, H.R. Doc. No. 94-661, at 3168 (1994) ("In order to save money, manpower and paper, [a Member] requested unanimous consent that the requirements of the Ramseyer rule be waived . . .").

[25] Elizabeth Rybicki, Cong. Research Serv., 96-305 GOV, Senate Committee Reports: Required Contents 2 (2008), *available at* http://www.judiciary.senate.gov/legislation/upload/CRS-ComReports.pdf.

[26] *About Congressional Reports—FDsys Help*, Gov't Printing Off., http://www.gpo.gov/help/index.html\#about_congressional_reports.htm (last visited July 6, 2012) ("The [Congressional Reports] collection for the current Congress is updated irregularly, as electronic versions of the documents become available."). Political considerations also affect the timing of when reports are filed and pub-

Without easy access to comparative prints, how are citizens expected to parse and understand such complicated bills? Ideally, software could help citizens *automatically* create comparative prints and see "redlined" changes to existing law. But for many reasons that we explore in this Chapter, it is impossible to apply such automated software to Congress today. The difficulties stem from the accumulated traditions and practices of Congress—a process built up over two hundred years, and one that has historically relied on human- rather than computer-processing. The amount and complexity of Congressional information being generated today far exceeds the processing capabilities (and patience) of most citizens. In other words, the cognitive costs of understanding legislative information are currently too high.

These costs are raised further by technical complexities in law codification, which has the effect of obscuring the location of laws. The authoritative version of some laws are found in the U.S. Code, while others are found in the Statutes at Large. Where laws are located is an arbitrary artifact of the codification process. This dichotomy leads to further complexities in legislative drafting, which introduces errors into the law and creates ambiguities in interpreting Congressional intent. Digital technologies could help remedy some of these problems, but this is only possible if Congress first modernizes its legislative process.

This Chapter is organized as follows: In §4.1, we introduce the U.S. Code and the current processes that create it. These laborious processes introduce "bugs" into the Code, which are categorized in §4.2. In §4.3, we propose a new structured approach to drafting and codification that significantly improves the efficiency of the legislative process and the clarity of Congress' formal written outputs. We discuss the practical barriers to implementing our proposed approach in §4.4. We conclude in §4.5.

4.1 The U.S. Code and Positive Law

In order to explain the opportunities for improving Congressional processes, we first need to explain how our federal laws are written, managed and published. Once Congress enacts a bill, and the President approves it, the bill becomes a law. The "enrolled" bill—which is the exact text passed by Congress—is sent to the Archivist of the United States for publication.[27] The Archivist assigns the bill a "public law number," such as Public Law 111-148, which refers to the 148th law passed by the 111th Congress.[28] The law is then published as a "session law" in the *United States Statutes*

---

lished. Notes from the Legislative Transparency Meeting at the Cato Institute (Jan 19. 2011) (on file with author).

[27] SULLIVAN, *supra* note 6, at 52.

[28] *Id.* at 52-53. Congress also passes private laws, which concern only one or a small group of citizens, rather than the general public. (Only two were passed in the 111th Congress.) This Chapter is focuses on public laws.

at Large—a chronological compilation of all laws passed by Congress. The Statutes constitute legal evidence of acts of Congress that are admissible in the courts.[29]

The Statutes at Large, while legal evidence, is not very useful for learning about the law. When looking at a statute, one cannot immediately determine whether it is still in effect, has been repealed, or has been modified by later statutes. The Statutes at Large also provide no help in consolidating the accumulated laws about a given topic—say, for all laws related to copyright. In order to do so, one would need to slog through all previous statutes passed by Congress, to find relevant provisions and manually apply subsequent amendatory provisions, one at a time. Therefore, to better aid understanding, Congress "codifies" the laws, by restating and rearranging them into easier-to-use subject matter titles.[30] The official codification of federal laws is the *United States Code*, which was first published by Congress in 1926.[31] Since 1974, the Office of the Law Revision Counsel (hereinafter OLRC)—an independent non-partisan office in the House of Representatives[32]—has been charged with "develop[ing] and keep[ing] current an official and positive codification of the laws of the United States."[33]

All of the "general and permanent" laws[34] are rearranged by the OLRC into 51 "titles" of the U.S. Code. Each title covers a different subject matter area.[35] For instance, Title 6 contains the laws related "Domestic Security," and Title 7 covers

---

[29] 1 U.S.C. 112 (2012).

[30] In general, codification stabilizes participatory democracies because it "promotes joint action by projecting the intentions of rulemakers into the future." Ober, *supra* note 4, at 212. As a reference point, "[c]odified Athenian legislation helped individual Athenians, and others subject to Athenian rules, to weight the likely costs and benefits of any given action and to be more confident in assessing the risks entailed by their own choices. When the rules of the game are specified and known, the game's players are in a position to make better choices. Yet, when those rules become ossified, or are exploited by strategic actors for socially unproductive purposes, organizational performance suffers." *Id.* at 213.

[31] *Detailed Guide to the United States Code Content and Features*, Off. L. Revision Couns., http://uscodebeta.house.gov/detailed_guide.xhtml (last visited July 6, 2012) [hereinafter *OLRC Guide*].

[32] *See* Office of the Law Revision Counsel, http://uscode.house.gov (last visited, July 6, 2012).

[33] 2 U.S.C. 285(a).

[34] 1 U.S.C. 204(a). The U.S. Code does not contain "[t]emporary laws, such as appropriations acts, and special laws, such as one naming a post office . . ." *Frequently Asked Questions and Glossary*, Off. L. Revision Couns., http://uscodebeta.house.gov/faq.xhtml (last visited July 6, 2012).

[35] For 83 years and from its inception, the U.S. Code only had 50 titles. Title 51, on National Commercial and Space Programs, was added in 2010 when "it became increasingly apparent that a distinct title . . . was needed." Rob Sukol, *Positive*

"Agriculture." Of course, many statutes passed by Congress do not cut cleanly into a single subject area—a statute may relate to both domestic security and agriculture, for example—so the OLRC may "classify" individual provisions, from a single statute, into multiple titles in the Code. The practice of classification is "a matter of opinion and judgment . . . [depending on where] the average user will be most likely to look for [a given provision]."[36] That opinion and judgment is exercised by a small team of 15 highly-trained OLRC attorneys, who make expert determinations about where enacted provisions should end up in the Code.[37] The OLRC is only required to publish the official version of the Code once every six years (with annual supplements), but today, it can usually classify a bill, and update the Code, by the time the bill gains presidential approval.[38]

As a legal research tool, the U.S. Code is incredibly more useful than the Statutes at Large. According to Tobias A. Dorsey, an assistant counsel at the OLRC, those in the legal profession "do not read [the Statutes at Large] anymore. We do not cite to them, we do not quote from them, and—the most recent development—we do not use them in statutory interpretation. . . . we read the United States Code instead."[39] Unlike the Statutes, the Code shows the current consolidated version of the law, by carrying out the amendments made to the law in the order they are passed, rather than simply reciting the original statutes passed by Congress.

But because the statutes are consolidated, the U.S. Code is not quite "real" law: The OLRC makes frequent editorial changes to "fit" statutory text into the U.S. Code.[40] Among other types of changes, internal section references within a statute are fixed to point to their new U.S. Code references.[41] Similarly, a relative date, like a deadline "one year after the enactment of this Act," is translated to its absolute date.[42] These editorial changes are not themselves acts of Congress, and mistakes happen during the consolidation process. Thus, the U.S. Code is only "prima facie"

Law Codification of Space Programs: The Enactment of Title 51, United States Code, 37 J. Space L. 1, 2 (2011).

[36] Charles J. Zinn, *Codification of the Laws*, 45 Law Libr. J. 2, 3 (1952).

[37] Ralph V. Seep, Statement to the Subcommittee on Legislative Branch of the House Committee on Appropriations 2 (Mar. 27, 2012), *available at* https://s3.amazonaws.com/assets.sunlightfoundation.com/policy/papers/Law\%20Revision\%20Counsel\%20Statement\%20FY\%202013\%20H\%20Leg\%20Branch\%20Approps\%202012-0327.pdf.

[38] *Id.* at 1-2.

[39] Tobias A. Dorsey, *Some Reflections on Not Reading the Statutes*, 10 Green Bag 2d 283, 284 (2007).

[40] *See OLRC Guide*, *supra* note 31.

[41] *Id.*

[42] *Id.*

evidence of the law,[43] meaning that if there is an inconsistency between the original provision printed in the Statutes and the restated provision in the Code, the version from the Statutes will prevail.[44] But Congress did not intend for the Code to remain "prima facie" law forever. The Code was supposed to be "a starting point for a title by title revision . . . [and] the enactment of revised titles into positive law, one by one."[45]

A "positive law" title is one which is itself enacted by Congress, making it "real" law. To enact a U.S. Code title into positive law, Congress needs to pass a "positive law codification" bill that simultaneously does two things.[46] First, the bill needs to repeal all of the provisions in the original Statutes that had been previously classified into the title. Once these provisions are repealed, the bill will enact the text of the Code title in its entirety into law. This makes *the title itself* an act of Congress, and gives it full legal authority.

The pace of positive law codification has been extremely slow. The issue is "so low profile that most people, including many members of Congress, have never heard of it."[47] Indeed, Congress took more than 20 years after the U.S. Code was established to enact its first title into positive law—it enacted four titles in 1947.[48] Since then, Congress has enacted additional titles every few years, but "generally did not regard revision work as a priority item."[49] Today, only about half of the Code titles (26 of

---

[43] 1 U.S.C. 204(a). When Congress established the U.S. Code in 1926, it could have designated the Code as binding law, but chose not do so. Congress may have hesitated because of a previous codification experience with the *Revised Statutes of 1874*. The Revised Statutes were largely considered a failure because they contained numerous errors and inaccuracies. *See* Erwin C. Surrency, *The Publication of Federal Laws: A Short History*, 79 Law Libr. J. 469, 478, 479 (1987); *see also* Mary Whisner, *The United States Code, Prima Facie Evidence, and Positive Law*, 101 Law Libr. J. 545, 549, 552 (2009).

[44] *See* Whisner, *supra* note 43, at 546-549.

[45] Michael J. Lynch, *The U.S. Code, the Statutes at Large, and Some Peculiarities of Codification*, 16 Legal Reference Services Q. 69, 71 (1997) (emphasis added).

[46] *Positive Law Codification*, Off. L. Revision Couns., http://uscodebeta.house.gov/codification/legislation.shtml (last visited July 6, 2012) [hereinafter *OLRC Positive Law*].

[47] Peter LeFevre, *Positive Law Codification Will Modernize U.S. Code*, Cong. Blog—The Hill (Sept. 28, 2010, 1:33 PM), http://thehill.com/blogs/congress-blog/judicial/121375-positive-law-codification-will-modernise-us-code.

[48] *See* Whisner, *supra* note 43, at 554.

[49] *Id.* at 554.

51) are positive law titles. The other half remain non-positive law titles, and are still "prima facie" evidence of the law.[50]

While positive law codification has not been at the top of Congress' legislative agenda, its has many important benefits.[51] When the OLRC prepares a codification bill, it makes "major improvements in the organization, clarity, and accessibility of the law."[52] The wording and the style of the title are made more consistent. Related provisions, which were previously classified far apart, are brought closer together. Sections are renumbered, to make citations less complicated[53] and to make room for future statutory growth.[54] Obsolete provisions—such as requirements for reports due decades in the past, or laws that automatically expired ("sunsetted" in legislative parlance) but were never explicitly repealed—are eliminated. Technical mistakes like "typographical errors, misspellings, and punctuation and grammar problems" are corrected.[55] All of these changes add clarity and compactness to the title, which are undoubtedly better for anyone using the Code. Throughout the process, the OLRC is careful only to make technical—rather than substantive—changes "to ensure that the restatement conforms to the understood policy, intent, and purpose of Congress in the original enactments."[56]

But the most significant impact of codification is that Congress is required to make "direct amendments" to positive law titles. Direct amendments propose precise changes to the "statutory text" in the title, *e.g.*, "Title 44, United States Code, is amended by inserting after chapter 35 the following: . . . ."[57] Since the entire statutory text in a positive law title has been enacted by Congress, only Congress is allowed to modify it. By contrast, in non-positive law titles, the statutory text

---

[50] The process of preparing a codification bill is deliberate, normally taking a year or more, and once introduced, may take many more years to be enacted. For example, the bill to enact Title 41 was first introduced in the 108th Congress in 2004. The bill was re-introduced in the 109th, 110th, and 111th Congresses, until it eventually passed seven years later, in 2011. *See* Seep, *supra* note 37, at 3; *see also Positive Law Codification, Title 41, United States Code*, Off. L. Revision Couns., http://uscodebeta.house.gov/codification/t41/index.html (last visited July 6, 2012).

[51] *See* Sukol, *supra* note 35, at 14-15.

[52] LeFevre, *supra* note 47.

[53] As one particularly egregious example, severe growth in conservation law has produced citations such as "16 U.S.C. 460zzz-7." *Id.*

[54] Sukol, *supra* note 35, at 19.

[55] *Id.* at 15.

[56] *OLRC Positive Law*, *supra* note 46.

[57] Lynch, *supra* note 45, at 80 (noting that "[i]f Congress does not specify exactly how [positive law] titles are to be changed, the [OLRC] will not make a change, no matter how obvious it appears.").

is edited at the discretion of the OLRC, which means that decisions about how the Code will change are often made *after the law is passed*.

The bifurcated nature of the Code makes a significant impact on legislative drafting. If a drafter wants to modify a topic covered by one of the 26 positive law titles, he should make an explicit change to the U.S. Code. But for all other topics, he should modify the underlying session law, or simply set out a free-standing legislative provision. Of course, for a complex piece of legislation that covers a wide range of topics, a bill will need to use both drafting styles. This greatly complicates the drafting process, and keeping the distinction straight requires years of legislative experience. If the drafter happens to use the wrong style, a bug is introduced into the U.S. Code.

4.2 Bugs in the U.S. Code

To many programmers, the U.S. Code and other legal codes strike a strong resemblance to computer source code. In the analogy, the Code is the source code trunk, and new laws enacted by Congress are sequential patches to the trunk. Drafting legislation, like writing code, is a very technical matter: Drafting style plays a key role in the Code's clarity, and correct syntax is necessary for the Code to parse properly after the proposed changes are incorporated. And like a complicated computer program that has been developed over many decades by numerous programmers, the Code has gathered cruft and bugs. As former Law Revision Counsel Peter LeFevre observes, the Code has accumulated "obsolete and redundant provisions, archaic and inconsistent language, and statutory errors," which is an "unavoidable result of 85 years of legislation."[58]

As with any labor-intensive human task, whether it be programming or legislative drafting, even the most skilled professionals will inevitably make mistakes. But many of these mistakes can be avoided, through smart uses of technology. For computer programmers, compilers and debuggers are indispensible tools for finding and fixing errors, as programs are being developed. Such tools cannot find all possible bugs, but using them can help make the end product more stable. The U.S. Code does not have analogous tools, and we contend in the next section that such tools would increase the clarity of the Code. However, because of the way the system works today, such tools cannot currently be applied.

The U.S. Code contains various kinds of bugs. As with computer code, bugs in the Code can be categorized roughly as *semantic* and *syntactic* errors.[59] A semantic error can be a *logic* error, where it is obvious from a statute's context and history that the drafter made a mistake (*e.g.*, the statute says "more" when it's clear that

---

[58] LeFevre, *supra* note 47.

[59] The distinction in programming is not always clear cut, especially in the case of interpreted languages. This is also true with the U.S. Code, depending on how strict one wants to be about the optimal precision of legal language."

the drafter really meant "less").[60] Other semantic errors can be *reference* errors (*e.g.*, when the Code refers to a part of the Code that doesn't exist) or *grammar* errors (*e.g.*, lowercase words that should be capitalized, or clearly misplaced punctuation marks).[61] But however obvious these errors are, semantic errors—and deliberate ambiguities in the language of the law—are the domain of the courts. Unless Congress passes a new law to clarify what it meant, it is up to judges to "fix" these semantic errors through statutory interpretation.

Other errors in the U.S. Code are syntactic, where the proposed changes cannot be cleanly executed within the existing statutory text. Syntax errors cause structural problems when the OLRC tries to incorporate them into the Code. One type of syntax error is a *misdirected provision*. These occur when a drafter adds a provision related to a topic covered by positive law, but fails to explicitly modify that title.[62] Here's an example: When Congress decided to give the U.S. Courts authority to charge user fees for PACER, as discussed in §3.1, it passed a statute that included the following free-standing provision:

> "The Judicial Conference shall hereafter prescribe reasonable fees . . . for collection by the courts . . . for access to information available through automatic data processing equipment."[63]

This provision clearly belongs in Title 28 on "Judiciary and Judicial Procedure." Because Title 28 is a positive law title, only Congress (and not the OLRC) is allowed to modify it, but Congress did not explicitly amend the title to insert the provision. If it had, the statute would have stated:

> "Title 28, United States Code, is amended by inserting after chapter X the following: "The Judicial Conference shall hereafter prescribe reasonable fees . . ."."

But the drafter made a mistake. When it came time for the OLRC to classify the provision into the Code, it couldn't simply add it arbitrarily into the statutory text

---

[60] These bugs are commonly known in the legal world as "scrivener's errors," which often require judges to look at legislative intent to avoid absurd results. The doctrine of scrivener's errors has a rich legal history, and is frequently cited as a contradiction to the textualist approach to legal interpretation. *See*, *e.g.*, Michael S. Fried, *A Theory of Scrivener's Error*, 52 Rutgers L. Rev. 589 (2000).

[61] To be clear, in the programming context, this is analogous to a grammar mistake in a string displayed to the user, rather than a syntax error in the grammar of the programming language. The U.S. Code has thousands of obvious grammar errors, which are helpfully noted by the OLRC in footnotes throughout the Code.

[62] Another rare type of misdirected provision happens when the drafter makes an explicit amendment to a non-positive law title, rather than the enacted statute. *See* Lynch, *supra* note 45, at 81.

[63] Pub. L. 102-104 §303(a).

of Title 28—only Congress is allowed to do that. The provision did not fit naturally into any of the non-positive law titles, so the OLRC was left with no better option than to add the provision to Title 28, but as a "statutory note."

Statutory notes are "second class" text that appear after each statutory text section in the Code. Notes are editorially added by the OLRC, and commonly contain metadata about the statutory text above them: effective dates, short titles, relevant regulations, congressional findings, and other miscellaneous information.[64] But when situations like this arise, statutory notes can also contain *actual provisions of law.* Adding laws to statutory notes is an editorial decision made by the OLRC. Notes "can consist of *as much as an entire act . . .* or as little as a clause."[65] And while laws can appear in statutory notes, the provisions are just as valid as the laws in statutory text.[66] Many readers of the Code likely do not realize that this division exists, and may miss important provisions of law that are buried in statutory notes.[67]

Laws in statutory notes are also less easy to work with than those in "first class" statutory text. Notes are not fully citable: The provision above is citable as "28 U.S.C. 1913 note." But even if a note contains an entire act—potentially with many pages of actual legal provisions—note citations do not distinguish among its various provisions. They are all simply a part of the "note." Notes also receive less descriptive metadata in the Code. For example, a provision in statutory text will receive "amendment notes," which describe the provision's legislative history.[68] Statutory notes don't receive amendment notes, which makes it more difficult to research how laws in notes evolve over time.[69]

Another type of syntax error is an *unexecutable provision.* These errors occur when Congress tries to strike words that no longer appear in the Code, or tries to insert language at a location that doesn't exist. In very minor cases, the OLRC will execute the change anyway "to reflect the probable intent of Congress," and will indicate that such a change was made. But in other cases, the OLRC will indicate that

---

[64] *See OLRC Guide*, *supra* note 31; *see also* Lynch, *supra* note 45, at 80 (stating that "[s]ections quoted in notes include "short title" designations, transition provisions, effective dates, funding provisions, and other matters which no one would contend ought to be included in the Code.").

[65] Sukol, *supra* note 35, at 8 (emphasis added).

[66] *Id.* at 8 (stating that "[a] provision of a Federal statute is a law whether the provision appears in the Code as a section text or as a statutory note, and even when it does not appear in the Code at all. The fact that a provision is set out as a note is merely the result of an editorial decision and has no effect on its meaning or validity.").

[67] *See* Lynch, *supra* note 45, at 80 (expressing that "[i]t is astonishing that laws of general significance such as these should be found in the United States Code only in the notes.

[68] *See OLRC Guide*, *supra* note 31.

[69] *Id.* .

the proposed change "could not be executed," and will simply restate the erroneous provision in the notes. The Code includes more than 3,000 unexecutable provisions that clutter the useful content that the notes provide.[70]

Moreover, *duplicate references* are syntax errors that cause two or more provisions to have the same citation. For example, Title 28 has two sections that are numbered 1932. Congress first passed an Act on April 26, 1996, adding a new section 1932 to the title.[71] Seven months later, it passed another Act that enacted a different provision, also section 1932.[72] Rather than clobbering the first provision, the OLRC decided that two sections 1932 would exist side-by-side, with a footnote indicating that two such numbered sections exist. But this solution does not ameliorate the ambiguity when "28 U.S.C. 1932" is cited. More egregiously, Title 26 (the "tax code") has *three* different provisions citable as §6104(6). There are at least 90 pairs (and one triple) of duplicate references in the Code.[73]

Many of these errors result from the manual nature of legislative drafting. Most bills introduced in Congress today are originally drafted by the Offices of the Legislative Counsel in the House and the Senate.[74] The non-partisan offices employ professional drafters who specialize in "clear, concise, and legally effective legislative language."[75] The offices use specialized drafting software called XMetaL[76] to help drafters conform to a standardized drafting style.[77] But during the legislative process, bills can go through significant changes, and bills are not systematically reviewed

---

[70] These provisions were counted by searching the statutory and editorial notes for the standard phrases used by the OLRC to indicate an unexecutable provision. The phrase "to reflect the probable intent of Congress" appeared 543 times in the 2009 edition of the Code. The phrase "could not be executed" appeared 2462 times.

[71] Omnibus Consolidated Rescissions and Appropriations Act, Pub. L. 104-134 §809(a) (1996).

[72] Federal Courts Improvement Act of 1996, Pub. L. 104-317 §403(a)(1) (1996).

[73] These pairs (and one triple) were found by searching all footnotes in the 2009 edition for "two.*have been enacted" (and "three.*have been enacted") and verifying by hand.

[74] Office of the Legislative Counsel, United States Senate, http://slc.senate.gov (last visited July 6, 2012) [hereinafter SOLC]. *See also* Office of the Legislative Counsel, U.S. House of Representatives, http://www.house.gov/legcoun (last visited July 6, 2012).

[75] SOLC, *supra* note 74.

[76] *Drafting Legislation Using XML at the U.S. House of Representatives*, U.S. House of Representatives, http://xml.house.gov/drafting.htm (last visited July 6, 2012).

[77] Manual on Drafting Style, H.R. Off. Legis. Couns., H.R. Doc. No. HLC 104-1 (1995), *available at* http://www.house.gov/legcoun/pdf/draftstyle.pdf.

for validity before they are passed. In the most extreme case, bills can be "drafted on the floor" in the Senate, where there are "virtually no rules . . . no chance to do legal research . . . [and] no 'adult supervision'."[78] Furthermore, bill drafts can originate outside of Congress, and are regularly written by the executive agencies, the White House or corporate lobbyists.[79] Because members of Congress can introduce a bill simply by dropping it into the chamber's "hopper,"[80] the style of bills, and the modifications they make, do not need to conform to any strict style.

Another source of errors is the gradual speed in which the U.S. Code is updated. Some bugs, like duplicate references, are classic concurrency bugs in computing. Either the second Act was drafted before the first Act was passed (and the second Act was not updated to reflect the changes made by the first), or the second Act was drafted using an out-of-date version of the Code. The latter case is possible because of delays between the time a statute is passed, and the time the Code is updated. Historically, the U.S. Code was only officially published once every six years, and although the Code is updated more frequently today, updates are still not instantaneous. Therefore, in order to maintain the most up-to-date version of the Code, each working copy needs to be independently updated. Errors in synchronization create out-of-date working copies, which contribute to downstream drafting errors.

Errors in drafting garble our ability to understand the U.S. Code. Laws are awkwardly crammed into statutory notes, which imposes a real cost on those trying to understand a section of the law. Readers are forced to look not only at the statutory text, but also at all of the statutory notes that follow the text. The amount of statutory notes is significant: Of the Code's 22 million words, about half are in statutory notes. In addition, the peculiarities of the Code's structure are not well understood by legal researchers and those in Congress, let alone the general public, and they lead to confusion and misunderstandings of the law. As law scholar Will Tress puts it, "[t]he United States Code we have today is a monumental, complex and confusing work, rooted in print technology and shaped by the struggles with codification over the last century and a half."[81]

The U.S. Code, and the legislative drafting process, has much to benefit from modern digital technologies. Like computer programs that can be tested (though far from perfectly) before release, changes to the U.S. Code could tested as they are being developed. Many of the syntactic errors described above could be detected statically, as either the proposed bill or the underlying Code changes. At "check-in" time—the instant before the bill is passed—legislators could verify that the bill is valid. This could significantly reduce the number of bugs that wind up in the Code, if only the legislative process could accommodate such features.

---

[78] Victoria F. Nourse & Jane Schacter, *The Politics of Legislative Drafting: A Congressional Case Study*, 77 NYU L. Rev. 575, 592 (2002).

[79] *Id.* at 587.

[80] Sullivan, *supra* note 6, at 8.

[81] Will Tress, *Lost Laws: What We Can't Find in the United States Code*, 40 Golden Gate L. Rev. 129, 162 (2010).

4.3 Designing a Structured U.S. Code

If the legislative drafting process were redesigned from scratch today, it would likely resemble the process of modern collaborative software development. Modern software development takes advantage of powerful tools that help developers collaborate to build complex pieces of stable software. Integrated development environments, for example, provide efficient interfaces that help developers write syntactically valid code. In addition, revision control systems are essential when large numbers of developers (like 535 members of Congress, thousands of their staff, and outside stakeholders) are simultaneously editing the same codebase (like the U.S. Code). They allow developers to work independently to craft new features, and to merge these updates together in a smooth and conflict-free way. They also track the sequential history of changes to the source code over time, which helps observers understand how the code has evolved. In this section, we postulate a clean slate design to the legislative process, using the lessons learned from software development, to demonstrate the potential for fundamental improvements to the U.S. Code.

In our design, every title of the U.S. Code is a positive law title, and all general and permanent laws appear in the Code. The structure of the Code is a strictly-enforced tree hierarchy, where the root of the Code is divided into titles, which are further divided into sections, subsections, paragraphs, subparagraphs, clauses, and subclauses. Each node in the tree contains a single text node and any number of child nodes. With this structured hierarchy, all text in the Code is uniquely addressable at the word-level, similar to the way that XPath can query an XML document. For example, 17 U.S.C. 512(b)(1)(C) can be referenced as `/17/512/b/1/C`, and a ¡substring of text be referenced by specifying the range of individual words: `/17/512/b/1/C/text@0:2` retrieves the first two words of the subparagraph.

With the above assumptions in place, the process of drafting legislation would be much simplified. Any general and permanent change to federal statutory laws would be made using a direct amendment to the Code. Drafters would no longer have to refer to individual acts in the Statutes at Large, and while enacted bills would still be compiled in the Statutes, their *effect* is always reflected in the Code itself. In turn, the Statutes essentially act as the chronological compilation of "check-ins" which records the revision history of the Code.

Drafters would acquire a new set of tools to compose legislation. Rather than manually describing in a bill how the Code should change, in descriptive English sentences, new tools would present drafters with a WYSIWYG (what-you-see-is-what-you-get) environment to directly modify the language of the Code, similar to using Microsoft Word with "track changes" enabled. Once the Code is modified to the drafter's liking, a bill that accurately describes the desired changes to the Code could be automatically generated.

A bill would be specified in machine-readable format, as a list of desired changes in the order to be executed. Each change takes the form of one of three possible commands:

- `insert` [U.S. Code reference] [string]

Example: `insert /17/118/b/3/text@31 ‘‘owners of’’`

This command inserts the string “owners of” beginning at the 31st word of 17 U.S.C. 118(b)(3).

- `delete` [U.S. Code reference]

  Example: `delete /17/119/b/4/C/text@18:37`

  This command deletes the substring in word range [18, 37) of 17 U.S.C. 119(b)(4)(C).

- `replace` [U.S. Code reference] [string]

  Example: `replace /17/111/d/2/text@74:83 ‘‘upon authorization’’`

  This command replaces the substring in word range [74, 83) of 17 U.S.C. 111(d)(2) with the string “upon authorization.”

While one goal of our approach is to make proposed bills more understandable, auto-generated bills would be (counterintuitively) less readable as a list of raw commands. However, by simply reversing the process in which the bill was auto-generated (*i.e.*, applying the bill to the Code), the effect of the bill within the full context of the Code can again easily be seen. Moreover, if one required a human-readable version of the bill, it would be simple to write a program to convert the machine-readable bill to English phrases. For example, the three example commands above could be automatically converted to:

```
Title 17, United States Code is amended--
    (1) in section 118(b) paragraph (3)
        (A) by inserting ‘‘owners of’’ before ‘‘copyright’’,
    (2) in section 119(b) paragraph (4)
        (A) by deleting in subparagraph (C) ‘‘withhold from
            distribution an amount sufficient to satisfy all
            claims with respect to which a controversy exists,
            but shall’’, and
    (3) in section 111(d) paragraph (2)
        (A) by replacing ‘‘in the event no controversy over
            distribution exists, or’’ with ‘‘upon authorization’’.
```

While generating human-readable versions of bills would be a straightforward task, we imagine that few people would actually want to read bills this way, given that comparative prints would be just as easy to generate.

When a member of Congress proposes a bill, she introduces the machine-readable version (perhaps by dropping a USB key with the proposed bill into the chamber’s “hopper,” or better yet, uploading the bill to an electronic hopper). These bills would be made available instantaneously, in real time online, for both Congressional staff and the public.

Throughout the deliberative process, automated programs would continuously check that proposed bills remain valid against the latest version of the Code. As

Congress enacts new bills, changes to the Code could potentially "break" pending bills that were drafted before the new enactments. Using operational transformations (OT),[82] programs would fix broken bills by rebasing its proposed changes onto the latest version of the Code.

To demonstrate, suppose that a U.S. Code provision, call it Title 17 section 555(a), initially states that: "The person shall not be liable for monetary damages." Representative A introduces a bill that would shield the person from all damages: `delete /17/555/a@7` (which deletes the word "monetary"). Meanwhile, Representative B introduces a separate bill that would additionally protect the person's employer: `insert /17/555/a@2 ‘‘or the person’s employer’’`. Rep. B's bill passes Congress, which changes the Code provision to: "The person or the person's employer shall not be liable for monetary damages." However, the passage of Rep. B's bill broke Rep. A's bill: Her `delete` command removes the 7th (zero-indexed) word—now the word "not"—which is clearly not what she intended. A program using OT could automatically help to fix her bill, with respect to Rep. B's enacted bill, to retain its original modification: `delete /17/555/a@11`.

While many syntactic fixes would be straightforward, these automated changes would still warrant human semantic review. In the above example, Rep. B's bill has slightly changed the meaning of Rep. A's bill by also shielding employers from nonmonetary damages. In some cases, programs could try to auto-detect newly introduced semantic problems, such as references to Code provisions that were moved or deleted, to help the drafter maintain the bill's intended meaning. Programs could also use OT algorithms to help drafters collaborate, such as automating the process of combining two proposed bills into a single piece of legislation that both drafters could support.

Drafting amendments to bills could be handled in a similar manner. Both bills and amendments could be given unique identifiers,[83] and an amendment (or an amendment to an amendment) could specify the identifier of the bill or amendment that it is immediately modifying. This would create a decentralized tree of proposed legislation, in which an amendment node is dependant on the other nodes in its path to the tree's root.[84] A proposed amendment in the tree could be kept up-to-date, even as other amendments affecting its ancestor nodes are voted on and passed.

Just before any vote is taken to pass a bill or amendment, legislators would verify that the proposal is syntactically valid. If it isn't, the vote should not take place until the proposal is fixed. Each chamber would need to bind itself to such a rule, which would ensure that the Code remains free from syntactic errors.

---

[82] *See generally*, Clarence A. Ellis & Simon Gibbs, *Concurrency Control in Groupware Systems*, 18 ACM SIGMOD Record, no. 2, June 1989, at 399.

[83] For example, by assigning each draft a randomly chosen large number with negligible probability of collision.

[84] The tree's root is always the U.S. Code. The root's children are bills, whose children are amendments, and so on.

4.3.1 Benefits of a Structured Approach

A redesigned process with these structural features offers three primary advantages.

First, all of the general and permanent laws passed by Congress would be available in one place, in the statutory text of the U.S. Code. Anyone trying to understand the law would no longer need to understand the intricacies of positive law codification, nor search the statutory notes and the Statutes at Large to gain a comprehensive understanding of a certain segment of law. Since new statutes are validated before passage, the possibility of enacting unexecutable amendments or other technical drafting mistakes is eliminated.

Second, it becomes much easier to understand Congress in real time, as soon as bills are introduced. The time-consuming process of redlining the Code to understand a bill now happens immediately and automatically. By lowering the manual labor required to understand Congressional activities, more people will be able to afford to track, and actively participate, in the legislative process. Machine-readable statutes would enable the development of point-in-time systems, that could display the version of the Code from any arbitrary date, rather than relying on the OLRC's publishing schedule.

Third, our approach makes legislative drafting far more efficient than it is today. Currently, a drafter needs to focus on not only the substantive legislative matter, but also the technicalities of drafting a proper bill that creates the intended textual changes in the law. The complexity of drafting is particularly acute with higher-order amendments to other pending provisions. Drafting tools made possible by our proposed design would relieve drafters of the burden of ensuring that insertions and deletions are properly described. This allows drafters to spend more time focusing on the substantive effects of their policy changes, and to collaborate in a more versatile and dynamic way with their colleagues. Furthermore, the art of drafting a proper bill is currently only understood by a select few professional drafters. With the new design, anyone can draft technically-sound changes to the Code, whether it be a member of Congress or an individual citizen. By opening up the drafting process, the cost of proposing policy ideas is significantly decreased, which could give rise to a robust marketplace for concrete legislative ideas.

4.3.2 A Proof-of-Concept

Because the processes for drafting and codification lack sufficient structure, Congress cannot fully realize all of these benefits today. However, we can build a proof-of-concept that—to a limited extent—demonstrates the value of our approach. In one direction, the proof-of-concept could be shown *prospectively*, by creating machine-readable versions of current bills pending in Congress, and automatically generating comparative prints of the U.S. Code. There are two main obstacles to the prospective approach. First, it is often impossible to create precise machine-readable versions of actual bills. Some bill provisions modify other statutes rather than modifying the Code directly, and other free-standing provisions don't specify how they will

eventually modify the Code.[85] Second, for bills that *could* be converted precisely into a machine-readable format (*i.e.*, those that only modify positive law titles), we would need to manually translate the natural language bill into its equivalent machine-readable form—a process that would be quite tedious.

Given these difficulties, we instead built a proof-of-concept that analyzes the Code *retrospectively*, by using available yearly snapshots of the Code as a starting point. By comparing two different snapshots of the same title, we can automatically generate machine-readable commands that are logically equivalent to the changes actually passed by Congress between the two snapshots. Since Congress only makes direct amendments to positive law titles, the proof-of-concept is more compelling for those titles, because the auto-generated commands actually map to provisions in bills previously passed by Congress.

The U.S. Code Parser

The OLRC publishes an unofficial, semi-structured version of the U.S. Code in XHTML format.[86] These files are created by the OLRC through a conversion process from the Code's legacy format—the Government Printing Office's "photocomposition codes"[87]—which specifies how documents should be typeset for physical printing. While the XHTML is geared toward online publication, the markup of the file still only describes the indentation and formatting of the text, rather than its logical components.

We developed a parser that restructures the XHTML of Title 17 (the positive law title on "Copyrights") into an XML format that implements the proposed tree structure described above.[88] We chose to parse Title 17 because its overall organization is relatively consistent. The parser uses the indentation and formatting cues from the XHTML file to figure out what the correct logical structure of the title should be. As the parser traverses the source document, it extracts the subdivision labels for each provision (*e.g.*, a label "(A)" that prefixes a subparagraph) and creates XML elements with the labels as tag names.[89] Based on the logical ordering and nesting of

---

[85] Free-standing provisions are editorially incorporated by the OLRC into the non-positive law titles of the U.S. Code (or in the case of a drafting error, into the statutory notes of a positive law title).

[86] Directory listing of the U.S. Code as XHTML files, http://uscode.house.gov/xhtml (last visited July 6, 2012).

[87] Elliot Chabot, *Specifications for Converting U.S. Code GPO Photocomposition Codes into XHTML*, Off. Chief Admin. Officer, U.S. House of Representatives (2001), *available at* http://voodoo.law.cornell.edu/lexcraft/uscode/docs/locod_xhtml.html.

[88] The Title 17 parsing packing is approximately 600 lines of Python code.

[89] Some labels could create ambiguous parsing situations. For example, the label "(v)" could represent either the Roman numeral for "5," or the alphabet letter "v." There could be edge cases in which a division has counted up to "(u)," and a

tags, the parser generates a structured XML version of Title 17, which can be queried by citation using XPath in order to fetch specific Code text.

Parsing a U.S. Code title is necessarily a semi-automated and somewhat brute force task. Each Code title may be organized differently, and the organization can even change within the same title, especially for positive law titles.[90] The reason is that the Code text for positive law titles is the product of direct amendments, therefore any single drafter can introduce arbitrary elements that do not conform with its surrounding text. Since drafting styles can vary by drafter, and recommended styles have changed over time, a title's organization can be difficult to predict. Furthermore, because certain parts of the Code have experienced unexpected growth, provisions have been crammed in arbitrary ways into already-small subdivisions of the Code. Thus, the Title 17 parser cannot automatically be applied to other titles of the Code, and will frequently encounter unforeseen anomalies that each need to be manually accounted for.[91] As a result of these obstacles, we did not extend the Title 17 parser to work on all of the other titles.

In addition, the parser does not attempt to structure statutory notes, since provisions included in notes do not follow any standard style. In many cases, the OLRC simply quotes the text from the underlying act, without any usable citations. But even in statutory text, some parts of the Code are difficult to cite precisely. For instance, at the beginning of Title 17, section 101 is an extensive block of indented definitions—spanning four printed pages—without granular citations to individual definitions.[92] While large, unstructured portions of the Code aren't necessarily problematic, they do make that portion of the Code less precise to reference, and more difficult to work with.

The difficulties involved in building a general purpose parser for the U.S. Code demonstrates the capriciousness and complexity of its overall structure. The disorga-

---

subdivision has counted up to roman numeral "(iv)." Then, when the "(v)" label appears, one would need to look at the indentation of the document to determine whether the provision is the next element of the division or the subdivision. In another example, the label "(aa)" is sometimes used as the next element after "(z)" in a long list (*e.g.*, 3 U.S.C. 136(aa)), but in other cases, "(aa)" is the first element in a new deeply-nested subdivision of "double-lettered" labels (*e.g.*, 2 U.S.C. 434(f)(3)(A)(i)(II)(aa)). Neither ambiguous situation occurs in Title 17, but it does suggest that the U.S. Code should be explicitly structured, so its citation logic is not dependant on its physical printed appearance.

[90] The organization of titles is "subdivided into some combination of smaller units such as subtitles, parts, chapters, divisions, sub- chapters, subparts, and sections, [but] not necessarily in that order." Sukol, *supra* note 35, at 7.

[91] As an illustrative example, Title 2 includes section numbers like "31a-2c"—a numbering scheme that was not contemplated by (nor necessary in) the Title 17 parser.

[92] To cite an individual definition, one would likely need to cite "the definition for 'Copyright owner' in 17 U.S.C. 101".

nization is the aftermath of more than half a century of arbitrary legislation, which imposed careless structural changes and errors into the Code. Our proposed design eliminates these flaws and ensure that the Code maintains a consistent, machine-parsable structure.

Automatic Bill Generator

To automatically create a machine-readable bill, we developed a separate *diff* utility that compares two versions of the same U.S. Code title. The utility takes as input two pre-parsed versions of Title 17 from different years (*e.g.*, from 2005 and 2006). The parser performs a tree-based comparison of the two versions, and outputs a "patch" that describes the transformation from the earlier version to the later one. The patch is a machine-readable bill that is logically equivalent to the provisions passed by Congress modifying Title 17 in the intervening time.

The tree-based comparison algorithm is relatively straightforward. Let A be the earlier XML version of the title, and B be the later version. For each input, the algorithm compiles a list of XPath addresses for all nodes in the document, in depth-first order, $X(A)$ and $X(B)$. Then, it computes a "change list" of *deleted* nodes, $C_R = X(A) - X(B)$, and a change list of *added* nodes, $C_A = X(B) - X(A)$. For each node in the intersection $X(A) \cap X(B)$, the algorithm performs a word-level *diff* to compare the text string in each corresponding node, and stores the differences in the change list of *modified* nodes C_M. Finally, the algorithm sorts the combined list $C_R \cup C_A \cup C_M$ in depth-first order, and outputs machine-readable commands—`insert`, `delete`, or `replace` as described earlier—for each change.

In 2006, Congress passed two bills that affected Title 17: Public Laws 109-181 and 109-303.[93] The machine-readable patch for Title 17, from 2005 to 2006, includes 107 commands that coincide precisely with the textual changes made by the two laws.[94]

The tree-based *diff* utility works on any two versions of the same U.S. Code title. Moreover, because the algorithm does not make any assumptions about the title's organization, the utility works on any U.S. Code title regardless of its organization, as long as the input is in a structured XML format.

Human-Readable Bills

Once we were able to generate machine-readable bills, we developed a translator program that automatically converts the bills from machine- to human-readable form, mimicking the style of bills actually passed by Congress. The program attempts to

---

[93] *U.S. Code Classification Table, 109th, 2nd Session*, Off. L. Revision Couns., *available at* http://uscode.house.gov/classification/tbl109cd_2nd.htm (last visited July 6, 2012).

[94] We verified this by hand, by mapping each command to specific legislative language in one of the two bills.

follow the general drafting style of the House Drafting Manual.[95] The example below shows the beginning of the Title 17 machine-readable bill between 2005 and 2006:

```
Title 17, United States Code is amended--

  (1) in section 111(d)--
    (A) in paragraph (2)--
      (i) by striking "in the event no controversy over
          distribution exists, or" and inserting "upon
          authorization"
      (ii) by striking "in the event a controversy over
          such distribution exists." after "Copyright
          Royalty Judges."
    (B) in paragraph (4)--
      . . .
```

The program groups together changes that affect the same tree node, and outputs a relatively compact bill whose hierarchy resembles the structured "change tree."

The program also needs to be careful that each English command describes a unique modification. To specify the exact location of a modification, bill drafters include adjacent words to create a unique context for the change. For example, a command may modify a paragraph "by inserting "world" after "hello"." However, if the paragraph contains two instances of "hello," the command is ambiguous as to which "hello" is referred to. Our translator program checks whether the straightforward version of the command is ambiguous, and if it is, it will seek one word at a time, before or after the location of the change, to find a sufficiently unique context for the change.

The inverse of this translator tool—to automatically transform human-readable bills to machine-readable ones—would be highly valuable, but practically impossible to build today for actual bills introduced in Congress. Even the most advanced natural language processing tools today likely can not precisely parse the sheer number of ways that legislative provisions and references can be expressed. What could be built is a far simpler version of that tool, which takes as input a human-readable bill that was generated by our original translator tool. This simplified inverse tool could be useful if a drafter wanted to make minor tweaks to a bill, and re-generate the machine-readable version. Alternatively, a drafter might want to add surrounding text, to describe the purpose of the bill, while still verifying that the effect of the bill remains unchanged. While this tool could be useful in these circumstances, we did not build a prototype.

These proof-of-concept tools demonstrate how parts of the structured legislative process might work. A representative could conceivably introduce a human-readable bill, output by our translator program, on the House floor today. She could create

---

[95] *See generally* H.R. Off. Legis. Couns., *supra* note 77.

this bill by directly modifying the structured Code title (positive law titles only), generating the machine-readable bill by comparing the modified title with the up-to-date title, and translating the bill into English. Correspondingly, a citizen could take the machine-readable bill and apply it to the up-to-date-title, to see the bill's effect in redlined form. These tools would transform both the way Congress works and the way citizens can understand Congress, but until Congress changes some of its practices, these benefits cannot be realized.

4.4 Practical Barriers to Implementation

Implementing our design would require a number of significant changes to the way that Congress works. The barriers to change are considerable, and it would take Congress enormous political will—and a strong grasp of technology—to make these changes. These aren't simply changes to the IT systems that Congress uses; these are fundamental changes to the way that bills and laws are prepared and passed. There are three necessary conditions to achieve a reasonable implementation of our proposal.

All U.S. Code titles must be positive law

Congress has taken 85 years to enact 26 of the 51 titles in the U.S. Code. The OLRC intends to prepare positive law codification bills for the remaining 24 titles,[96] but unless Congress—and more specifically, the Committee on the Judiciary in the House—prioritizes positive law codification, progress will continue to be slow. The Committee needs to be better educated about the benefits of codification, and the OLRC needs more resources to more quickly prepare the remaining codification bills. Some of the outstanding non-positive law titles may be extremely complicated to revise, especially Title 42 on "The Public Health and Welfare," which is notorious for containing a grab bag of miscellaneous laws.[97]

To make matters more complicated, once the remaining 24 titles are enacted, Congress will need to revise and enact all of the titles Code, one more time, to correct existing errors in today's positive law titles. In the long run, Congress will need to "refactor" the Code every once in a while, because of semantic cruft that will inevitably build up over time. Repealed laws might leave empty spaces in the Code, or leave provisions that are related far apart. Or, drafters will make semantic errors that clutter the Code, that should be clarified directly. As suggested by Professor

---

[96] There are only 24 (not 25) other titles since Title 34 is empty. In 1956, all of the provisions in Title 34 on the "Navy" were repealed and moved to Title 10 on the "Armed Forces." 34 U.S.C. (2012).

[97] Title 42 itself accounts for approximately 30% of the size of the entire U.S. Code, comprising 7.6 million words of statutory text and notes, out of 22 million.

Tress, the OLRC could draft "annual corrective bills" that makes technical revisions to improve the Code's clarity.[98]

The structure of bills and the U.S. Code must be enforced

The Constitution permits the House and the Senate to each "determine the Rules of its Proceedings."[99] As such, once the U.S. Code is all positive law, each chamber would need to establish new rules to require that bills conform to the defined machine-readable style. Bill that do not validate against the current version of the Code should not be passed. Moreover, the structure of the Code also needs to be standardized and enforced. Members can currently draft and introduce bills in any style or format,[100] which has caused various titles of the U.S. Code to have different structural hierarchies.[101] Arbitrary structural elements make parts of the Code ambiguous to cite, and they make the Code more laborious to parse programmatically. Only by setting limits on how bills can change the Code will the Code remain structurally sound.

Drafters would still be able to make vague semantic changes to the law. For example, a statute could state that "all laws inconsistent herewith are hereby repealed,"[102] rather than specifically repealing certain provisions of the Code. Similarly, laws could take effect or sunset when a certain measurable threshold is reached, rather than on a fixed calendar date. These ambiguities in the law are not addressed by our design, and it would be up to a judge to decide how these provisions are interpreted.

Temporary laws must be included in the U.S. Code

Up until this point, we have focused on the "general and permanent" laws passed by Congress—that is, the laws that wind up in the Code. However, many important laws are never codified, and thus can only be located in their original form in the Statutes at Large. Temporary laws, such as appropriations acts, may include provisions that are *de facto* permanent, even though they never appear in the Code.[103] The most high-profile example is the so-called Hyde Amendment, which prohibits the federal government from funding abortions. The Amendment has been a "rider" on appropriations acts every year since 1977—included year after year in temporary law—which makes it essentially permanent.[104]

---

98 *See* Tress, *supra* note 81, at 159.

99 SULLIVAN, *supra* note 6, at 3.

100 In the extreme case, a member could introduce a bill by scribbling a legislative provision on a napkin and dropping it in the "hopper."

101 Sukol, *supra* note 35, at 7.

102 *See* Zinn, *supra* note 36, at 4.

103 *See generally*, Tress, *supra* note 81.

104 *Id.* at 155.

Even as far back as 1952, the OLRC has had "a great deal of trouble with the codification of permanent laws that are contained in appropriation acts."[105] When the Code was first created, Congress decided that only permanent laws would be included. At that time, the Code was published every six years, so including single-year provisions would not have made much sense. But today, there is no reason why temporary laws could not be included in the Code. Appropriations bills could be codified immediately after enactment and scheduled for automatic removal after one year. In general, bills that have sunrise and sunset provisions could be automatically scheduled for addition and removal. The temporal aspect of bills adds a second layer of required consistency checking: New enactments must not conflict with already-enacted future laws that are scheduled to sunrise, and they must not reference any provisions that are due to sunset, unless the references also sunset before that time. Validation programs could automatically check for these situations and notify drafters if temporal conflicts would arise.

4.5 Conclusion

On paper, Congress is the most transparent branch of the federal government, but its complexities render much of its public processes opaque to the general public.[106] Anyone can retrieve bills online in near real time, and the U.S. Code and the Statutes at Large are widely available in electronic formats. But trying to convert this slew of important documents into a body of useful knowledge is still a perplexing and time-consuming challenge. Even trying to locate laws in the U.S. Code is a surprisingly complicated and confusing problem.

With modern digital technologies, we can imagine a Congress that works far more efficiently, precisely, and transparently than the one we have today. However, current Congressional processes are ossified in the paper age, and they are running on out-dated assumptions about the information processing capacity of its own institution, and of private individuals. Until Congress modernizes its drafting and codification practices—starting with the full enactment of the U.S. Code—it will not be able to accommodate the most transformative opportunities that digital technologies have to offer.

---

[105] Zinn, *supra* note 36, at 3.

[106] To be sure, much of the work in Congress is deliberately private, with behind-the-scenes negotiations and confidentially circulated drafts, but that is a separate issue. *See generally* Oleszek, *supra* note 7.

Chapter 5

The New Ambiguity of "Open Government"

We now have tools that previous generations of open government advocates couldn't even dream of. . . . But of course, technology isn't some kind of magic wand.

Hillary Clinton, 2012[1]

The Internet's power to make government information more available and useful has, in the last several years, become a topic of keen interest for citizens, scholars, and policymakers alike. In the United States, volunteers and activists have harnessed information that the government puts online in key domains, ranging from the federal legislative branch to local city services, and have created dynamic new tools and interfaces that make the information dramatically more useful to citizens. These new tools have sparked significant academic and popular interest and have begun to prompt a fundamental shift in thinking: Policymakers have begun to consider not only the citizens who may ultimately benefit from government information, but also the third parties who can play a valuable mediating role in getting the information to citizens.

The primary concrete result of this trend is that governments have made a growing range of public sector data available in machine-processable electronic formats that are easier for others to reuse. Information that enhances civic accountability, including pending congressional legislation and federal regulations, is indeed more readily available. But more mundane and practical government information, from bus schedules to restaurant health inspection data, is also being provided in friendlier formats. Such data can be used to improve quality of life and enhance public service delivery, but may have little impact on political accountability.

Recent policy initiatives that promote or reinforce this trend have been described as "open government" projects. These initiatives usually include the provision of

[1] *Remarks at the Open Government Partnership Opening Session*, U.S. Department of State (Apr. 17, 2012), http://www.state.gov/secretary/rm/2012/04/188008.htm.

reusable data as one among a range of steps designed to increase overall governmental transparency. For example, President Obama's Open Government Directive, which was designed to implement the new administration's overall "principles of transparency, participation and collaboration,"[2] instructed executive branch agencies, inter alia, to "publish information online in an open format . . . . An open format is one that is platform independent, machine readable, and made available to the public without restrictions that would impede the re-use of that information."[3] Similarly, the multilateral Open Government Declaration,[4] signed by the United States and seven other countries in September 2011,[5] situates these new technologies of data sharing in the context of political accountability.[6] It begins with an acknowledgment that "people all around the world are demanding more openness in government."[7] Among their promises, the signatories commit to "provide high-value information, including raw data, in a timely manner, in formats that the public can easily locate, understand and use, and in formats that facilitate reuse."[8]

These new "open government" policies have blurred the distinction between the technologies of open data and the politics of open government. Open government and open data can each exist without the other: A government can be an open government, in the sense of being transparent, even if it does not embrace new technology (the key question is whether stakeholders know what they need to know to keep the system honest).[9] And a government can provide open data on politically neutral topics even as it remains deeply opaque and unaccountable. The Hungarian cities of Budapest and Szeged, for example, both provide online, machine-readable transit schedules,[10]

---

[2] Peter R. Orszag, Exec. Office of the President, Memorandum No. M-10-06, Open Government Directive 1 (2009), *available at* http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf.

[3] *Id.* at 2.

[4] *Open Government Declaration*, Open Gov't Partnership (Sept. 2011), http://www.opengovpartnership.org/sites/www.opengovpartnership.org/files/page_files/OGP_Declaration.pdf.

[5] *See* Maria Otero, *On Open Government*, Open Gov't Partnership (Sept. 19, 2011), http://www.opengovpartnership.org/news/open-government.

[6] *See generally Open Government Declaration*, *supra* note 4 (committing to principles related to human rights and good governance and recognizing the opportunities that new technologies offer).

[7] *Id.* at 1.

[8] *Id.*

[9] In the extreme, important political disclosures could be "open government" data even if they were chiseled on stone tablets.

[10] See *List of Publicly-Accessible Transit Data Feeds*, GoogleTransitDataFeed Project, https://code.google.com/p/googletransitdatafeed/wiki/PublicFeeds (last updated June 5, 2012), for a list of more than 150 transit agencies worldwide

allowing Google Maps to route users on local trips. Such data is both open and governmental, but has no bearing on the Hungarian government's troubling lack of accountability. The data may be opening up, but the country itself is "sliding into authoritarianism."[11]

The popular term "open government data" is, therefore, deeply ambiguous—it might mean either of two very different things. If "open government" is a phrase that modifies the noun "data," we are talking about politically important disclosures, whether or not they are delivered by computer. On the other hand, if the words "open" and "government" are separate adjectives modifying "data," we are talking about data that is both easily accessed and government related, but that might or might not be politically important. (Or the term might have a third meaning, as a shorthand reference to the intersection of data meeting both definitions: governmental data that is *both* politically sensitive and computer provided.)

In this Chapter, we acknowledge that this ambiguity may sometimes be beneficial, but ultimately argue that the term "open government" has become too vague to be a useful label in most policy conversations. Open data can be a powerful force for public accountability—it can make existing information easier to analyze, process, and combine than ever before, allowing a new level of public scrutiny. At the same time, open data technologies can also enhance service delivery in any regime, even an opaque one. When policymakers and the public use the same term for both of these important benefits, governments may be able to take credit for increased public accountability simply by delivering open data technology.

In place of this confusion, we offer a stylized framework to consider each of these two questions independently. One dimension describes technology: How is the disclosed data structured, organized, and published? We describe the data itself as being on a spectrum between *adaptable* and *inert*, depending on how easy or hard it is for new actors to make innovative uses of the data. The other dimension describes the actual or anticipated benefits of the data disclosure; the goals of disclosure run on a spectrum between *service delivery* and *public accountability*. This is admittedly a simplification of reality: In practice, many disclosures serve both objectives. However, it is common for one of the two motives to predominate over the other, and we believe this provides a useful starting point for thinking about the competing goals of disclosure.

In Figure 5.1, the vertical axis describes the data itself, and the horizontal axis describes the extent to which service delivery or public accountability predominates as a goal or anticipated result of the disclosure. Along the vertical dimension, there is broad political consensus in favor of adaptable data; but, horizontally, there are

---

that provide their schedule data online to the public, using a standard called the General Transit Feed Specification (GTFS).

[11] Kim Lane Scheppele, *Hungary's Constitutional Revolution*, N.Y. Times Blogs—Paul Krugman, Conscience of a Liberal (Dec. 19, 2011, 10:31 AM), http://krugman.blogs.nytimes.com/2011/12/19/hungarys-constitutional-revolution.

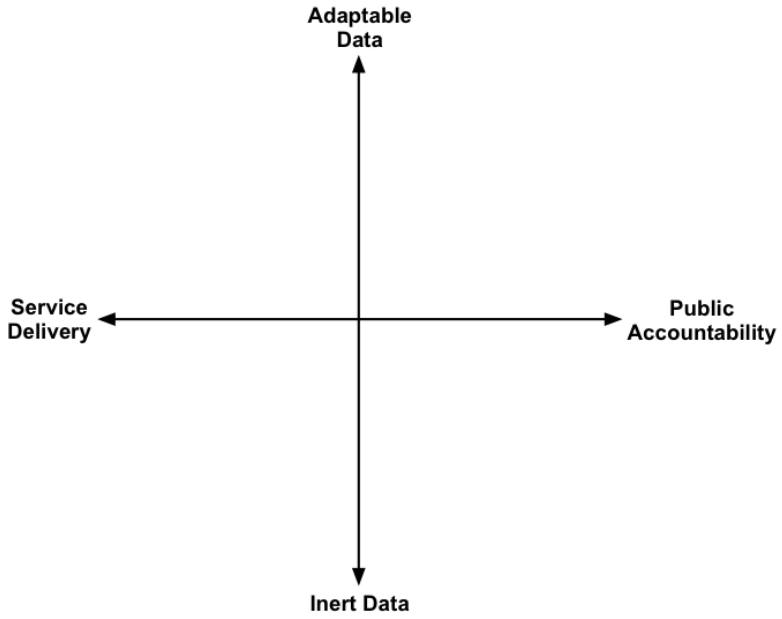

Figure 5.1: Conceptual framework separating the technologies of open data (vertical) from the politics of open government (horizontal).

differences of opinion about the relative political importance of service delivery and public accountability as end goals for public disclosure. (Our discussion in §5.3, below, illustrates these dimensions by populating the graph with examples of concrete public policies.)

We have organized our discussion as follows: §5.1.1 explains the conceptual origins of the relatively modern idea of open government as a public policy, starting with the first recognized use of the term in the mid-twentieth century. The phrase is of fairly recent vintage, but it reflects a particular perspective on the issues it describes—and it was well established before the Internet came into being. §5.1.2, correspondingly, explores the conceptual roots of open data, an idea that has always included, but has always applied far beyond, the kinds of information associated with civic transparency. §5.2 follows the story forward in time, as these concepts begin to merge and give rise to the ambiguous idea of open government data, and details some of the confusions that have ensued in the wake of this ambiguity. §5.3 describes our alternative proposal, which differentiates the widely shared goal of adaptable data from the more controversial choice between enhanced service delivery and enhanced public accountability as the end goals of disclosure.

5.1 Conceptual Origins

Open government and open data each have rich conceptual histories with independent origins. These histories are indispensable tools for understanding the current debate.

5.1.1 Conceptual Origins of Open Government

The idea of open government, as a synonym for public accountability, is part of the peacetime dividend that America reaped after the Second World War. After the war ended, the federal government was left in a state of relative opacity. Having grown accustomed to wartime information restrictions, the federal workforce was "fearful of Cold War spies, intimidated by zealous loyalty investigators within and outside of government, and anxious about" workforce reductions following the war.[12] As a result, "the federal bureaucracy generally was not eager to have its activities and operations disclosed to the public, the press, or other governmental entities."[13]

The opacity surrounding World War II was not, as wartime opacity might be today, a deviation from a clearly established statutory requirement of federal government transparency. Instead, prior to World War II, the key federal law controlling disclosure of government information was the archaic Housekeeping Statute of 1789,[14] which gave "[g]overnment officials general authority to operate their agencies" and withhold records from the public.[15] The Administrative Procedure Act of 1946,[16] while it did contain a general requirement of access to public records, empowered agencies to restrict access "in the public interest," with or without "good cause found"—a faint precursor of the robust justificatory requirements and procedural assurances of modern administrative law.[17]

The period from 1945 to 1955 was a "crucial decade" of early pressure toward greater openness, driven in part by the American Society of Newspaper Editors

---

[12] Harold C. Relyea & Michael W. Kolakowski, Cong. Research Serv., 97-71 GOV, Access to Government Information in the United States, at CRS-2 (2007), *available at* http://www.dtic.mil/dtic/tr/fulltext/u2/a470219.pdf.

[13] *Id.*

[14] *See id.* The Housekeeping Act, ch. 14, 1 Stat. 68 (1789), was first codified at 5 U.S.C. §22. *See generally* 26A Charles Alan Wright et al., Federal Practice and Procedure §5682 (1992) (describing the case law of the housekeeping privilege, which has sometimes been asserted as a basis for executive branch resistance to judicial subpoenas). In 1958, Congress amended the statute to reflect an increasing interest in transparency, adding the sentence, "This section does not authorize withholding information from the public or limiting the availability of records to the public." Act of Aug. 12, 1958, Pub. L. No. 85-619, 72 Stat. 547 (codified as amended at 5 U.S.C. §301).

[15] *See* H.R. Rep. No. 89-1497, at 2-3 (1966), *reprinted in* 1966 U.S.C.C.A.N. 2418, 2419.

[16] Pub. L. No. 79-404, 60 Stat. 237 (1946) (current version at 5 U.S.C. §§551-559, 3105, 7521 (2006)).

[17] *See* Relyea & Kolakowski, *supra* note 12, at 2.

(ASNE).[18] In 1953, ASNE commissioned a report, prepared by a prominent newspaper attorney named Harold Cross, titled *The People's Right to Know: Legal Access to Public Records and Proceedings*.[19] The report's foreword noted that Cross had "written with full understanding of the public stake in open government"[20] —one of the earliest known uses of the term. The report became "the Bible of the press and ultimately a roadmap for Congress regarding freedom of information,"[21] and it served as "a call to battle . . . aimed primarily at the needs of news editors and reporters."[22]

In 1955, the U.S. Congress created the Special Subcommittee on Government Information, also known as the Moss Committee,[23] which incubated the legislation that became the Freedom of Information Act a decade later.[24] Wallace Parks, who served as counsel to the subcommittee,[25] gets credit as the first to expound on the term "open government" in print, thanks to his posthumous 1957 article, *The Open Government Principle: Applying the Right to Know Under the Constitution*.[26] Parks does not explicitly define the term "open government" in the article (in fact, he uses the phrase just four times in twenty-two pages), but his usage makes clear that he sees open government as a matter of accountability:

> From the standpoint of the principles of good government under accepted American political ideas, there can be little question but that *open government* and *information availability* should be the general rule from which exceptions should be made only where there are substantial rights, interests, and considerations requiring secrecy or confidentiality and these are

---

[18] George Penn Kennedy, Advocates of Openness: The Freedom of Information Movement 17-19 (Aug. 1978) (unpublished Ph.D. dissertation, University of Missouri-Columbia) (on file with author).

[19] *Id.* at 31.

[20] James S. Pope, *Foreword* to Harold L. Cross, The People's Right to Know: Legal Access to Public Records and Proceedings, at ix (1953). The *Foreword* was written in October 1952. Pope was the chairman of the American Society of Newspaper Editors' Committee on Freedom of Information and was later the society's president.

[21] Michael R. Lemov, People's Warrior: John Moss and the Fight for Freedom of Information and Consumer Rights 49 (2011).

[22] Kennedy, *supra* note 18, at 31-32.

[23] Congressman John E. Moss, a Democrat from California, chaired the Special Subcommittee on Government Information within the House Committee on Government Operations. *See* Lemov, *supra* note 21, at 50.

[24] *See* Kennedy, *supra* note 18, at 63.

[25] *See* Lemov, *supra* note 21, at 51.

[26] Wallace Parks, *The Open Government Principle: Applying the Right to Know Under the Constitution*, 26 Geo. Wash. L. Rev. 1 (1957).

> held by competent authority to overbalance the *general public interest in openness and availability.*[27]

Parks's thinking, and perhaps his choice of words,[28] was part of a long campaign of legislative pressure that would culminate with the passage of the Freedom of Information Act (FOIA)[29] in 1966. Although President Lyndon B. Johnson "hated the very idea of journalists rummaging in government closets, hated them challenging the authorized view of reality, [and] hated them knowing what he didn't want them to know,"[30] he nonetheless signed the FOIA bill, professing "a deep sense of pride that the United States is an open society in which the people's right to know is cherished and guarded."[31]

Over the next several decades, policy stakeholders used the term "open government" primarily as a synonym for public access to previously undisclosed government information. When Congress amended FOIA in 1974,[32] it noted that "[o]pen government has been recognized as the best insurance that government is being conducted in the public interest."[33] Similarly, the Privacy Act of 1974 aimed to achieve the ideals of "accountability, responsibility, legislative oversight, and open government" together, while respecting citizen privacy in government-held information.[34] Congress also considered open-meeting laws—like the Government in the Sunshine Act,[35] which threw open the doors of federal agency meetings—to be under the umbrella of open

---

[27] *Id.* at 4 (emphasis added).

[28] Parks may actually owe this famous turn of phrase to one of his editors: According to a footnote, Parks passed away unexpectedly eight months before his article was published, and we have found no further record to describe his editors' role in putting the piece together. *See* Parks, *supra* note 26, at 1 n.*.

[29] Pub. L. No. 89-487, 80 Stat. 250 (1966). For a history of the passage of the Act, see generally LEMOV, *supra* note 21, at 53-72.

[30] Bill Moyers, *Is This a Private Fight or Can Anyone Get in It?*, COMMON DREAMS (Feb. 15, 2011), https://www.commondreams.org/view/2011/02/15-7.

[31] *Statement by President Lyndon B. Johnson Upon Signing Pub. L. 89-487 on July 4, 1966*, *in* ATTORNEY GENERAL'S MEMORANDUM ON THE PUBLIC INFORMATION SECTION OF THE ADMINISTRATIVE PROCEDURE ACT (1967), *available at* http://www.justice.gov/oip/67agmemo.htm.

[32] Act of Nov. 21, 1974, Pub. L. No. 93-502, 88 Stat. 1561 (1974) (amending 5 U.S.C. §552).

[33] S. REP. NO. 93-854, at 1 (1974).

[34] S. REP. NO. 93-1183, at 1 (1974).

[35] Pub. L. No. 94-409, 90 Stat. 1241 (1976) (codified as amended at 5 U.S.C. §552(b) (2006)).

government.[36] As the case law of FOIA and related statutes developed through the 1970s and 1980s, federal court decisions began to use the term "open government" as well, likewise referring to governmental transparency.[37]

5.1.2 Conceptual Origins of Open Data

The Internet holds obvious promise as a tool for sharing more data, more widely, than has ever been possible before. Across a wide range of technical fields, the adjective "open" has become a powerful, compact prefix that captures information technologies' transformative potential to enhance the availability and usefulness of information.

Parallel explorations of the possibilities have been unfolding in a number of areas, accelerating in tandem with the growing uptake of the Internet. For example, the Open Access movement aims to make peer-reviewed scientific literature freely available online.[38] The Open Educational Resources campaign seeks to create digital repositories of free learning materials to support global access to knowledge.[39] Open technological standards create pools of patent rights, relieving individual innovators of the need to negotiate patent licenses.[40] The Creative Commons system of copyleft licenses, which makes it easier for creative artists to share and reuse each other's

---

[36] *See* H.R. Rep. No. 94-880, pt. 1, at 39, 1976 U.S.C.C.A.N. 2183, 2210 (considering how well the bill "[b]alanc[es] these three goals . . . (1) open government (2) cutting costs of government and (3) discouraging undue litigation . . .").

[37] *See*, *e.g.*, Bast v. U.S. Dept. of Justice, 665 F.2d 1251, 1253 (D.C. Cir. 1981) ("[T]he importance attributed by Congress to open government is clear, and the Act is designed to resolve most doubts in favor of public disclosure."); Rocap v. Indiek, 539 F.2d 174, 180 (D.C. Cir. 1976) ("[B]y enacting the Freedom of Information Act, Congress determined that the benefits to be derived from open government' outweighed the costs . . . ."); Mobley v. IRS, No. C 77-1693 WWS, 1968 WL 1747, at *6 (N.D. Cal. June 14, 1978) ("[Plaintiffs] have established their right to see what information the IRS has collected on them and thereby affirmed one of the express policies of the FOIA, the right to open government.").

[38] *See*, *e.g.*, Peter Suber, *Open Access to the Scientific Journal Literature*, 1 J. Biology 1 (2002), *available at* http://www.earlham.edu/~peters/writing/jbiol.htm.

[39] *See generally About*, Open Educ. Resources Commons, http://www.oercommons.org/about (last visited June 8, 2012) (explaining that OER Commons "provide[s] support for and build[s] a knowledge base around the use and reuse of open educational resources").

[40] *See*, *e.g.*, Laura DeNardis, *Open Standards and Global Politics*, 13 Int'l J. Comm. L. & Pol'y 168 (2009).

work, aims toward "an Internet full of open content, where users are participants in innovative culture, education, and science."[41]

Similarly, a programmer's decision to release her software under an "open source" license means that the program's source code will be freely available to its users.[42] The phrase "open source" has also, more broadly, become shorthand for the collaborative innovation strategy that underlies many open source software projects—an ethos in which anyone can contribute, abundant scrutiny can help to find and resolve bugs,[43] and a community of creators can take pride in a useful, freely available end product.[44]

---

[41] *About*, Creative Commons, https://creativecommons.org/about (last visited June 8, 2012).

[42] More practically, however, the definition of "open source" from the Open Source Initiative includes a number of other criteria, including redistribution, licensing, and nondiscrimination requirements. *See The Open Source Definition*, Open Source Initiative, http://opensource.org/docs/osd (last visited June 8, 2012).

[43] As Linus Torvalds—creator of the Linux operating system—famously remarked, "Given enough eyeballs, all bugs are shallow." Eric Steven Raymond, *The Cathedral and the Bazaar (Version 3.0)*, CATB.ORG (2000), http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar (internal quotation marks omitted).

[44] Not all open source software is free software, and the usage of these terms has been subject to significant philosophical debate. *See*, *e.g.*, Richard Stallman, *Why Open Source Misses the Point of Free Software*, GNU Operating Sys., https://www.gnu.org/philosophy/open-source-misses-the-point.html (last updated May 18, 2012). Several of the most widely used open source regimes, such as the GNU General Public License (GPL), actually impose additional, stringent conditions; most importantly, the GPL imposes the condition that modified versions of the software must be distributed on the same permissive and noncommercial terms as the original. *See GNU General Public License*, GNU Operating Sys. (June 29, 2007), http://www.gnu.org/licenses/gpl.html. In the license's preamble, the GPL's authors state, "When we speak of free software, we are referring to freedom, not price." *Id.* Other licenses, such as the BSD License, simply require that source code be made available, without restricting commercialization. *See The BSD 2-Clause License*, Open Source Initiative, http://www.opensource.org/licenses/bsd-license.php (last visited June 8, 2012). And still others, such as the Microsoft Public License (MPL), require that if the source code for a licensed program is distributed at all, it must be distributed in full and must be freely available to reuse. These licenses, however, do not require that the source code be distributed—thus allowing for anyone to build commercial, closed-source software that incorporates the MPL-licensed components. *See Open Source Licenses: Microsoft Public License*, Microsoft, http://www.microsoft.com/en-us/openness/licenses.aspx\#MPL (last visited June 8, 2012).

Across each area, there is a common thread: When many individuals or groups are able to access information themselves and interact with it on their own terms (rather than in ways prescribed by others), significant benefits can accrue. Each of these movements is focused on certain classes of information, and each one leverages new technology to make that information more freely and readily available and useful.

The label "open," as applied to various kinds of information, thus inherits both a technological and a philosophical meaning. At a technological level, the term suggests using computers to handle information efficiently in place of manual human processing, greatly extending the range of logistically feasible ways in which information can be used. The extent to which this is possible often turns on technical details, as computers can more readily transform information that is provided in standard, structured formats.

Philosophically, the term suggests participation and engagement—all the people who might benefit from information can share and reuse it in a democratized, accessible way. This implies an absence of legal barriers to innovative new projects, and a larger cultural enthusiasm for innovative and sometimes unexpected developments.[45]

The label "open data" combines both senses of the word "open"—both the term's technological meaning and its philosophical meaning—with a focus on raw, unprocessed information that allows individuals to reach their own conclusions. Before its civic uses, scientists used the term to refer to raw, unprocessed scientific data.

The earliest appearance of the term "open data" in a policy context appears to come from science policy in the 1970s: When international partners helped NASA operate the ground control stations for American satellites, the operative international agreements required those partners to adopt an "open-data policy comparable to that of NASA and other U.S. agencies participating in the program, particularly with respect to the public availability of data."[46] The agreements also required that data be made available to NASA "in the NASA-preferred format."[47]

---

[45] *See*, *e.g.*, The Power of Open, http://thepowerofopen.org (last visited June 8, 2012).

[46] Memorandum of Understanding on Remote Sensing, U.S.-It., May 9, 1974, 26 U.S.T. 3078, 3080 [hereinafter U.S.-It. MOU]. Between 1973 and 1975, the United States concluded similar agreements with a number of other countries. *E.g.*, Memorandum of Understanding on Remote Sensing, U.S.-Chile, Sept. 8, 1975, 26 U.S.T. 3040; Memorandum of Understanding on Remote Sensing, U.S.-Zaire, Jan. 31, 1975, 26 U.S.T. 1699 [hereinafter U.S.-Zaire MOU]; Memorandum of Understanding on Remote Sensing, U.S.-Iran, Oct. 29, 1974, 26 U.S.T. 2936; Memorandum of Understanding on Remote Sensing, Apr. 6, 1973, 24 U.S.T. 897. The language varied slightly from one agreement to the next, but each further assigned to a local partner research organization the responsibility to "ensure unrestricted public availability" of the data "at a fair and reasonable charge based on actual cost." U.S.-Zaire MOU, *supra*, at 1703.

[47] U.S.-It. MOU, *supra* note 46, at 3079.

Later, a 1995 National Academy of Sciences report titled *On the Full and Open Exchange of Scientific Data* elaborated on the idea of sharing data from environmental monitoring satellites, perhaps reflecting its shared lineage with those earlier NASA agreements: "International programs for global change research and environmental monitoring crucially depend on the principle of full and open exchange . . . . Experience has shown that increased access to scientific data, information, and related products has often led to significant scientific discoveries and the opportunity for educational enhancement."[48]

The term "open data" has also appeared in the life sciences context, principally in relation to genetic data. A feature on Jim Kent, the graduate student whose programming work allowed the publicly funded Human Genome Project to finish its work before competing private efforts did, said in part: "Kent's work illustrates the need to think about more than just open source code; in the scientific community there is a growing awareness of the importance of open data."[49]

5.2 "Open Government" Meets "Open Data"

5.2.1 Early Roots of the Convergence

Government data started going online almost as soon as the Internet opened to individual users in the early 1990s. The earliest pioneer was Jim Warren, a sixties radical from Silicon Valley. Warren was well known as the founder of the West Coast Computer Faire, one of the first venues to showcase the personal computer.[50] He was also known as an open government activist, but his particular flavor of transparency had a high-tech twist.[51] In 1993, he "show[ed] California Assembly Member Debra Bowen how public access to state legislative records could be accomplished via the Internet

---

[48] *On the Full and Open Exchange of Scientific Data*, Nat'l Res. Council (Apr. 3, 1995), http://www.nap.edu/readingroom.php?book=exch\&page=summary.html.

[49] Bruce Stewart, *Keeping Genome Data Open: An Interview With Jim Kent*, O'Reilly (Apr. 5, 2002), http://www.oreillynet.com/pub/a/network/2002/04/05/kent.html.

[50] The first West Coast Computer Faire was held in San Francisco in 1977—and it was, at the time, the world's biggest computer trade show. It was at there that Steve Jobs and Steve Wozniak first launched the Apple II personal computer. *See Triumph of the Nerds: The Television Program Transcripts: Part 1*, PBS, http://www.pbs.org/nerds/part1.html (last visited June 8, 2012).

[51] *See, e.g.*, Peter H. Lewis, *Cyberspace Prophets Discuss Their 'Revolution' Face to Face*, N.Y. Times, Aug. 23, 1995, http://www.nytimes.com/1995/08/23/us/cyberspace-prophets-discuss-their-revolution-face-to-face.html (describing Warren as "an advocate for open government").

at low cost and high benefit to the public."[52] Bowen introduced A.B. 1624[53] in March 1993, and Warren "single-handedly launched a crusade to ensure the bill's passage," which succeeded later that year.[54] California became "the first state in the nation to put its legislative information, voting records, and state laws online."[55] Following California's lead, open government advocates in at least a dozen other states began to push similar grassroots proposals.[56]

At the federal level, when the Republicans gained control of Congress in 1994, they enjoyed a fresh opportunity to overhaul that body's infrastructure—the first such opportunity since widespread public use of the Internet began. The website THOMAS, launched in 1995, provided public access to proposed legislation, directory information about members and committees, and daily hearing schedules, among other useful documents.[57] Although today discussions of open government in Congress often begin with THOMAS, the website was not clothed in the language of "open government" at its launch.[58] Before the convergence, "open government" referred narrowly to the initial release of previously undisclosed government information or the effort to get such information released. At its inception, THOMAS simply increased the accessibility of congressional work that was already publicly available.[59] While this increase in accessibility was dramatic, it arguably did not fall within the then-current meaning of the term "open government," because it did not disclose any previously unavailable material.

THOMAS was not what would now be called an open data project either, because the information it provided was accessible only via a government-supplied interface.

---

[52] *Jim Warren*, http://www.svipx.com/pcc/PCCminipages/z2854bc4b.html (last modified May 15, 2001).

[53] 1993 Cal. Stat. 7095.

[54] *See* Press Release, Playboy Found., Computer Columnist and Open-Government Activist Jim Warren to Receive 1994 Hugh M. Hefner First Amendment Award (Oct. 14, 1994), *available at* http://cu-digest.org/CUDS6/cud6.91. For a first-hand account of the battle to pass A.B. 1624, see Interview by Russell D. Hoffman With Jim Warren (June 6, 1995) (transcript available at http://www.animatedsoftware.com/hightech/jimwarre.htm).

[55] *California Legislature Marks 10 Years Online*, Gov't Tech. (Jan. 22, 2004), http://www.govtech.com/e-government/California-Legislature-Marks-10-Years-Online.html.

[56] *See* Jim Warren, *A Once-in-a-Lifetime Opportunity for Real Citizen Access to Government*, Internet Gazette & Multimedia Rev. (Jan. 1995), http://www.kenmccarthy.com/archive/gazette/ig4.html.

[57] *See* Guy Lamolinara, *Congress on the Internet: New Web Server Organizes Online Information*, Libr. Congress (Jan. 23, 1995), http://www.loc.gov/loc/lcib/9502/thomas.html.

[58] *See id.*

[59] *See id.*

The website was designed to serve the needs of citizens—not to open the door for third parties to innovate. By contrast, although they may not have used the term "open data," several other key government offices have long pursued open data policies, providing key data online in machine-readable formats that (unlike THOMAS) facilitate third-party analysis and reuse. The greatest example may be the U.S. Census, which was providing public information through Census.gov as early as 1996.[60]

The first major project to take advantage of open data for an open government purpose—that is, to make data machine readable and accessible in order to promote government transparency and accountability—was OpenSecrets.org, a website that allows users to search and analyze campaign finance disclosures.[61] It launched in 1998 under the auspices of the Center for Responsive Politics, combining government data with third-party innovation. From the beginning, the website aimed to let users adapt the data to their own purposes. On the site's early home page, its creators explained that they planned on "expanding the interactivity of the site, making it possible for you to ask your own questions—how much did the tobacco industry give in the last election, for example, or where does your congressman rank in dollars from labor unions, defense contractors, or phone companies."[62] True to that promise, the site quickly emerged as a powerful and popular tool for members of the public, researchers, and journalists—a role it still enjoys today.

GovTrack.us, a website launched in 2004 as a side project of then-graduate student Joshua Tauberer,[63] was a landmark in the convergence of open government and open data.[64] It focused on the same core information as THOMAS: legislative data about Congress. The website included bills, votes, biographical information on members, and reusable digital maps of congressional districts, and it offered new functionality beyond THOMAS's own for people to search, sort, and monitor legislation of interest to them.

The data in THOMAS was not freely available in bulk at the time of GovTrack's launch—instead, Tauberer had to painstakingly write computer code to systematically scrape and reassemble the data in THOMAS. But once he had reassembled the data for his own use, Tauberer did not keep it to himself. Instead, he made it freely

---

60 *See, e.g., U.S. Census Bureau Home Page*, Internet Archive (Dec. 14, 1996), http://web.archive.org/web/19961227012639/http://www.census.gov.

61 Admittedly, OpenSecrets.org did not itself use the language of "open data" or "open government." *See* Joshua Tauberer, *Big Data Meets Open Government*, *in* Open Government Data (2012), *available at* http://opengovdata.io/2012-02/page/1/big-data-meets-open-government.

62 *Center for Responsive Politics, Open Secrets Interactive Home Page*, Internet Archive (Jan. 10, 1998), http://web.archive.org/web/19980110220043/http://opensecrets.org.

63 *See About GovTrack.us*, GovTrack.us, http://www.govtrack.us/about.xpd (last visited June 8, 2012).

64 *See supra* §2.1.

available, both in bulk and through an application programming interface (API) so that other websites could dynamically access his database and provide up-to-the-minute legislative information themselves, in whatever format or context they judged best. A partial inventory on GovTrack lists at least thirty current and former online projects that rely on GovTrack's data, including prominent sites like OpenCongress and MAPLight.org.[65]

Well into the 2000s, however, the concept of "open government" among public officials still centered on fresh disclosures, rather than improved access to already-public data. The Honest Leadership and Open Government Act of 2007[66] dealt with requirements related to lobbying waiting periods and disclosures, earmark requests, and gifts to Congress. That same year, another law with a similar title, the OPEN Government Act of 2007,[67] modified FOIA's fee structure and established an ombudsman to oversee FOIA's processes. Neither of these bills approached "open government" in the technologically innovative mode of sites like GovTrack.

5.2.2 "Open Government" Becomes a Label for Both Technological Innovation and Political Accountability

In recent years, participants in the policy debate—first in the United States, and then internationally—began to use the term "open government" in a more ambiguous way.

President Obama and his team, both during the campaign and in government, have shown a major commitment to both open government and open data—and they have also been the leading force behind the conceptual merger of the two ideas.

On the campaign trail, then-Senator Obama promised to "restore the American people's trust in their government by making government more open and transparent,"[68] responding in part to his predecessor's perceived lack of transparency. At the same time, the technology and Internet industries based in Silicon Valley served as a key source of financial and logistical support for the campaign, both through their own financial contributions and by helping to build a record-setting, web-based fundraising machine.[69] Obama was no stranger to the power of the Internet: As a Senator, he sponsored the legislation that established USASpending.gov, an online portal that gave Internet users an unprecedented degree of insight into the federal

---

[65] *See Sites That Use GovTrack Data*, GovTrack.us, http://www.govtrack.us/downstream.xpd (last visited June 8, 2012).

[66] Pub. L. No. 110-81, 121 Stat. 735.

[67] Pub. L. No. 110-175, 121 Stat. 2524 (codified in scattered sections of 5 U.S.C.).

[68] *Agenda: Ethics*, CHANGE.GOV, http://change.gov/agenda/ethics_agenda (last visited June 8, 2012).

[69] *See* Joshua Green, *The Amazing Money Machine: How Silicon Valley Made Barack Obama This Year's Hottest Start-Up*, Atlantic Monthly, June 2008, http://www.theatlantic.com/magazine/archive/2008/06/the-amazing-money-machine/6809.

budget.[70] His background as a grassroots organizer also helped him appreciate the power of online networking to connect his supporters with the campaign and with each other.

Alongside their specific policy impulse toward transparency, therefore, the candidate and campaign harbored a powerful, if general, sense that Internet technologies could open doors for innovation, efficiency, and flexibility in government. In effect, this was a commitment to open data. "From a policy standpoint, there [were] many reasons for tech-minded types to support Obama, including his pledge to establish a chief technology officer for the federal government and to radically increase its transparency by making most government data available online."[71] The campaign itself embraced a data-driven approach to its fundraising appeals, rigorously tested alternative fundraising and outreach messages, and devolved to its supporters a significant degree of autonomy in interacting with their friends to build support.[72]

The Obama transition team created a high-level working group on technology and innovation, alongside similar working groups on economics, national security, health care, and other major issues.[73] The group had an ungainly name but an endearing acronym: the Technology, Innovation & Government Reform Policy Working Group, or TIGR (pronounced like Tigger, the friendly tiger from Winnie the Pooh). The group's charter was to help prepare the incoming administration to implement its Innovation Agenda, which included a range of proposals to

> create a 21st century government that is more open and effective; [that] leverages technology to grow the economy, create jobs, and solve our country's most pressing problems; [that] respects the integrity of and renews our commitment to science; and [that] catalyzes active citizenship and partnerships in shared governance with civil society institutions.[74]

This charter was squarely focused on technological innovation rather than on civic accountability.[75]

---

[70] *See* Federal Funding Accountability and Transparency Act of 2006, Pub. L. No. 109-282, 120 Stat. 1186 (codified at 31 U.S.C. §6101).

[71] Green, *supra* note 69.

[72] *See* Daniel Kreiss, Taking Our Country Back: The Crafting of Networked Politics From Howard Dean to Barack Obama (forthcoming 2012), *available at* http://danielkreiss.files.wordpress.com/2010/05/kreiss_takingourcountryback1.pdf.

[73] *See Policy Working Groups*, Change.gov, http://change.gov/learn/policy_working_groups (last visited June 8, 2012).

[74] *Id.*

[75] *See id.* Reflecting this focus, the group's three leaders were former FCC official Blair Levin, Google.org executive Sonal Shah, and Julius Genachowski, whom Obama would later appoint as his FCC chairman. The group included the future

Meanwhile, the communities of technological and political openness had continued to merge outside of government. A key meeting took place in the San Francisco Bay Area a year before the transition team's work.[76] The recommendations drawn up by attendees at the meeting speak in merged terms of "open government data":

> This weekend, 30 open government advocates gathered to develop a set of *principles of open government data*. The meeting . . . was designed to develop a more robust understanding of why open government data is essential to democracy.
>
> . . . .
>
> . . . The group is offering a set of fundamental principles for open government data. By embracing [these] eight principles, governments of the world can become more effective, transparent, and relevant to our lives.
>
> . . . .
>
> Government data shall be considered open if it is made public in a way that complies with the principles below.[77]

---

leaders of what would become the administration's Open Government Initiative: Beth Noveck (who would go on to lead these efforts as Deputy Chief Technology Officer for Open Government) and Vivek Kundra (who would go on to serve as the Chief Information Officer). *See* Jesse Lee, *Transparency and Open Government*, White House blog: Open Gov't Initiative (May 21, 2009, 1:00 PM), http://www.whitehouse.gov/blog/09/05/21/Opening. Noveck is a law professor who has long studied innovative ways to use technology to enhance the governance process. She orchestrated a pilot project for citizens to assist patent examiners in locating prior art and wrote a series of articles on technology-mediated governance. *See* Beth Simone Noveck, *"Peer to Patent": Collective Intelligence, Open Review, and Patent Reform*, 20 Harv. J.L. & Tech. 123 (2006). At the time of the Obama administration's transition, she was finishing a book on technology and governance. *See* Beth Simone Noveck, Wiki Government: How Technology Can Make Government Better, Democracy Stronger, and Citizens More Powerful (2009).

[76] *See* Memorandum From Carl Malamud, Public.Resource.Org, to Attendees of Open Government Working Group Meeting (Oct. 22, 2007), https://public.resource.org/open_government_meeting.html. Malamud (a longtime advocate of putting government data online who led a successful effort to make the SEC filings of public companies freely available online) and Tim O'Reilly (a prominent Silicon Valley publisher and investor) organized the meeting; it received sponsorship from the Sunlight Foundation, Google, and Yahoo. *Id.*

[77] *Request for Comments: Open Government Data Principles*, Public.Resource.Org (Dec. 8, 2007), https://public.resource.org/8_principles.html (emphasis added).

The language here is telling: Participants understood themselves as "open government advocates," but the principles they produced specify circumstances under which "[g]overnment *data* shall be considered open" (emphasis added), rather than government itself. The eight principles, which include completeness, timeliness, and freedom from license restrictions, are requirements that attach to disclosures, not to regimes.[78] It may be true in some sense that a regime becomes more open whenever it provides additional open data, even for mundane and apolitical topics,[79] but it is easy to imagine that a closed regime might disclose large amounts of data conforming to these eight requirements without in any way advancing its actual accountability as a government.[80]

There was also an emerging scholarly literature on the benefits that government might enjoy from fuller use of the Internet, encompassing but reaching well beyond technology-driven enhancements of public accountability. Beth Noveck, who played a leading role in the Obama administration's open government initiatives, wrote a book in this vein arguing not only for transparency, but also for new modes of "collaborative participation" that leverage citizens' expertise.[81] We ourselves made similar arguments in our paper, *Government Data and the Invisible Hand*.[82] There, we advocated for the release of machine-readable, structured government data to help close "the wide gap between the exciting uses of Internet technology by private parties, on the one hand, and the government's lagging technical infrastructure, on the other."[83]

On President Obama's first day in office, he issued two memoranda that dealt with "open government," using the term to refer both to increased transparency and to technological innovation. The first, a memorandum on the Freedom of Information Act,[84] was designed to encourage agencies to be more responsive to FOIA requests. It stated that FOIA

---

[78] *See id.* The remaining five criteria are that the data be primary, accessible, machine processable, nondiscriminatory, and nonproprietary.

[79] *See infra* §5.3.

[80] An electronic release of the propaganda statements made by North Korea's political leadership, for example, might satisfy all eight of these requirements and might not tend to promote any additional transparency or accountability on the part of the notoriously closed and unaccountable regime.

[81] Noveck, *supra* note 75, at 19.

[82] *See supra* §2, *adapted from* David Robinson, Harlan Yu, William P. Zeller, & Edward W. Felten, *Government Data and the Invisible Hand*, 11 Yale J.L. & Tech. 160 (2009).

[83] *Id.* at 161.

[84] Presidential Document, Memorandum of January 21, 2009, Freedom of Information Act, 74 Fed. Reg. 4683 (Jan. 26, 2009), *available at* http://www.whitehouse.gov/the_press_office/Freedom_of_Information_Act [hereinafter FOIA Memo].

> encourages accountability through transparency [and] is the most prominent expression of a profound national commitment to ensuring an *open Government*. . . .
> . . . .
> All agencies should adopt a presumption in favor of disclosure, in order to renew their commitment to the principles embodied in FOIA, and to usher in a *new era of open Government*.[85]

The creators of FOIA, as described above, had political objectives, not technological ones, and this memorandum focuses squarely on those political goals—transparency and accountability.[86] The word "innovation" does not appear, and technology earns a mention not as an end itself, but rather as one of the key means of achieving the political objective: "All agencies should use modern technology to inform citizens . . . . [Future Office of Management and Budget (OMB) guidance should] increase and improve information dissemination to the public, *including through the use of new technologies*."[87]

The second memorandum, on Transparency and Open Government,[88] took a much broader view. Whereas the FOIA memorandum suggested that a "new era of open Government" could be achieved through the transparency that FOIA compliance entails,[89] the Open Government memorandum treated transparency as just one among a trio of goals, setting out in separate paragraphs that an open government is transparent, participatory, and collaborative.[90] Transparency was just *one* of the features of open government, and public trust was just *one* of the benefits: "We will work together to ensure the public trust and establish a system of transparency, public participation, and collaboration. Openness will strengthen our democracy and promote efficiency and effectiveness in Government."[91]

The new administration thus began to move toward a broader conception of open government than had existed before—one that drew on the technological and philosophical commitments to innovation that the word already carried in technical circles. The president's memoranda set the stage for the Open Government Directive and the Initiative that were to follow. Being accountable was just one part of what made a

---

85 *Id.* at 4683 (emphasis added).

86 *See supra* §5.1.1.

87 FOIA Memo, 74 Fed. Reg. at 4683 (emphasis added).

88 Presidential Document, Memorandum of January 21, 2009, Transparency and Open Government, 74 Fed. Reg. 4685 (Jan. 26, 2009), *available at* http://www.whitehouse.gov/the_press_office/Transparency_and_Open_Government [hereinafter Transparency and Open Government Memo].

89 FOIA Memo, 60 Fed. Reg. at 4683.

90 *See* Transparency and Open Government Memo, 74 Fed. Reg. at 4685.

91 *Id.*

government "open"—participatory or collaborative measures that enhanced efficiency or effectiveness might equally claim to be making the government more "open."

The central practical mandate of the Open Government Directive,[92] issued eleven months later, was an open data requirement, not a political transparency requirement: The directive required agencies to "publish online in an open format at least three high-value datasets" via the new federal data portal at Data.gov.[93] High value, in turn, did not necessarily mean data that would "increase agency accountability and responsiveness [or] improve public knowledge of the agency and its operations."[94] Aside from making the agency more transparent or accountable, data might also be high value if it would "further the core mission of the agency"[95] in some way, or "create economic opportunity."[96]

Predictably, agencies responding to this mandate have tended to release data that helps them serve their existing goals without throwing open the doors for uncomfortable increases in public scrutiny. In many cases, agencies published datasets on Data.gov that were already available in other online locations.[97] While agencies packaged some of these datasets into more usable machine-readable formats, critics questioned how these disclosures added to the public's "insight into agency management, deliberations, or results."[98] Critics saw the repackaging of old information as providing only "marginal value" and urged the government to make available "public data that holds an agency accountable for its policy and spending decisions."[99] A broader study of Data.gov in 2011 noted a significant downward trend in agency dataset publication over the site's first year.[100] It concluded that most federal agencies "appear[ed] to cooperate with the program while in fact effectively ignoring it," and that Data.gov had become "the playground for a tiny group of agencies."[101]

---

92 Orszag, *supra* note 2.

93 *Id.* at 2.

94 *Id.* at 7.

95 *Id.* at 7-8.

96 *Id.* at 8.

97 *See* Bill Allison, *Surveying the First Fruits of the Open Government Directive*, Sunlight Found. Reporting Group (Jan. 25, 2010, 5:48 PM), http://reporting.sunlightfoundation.com/2010/data-gov-opinion.

98 Jim Harper, *Grading Agencies' High-Value Data Sets*, Cato@Liberty (Feb. 5, 2010, 12:27 PM), http://www.cato-at-liberty.org/grading-agencies-high-value-data-sets.

99 Letter From Gary Bass, Exec. Dir., OMB Watch, et al., to Vivek Kundra, Fed. Chief Info. Officer (Feb. 3, 2010), *available at* http://www.ombwatch.org/files/info/Kundra-HVD_letterFinal.pdf.

100 Alon Peled, *When Transparency and Collaboration Collide: The USA Open Data Program*, 62 J. Am. Soc'y for Info. Sci. & Tech. 2085, 2088 (2011).

101 *Id.* at 2085, 2088.

Even as the administration's political momentum for its Open Government Initiative waned, local and state governments began to adapt these ideas for their own purposes. From New York to San Francisco, city and state leaders launched new websites devoted to sharing public data, often describing them as "open data" projects.[102] But the rhetoric among localities was more focused on service delivery than on accountability. City leaders in particular put an emphasis on improving communities through better services. San Francisco mayor Gavin Newsom expressed his hope that DataSF.org would "stimulate local industry, create jobs and highlight San Francisco's creative culture and attractiveness as a place to live and work," and only briefly acknowledged the possibility for greater accountability.[103]

Meanwhile, similar ideas have gained momentum internationally, reflecting other nations' growing recognition of the new technological realities. The European Union's 2003 Directive on the Re-use of Public Sector Information instructed that "[w]here possible, documents shall be made available through electronic means,"[104] and the EU now operates a website and program to encourage member states to develop their own national data portals.[105] Independent efforts were underway in the United Kingdom by 2007,[106] leading to the creation in 2008 of a "Power of Information

---

[102] *See*, *e.g.*, *About*, DATA.CA.GOV, http://www.data.ca.gov/about (last visited Apr. 17, 2012) ("The State of California was one of the first states to launch an open data repository. Data.ca.gov was designed to provide a single source of raw data in the state. By posting state government data in raw, machine-readable formats, it can be reformatted and reused in different ways, allowing the public greater access to build custom applications in order to analyze and display the information."); NYC OPENDATA, http://nycopendata.socrata.com (last visited June 8, 2012) ("The data sets are now available as APIs and in a variety of machine-readable formats, making it easier than ever to consume City data and better serve New York City's residents, visitors, developer community and all[.]"); *Open Data*, TEXAS.GOV, http://www.texas.gov/en/Connect/Pages/open-data.aspx (last visited June 8, 2012) (displaying rural-health, school-performance, and other data for the state of Texas).

[103] *See* Gavin Newsom, *San Francisco Opens the City's Data*, TECHCRUNCH (Aug. 19, 2009), http://techcrunch.com/2009/08/19/san-francisco-opens-the-city\%E2\%80\%99s-data.

[104] Directive 2003/98, of the European Parliament and of the Council of 17 November 2003 on the Re-use of Public Sector Information, art. 3, 2003 O.J. (L 345) 94, *available at* http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:345:0090:0096:EN:PDF.

[105] *See* EUR. PUB. SECTOR INFO. PLATFORM, http://epsiplatform.eu (last visited June 8, 2012).

[106] *See* ED MAYO & TOM STEINBERG, THE POWER OF INFORMATION: AN INDEPENDENT REVIEW (2007), *available at* http://www.epractice.eu/files/media/media1300.pdf; *see also* CHANCELLOR OF THE DUCHY OF LANCASTER, THE

Task Force" to explore the benefits of adaptable government data.[107] Data.gov.uk, launched in October 2009, appears to have been the first site of its kind outside the United States.[108]

A new multilateral initiative, instigated by the United States, has dramatically accelerated the spread of these ideas over the past year. In October 2010, President Obama addressed the United Nations General Assembly and urged member states:

> In all parts of the world, we see the promise of innovation to make government more open and accountable. And now, we must build on that progress. And when we gather back here [in 2011], we should bring specific commitments to promote transparency; to fight corruption; to energize civic engagement; and to leverage new technologies so that we strengthen the foundation of freedom in our own countries, while living up to ideals that can light the world.[109]

Following up on this idea, the U.S. State Department organized a series of meetings leading to what became the multilateral Open Government Partnership (OGP).[110] As conditions of entry into the OGP, prospective member countries are required to meet a minimum set of standards that are based on traditional contours of government accountability: timely publication of essential budget documents, an "access-to-information" law that allows the public to obtain key government information, anticorruption disclosure requirements for public officials, and measures to promote citizen participation and engagement.[111] These factors are fundamentally political, so the "open government" goals of the OGP initially appear to be centered on public accountability.

Government's Response to The Power of Information: An Independent Review by Ed Mayo and Tom Steinberg (2007), *available at* http://www.official-documents.gov.uk/document/cm71/7157/7157.pdf.

[107] *See About the Taskforce*, Power Info. Taskforce, http://powerofinformation.wordpress.com/about (last visited June 8, 2012).

[108] *See* Tim Davies, Open Data, Democracy, and Public Sector Reform: A Look at Open Government Data Use From Data.gov.uk (2010), *available at* http://practicalparticipation.co.uk/odi/report/wp-content/uploads/2010/08/How-is-open-government-data-being-used-in-practice.pdf.

[109] Press Release, White House Office of the Press Sec'y, Remarks by the President to the United Nations General Assembly (Sept. 23, 2010), http://www.whitehouse.gov/the-press-office/2010/09/23/remarks-president-united-nations-general-assembly.

[110] *See, e.g., Working Agenda for Open Government Partnership: An International Discussion Meeting of July 12, 2011*, State.gov, http://www.state.gov/documents/organization/167614.pdf (last visited June 8, 2012).

[111] *See* OGP Minimum Eligibility Criteria, Open Gov't Partnership, http://www.opengovpartnership.org/eligibility (last visited June 8, 2012).

However, the Open Government Declaration that OGP member countries sign takes a broader approach toward "openness," as signatories commit to "seeking ways to make their governments more transparent, responsive, accountable, and effective."[112] In addition to transparency and accountability, OGP member countries promise to "uphold the value of openness in our engagement with citizens to improve services, manage public resources, promote innovation, and create safer communities."[113] Thus, the stated goals of the OGP include making governments both more efficient and more accountable, and it remains to be seen how much focus each of these disparate goals will receive. By casting a wide net, the OGP has received the "open government" pledges of more than 55 countries,[114] including historically closed regimes like Russia.[115] The practical impact of such pledges remains to be seen.

The framing value of "open government" has not gone unnoticed in the private sector, either: A growing list of companies have repackaged their government-oriented information technology products under this attractive new label. Microsoft, for example, has created an "Open Government Data Initiative," which promotes the use of Microsoft's Windows Azure online platform as a technological underpinning for open data efforts.[116] Adobe is best known in the government data context as the creator of the PDF document format, which is the baseline digital format for scanned paper documents (and which, like paper, tends to be difficult for downstream innovators to reuse). Notwithstanding the frustrations associated with the PDF format, however, the company undertook a major federal government marketing campaign in 2009 under the tagline "Adobe Opens Up," triggering consternation among some activists.[117] One company, Socrata, has even dedicated itself exclusively to the gov-

---

[112] *Open Government Declaration*, *supra* note 4, at 1.

[113] *Id.*

[114] *See* Maria Otero, *How the Open Government Partnership Can Reshape the World*, Guardian Prof'l—Open Gov't Brasilia 2012 (May 11, 2012, 3:30 AM), http://www.guardian.co.uk/public-leaders-network/blog/2012/may/11/open-government-partnership-reshape-world ("55 countries have committed to taking steps towards openness through OGP.").

[115] *See* Russia, Open Gov't Partnership, http://www.opengovpartnership.org/countries/russia (last visited June 8, 2012).

[116] *See What Is the Open Government Data Initiative?*, Microsoft, http://www.microsoft.com/industry/government/opengovdata/Default.aspx (last visited June 8, 2012).

[117] *See* Clay Johnson, *Adobe Is Bad for Open Government*, Sunlight Labs Blog (Oct. 28, 2009, 12:57 PM), http://sunlightlabs.com/blog/2009/adobe-bad-open-government ("They've spent what seems to be millions of dollars wrapping buses in DC with Adobe marketing materials all designed to tell us how necessary Adobe products are to Obama's Open Government Initiative. . . . Here at the Sunlight Foundation, we spend a lot of time with Adobe's products—mainly trying to reverse the damage that these technolo-

ernmental open data market, with a "Customer Spotlight" on its website that touts its product's adoption by Data.gov, Medicare, the State of Oregon, and the cities of Chicago and Seattle.[118] These businesses have an incentive to sell open data technologies for the broadest range of governmental uses; their decision to brand their efforts in terms of "open government" is powerful evidence of how vague the term has become.

5.2.3 Assessing the Merger

Taken together, these developments have caused a major change in the conceptual landscape: "Open government" policies no longer refer to those that only promote accountability. New modes of citizen engagement and new efficiencies in government services now share the spotlight with the older goal of governmental accountability, which once had this felicitous phrase all to itself.

The shift has real-world consequences, for good and for ill: Policies that *encourage* open government now promote a broader range of good developments, while policies that *require* open government have become more permissive. A government can now fulfill its commitment to be more "open" in a wider variety of ways, which makes such a promise less concrete than it used to be. Whether used as a campaign slogan, in a speech or policy brief, or in a binding national or international policy instrument, the phrase "open government" no longer has the clarity it once had. Existing documents and historical arguments that refer to open government may have lost some of their precision, becoming more ambiguous in retrospect than they were when first authored.

This new ambiguity might be helpful: A government could commit to an open data program for economic reasons—creating, say, a new online clearinghouse for public contracting opportunities—only to discover that the same systems make it easier for observers to document and rectify corruption. In any case, there is much to like about economic opportunity, innovation, and efficiency, and a convenient label could be a good way of promoting them all. Also, the new breadth of the "open

---

gies create when government discloses information. . . . As ubiquitous as a PDF file is, often times they're non-parsable by software, unfindable by search engines, and unreliable if text is extracted."); *see also* Chris Foresman, *Adobe Pushes Flash and PDF for Open Government, Misses Irony*, Ars Technica (Oct. 30 2009, 8:58 AM), http://arstechnica.com/tech-policy/news/2009/10/adobe-pushes-flash-and-pdf-for-open-government-misses-irony.ars ("[W]e can't help but notice how the entire site—designed in [a proprietary Adobe format called] Flash—is practically inaccessible. . . . Wrapping all publicly accessible information in proprietary formats is neither a good nor complete solution. Providing documents in PDF form, or augmenting a website with additional Flash content is certainly useful. However, the goal of open government would be better served using open standards, like HTML, XML, JSON, ODF, and other formats that are both accessible and machine-readable.").

[118] *See* Socrata, http://www.socrata.com (last visited June 8, 2012).

government" label creates a natural cognitive association between civic accountability and the Internet, which may be for the best. Accountability policies that embrace the Internet are often a great deal more effective than those that do not. (It might even make sense to say that if a government is not transparent through the Internet, it is effectively not transparent at all.[119])

But this shift might also allow government officials to placate the public's appetite for accountability by providing less nourishing, politically low-impact substitutes. If the less specific idea of "open government" displaces accountability as the conceptual focus of public reform efforts, less accountability may be achieved.[120]

In April 2011, in response to criticism that its Open Government Initiative was not doing enough for transparency and accountability, the Obama administration launched a new site on "Good Government."[121] The new site focuses on harder-edged issues like shutting down superfluous federal buildings, publicizing the White House visitor logs, and strengthening ethics rules that restrict the lobbying activities of former administration staff.

Meanwhile, the Open Government Initiative and Data.gov appear to be focusing more and more on technological innovation and service delivery. Beth Noveck, who launched and led the program as the U.S. Deputy Chief Technology Officer (CTO) for Open Government, has returned to private life; her successor, Chris Vein, is described instead as the Deputy CTO for Government Innovation, a title seemingly more appropriate to Data.gov's accomplishments.[122]

Noveck herself now regrets the decision to adopt "open government" as the umbrella term for Internet technologies' transformative potential in the public sector:

---

[119] The Sunlight Foundation, a key actor in this area, goes so far as to say it is "committed to improving access to government information by making it available online, indeed redefining public' information as meaning online.' " *Our Mission*, Sunlight Found., http://sunlightfoundation.com/about (last visited June 8, 2012).

[120] *See* Jennifer Shkabatur, *Transparency With(out) Accountability: Open Government in the United States*, 31 Yale L. & Pol'y Rev. (forthcoming 2013) (manuscript at 3-4), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2028656 ("[C]urrent [U.S.] transparency policies do not strengthen public accountability. . . . The existing architecture of online transparency allows [federal] agencies to retain control over regulatory data and thus [to] withhold information that is essential for public accountability purposes; prioritizes quantity over quality of disclosures; and reinforces traditional barriers of access to information. Hence, although public accountability is the raison d'être of online transparency policies, they largely fail to improve it.").

[121] *21st Century Government*, Whitehouse.gov, http://www.whitehouse.gov/21stcenturygov (last visited June 8, 2012).

[122] *OSTP Leadership & Staff*, Whitehouse.gov, http://www.whitehouse.gov/cto (last visited June 8, 2012).

> [T]he White House Open Government Initiative that I directed and the Open Government Directive . . . were never exclusively about making transparent information about the workings of government. . . .
> . . . .
> In retrospect, "open government" was a bad choice. It has generated too much confusion. Many people, even in the White House, still assume that open government means transparency about government.[123]

Instead, she writes, the term was meant to be "a shorthand for open innovation or the idea that working in a transparent, participatory, and collaborative fashion helps improve performance, inform decisionmaking, encourage entrepreneurship, and solve problems more effectively. By working together as [a] team with government in [a] productive fashion, the public can . . . help to foster accountability."[124] She suggests that the new White House structure, with separate focuses for transparency and for public sector innovation, may be more effective.[125]

Notwithstanding a possible change of heart at the White House, however, the ambiguity of open government remains alive and well in the international sphere. In some foreign countries, the need for public accountability is far more acute, and the opportunity cost of deprioritizing it may be far greater. One of the clearest statements of this view comes from Nathaniel Heller, who directs an NGO called Global Integrity and was a key participant in the creation of the Open Government Partnership. He raised the question after Kenya launched an open data website:

> The obvious explanation (in my mind) for why "open data" gets so much attention in the context of "open government" is that it is the sexiest, flashiest reform of the bunch. It's much cooler (and frankly less politically controversial) for any government to put government health databases online . . . than it is for the same government to provide greater transparency around the financing of political parties in the country. . . .
> . . . [O]pen data [may provide] an easy way out for some governments to avoid the much harder, and likely more transformative, open government reforms that should probably be higher up on their lists. . . .
> . . . [W]hen I see the Kenyan government's new open data portal . . . I can only wonder whether the time, expenses, and political capital devoted to building that website were really the best uses of resources. To vastly understate the problem, Kenya has a range of governance and open gov-

---

[123] Beth Simone Noveck, *Defining Open Government*, Cairns Blog (Apr. 14, 2011, 12:57 PM), http://cairns.typepad.com/blog/2011/04/whats-in-a-name-open-gov-we-gov-gov-20-collaborative-government.html.

[124] *Id.*

[125] *Id.*

> ernment challenges that go far beyond the lack of a website where citizens (many of whom are not online) can chart government datasets.[126]

The common thread to these observations is that "open government" is vogue but vague, an agreeable-sounding term with an amorphous meaning. We need better conceptual and linguistic tools, both for keeping governments honest and for exploring the transformative potential of information technologies in civic life.

To some ears, the idea of "open government data" has also developed a more threatening cast. Wikileaks, first launched in 2008, has created what some call "involuntary transparency,"[127] reshaping the conversation over leaks of secret government information to the press.[128] In earlier instances such as the Pentagon Papers, secret government documents reached a single journalist or a small group of journalists, and the public gained access not directly to the secret information itself but instead to the finished journalistic product.[129] The raw material was summarized, adapted, or otherwise filtered before it reached the masses, and sometimes it included changes that reflected the requests of incumbent government officials. Now, however, Wikileaks has made a series of large-scale disclosures of secret government information readily available to individual members of the public, often with little or no redaction of sensitive information. The site has provoked complaints from sources as diverse as the U.S. Department of Defense and Amnesty International, particularly after a trove

---

[126] Nathaniel Heller, *Is Open Data a Good Idea for the Open Government Partnership?*, Global Integrity Commons (Sept. 15, 2011, 12:41 PM), http://www.globalintegrity.org/blog/open-data-for-ogp.

[127] *See* Shkabatur, *supra* note 120, at 37-41 (defining and discussing a category of "involuntary transparency"); *see also* Andy Greenberg, *WikiLeaks' Julian Assange Wants to Spill Your Corporate Secrets*, Forbes, Nov. 29, 2010, http://www.forbes.com/sites/andygreenberg/2010/11/29/wikileaks-julian-assange-wants-to-spill-your-corporate-secrets ("Admire Assange or revile him, he is the prophet of a coming age of involuntary transparency. . . . Long gone are the days when Daniel Ellsberg had to photocopy thousands of Vietnam War documents to leak the Pentagon Papers. Modern whistleblowers, or employees with a grudge, can zip up their troves of incriminating documents on a laptop, USB stick or portable hard drive, spirit them out through personal e-mail accounts or online drop sites—or simply submit them directly to WikiLeaks.").

[128] *See* Curt Hopkins, *ReadWriteWeb's Comprehensive WikiLeaks Timeline (UPDATED)*, ReadWriteWeb (Dec. 29, 2010, 7:02 PM), http://www.readwriteweb.com/archives/readwritewebs_wikileaks_timeline.php.

[129] For a review of the Pentagon Papers case, written in light of the WikiLeaks events, see Tom Kiely, *Pentagon Papers: National Security and Prior Restraint*, 20 HISTORIA 138 (2011), *available at* http://castle.eiu.edu/historia/archives/2011/2011Hostetler.pdf.

of 250,000 unredacted documents—apparently released by accident—put the lives of some foreign supporters of U.S. policy at risk.[130]

But even for voluntary government disclosures, increased privacy risk may be a fundamental objection to these new technologies: The more easily disparate sources of information can be analyzed, combined, and cross-referenced, the greater the chance that previously pseudonymous information can be tied to the identities of particular real people.[131] On the other hand, a rush to limit adaptability to reduce the risk of privacy harms could create a "tragedy of the data commons," in which privacy fears foreclose valuable new insights into public issues.[132]

"Mosaic" risks in national security present an analogous problem: Even if it is not sensitive when considered in isolation, a release of seemingly innocuous data may become useful to America's adversaries if it can be combined to yield sensitive inferences about America's defense and intelligence posture.[133]

Our goal here is not to take a position as to the salience or implications of these risks but rather simply to point out that they can complicate the cost-benefit calculus of the governmental "open data" trend.

5.3 Our Proposal for a Clearer Framing

Clearer language is possible, and it will serve everyone well.

From civic accountability to transit data to health statistics, online disclosures of government data across the world share one exciting feature: They are far more *adaptable* than ever before. Statistics can be mapped, schedules automated, disparate trends cross-referenced, and useful information localized and personalized to a his-

---

[130] *See* Gloria Goodale, *Who Released the Trove of Unredacted WikiLeaks Documents?*, Christian Sci. Monitor, Sept. 1, 2011, http://www.csmonitor.com/USA/2011/0901/Who-released-the-trove-of-unredacted-WikiLeaks-documents.

[131] *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. Rev. 1701, 1701 (2010) ("Computer scientists have recently undermined our faith in the privacy-protecting power of anonymization, the name for techniques that protect the privacy of individuals in large databases by deleting information like names and social security numbers.").

[132] *See* Jane Yakowitz, *Tragedy of the Data Commons*, 25 Harv. J.L. & Tech. 1, 3-4 (2011) ("[P]roposals that inhibit the dissemination of research data dispose of an important public resource without reducing the privacy risks. . . . [I]t is in fact the research data that is now in great need of protection. People have begun to defensively guard anonymized information about themselves. We are witnessing a modern example of a tragedy of the commons.").

[133] *See* David E. Pozen, Note, *The Mosaic Theory, National Security, and the Freedom of Information Act*, 115 Yale L.J. 628 (2005) (drawing attention to the growing use of mosaic claims to deny FOIA requests in the wake of the September 11, 2001 terrorist attacks).

torically unprecedented extent. Online data—particularly if it is structured, machine readable, and available for interested users to download in bulk—can be more readily adapted to new formats, new uses, and new combinations than ever before. Adaptability is independent of subject matter: Any subject—including transit, regulation, schools, crime, or housing—can be a source of data, and that data may be more or less adaptable depending on the format in which it is gathered and presented.

Offline data is very different: They gather dust in filing cabinets, often disorganized and disregarded. An obscure bit of information remains apart from the handful of people who might really benefit from knowing it because it would cost too much to search, sort, or reorganize. Offline data, though available in principle, is physically and psychologically heavy, encumbered by brick and mortar logistics, and tucked away in rooms with limited opening hours. Offline data is *inert*.

Public disclosures thus occupy a spectrum, from the most adaptable data to the most inert. Adaptability may depend on not only the format of the data itself but also on the prevalence and cost of the human and technological capital necessary to take advantage of it.

Disclosures also vary in a second dimension: They differ markedly in their actual or anticipated impact. A machine-readable bus schedule aims to promote convenience, commerce, and a higher quality of life—it enhances *service delivery*. Core political data, such as legislative or campaign finance information, serves a more purely civic role, enhancing *public accountability*. Disclosures of public contracting opportunities play a dual role, potentially enhancing both economic opportunity and public integrity.[134]

Figure 5.2 displays this conceptual model and gives several examples. The vertical axis describes the data itself, in terms of its degree of adaptability. The lateral axis is a continuum from purely pragmatic to purely civic disclosures.

5.4 Conclusion

The vagueness of "open government" has undercut its power. Separating technological from political openness—separating the ideal of adaptable data from that of accountable politics—will make both ideals easier to achieve. Public servants can more readily embrace open data, and realize the full range of its benefits, when it is separated from the contentious politics of accountability. At the same time, political reformers—no longer shoehorned together with technologists—can concentrate their efforts on political accountability, whether or not they rely on new technology. And governments will be less likely to substitute technology initiatives for hard political change.

---

[134] Admittedly, the lateral distinction is a simplication that broadly distinguishes the diverse motivations within open government. Service delivery and public accountability illustrate differences in the political sensitivity of data and whether governments are likely to cooperate with (or oppose) disclosure.

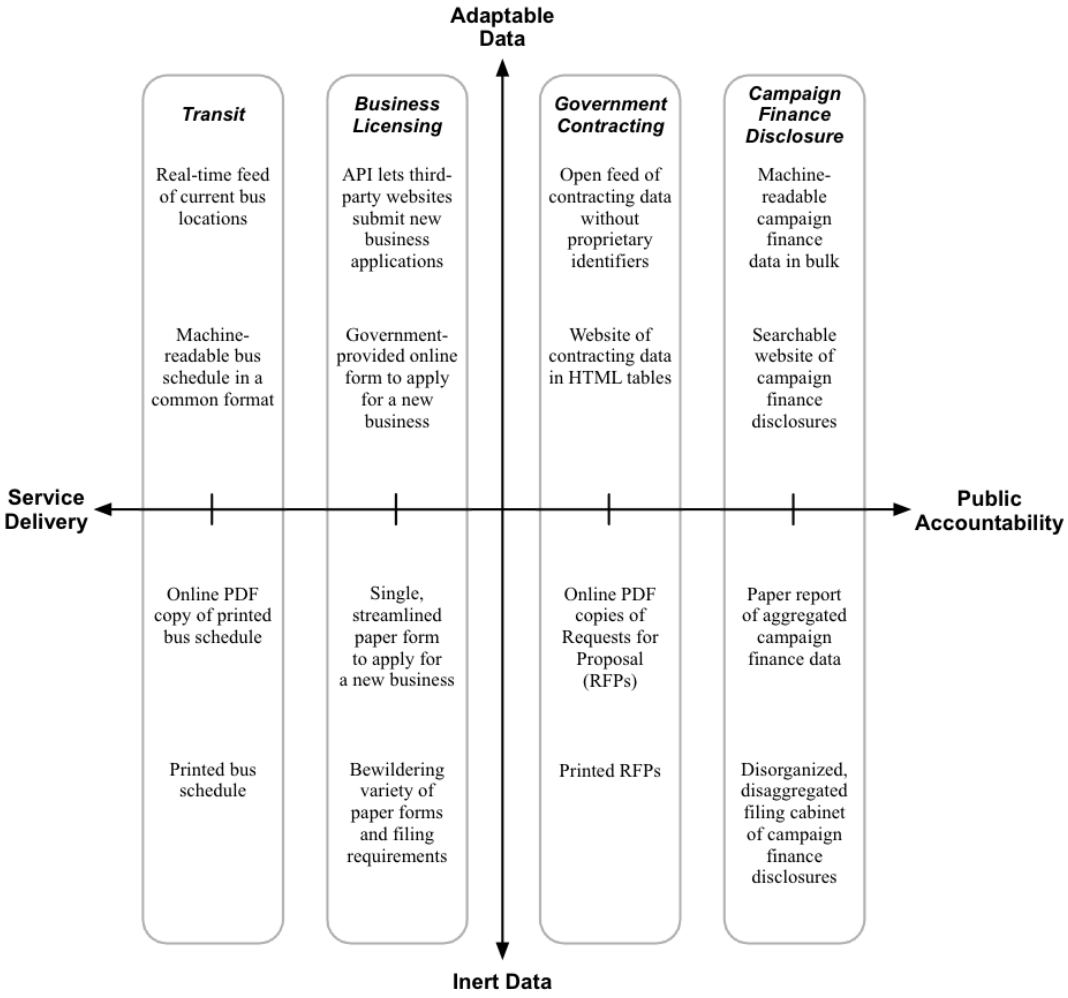

Figure 5.2: Conceptual framework filled with several examples.

Chapter 6

Conclusion

As a political disinfectant, silicon beats sunlight hands down.

Lawrence Lessig, 2009[1]

Governments today have a unique opportunity to radically transform how they interface with their citizens. In this dissertation, we have described how digital technologies have upset government's old-fashioned assumptions about how its information is managed and communicated. Governments can now host vast quantities of public data online at a very low cost, and citizens can instantaneously retrieve the data and discover innovative new ways to use them.

As we have seen, software plays a key mediating role in the interaction between governments and citizens. In the U.S. federal courts, the government's policies surrounding PACER obstruct the widespread dissemination of public court records. We have shown how the design and implementation of RECAP can mitigate the adverse effects of poorly-conceived public policies. Our technical efforts have immediately improved public access to federal court records despite the U.S. Courts' reluctance to dismantle the PACER paywall. By building a public repository, we have demonstrated how even modest increases in accessibility can lead to unexpected downstream civic benefits. The existence of RECAP, and its adoption by thousands of users, serves as a useful policy lever to advocate for positive changes to the PACER policy in front of both the Courts and Congress. However, the long-term goal is to make RECAP obsolete, since our goal is to induce the government to eliminate user fees for public records, rather than running a parallel public mirror that mimics the official repository.

We have also examined how access to information does not guarantee an open and understandable process. From the halls of the U.S. Congress, information is available, but it is difficult to understand. We have shown how idiosyncrasies in the legislative drafting and codification processes have resulted in the complicated and sometimes

[1] Lawrence Lessig, *Against Transparency*, New Republic (Oct. 9, 2009), *available at* http://www.tnr.com/print/article/books-and-arts/against-transparency

definitive body of laws we call the U.S. Code. But things need not be this way: The legislative process is rather similar to the process of software development, and accordingly, the U.S. Code could be managed with systems not unlike revision control for source code. We imagined how a structured legislative design might look, and built software prototypes to demonstrate the efficiency and transparency gains of such a design. While implementing the design would require substantial changes to current legislative procedures, we hope this study entices Congress to begin modernizing its processes.

While this dissertation concentrates on the policies and processes of the U.S. federal government, the methods we used can be exported to other policy contexts. At this new intersection between computer science and public policy, we have presented two prominent examples of how carefully designed technical work can influence traditionally non-technical issue areas.

6.1 Future Work

This specific issues addressed by this work represent just the tip of the iceberg of open government problems that would benefit from attention from computer scientists. One particularly important and challenging problem is the redaction of sensitive information in natural language text. Government data often includes snippets of personal data, or information related to national security, business secrets, or other sensitive topics. Often, important government data is held back from the public because of the high cost of manual redaction. Better automated methods using machine learning techniques could help government officials identify, or at least prioritize, potentially problematic documents. Theoretical work in differential privacy, and practical systems for selective disclosure, could aid the government in balancing openness with privacy and confidentiality.

Other areas of open government could similarly benefit. To help increase public participation efforts, computer scientists could develop new algorithmic strategies to more efficiently aggregate public sentiment into fair public policy priorities.[2] How to optimally implement these strategies, on a variety of devices and in different social settings, would require expertise in human-computer interaction. Related HCI expertise may help governments create more structured data through improved markup interfaces, or make it easier for governments to make its data more accessible to disabled citizens. Research in databases and information retrieval could assist government archivists retain and manage massive amounts of digital government data.

---

[2] *See, e.g.*, Matthew J. Salganik & Karen E.C. Levy, *Wiki Surveys: Open and Quantifiable Social Data Collection* (2012) (unpublished working paper), *available at* http://arxiv.org/abs/1202.0500.

And in our elections, the information security properties of electronic voting systems underpin public accountability and trust in election results.[3]

Many other problems will be of interest to computer scientists, and surely many others have yet to be discovered.

6.2 Final Remarks

At the surface, the role of computer scientists in the field of open government policy may not be obvious. But as this dissertation illustrates, the deeper we dive into the policy issues in open government, the more technical problems we find that require novel solutions based in computer science. Like other applied computer science fields, this research is interdisciplinary. It necessitates a deep understanding of the specific policy problem—who controls the policy, who the policy affects, its historical context, and other important factors—before computing techniques can be successfully applied. If computer scientists can continue to bridge the gap to open government and public policy in general, we can make a real and meaningful impact on the quality of our democracy.

---

[3] *See, e.g.*, Joseph A. Calandrino et al., *Source Code Review of the Diebold Voting System* (2007) (report commissioned as part of the California Secretary of State's Top-To-Bottom Review of California voting systems).

Selected Bibliography

Chapter 1: Introduction

[1] David Robinson, Harlan Yu, William P. Zeller, and Edward W. Felten. Government Data and the Invisible Hand. *Yale Journal of Law and Technology*, 11:160, 2009.

[2] Harlan Yu and David G. Robinson. The New Ambiguity of "Open Government". *UCLA Law Review Discourse*, 59:178, 2012.

[3] Harlan Yu and Stephen Schultze. Using Software to Liberate U.S. Case Law. *ACM XRDS*, 18:12, 2011.

Chapter 2: Government Data and the Invisible Hand

[1] David M. Blei and John D. Lafferty. A Correlated Topic Model of Science. *Annals of Applied Statistics*, 1:17, 2007.

[2] Jerry Brito. Hack, Mash & Peer: Crowdsourcing Government Transparency. *Columbia Science and Technology Law Review*, 9:119, 2007.

[3] Cass R. Sunstein, Office of Management and Budget, Executive Office of the President. Disclosure and Simplification as Regulatory Tools, 2011.

[4] Clay Johnson III, Office of Management and Budget, Executive Office of the President. Memorandum No. M-05-04. Policies for Federal Agency Public Websites, 2004.

[5] Cynthia Farina and Sally Katzen. Achieving the Potential: The Future of Federal e-Rulemaking. *Section of Administrative Law and Regulatory Practice, American Bar Association*, 2008.

[6] Roy Thomas Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. PhD thesis, University of California, Irvine, 2000.

[7] Open Government Working Group. 8 Principles of Open Government Data. http://www.opengovdata.org/home/8principles, 2007.

[8] Kristin Adair, et al. File Not Found: Ten Years After E-FOIA, Most Federal Agencies Are Delinquent. *National Security Archive*, 2007.

[9] John Markoff. A Quest to Get More Court Rulings Online, and Free. *The New York Times*, August 20, 2007.

[10] National Institute of Standards and Technology. FIPS PUB No. 186-2. Digital Signature Standard (DSS), 2000.

[11] Office of Management and Budget, Executive Office of the President. Expanding E-Government: Partnering for a Results Oriented Government, 2004.

[12] Peter R. Orszag, Office of Management and Budget, Executive Office of the President. Memorandum No. M-10-06, Open Government Directive, 2009.

[13] Joshua Tauberer. Case Study: GovTrack.us. In Daniel Lathrop and Laurel Ruma, editors, *Open Government: Collaboration, Transparency, and Participation in Practice*, page 201. 2010.

Chapter 3: RECAP: Turning PACER Around

[1] Administrative Office of the U.S. Courts. Understanding the Federal Courts, 2003.

[2] Administrative Office of the U.S. Courts. Next Generation CM/ECF: Additional Functional Requirements Group Final Report. http://www.uscourts.gov/uscourts/FederalCourts/Publications/ASFRG-Final-Report.pdf, 2012.

[3] Grayson Barber and Frank L. Corrado. How Transparency Protects Privacy in Government Records. http://ssrn.com/abstract=1850786, 2011.

[4] Christina L. Boyd, David A. Hoffman, Zoran Obradovic, and Kosta Ritovski. Building a Taxonomy of Litigation: Clusters of Causes of Action in Federal Complaints. Temple University Legal Studies Research Paper No. 2012-23, 2012.

[5] Amanda Conley, Anupam Datta, Helen Nissenbaum, and Divya Sharma. Sustaining Privacy and Open Justice in the Transition to Online Court Records: A Multidisciplinary Inquiry. *Maryland Law Review*, 71:772, 2012.

[6] James Grimmelmann. Copyright, Technology, and Access to the Law: An Opinionated Primer. http://james.grimmelmann.net/essays/CopyrightTechnologyAccess, 2008.

[7] George R. Grossman. *Legal Research: Historical Foundation of the Electronic Age*. Oxford University Press, 1994.

[8] Woodrow Hartzog and Frederic D. Stutzman. The Case for Online Obscurity. *California Law Review*, 101, 2013.

[9] Edwin R. Keedy. Ignorance and Mistake in the Criminal Law. *Harvard Law Review*, 22:75, 1908.

[10] Peter W. Martin. Online Access to Court Records – From Documents to Data, Particulars to Patterns. *Villanova Law Review*, 53:855, 2008.

[11] Peter W. Martin. Rewiring Old Architecture: Why U.S. Courts Have Been So Slow and Uneven in Their Take-up of Digital Technology. Cornell Law Faculty Working Papers No. 84, 2011.

[12] Beth Simone Noveck. "Peer to Patent": Collective Intelligence, Open Review, and Patent Reform. *Harvard Journal of Law and Technology*, 20:123, 2006.

[13] PACER Service Center. Public Access to Court Electronic Records—User Manual. http://www.pacer.gov/documents/pacer.txt, 1998.

[14] The Honorable Ronald Leighton, et al. Panel Three: Implementation—What Methods, If Any, Can Be Employed To Promote the Existing Rules' Attempts to Protect Private Identifier Information From Internet Access? *Fordham Law Review*, 79:45, 2011.

[15] Joseph Turow, Lauren Feldman, and Kimberly Meltzer. Open to Exploitation: American Shoppers Online and Offline. Departmental Papers, Annenberg Public Policy Center, University of Pennsylvania, 2005.

[16] Peter Winn. On-Line Access to Court Records. Unpublished manuscript presented at the Privacy Law Scholars Conference. http://docs.law.gwu.edu/facweb/dsolove/PLSC/PLSC-Papers/Winn-Peter.pdf, 2008.

[17] Peter A. Winn. Judicial Information Management in an Electronic Age: Old Standards, New Challenges. *Federal Courts Law Review*, 3:135, 2009.

Chapter 4: Debugging the United States Code

[1] Richard S. Beth. How Bills Amend Statutes, Congressional Research Service Report No. RS20617, 2008.

[2] Christopher M. Davis. The Legislative Process on the House Floor: An Introduction, Congressional Research Service Report No. 95-563, 2010.

[3] Lewis Deschler. Deschler's Precedents of the United States House of Representatives, H.R. Doc. No. 94-661, 1994.

[4] Tobias A. Dorsey. Some Reflections on Not Reading the Statutes. *Green Bag 2d*, 10:283, 2007.

[5] Clarence A. Ellis and Simon Gibbs. Concurrency Control in Groupware Systems. *ACM SIGMOD Record*, 18(2):399, 1989.

[6] Michael S. Fried. A Theory of Scrivener's Error. *Rutgers Law Review*, 52:589, 2000.

[7] Karen L. Haas. Rules of the House of Representatives, One Hundred Twelfth Congress, 2011.

[8] Michael J. Lynch. The U.S. Code, the Statutes at Large, and Some Peculiarities of Codification. *Legal Reference Services Quarterly*, 16:69, 2007.

[9] Matthew McGowan. Senate Manual, One Hundred Twelfth Congress, 2011.

[10] Victoria F. Nourse and Jane Schacter. The Politics of Legislative Drafting: A Congressional Case Study. *NYU Law Review*, 77:575, 2002.

[11] Barack Obama. Presidential Document, Memorandum of January 21, 2009, Freedom of Information Act. *Federal Register*, 74:4683, 2009.

[12] Josiah Ober. *Democracy and Knowledge: Innovation and Learning in Classical Athens*. Princeton University Press, 2008.

[13] Walter J. Oleszek. Congressional Lawmaking: A Perspective on Secrecy and Transparency, Congressional Research Service Report No. R42108, 2011.

[14] Norman J. Ornstein, Thomas E. Mann, and Michael J. Malbin. *Vital Statistics on Congress 2008*. Brookings Institution Press, 2008.

[15] Elizabeth Rybicki. Senate Committee Reports: Required Contents, Congressional Research Service Report No. 96-305 GOV, 2008.

[16] Rob Sukol. Positive Law Codification of Space Programs: The Enactment of Title 51, United States Code. *Journal of Space Law*, 37:1, 2011.

[17] John V. Sullivan. How Our Laws Are Made, H.R. Doc. No. 110-49, 2007.

[18] Erwin C. Surrency. The Publication of Federal Laws: A Short History. *Law Library Journal*, 79:469, 1987.

[19] The Office of the Legislative Counsel, U.S. House of Representatives. House Legislative Counsel's Manual on Drafting Style, H.R. Doc. No. HLC 104-1. 1995.

[20] Will Tress. Lost Laws: What We Can't Find in the United States Code. *Golden Gate Law Review*, 40:129, 2010.

[21] Mary Whisner. The United States Code, Prima Facie Evidence, and Positive Law. *Law Library Journal*, 101:545, 2009.

[22] Charles J. Zinn. Codification of the Laws. *Law Library Journal*, 45:2, 1952.

Chapter 5: The New Ambiguity of "Open Government"

[1] Laura DeNardis. Open Standards and Global Politics. *International Journal of Communications Law and Policy*, 13:168, 2009.

[2] George Penn Kennedy. *Advocates of Openness: The Freedom of Information Movement*. PhD thesis, University of Missouri-Columbia, 1978.

[3] Daniel Kreiss. *Taking Our Country Back: The Crafting of Networked Politics From Howard Dean to Barack Obama*. Oxford University Press, 2012.

[4] Michael R. Lemov. *People's Warrior: John Moss and the Fight for Freedom of Information and Consumer Rights*. Fairleigh Dickinson, 2011.

[5] Ed Mayo and Tom Steinberg. The Power of Information: An Independent Review, 2007.

[6] Beth Simone Noveck. *Wiki Government: How Technology Can Make Government Better, Democracy Stronger, and Citizens More Powerful*. Brookings Institution Press, 2009.

[7] Barack Obama. Presidential Document, Memorandum of January 21, 2009, Transparency and Open Government. *Federal Register*, 74:4685, 2009.

[8] Chancellor of the Duchy of Lancaster. The Government's Response to The Power of Information: An Independent Review by Ed Mayo and Tom Steinberg, 2007.

[9] Paul Ohm. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, 57:1701, 2010.

[10] Wallace Parks. The Open Government Principle: Applying the Right to Know Under the Constitution. *George Washington Law Review*, 26:1, 1957.

[11] Alon Peled. When Transparency and Collaboration Collide: The USA Open Data Program. *Journal of the American Society for Information Science and Technology*, 62:2085, 2011.

[12] Peter R. Orszag, Office of Management and Budget, Executive Office of the President. Memorandum No. M-10-06. Open Government Directive, 2009.

[13] James S. Pope. *Foreword to Harold L. Cross, The People's Right to Know: Legal Access to Public Records and Proceedings*, page ix. Columbia University Press, 1953.

[14] David E. Pozen. The Mosaic Theory, National Security, and the Freedom of Information Act. *Yale Law Journal*, 115:628, 2005.

[15] Eric Steven Raymond. *The Cathedral and the Bazaar (Version 3.0)*. Thyrsus Enterprises, 2000.

[16] Harold C. Relyea and Michael W. Kolakowski. Access to Government Information in the United States, Congressional Research Service Report No. 97-71 GOV. 2007.

[17] David Robinson, Harlan Yu, William P. Zeller, and Edward W. Felten. Government Data and the Invisible Hand. *Yale Journal of Law and Technology*, 11:160, 2009.

[18] Jennifer Shkabatur. Transparency With(out) Accountability: Open Government in the United States. *Yale Law and Policy Review*, 31, 2013.

[19] Peter Suber. Open Access to the Scientific Journal Literature. *Journal of Biology*, 1:1, 2002.

[20] Joshua Tauberer. *Open Government Data*. http://opengovdata.io, 2012.

[21] The European Parliament and the Council of the European Union. Directive 2003/98, Re-use of Public Sector Information, 2003.

[22] Charles Alan Wright and Arthur Raphael Miller. *Federal Practice and Procedure*. Thomson West, 1992.

[23] Jane Yakowitz. Tragedy of the Data Commons. *Harvard Journal of Law and Technology*, 25:1, 2011.

Chapter 6: Conclusion

[1] Matthew J. Salganik and Karen E.C. Levy. Wiki Surveys: Open and Quantifiable Social Data Collection. http://arxiv.org/abs/1202.0500, 2012.

[2] Joseph A. Calandrino, Ariel J. Feldman, J. Alex Halderman, David Wagner, Harlan Yu, and William P. Zeller. Source Code Review of the Diebold Voting System. http://www.sos.ca.gov/voting-systems/oversight/top-to-bottom-review.htm, July 20, 2007. Report commissioned as part of the California Secretary of State's Top-To-Bottom Review of California voting systems.