UIC John Marshall Journal of Information Technology & Privacy Law

Volume 24
Issue 4 *Journal of Computer & Information Law - Summer 2006*

Article 5

Summer 2006

Cleaning Metadata on the World Wide Web: Suggestions for a Regulatory Approach, 24 J. Marshall J. Computer & Info. L. 531 (2006)

Marcel Gordon

ARTICLES

CLEANING METADATA ON THE WORLD WIDE WEB: SUGGESTIONS FOR A REGULATORY APPROACH

MARCEL GORDON

Where is the Life we have lost in living?
Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?
T. S. Eliot, Choruses from "The Rock" 1934.

Results 1 - 10 of about 9,310,000,000.
Google, 2006.

I. INTRODUCTION

The World Wide Web has made an enormous amount of information available to millions of users around the world. Web pages, knit together by the ubiquitous hyperlink, form a dazzling multiverse of facts and opinions, polemic and propaganda, truth and lies. To some, the brilliance of the Web lies in its lack of hierarchy, its openness, its egalitarian design, its universality, its freedom. It is a platform for the greatest dreams or the smallest gestures. To others, however, those same characteristics reduce the Web to a rabble, a noisy crowd of shiny-shoed hucksters that has overtaken the old-time county fair.[1] The reality is that what you see in the Web depends less on who you are than on what you are looking for. That is the issue at the core of this paper: helping users of the Web find the Web that they are seeking.

As the opening sentence of this paper implies, one use of the World Wide Web is as a source of information. If this is not already the most important use of the Web, such a use clearly has vast potential. For users seeking information, the Web is almost limitless in its breadth of

1. *ACLU v. Reno*, 521 U.S. 844, 853 (1997) (stating, "[t]he Web is thus comparable, from the readers' viewpoint, to both a vast library including millions of readily available and indexed publications and a sprawling mall offering goods and services").

topics and number of sources. To these users, the most important quality is the ability to locate the resources which are relevant to them. Peter Morville has termed this characteristic *findability*,[2] and information architects have long recognized its importance. Presently, it can be very difficult and time-consuming, in comparison to traditional information sources such as databases or libraries, to locate appropriate information on the Web. Even the best search engines return an enormous number of results, leaving users to filter through hundreds of pages to determine their relevance and authenticity. In the end, the Web's lack of findability limits its usefulness. However, this paper argues that a regulatory system focused on improving metadata could, among other things, greatly increase the findability of the Web.

Metadata is information about other information. It is everywhere in modern, connected life: title and author information in library catalogues; date of creation and change tracking data in word processing documents; title, date and sender fields in e-mails; and so forth. Metadata also helps us identify, organize, locate and assess information. On the Web, however, metadata is under-utilized. One reason is the ubiquity of the Web: the Web is a platform for enterprises of all types, and metadata is not useful to all of them. More significant, though, is that the freedom of the Web led to the abuse of metadata. Devious content creators seeking increased traffic added misleading metadata to their pages, describing every page with long lists of popular keywords. Search engines responded by virtually ignoring metadata. Removing misleading metadata, or cleaning the Web's metadata, would restore its usefulness as an organizational tool, improving the accessibility of information on the Web. Furthermore, of late there has been a renewed interest in metadata on the Web, with emergent technologies taking advantage of pockets of credible metadata. A regulatory system that ensures clean metadata could be the foundation for a new generation of Web-based tools.

While much has been written regarding the unauthorized use of trademarks in metadata,[3] the more general, and much larger, problem of

2. Peter Morville, *Ambient Findability*, 4 (O'Reilly 2005) (providing three definitions for findability: a. the quality of being locatable or navigable; b. the degree to which a particular object is easy to discover or locate; c. the degree to which a system or environment supports navigation and retrieval).

3. *See Infra* pt. II (Academics seem to have seized on the area as it is one in which existing law can be applied, *mutatis mutandis*, in order to solve a problem unique to the Internet. Interestingly, however, the trademark issue is something of a storm in a teacup because of the low impact that metadata has on search engine results. The solution of the problem addressed in this paper – that of unclean metadata in general – would rejuvenate the issue.); *see e.g.* F Gregory Lastowka, *Search Engines, HTML and Trademarks: What's the Meta For?* 86 Va. L. Rev. 835 (2000); Veronica Tucci, *The Case of the Invisible Infringer: Trademarks, Metatags and Initial Interest Confusion*, 5 J. Tech. L. & Pol'y 2 (2000);

unclean metadata is yet to be confronted from a legal perspective. This paper begins, in Part II, by considering the various beneficial applications which clean metadata presently has and those which it could have in the future. Part III then considers the challenges faced in creating an effective regulatory system to deal with unclean metadata. Part IV looks to regulatory systems which deal with similar problems for inspiration, examining the strengths and weaknesses of each strategy. The systems examined are representative and not exhaustive; the aim is to draw out the key challenges faced and possible solutions. Part V incorporates the strengths of the various approaches in order to propose a workable system, setting it out in detail, and lastly, Part VI offers ideas to deal with potential difficulties in implementation and enforcement.

II. THE APPLICATIONS OF CLEAN METADATA

A. METADATA AND CLEANNESS

Metadata is information about information.[4] More technically, metadata is "structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities."[5] To create metadata, characteristics such as title, author, and date of publication are collected for information sources. These pieces of information may to some extent be a part of the source itself (the title, for example, is arguably part of a novel) but they also describe it. Most importantly, this information can be used to organize a collection of sources. The use of metadata is the difference between managing information and merely collecting it.

Clean metadata is a key concept in this paper. Metadata is only useful insofar as it accurately describes the resource to which it pertains. In order to provide a working definition, clean metadata is metadata which

Rachael Jane Posner, *Manipulative Metatagging, Search Engine Baiting, and Initial Interest Confusion*, 33 Colum. J.L. & Soc. Probs. 439 (2003); Joseph T. Kucala Jr., *Putting the Meat Back in Meta-tags!* 2001 J.L. Tech.& Pol'y 129 (2001).

4. Much of the academy uses the term 'meta-tags' (sometimes 'meta tags' or 'metatags') to refer to metadata on the Web. Most of the Web is made up of documents defined in HTML, a markup language which uses tags for formatting, and in HTML metadata is defined using tags, leading to the use of that term. However, documents can be published on the Web in any format whatsoever, and popular formats such as Adobe's PDF and Microsoft Word's document format are indexed by modern search engines. Use of the term 'metatags' excludes metadata in these kinds of documents, and so the more general 'metadata' is preferable.

5. Association for Library Collections and Technical Services Committee on Cataloging: Description and Access Task Force on Metadata, *Summary Report*, http://www.libraries.psu.edu/tas/jca/ccda/tf-meta3.html (accessed March 16, 2006); *see also* Lars Marius Garshol, *Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all*, 30 J. Info. Sci. 378, 379 (2004) (stating, "information about a set of data in a particular representation" or "any statement about an information resource").

is not false, inaccurate, or misleading. Undoubtedly, there is scope for disagreement as to whether any particular metadata is inaccurate or misleading; that debate is a practical one which is taken up later in the paper.

B. METADATA ON THE WEB

Tim Berners-Lee, recognized as the inventor of the Web, included metadata as part of his original proposal.[6] Berners-Lee envisaged the Web as a collection of organized, linked ideas.[7] The proposed metadata was not only information on the identity of documents, termed 'identity metadata,' but also on the organization of documents referred to as'relational metadata.' It soon became clear, however, that the flexibility and simplicity of the Web was more important to its early users than the maintenance of a strict structure. While linking caught on, due to its usefulness as a practical tool rather than its value as metadata, the other, explicit relational metadata did not, and identity metadata was generally limited to titles.[8]

As the Web grew in popularity, the use of metadata was heavily influenced by the interaction of two innovations. First, search engines began to map the Web, introducing what is now the Web's primary interface.[9] Second, commercial interests appeared on the Web. Search engines, recognizing that metadata could be used to organize the Web, began to use identity metadata such as titles and keywords in assessing the relevance of a page to a given query. Metadata's importance skyrocketed. In response, commercial Web site operators supplied false metadata about their pages to ensure that they appeared more frequently in search results and attracted more users.[10] As a result, metadata became untrustworthy and effectively useless,[11] depriving the

6. Tim Berners-Lee, *Information Management: A Proposal*, http://www.w3.org/History/1989/proposal.html (accessed March 23, 2006) (explaining that, "[i]n practice, it is useful for the system to be aware of the generic types of the links between items (dependences, for example), and the types of nodes (people, things, documents) without imposing any limitations").

7. Tim Berners-Lee, *Spinning the Semantic Web* xiii-xiv (MIT Press 2005).

8. Tim Berners-Lee and Dan Connolly, *Hypertext Markup Language 2.0*, http://ftp.ics.uci.edu/pub/ietf/html/rfc1866.txt (accessed March 16, 2006) (The first HTML standard, HTML 2.0, included <isindex>, <link>, <title> and <meta> tags. It was issued by the World Wide Web Consortium ('W3C') in 1996, but "roughly corresponds to the capabilities of HTML in common use prior to June 1994.").

9. Mei Kobayashi and Koichi Takeda, *Information Retrieval on the Web*, ACM Computing Surveys 144, 146 (June 2000).

10. *See e.g.* Ira Nathenson, *Internet Infoglut and Invisible Ink: Spamdexing Search Engines with Meta Tags*, 12 Harv. J. Law & Tech. 43, 61-66 (2000).

11. *See e.g.* Danny Sullivan, *How to Use HTML Meta Tags*, http://searchenginewatch.com/webmasters/article.php/2167931 (accessed April 12, 2006); Jin Zhang and Alexandra

Web of this powerful information management tool.

C. Impact of Clean Metadata in the Search Space

Search – the selection of resources from across the Web based upon a user's query – is now the domain of massive corporations such as Google, Yahoo! and Microsoft. This is a testament to two things: the importance of search to the Web; and the difficulty of searching effectively. Due to the lack of credibility of author provided metadata, search engines use sophisticated algorithms to produce their own metadata. One of Google's key innovations was to use links, created primarily for human navigation, to describe the relationships between different sites, that is, as metadata.[12] Each search engine manages an enormous repository of metadata generated from analysis of the Web.

If the Web's metadata was clean, some of this effort could be saved. In effect, the cost of metadata generation would be distributed amongst individual content creators and the metadata would be shared among all search engines. The Web would be the repository for its own metadata.

This would not be of great consequence for the current generation of search engines, as other elements of their operation – crawling the Web, storing the information gathered and indexing it for use – are considerably more arduous than metadata generation. However, it is potentially important for other search applications, particularly smaller scale applications. For example, a personal search engine could be instructed to crawl outward from a particular Web page, making an annotated map of the results. A personal search engine could do this without having to generate and store its own content-related metadata, decreasing the cost in computing resources and removing the dependence on expensive development and closely guarded algorithms for generating accurate metadata.

The latter point is the most important in terms of the cost of operation of a modern day search engine. While the physical resources needed to run a search engine would not be substantially decreased, the barrier to entry in terms of intellectual property would be lowered. This means that a team of brilliant technicians and years of research and development would no longer be prerequisites to the operation of a search engine. Undoubtedly, search engines would still seek to distinguish themselves through the analysis that they perform on the metadata, requiring both engineers and effort; however, the level of performance of

---

Dimitroff, *Internet Search Engines' Response to Metadata Dublin Core Implementation*, 30 J. Info. Sci. 310 (2004) (explaining that it appears that metadata may have some minor influence on search engine results).

12. Google, *Our Search: Google Technology*, http://www.google.com/technology/ (accessed March 16, 2006).

basic search – search without doing anything outside of the public domain – would be substantially increased.

Furthermore, there are prospects of a reduction of computational resource requirements. Guaranteed clean metadata generated by the Web site author lends itself to a reorientation of the search engine model. Rather than proactively analyzing the Web in order to keep a metadata repository up to date, a search engine could rely upon submissions from those who wish to be indexed, whether from the Web in general or targeting particular communities of interest.[13] This removes the need to constantly crawl the Web. While this model may not be suitable for all search engines or applications, it does offer a potentially cheaper alternative.

The efficiency of search engine use would also increase, mainly due to increased accuracy of results. Most obviously, the metadata upon which the response is based would be generated by human understanding rather than by computer analysis. Additionally, the metadata could be structured according to standards.[14] The Dublin Core Metadata Initiative has seen the development of a standard system of metadata for information resources.[15] Its widespread adoption would allow very accurate search on common criteria. It could also make any type of resource accessible using search, including those such as video and audio, which are not amenable to current search engine methods. Metadata-based schemes for privacy and copyright information have also been proposed.[16]

Given that search is so important to the Web, any development which improves its effectiveness will produce enormous benefits. While search engines have improved considerably, they still struggle to provide relevant results in many cases.[17] Clean, structured metadata would bring the accuracy of Web search closer to that of searches in a library catalogue or database. For users, more efficient search means less time

13. The idea of community-based search engines ties in with the focus on structure, *infra* pt. II (Democracy, Freedom and Choice). If academics, for instance, submit structured information on research published on the Web, it can be searched according to various criteria useful in that domain. Such a proposal raises the question: why can't this be done already? The difficulty is that it is presently impossible to automatically verify the accuracy of a submission. If the submission can be checked against the metadata on the page itself, and the metadata on the page is guaranteed to be accurate, then the submission can very easily be verified.

14. Nathenson, *supra* n. 10, at 136.

15. Dublin Core Metadata Initiative, *Dublin Core Metadata Element Set, Version 1.1: Reference Description*, http://dublincore.org/documents/dces/ (accessed March 16, 2006) (stating "[i]f the Internet is a library, then it is one with all the books scattered on the floor").

16. Tim Berners-Lee, *supra* n. 7, at xvii.

17. *See* Nathenson, *supra* n.10, at 51-57 (discussing the so-called infoglut).

spent searching and more relevant information obtained. The time savings alone are potentially enormous; the consequential benefits of improved information are incalculable.

D. Democracy, Freedom, and Choice

More fundamentally, clean metadata holds value for a democratic society. The well-known arguments which value diversity in media ownership applies even more compellingly to the role of search engines on the Web. Whereas we generally look to traditional media for a limited range of information, in particular, news and current affairs, and our concern for censorship generally extends only to the reporting of events and the formation of public opinion, we use the Web for a far broader and more diverse range of purposes. Controlling access to the Web means controlling access not only to news, but to marketplaces, community forums, public records, and all manner of basic information resources.

Search is the most important interface to the Web, especially in the discovery of new information and resources. However, large American corporations own most of the metadata that is needed to discover new information and resources effectively. Functionality, access, and accuracy are all subject to commercial interest. Presently, access is free and relatively open; the process, however, is far from transparent. Google, the current market leader, has generally defended itself based on the objectivity of its results and its informal motto, "Don't be evil."[18] However, its capitulation to China's demands for censorship has caused considerable controversy[19] and its recent partnership with AOL, a subsidiary of media giant Time-Warner, raise difficult issues regarding independence.[20] Google has also been subjected to lawsuits based on the measures it takes against sites which seek to manipulate search results.[21] Whatever the track record of the search superpowers, the problem remains. By controlling the repositories of metadata needed in order

18. Google, *Google Code of Conduct*, http://investor.google.com/conduct.html (accessed March 16, 2006).

19. *See e.g. A setback for free speech in China Google's agreement to censorship sacrifices its ideals*, Financial Times (London), 16 (Jan. 26, 2006); Mark Ellis and David Edwards, *Goog or Evil? Fury as Google agrees to censorship in deal with Chinese leaders*, The Daily Mirror (London), 16 (Jan. 26, 2006); Lester Haines, *The Register, Google pulls 'we don't censor' statement*, http://www.theregister.co.uk/2006/01/27/google_doesnt_censor/ (accessed March 16, 2006); *see also* Jonathan Zittrain and Benjamin Edelman, *Berkman Center for Internet and Society, Harvard Law School, Documentation of Internet Filtering Worldwide*, http://cyber.law.harvard.edu/filtering/ (accessed March 31, 2006) (regarding Google's filtering in other areas).

20. *See e.g.* Emiliya Mychasuk and Kate Mackenzie, *Speak no evil Issue of the Week: Google Deal with AOL*, Financial Times (London) (Dec. 19, 2005).

21. *See e.g.* Michael Liedtke, *Web site files complaint against Google*, Associated Press (March 17, 2006).

to search the Web, these companies effectively control access to the Web itself, without any accountability or obligation to the public; worse, they are bound to pursue their own financial interests.[22]

Additionally, these corporations are subject to the control of the government of the United States.[23] The creation of an alternative source of clean metadata would make other nations less dependent upon U.S.-owned search infrastructure, increasing the stability and security of local information economies. It would also potentially create a broader distribution of the economic benefits derived from the Web. For users, it would ensure that guarantees of values such as privacy conformed to local standards and that records of their activities were not at the behest of a foreign government. Indeed, the European Union has proposed the development of a competitor to Google for these very reasons.[24]

Even if these power relations do not raise any concern for the reader – perhaps the powers-that-be are presently benign enough to placate any latent concerns – the opportunity to make search a more affordable and accurate task is still valuable. Even if choice is not important for reasons of avoiding undue influence from one party or protecting privacy, it is nevertheless attractive in its own right. As consumers of search services, we prefer the ability to choose the way in which we search. Distribution of clean metadata and the consequent reduction in cost of the provision of search raises the possibility of increased choice in the search marketplace.

E. New Tools, New Dimensions

A new generation of metadata based tools has emerged on the Web in recent years.[25] One promising technology is folksonomy.[26] Where pre-

22. *See*, Nathenson, *supra* n. 10, at 51 ("What's important - increasingly important [in the Information Economy] - is the process by which you figure out what to look at. This is the beginning of the real and true economics of information - not who owns the books, who prints the books, who has the holdings. The crux today is access, not holdings. And not even access itself but the signposts that tell you what to access - what to pay attention to. In the Information Economy everything is plentiful - except attention."); *see also* Jack Goldsmith and Tim Wu, *Who Controls the Internet?* 74-76, 95-96 (Oxford University Press 2006) (discussing Google's compliance with requests to remove sites from its results and Yahoo and MSN's filtering for the Chinese government).

23. Steven Levy, *Technology: Searching for Searches; The government is demanding millions of your queries. AOL, Yahoo and Microsoft have coughed up. Google is resisting.*, Newsweek, 34 (Jan. 30, 2006).

24. *See e.g. Attack of the Eurogoogle*, The Economist (March 11, 2006); Kevin O'Brien, *Europeans weigh plan for search engine; Challenge to Google may get 2 billion*, International Herald Tribune, 1 (Jan. 18, 2006).

25. Examples include flickr (www.flickr.com), del.icio.us (http://del.icio.us/), and Technorati (www.technorati.com).

26. Morville, *supra* n.2, 134-41.

viously only the implicit metadata value of a link was utilized (as by Google, for instance), folksonomy allows users to add additional metadata, called tags, to links. Folksonomy has proven to be a popular way of annotating the Web because of its inherent flexibility. Users have built extensions, mixing metadata and data from various sources to produce innovative tools and perspectives.[27] Another technology, known as microformats, seeks to pave the cowpaths, and develope standards based on popular uses of metadata.[28] With standards in place, this metadata can then be used by a variety of applications. These new technologies highlight the potential of clean metadata beyond general search.

Stepping into the speculative, the World Wide Web Consortium's Semantic Web proposal highlights the potential usefulness of clean metadata and commoditized search.[29] The Semantic Web seeks to add meaning to the Web that can be understood by machines through the use of semantic mark-up. The most exciting visions of the Semantic Web invoke autonomous, intelligent agents acting as personal assistants.[30] These agents would be able to assist in basic tasks such as booking plane tickets and organizing appointments by matching information available on the Web with personal information such as a diary. While all semantic mark-up is arguably metadata, certain standard metadata would form the infrastructure upon which the Semantic Web is built.[31] This core metadata would have to be clean. In addition, the availability of commoditized search is important to the Semantic Web, as search would be the basic means of navigation for agents gathering information.

It is impossible to explore all the various possible uses of clean metadata. The crucial point is that once a system for guaranteeing clean metadata has been established, it can be useful for various applications. Search is the most obvious and immediate beneficiary, but the possible applications are literally infinite. They are constrained only by imagination and enterprise. When a platform is established which is not preferential as to how it ought to be used, the possibilities for its exploitation by the world's combined ingenuity are endless. This characteristic, known as application neutrality, is what has made the Internet, and within it the Web, so successful. Rather than constraining the evolution

27. *See e.g.* Programmable Web: Web 2.0 Mashup Centre, http://www.programmableweb.com/mashups (last updated Aug. 31, 2007) (discussing such extensions, known as mashups).

28. Microformats, *About microformats*, http://microformats.org/about/ (accessed March 19, 2006).

29. W3C Technology and Society Domain, *Semantic Web,* http://www.w3.org/2001/sw/ (accessed March 19, 2006).

30. Tim Berners-Lee, James Hendler & Ora Lassila, *The Semantic Web,* 284(5) Scientific American 34 (2001).

31. Morville, *supra* n. 2, at 125.

of technology or having to be adapted or updated to remain useful, a neutral platform allows technology to evolve freely without any fear of obsolescence. In this way, a scheme to guarantee clean metadata represents an investment in the public infrastructure of the Web. Web users can build whatever they please upon it, be it a shopping mall or a library, now and in the future.

F. REGULATION'S ROLE

The Web has a unique problem with unclean metadata. Traditional applications of metadata involve closed systems, or systems over which some central authority has control. The Web's open nature exposes a lack of authority and an opportunity to profit from dishonest metadata. The past decade has seen the rise and fall of metadata on the Web, and no technical solution alleviating its chronic lack of credibility. Functionally, clean metadata can bring new efficiency to search, and expose the Web as a whole to new metadata-based technologies. Making clean metadata part of the public infrastructure of the Web accords with the Internet's fundamental end-to-end architecture.[32] As Berners-Lee acknowledges, "if we can make of the Web something decentralized and of great simplicity, we must be prepared to be astonished at whatever might grow out of that new medium."[33]

There has been much commentary on the perceived incompatibilities of governmental regulation and the Internet, both popular[34] and academic.[35] However, just as it does offline, regulation can create authority

32. *See* Stefan Bechtold, *ICANN Governance: Governance in Namespaces*, 36 Loy. L.A. L. Rev. 1239, 1292-94 (2003) ([T]he [end-to-end] argument claims that as much intelligence as possible should reside at the 'edges' of the network, that is, at applications running on networked computers, not in the network itself. . .[b]y decentralizing control, [end-to-end] enables decentralized innovation. . .[and] [n]etwork architectures that violate the [end-to-end] design principle tend to build 'complex function into a network [which] implicitly optimizes the network for one set of uses while substantially increasing the cost of a set of potentially valuable uses that may be unknown or unpredictable at design time'" (footnotes omitted)); *see also*, Timothy Wu, *Application Centered Internet Analysis*, 85 Va. L. Rev. 1163, 1192-93 (1999); Jerome Saltzer, David Reed, and David Clark, *End-to-End Arguments in System Design*, 4 ACM Transactions on Computer Systems 2, 277 (1984) (providing the end-to-end principle).

33. Berners-Lee, *supra* n. 7, at xxii.

34. *See e.g.* John Perry Barlow, *A Declaration of the Independence of Cyberspace*, http://homes.eff.org/~barlow/Declaration-Final.html (accessed April 2, 2006); Lawrence Lessig, *Code and Other Laws of Cyberspace* 4-5 (Basic Books 1999).

35. *See e.g.* David Johnson and David Post, *Law and Borders - The Rise of Law in Cyberspace*, 48 Stan. L. Rev. 1367 (1996); Neil Netanel, *Cyberspace Self-Governance: A Skeptical View from Liberal Democratic Theory*, 88 Cal. L. Rev. 395 (2000); Alfred Yen, *Western Frontier or Feudal Society?: Metaphors and Perceptions of Cyberspace*, 17 Berkeley Tech. L. J. 1207 (2002).

and remove the incentive for abuse, to the benefit of the public.[36] Law can be used as a tool for empowerment, an organizing force. Its intervention, rather than restricting the potential of the Internet, may in fact allow that potential to be fully realized.[37] As Lawrence Lessig has pointed out, the Internet is already governed by code; government regulation is merely a more overt form of intervention.[38] Indeed, law may be necessary to protect the public interest against code favoring private concerns.[39] Given the scale of the possible benefits, both concrete and speculative,[40] metadata is an area ripe for regulatory intervention. The challenge is to identify the most appropriate regulatory scheme. Part III considers the difficulties which need to be taken into account.

III. DIFFICULTIES IN THE REGULATION OF METADATA

The regulation of metadata involves a number of challenges. The most significant hurdle to effective regulation is enforcement. Enforcement is made difficult by the nature of the Internet- the communication medium upon which the Web is built[41] -and the unusual nature of the harm sought to be regulated. Even if these considerations can be overcome, the existence of a culture of non-compliance could compromise regulatory efforts. Beyond enforcement, two considerations ought to shape the formulation of a regulatory system. First, the ideal solution would be in harmony with the open, end-to-end, decentralized nature of the Web. Second, it would possess the systemic characteristic extensibility valued by computer scientists.[42]

A. LAW AND THE INTERNET

Any discussion of law and the Internet must inevitably deal with the

36. Goldsmith and Wu, *supra* n. 22, at 140-42; Lawrence Lessig, *The Law of the Horse: What Cyberlaw Might Teach*, 113 Harv. L. Rev. 501, 505-06 (1999).

37. Thomas Nachbar, *Paradox and Structure: Relying on Government Regulation to Preserve the Internet's Unregulated Character*, 80 Minn. L. Rev. 215, 215-18 (2000); Goldsmith and Wu, *supra* n. 22, at 44-45.

38. *See generally* Lessig, *supra* n. 34.

39. R. Polk Wagner, *On Software Regulation*, 78 S. Cal. L. Rev. 457, 508 (2005).

40. Nachbar, *supra* n. 37, at 218 (recognizing that "the structure we use for arranging [the Web's] content may have a greater impact on society than the underlying content itself").

41. *See* Internet.com, *Internt*, http://webopedia.internt.com/TERM/I/internt.html (last accessed Oct. 15, 2007) (The name 'Internet' comes from the fact that it is a network of networks. The Internet supports the transmission of data from any point to any other via a number of different paths based on the Internet Protocol. The Web is a particular use of that capacity – an application – governed by a set of communication and document format standards (HTTP and HTML, amongst others).).

42. *See infra*, pt. III, (discussing *Maintaining the Tao of the Web*).

issue of jurisdiction.[43] The problem, in short, is that the Internet is global while municipal legal systems are not. The Internet crosses geographical boundaries indiscriminately, meaning that creators and recipients of content can be in different jurisdictions.[44] This limits the effectiveness of regulation, leaving aside the question of the legitimacy of trying to do so,[45] because a government has no capacity to enforce its law in foreign jurisdictions. Hence, content providers in other jurisdictions cannot be controlled, and providers in the regulating jurisdiction can simply relocate their content without prejudice.

Others have proposed ideal solutions to the conundrum of jurisdiction.[46] This paper seeks a pragmatic and immediate way in which legislation can be made enforceable, and hence, effective. A regulatory scheme must either establish a means of enforcement against persons in other jurisdictions, or circumscribe its ambit to the area in which it can already ensure that its legislation will be enforced. The former approach has proven ineffective in the past when, even when some degree of international consensus has been reached in order to enable a global response to particular issues, the results have been underwhelming.[47] Concerted international action is rare, and the usual outcome is compliance in a small number of countries with a significant interest undone by safe havens outside of those jurisdictions.[48] The latter approach, limiting the regulation to one's own jurisdiction, can be difficult to achieve without compromising the aims of the regulation. For instance, a United States law on clean metadata limited to hosts in the United States would – on its own – do little to alleviate the problem.

Furthermore, the architecture of the Internet places constraints on the legal options which might otherwise be available, and those constraints must be taken into account. Identifying and locating those who

43. Michael Geist, *Internet Law in Canada*, 41 (Captus Press 2000); Stuart Biegel, *Beyond Our Control? Confronting the Limits of Our Legal System in the Age of Cyberspace* 112-14 (MIT Press 2003).

44. Johnson and Post, *supra* n. 35, at 1368-76.

45. *See* Johnson & Post, *supra* n. 35, at 1369-70; Biegel, *supra* n. 43, at 111; Jack Goldsmith, *Against Cyberanarchy*, 65 U. Chi. L. Rev. 1199, 1239-44 (1998).

46. Johnson & Post, *supra* n. 35 (The authors famously advocated "taking cyberspace seriously," whereby national governments would relinquish sovereignty over certain areas to organic, Internet-developed authorities. While it is certainly arguable on the merits that Internet-based governance of the Internet is the best solution, such a proposal provides very little practical guidance as to how effective regulation can be produced.).

47. Miriam Miquelon-Weismann, *The Convention on Cybercrime: A Harmonized Implementation of International Penal Law: What Prospects For Procedural Due Process?*, 23 John Marshall J. Computer & Info. L. 329, 351-61 (2005); Goldsmith & Wu, *supra* n. 22, at 65-67.

48. The majority of these countries do not actively seek to become safe havens, they are simply used that way because of a lack of motivation and/or capacity to police the behavior in question.

are not complying with a regulatory scheme can be difficult due to the open architecture of the Internet.[49] This makes enforcement problematic, even within a sophisticated jurisdiction such as the United States.[50] In jurisdictions with less legal and technological infrastructure and resources, it can render enforcement impossible.

B. Traditional Methods of Enforcement

Civil actions are brought by individual citizens. While the problem of jurisdiction may be less severe for civil action than criminal prosecution, the problems of identification and enforcement of judgments present an even greater hurdle to private parties. Moreover, the diffuse nature of the harm in question makes civil remedies in general unattractive as no single party has sufficient individual interest to pursue them. The harm suffered by any individual user is small, yet there is a very significant cumulative harm to a large group of people. While representative actions such as class actions exist to address this very problem, in the case of unclean metadata, the source of harm is also diffuse. Representative actions are therefore unworkable. The only solution is to leave enforcement to a central party, such as the government. If the key difference between government-imposed sanctions and civil actions is that civil actions are brought by individuals, the problem of diffuse harm seems to render them irremediable and unsuitable for the problem at hand.

The traditional form of government sanction is the criminal penalty, enforced through the criminal justice system. Governments also impose civil penalties, enforced without the cost and rigor of the criminal justice system. Occasionally, where they can be effective, governments impose in-kind penalties. Where regulation has granted some special status or right, that right may be taken away for failure to obey the conditions attached to it. Examples include the de-registration of doctors and the suspension of drivers' licenses. These latter types of penalties are generally either reinforced by criminal penalties in the case of persistent infringement or layered together in a regulatory system, with the particular penalty imposed depending on the seriousness of the breach in question. Government imposed penalties are sensitive to the architectural difficulties of jurisdiction and anonymity identified above. However, they deal well with the problem of diffuse harm, as the government acts in the public interest to prevent the harm and the government is less sensitive to the issue of cost.

---

49. Biegel, *supra* n. 43, at 112-13.

50. United States Department of Justice, *The Electronic Frontier: The Challenge of Unlawful Conduct Involving the Use of the Internet. A Report of the President's Working Group on Unlawful Conduct on the Internet*, http://www.usdoj.gov/criminal/cybercrime/unlawful.htm (Mar. 2000).

Government imposed sanctions appear to be the only option, and significant issues need to be resolved in order for them to be effective. However, even if the problems identified above can be overcome, the existence of a culture of non-compliance would be fatal to regulatory efforts.

C. CULTURES OF NON-COMPLIANCE

If enough people choose to disobey the law, it ceases to have effect.[51] Where a general inclination to disobey the law, or a particular law, is present, it might be said that a culture of non-compliance exists. The running battle between consumers and record labels has highlighted the problem of enforcing a legal regime against a culture of non-compliance.[52] Despite the clarity and universality of intellectual property laws, the music industry is struggling to protect its copyrights. The issue is a complex one; what is clear, however, is that there is significant intellectual and public opposition to the idea that the copying of music over the Internet ought to be proscribed, at least in light of the music industry-sanctioned alternatives.[53] Copyright exists for the public benefit, sacrificing the liberty to deal with certain materials in certain ways in order to encourage creativity and innovation, yet the public remains unconvinced that the balance that has been struck is, in light of the Internet, the correct one. The result has been widespread disobedience,[54] compromising the effectiveness of the legislation. The establishment of a culture of non-compliance is made more likely on the Internet by the problems of jurisdiction and anonymity discussed above.

Regulation prohibiting unclean metadata is vulnerable to the same problem. As is the case with the music industry's attempt to prosecute copyright violation, the regulation of metadata would seek to stop an activity that is widespread and tolerated. Worse, the publishing of unclean metadata is not illegal offline, nor is it analogous to an activity which is illegal offline, unlike breach of copyright law.[55] Unclean metadata is a

51. Biegel, *supra* n. 43, at 101-07 (considering prohibition, marijuana and race riots as examples).

52. Geist, *supra* n. 43, at 517-583; Biegel, *supra* n. 43, at xvii-xix, 73-76, 279-91; Goldsmith and Wu, *supra* n. 22, at 105-25.

53. *See* Lawrence Lessig, *Free Culture: How Big Media Uses Technology and the Law toLock Down Culture and Control Creativity* (Penguin Press 2004) (discussing intellectual objections to the present state of copyright law); Timothy Wu, *When Code Isn't Law*, 89 Va. L. Rev. 679, 685, 722-6 (2003) (regarding the public); John Perry Barlow, *The Next Economy of Ideas*, Wired Magazine, 8.10 (Oct. 2000) (available at http://www.wired.com/wired/archive/8.10/download.html).

54. Biegel, *supra* n. 43, at xi-xii, 116-17.

55. It is, however, generally seen to be a dishonest practice. In some limited circumstances the use of false or misleading metadata might be compared to false or misleading advertising, but the relationship between the offence and the profit is much less direct.

problem, through lack of information, to which the general public might well be apathetic. Furthermore, based on libertarian ideals, some activists oppose any regulation, especially regulation which touches upon content. Authoring content on the Web is an area in which an enormous number of people are involved, making non-compliance among only a sub-section of those it seeks to control potentially overwhelming. Any regulation must, therefore, pay careful regard to the balance which it strikes between the public benefit produced by the creation of a system of clean metadata and the cost imposed upon individual authors.

D. Freedom of Speech

Regulating metadata engages the sensitive subject of freedom of speech. While it is not possible to address the specific requirements of each jurisdiction, a few salient points can be made. Jurisprudence on freedom of speech generally involves some question of reasonableness, of proportionality between the harm sought to be avoided, and the measure introduced to avoid it.[56] A measure would require disclosure of identification, assessment of metadata by a third party, or limited publication to certain types of persons might well be considered disproportionate, especially in light of the Web's present near-absolute freedom and the breadth of purposes which it incorporates. The key is to ensure not just that the benefits of the scheme outweigh the costs imposed, but to ensure that the scheme is the least restrictive alternative which produces those benefits.

E. Maintaining the Tao of the Web

In advocating the use of regulation to ensure clean metadata on the Web, the value of democratized, commoditized search and the potential for innovation built upon an end-to-end architecture are stressed.[57] Although replacing one bottleneck or centralizing influence with another may have some net benefit, it fails to realize important benefits associated with decentralization.[58] Hence, it is highly desirable that metadata is regulated in accordance with these ideas in order to procure the maximum advantage from clean metadata. This is a difficult goal, because government regulation has a centralizing effect;[59] decentralized regulation is almost an oxymoron. Nevertheless, centralization can be used as a comparative criterion. A system that requires the vetting of all Web

56. *See e.g. Ashcroft v ACLU*, 122 S. Ct. 1700 (2002); *Lange v Austrl. Broad. Corp.* (1997) 189 CLR 520, 567 (in Australia).

57. *See supra*, pt. II, Regulations Role; *see also* Nachbar, *supra* n. 37, at 51-59.

58. *See supra*, pt. I, Impact of Clean Metadata in the Search Space (discussing how replacing private authority with public authority may be worthwhile on the whole).

59. Bechtold, *supra* n. 32, at 1317.

pages by a single, central authority is less attractive, in terms of this criterion, than one which allows vetting by a number of different authorities.

Furthermore, computer science values a systemic characteristic known as extensibility. Extensibility is the facility with which a system can be extended to allow new uses without affecting the original functionality.[60] This is relevant in two respects. First, the regulatory system introduced ought not to limit the existing functionality of the Web, or the Internet, for that matter. Second, the system which is introduced should itself be extensible, such that it can be used as a platform for other innovations should the need arise. Extensibility is also comparative, with the status quo as a benchmark. While identifying the least restrictive regulatory scheme is not difficult once schemes have been proposed, extensibility in the second sense may rest upon the details of implementation, making assessment across high-level models difficult.

Clearly, there are a number of significant difficulties which need to be overcome. The nature of the Internet and the problem at hand favor government imposed sanction over private action, but also pose significant impediments to effective enforcement. Furthermore, any regulation would need to be framed in such a way as to avoid falling victim to a culture of non-compliance. If these considerations can be accounted for and options remain, decentralization and extensibility are relevant criteria for assessing which alternative is the most attractive. Part IV looks at models found in other areas of law for inspiration in addressing these challenges.

IV. WORKING MODELS FOR THE REGULATION OF METADATA

As the adage counsels, nothing teaches better than experience. As such, existing regulatory schemes are a logical place to look in considering how regulation can produce clean metadata on the Web. However, before doing so it is useful to consider some more fundamental ideas regarding regulation of the Internet. The question of how law should deal with the Internet has been a source of considerable difficulty for legal scholars, and the melding of a number of different perspectives exhorts both caution and creativity. With this in mind, the current approaches to regulating spam and cybersquatting are considered. A number of points emerge, most importantly the effectiveness and the limits of in-kind regulation based on the domain name system.

60. *Oxford English Dictionary* vol. 5, 596 (James A.H. Murray et al. eds., 2d ed., Carendon Press 1989) (A system design principle that takes into consideration future growth.).

A. Regulation and the Internet

The derisory comparison of the study of 'the law of cyberspace' to 'the law of the horse' appears to suggest that the Internet can be understood, and ought to be approached, entirely within the foundations provided by existing doctrines of law.[61] Clearly, there is merit in the extension of known principles to new circumstances – centuries of compounded wisdom can illuminate seemingly unprecedented controversies with surprising clarity. As Benjamin Cardozo noted, "[w]e fancy ourselves to be dealing with some ultramodern controversy, the product of the clash of interests in an industrial society. The problem is laid bare, and at its core are the ancient mysteries crying out for understanding."[62]

However, while Easterbrook certainly makes a point about the usefulness of general principles, he does not deny the novelty of the Internet. Instead, he points out two threats to the credibility of legal scholarship focused on the Internet's novelty. The first is the "risk of multidisciplinary dilettantism," where legal analysis based on an incomplete or inaccurate understanding of the technology involved leads to "the worst of both worlds."[63] The second, Easterbrook's "principal conclusion," is a warning against applying doctrines to cyberspace which are not fully understood in their present context.[64] Easterbrook asks, regarding intellectual property, "[i]f we are so far behind in matching law to a well-understood technology such as photocopiers. . .what chance do we have for a technology such as computers?"[65] If we cannot explain the functioning of a doctrine in the familiar, offline world, then we have no basis on which to judge how it ought to be modified in order to best regulate cyberspace.

At the root of these arguments lies a significant supposition. If a deep knowledge of the technology and a thorough understanding of the rationales informing the doctrines are necessary, then it must be because some decision needs to be made which is best made in an informed manner. Implicit in Easterbrook's criticism is a belief that the application of existing legal principles to the Internet must be on a thoughtful, reasoned basis. That is, while recognizing the importance of an approach to law based on general principles, Easterbrook recognizes that the new environment of the Internet mandates an active consideration of their underlying justifications. This is not, therefore, merely another set of facts

61. *See* Frank Easterbrook, *Cyberspace and the Law of the Horse*, 1996 U. Chi. Legal F. 207 (2006); Lessig, *supra* n. 36 at 501-02 (discussing the same perspective).

62. Benjamin N. Cardozo, *The Paradoxes of Legal Science* (Columbia Univ. Press 1928).

63. Easterbrook, *supra* n. 61, at 207.

64. *Id.* at 208-209.

65. *Id.* at 210.

to be dealt with under existing law; something more is needed. The Internet needs its own regulatory balance, struck in light of its own special circumstances, and as Easterbrook stresses, that balance should be found with thoughtful, cautious, and informed steps.[66]

In this sense, then, there really is a cyberspace. It is not a jurisdiction disjoint from the familiar one, in which we have to recreate law as we know it. It is a new place in the sense that we are dealing with Cardozo's "ancient mysteries" in a fresh context.[67]

This is not to say that some, even the majority, of existing principles will not be applicable, becasue existing law provides an essential foundation for much of what is already possible on the Internet.[68] However, as Lessig stresses in considering constitutional dilemmas caused by new technology, existing doctrines must be actively translated to the Internet, when the principles which inform them are relevant.[69] Wagner highlights one novel consideration, stating that "it is not quite as simple as property rules versus liability rules; the recognition of the participation of software regulation demands a more nuanced approach."[70] Furthermore, this process must be richer than a simple online versus offline dichotomy might suggest. The Internet supports a diverse range of applications, each with different architectures, and so in each case the regulatory balance requires distinct consideration.[71] A deep and accurate appreciation of the technology involved is necessary to avoid, appropriating Easterbrook's equine theme, making an ass of oneself.

While existing legal structures are a logical place to start in searching for a regulatory solution to the problem of unclean metadata, the assumptions which underlie them need to be carefully considered before they can be translated to the Web. What is being sought in considering existing regulatory models is not necessarily a working example which can be applied to solve the problem, or a place in the existing legal structure into which a remedy can be slotted - although either would obviously be welcome discoveries. Instead, the solution can be written on a blank page; anything is possible. The regulatory models which are considered below are considered as inspirations, a kind of collective experience of Internet regulation. All raise salient points and illuminate elements of a potential solution.

---

66. *Id.*

67. Chris Reed, *Internet Law Text and Materials* 307 (Cambridge University Press 2004); Goldsmith, *supra* n. 45, at 1201, 1250.

68. Goldsmith & Wu, *supra* n. 22, at 140-42; Easterbrook, *supra* n. 61, at 210.

69. Lessig, *supra* n. 34, at 111-21.

70. Wagner, *supra* n. 39, at 510.

71. Wu, *supra* n. 32, at 1163-65.

B. E-MAIL SPAM AND CAN-SPAM

The word 'spam' is generally used to describe unsolicited electronic messages distributed in bulk. E-mail spam is a significant problem on the Internet. Estimates of its cost –productivity lost in dealing with unwanted e-mails, infrastructure costs incurred in moving, storing and managing a greater volume of e-mail, and countermeasures used to reduce its impact – are staggering. The consensus is that spam costs at least $10 billion per year.[72] That is without considering questions of privacy, annoyance or the variety of malicious purposes for which it is used, such as fraud and identity theft. E-mail employs a very open architecture that lacks strict authentication mechanisms,[73] and make the bulk sending of e-mail very difficult to prevent.

E-mail spam shares much in common with the problem of unclean metadata. It is an abuse of a network, imposing costs upon the public for private benefit. Indeed, the use of unclean metadata has been called "spamdexing."[74] The harm is similarly diffuse. The same private benefit and self-promotion is involved, and the regulation of spam presents similar, though more direct, concerns regarding freedom of speech, jurisdiction, and anonymity.

Presently, spam differs from unclean metadata in that each spammer (a person who sends spam) is individually responsible for a significant amount of harm. By contrast, at present each publisher of unclean metadata is only responsible for a small amount of harm, as the greater part of the harm is caused by the general inutility of metadata due to widespread abuse. However, if metadata were generally clean, and

---

72. Dominic-Chantale Alepin, *Opting-Out: A Technical, Legal and Practical Look at the CAN-Spam Act of 2003*, 28 Colum. J. L. & Arts 41, 42 (2004) ($9.4 billion world-wide); Erin Elizabeth Marks, *Spammers Clog In-boxes Everywhere: Will the CAN-Spam Act of 2003 Halt the Invasion?* 54 Case W. Res. 943, 944 (2004) ($20.5 billion world-wide); Reagan Smith, *Eliminating the Spam From Your Internet Diet: the Possible Effects of the Unsolicited Commercial Electronic Mail Act of 2001 on Junk E-Mail*, 35 Tex. Tech L. Rev. 411, 412 (2003) ($10 billion in the U.S.); Richard Warner, *Spam and Beyond: Freedom, Efficiency, and the Regulation of E-Mail Advertising*, 22 John Marshall J. Computer & Info. L. 141, 141 (2003) ($20 billion world-wide); Adam Mossoff, *Spam – Oy, What a Nuisance!* 19 Berkeley Tech. L. J. 625, 628 (2004) (up to $198 billion by 2007); Elizabeth Alongi, *Has the US Canned Spam?* 46 Ariz. L. Rev. 263, 263 (2004) ($11.9 billion in the U.S. and Europe in 2002); Jordan Blanke, *Canned Spam: New State and Federal Legislation Attempts to Put a Lid On It*, 7 Comp. L. Rev. & Tech. J. 305, 305 (2004) ($10 to $87 billion in the U.S. in 2003).

73. Warner, *supra* n. 72, at 147-48; Fed. Trade Commn., *National Do Not Email Registry A Report to Congress*, 3-12, http://www.ftc.gov/reports/dneregistry/report.pdf (accessed April 9, 2006) (informing Congress that the FTC will not set up a "do not e-mail" registry, despite being empowered to do so by CAN-SPAM, because such a registry would be ineffective or even counter-productive).

74. Nathenson, *supra* n. 10.

therefore trusted, one person might individually inflict a significant amount of harm by publishing unclean metadata.

In 2003 Congress enacted the *Controlling the Assault of Non-Solicited Pornography and Marketing Act* ("CAN-SPAM"),[75] and on January 1, 2004 it came into effect at the expense of a number of states' legislative schemes.[76] CAN-SPAM's substantive provisions, particularly its use of an opt-out rather than opt-in scheme, have been criticized as legitimizing rather than preventing spam.[77] Unfortunately, it is hard to separate the difficulties caused by the Act's substantive provisions from the problems of enforcement. Nevertheless, CAN-SPAM is instructive in three ways.

First, CAN-SPAM relies upon traditional enforcement mechanisms. It makes certain spam-related behavior criminal, which includes sending e-mails with falsified header information and harvesting e-mail addresses,[78] and subjects others to considerable civil penalties based on either actual loss or statutory damages.[79] CAN-SPAM removes the rights of individuals to bring civil actions against spammers, instead granting powers of enforcement to the Federal Trade Commission ("FTC"), State Attorney Generals, and Internet Service Providers.[80] Clearly, legislators felt that making a civil action available to the general public would be inappropriate, and instead granted enforcement authority to particular persons on behalf of the public at large. The designation of public persons such as State Attorney Generals and the FTC allows the exercise of appropriate discretion in the enforcement of the legislation, mitigating any unintended side effects which the legislation might have on its face. The FTC claims that CAN-SPAM has helped reduce the amount of spam sent,[81] and there have been over fifty actions brought under the

75. Pub. L. No. 108-187, 117 Stat. 2699-2719 (2003).

76. 15 U.S.C. § 7707 (2006).

77. *See* Alongi, *supra* n. 72, at 287-88 (Critics have suggested that the word 'can' in the act's title should be read as 'to be able to' rather than 'to throw in the trash.').

78. 15 U.S.C § 7704 (2006).

79. 15 U.S.C § 7706 (f)-(g) (2006).

80. 15 U.S.C § 7706 (2006); *see* 15 U.S.C. § 7706 (b) (2006) (Other federal agencies and parties, such as banks, can take action against particular types of spammers, such as those undertaking phishing or other fraudulent activities.).

81. Fed. Trade Commn., *Effectiveness and Enforcement of the CAN-SPAM Act: A Report to Congress*, 7-8, 23, http://www.ftc.gov/reports/canspam05/051220canspamrpt.pdf (accessed April 9, 2006) (However, the FTC's report casts a decrease in the percentage of e-mail which is spam as a sign that progress is being made, despite the report's own source indicating that the actual volume of spam has increased.); MX Logic, *MX Logic Reports Spam Accounts for 67 Percent of All Email in 2005*, 9-13, http://www.mxlogic.com/news_events/press_releases/09_22_05_SpamStats.html (accessed April 9, 2006) (The progress is made is more accurately described as a slowing of the growth of spam to below the rate of the growth of legitimate e-mail. Also, the report acknowledges that much of the progress is attributable to other factors such as advances in anti-spam technology.); *see*

scheme.[82]

However, while the prosecutions evince some limited success within the United States, CAN-SPAM demonstrates the difficulties involved in using national legislation to address a global problem.[83] The amount of spam sent to Internet users has in fact increased since CAN-SPAM came into effect.[84] Those involved in the sending of spam recognize the importance of jurisdictional impediments to prosecution.[85] Recognizing the limits of existing means of international enforcement, the FTC proposed and Congress passed the US SAFE WEB Act to extend its own capacity to cooperate with other agencies and pursue offenders overseas.[86] The lack of significant progress, despite considerable International Telecommunication Union and Organization for Economic Co-operation and Development attention to the problem, highlights the difficulty of achieving concerted and effective international action.[87] At this point, CAN-SPAM's history does not bode well for the capacity to eliminate metadata through regulation enforced through these traditional avenues.

Second, CAN-SPAM demonstrates that technical problems of this nature are potentially of sufficient political significance to compel legislation. Laypersons understand the problem of spam, whether by analogy from other forms of direct marketing or through experience; the problem

---

also Anne Broache, *FTC says federal spam law has worked*, http://news.com.com/FTC+says+Federal+spam+law+has+worked/2100-1028_3-6003071.html (April 9, 2006).

82. Fed. Trade Commn, *supra* n. 81, at ii.

83. ' *Id.* at 25 (stating "[t]hese obstacles are formidable, and in some instances, insurmountable").

84. MX Logic, *supra* n. 81.

85. *See e.g.* hostbp.com, *Overseas Bulletproof Web Hosting*, http://hostbp.com/BulletProof.htm (accessed April 9, 2006) ("Another reason we do this [allow the sending of spam] is that we put your [Web site] in our overseas servers where the local law will protect your [Web sites] should not be shut down by any reason (sic). No illegal content is accepted in our servers."). This Web site may, due to its nature, no longer be available. A copy is on file with the author. A search for terms such as 'bulletproof hosting' should reveal similar Web sites.

86. Pub. L. No. 109-455, 120 Stat. 3372 (2006) (*The Undertaking Spam, Spyware, and Fraud Enforcement with Enforcers beyond Borders Act of 2005*,introduced by Senator Gordon H Smith, OR); Fed. Trade Commn., *The US SAFE WEB Act: Protecting Consumers from Spam, Spyware, and Fraud*, http://www.ftc.gov/reports/ussafeweb/USSAFEWEB.pdf (accessed April 9, 2006) (The Act is intended "to address the challenges posed by globalization of fraudulent, deceptive, and unfair practices" by, among other things, improving its capacity for cooperation with its foreign counterparts and strengthening its participation in international projects.). On Oct. 10, 2007 the FTC announced the first law enforcement action brought using the US SAFE WEB Act and by sharing information with foreign partners.

87. International Telecommunication Union, *ITU Activities on Countering Spam*, http://www.itu.int/osg/spu/spam/ (accessed April 27, 2006); Organization for Economic Co-operation and Development, *OECD Work on Spam*, http://www.oecd.org/topic/0,2686,en_2649_22555297_1_1_1_1_37441,00.html (accessed April 27, 2006).

of unclean metadata should be similarly comprehensible.[88] Furthermore, CAN-SPAM demonstrates that the U.S. Congress is prepared to intervene on the Internet in the public interest, and not simply defer to the balance established by the technological efforts of either side. It supports the understanding that although regulation of Internet-related behavior can be troublesome, "regulation need not be perfect to be effective – that regulation works through transaction cost rather than hermetic seal."[89]

Finally, CAN-SPAM prohibits the use of deceptive subject headings. A subject heading is misleading if, based on an objective test, it would mislead the recipient.[90] This is useful because it introduces a precedent – tested in at least eight cases[91] – with regard to freedom of speech concerns. It also highlights the unacceptability of deceptive online marketing practices. It is only a small step from prohibiting the use of deceptive subject headings in e-mail to prohibiting the use of deceptive metadata on the Web.

C. Cybersquatting and the Anti-Cybersquatting Consumer Protection Act

Cybersquatting has been defined as "the deliberate, bad-faith, and abusive registration of Internet domain names in violation of the rights of trademark owners."[92] The initial response to cybersquatting was to bring actions under the rubric of trademark protection: direct trademark infringement, unfair competition, dilution, contributory infringement, and vicarious liability.[93] However, these actions were not overly successful as important elements required to establish liability were not present in cases of cybersquatting.[94] In addition, the problems raised by jurisdiction, such as comity and enforcement of judgments, compromised actions which might otherwise have been successful.[95]

The response of the United States' Congress was to enact legislation adapting trademark actions to make them more effective against cybersquatting. That legislation is the *Anti-Cybersquatting Consumer Protec-*

88. The problem of unclean metadata was summarized aptly by a colleague with the question: "You know how when you search for anything on the Internet (sic) you get porn?" The comprehension is intuitive; the challenge is in convincing people that the problem is not "just the way the Web is" but rather is a result of misbehavior, and is potentially remediable through legislation.

89. Wu, *supra* n. 32, at 1195.

90. 15 U.S.C. § 7706 (a)(2).

91. Fed. Trade Comm'n, *supra* n. 81, at A-10.

92. Brian Holland, *Tempest in a Teapot or Tidal Wave? Cybersquatting Rights and Remedies Run Amok*, 10 J. Tech. L. & Pol'y 301, 307 (2005).

93. *Id.* at 311.

94. *Id.* at 311-15; Milton Mueller, *Ruling the Root* 115 (MIT Press 2002).

95. Holland, *supra* n. 92, at 315-16.

tion Act ("ACPA"),[96] enacted in 1999. In addition to tailoring the substantive requirements of trademark infringement actions to cybersquatting, the ACPA also addresses the problem of jurisdiction. An action *in rem* is allowed where personal jurisdiction over the owner of the domain cannot be established. In that case, the remedies available, which normally include statutory damages as well as traditional trademark remedies of damages and injunctions, are limited to an order for the forfeiture or cancellation of the domain name.[97] This in-kind remedy provides very effective relief for plaintiffs, as its execution takes little more than a few keystrokes, but it relies on U.S. control of a significant part of the domain name system.

The domain name system ("DNS") is a look-up mechanism which translates memorable names such as 'walmart.com' to Internet protocal ("IP") addresses. IP addresses are numeric addresses used to identify computers on the Internet.[98] Ultimate responsibility for the domain name system rests with the Internet Corporation for Assigned Names and Numbers (ICANN), a private non-profit company.[99] The U.S. government exercises *de jure* control over ICANN. In addition to ICANN being a California corporation, ICANN's authority comes from a memorandum of understanding between itself and the U.S. government.[100] This relationship has been the source of considerable international controversy[101] and academic commentary.[102]

As the U.S. government is capable of regulating ICANN, it can address the issue of cybersquatting in domains under the control of ICANN through direct intervention by directing that orders made under the ACPA be enforced, in other words, that ICANN carry out the cancellation or forfeiture of the domain names in question. However, not all do-

96. Pub. L. No. 106-113, 113 Stat. 1536 (1999).

97. 15 U.S.C. § 1125(d)(2)(D)(I) (2006).

98. ICANN, *What is the Domain Name System?* http://www.icann.org/faq/#dns (accessed April 7, 2006).

99. A Michael Froomkin and Mark Lemley, *ICANN and Antitrust*, U. Ill. L. Rev. 1, 2 (2003); Holland, *supra* n. 92, at 306; ICANN, *What is ICANN?* http://www.icann.org/faq/#WhatisICANN (accessed April 7, 2006).

100. National Telecommunications and Information Administration, *Memorandum of Understanding Between the Department of Commerce and the Internet Corporation for Assigned Names and Numbers (ICANN)*, http://www.ntia.doc.gov/ntiahome/domainname/icann.htm (accessed April 7, 2006).

101. Editorial, *Internet Control*, Business Standard, 11 (Nov. 18, 2005); Andy Sullivan, *Digital divide a focus at close of Net summit*, Reuters News (Nov. 19, 2005); Matt Moore, *Tech summit ends with Internet control in question despite agreement*, Associated Press Newswires (Nov. 19, 2005); Goldsmith & Wu, *supra* n. 22, at 169-71; Mueller, *supra* n. 94, at 223.

102. *See generally*, Mueller, *supra* n. 94; Froomkin & Lemley, *supra* n. 99; John Palfrey, *The End of the Experiment: How ICANN's Foray into Global Internet Democracy Failed*, 17(2) Harv. J. L. & Tech. 409, 410-11 (2004).

mains are under the direct control of ICANN. Domains come in two types: global domains such as .com, .net and .org (known as generic top level domains or gTLDs); and country-specific domains such as .com.au, .co.uk and .org.nz (known as country code top level domains or ccTLDs). While control of the former still rests with ICANN, control of the latter has been ceded to local authorities in each country.

While ICANN has a legitimate role in ensuring compliance with technical standards, these local authorities would resist attempts to force compliance with U.S. law.[103] Technically, ICANN could withdraw the delegation of power to a country's authority,[104] but that is not a politically viable option – it would result in the breakdown of the entire domain name system and compromise the operation of the Internet.[105]

The importance of gTLDs means that the orders available under the ACPA are potentially effective remedies. An action by a U.S. company against a cybersquatter, whether the cybersquatter is subject to U.S. jurisdiction or not, is most likely to involve a gTLD because these domains are both the most common[106] and the most commonly used by U.S. companies. However, the orders provided for by the ACPA are of no effect against cybersquatters holding ccTLDs outside the United States. Moreover, the U.S. is unique in its ability to regulate gTLDs. No other country has that privilege; if the same legislation was enacted in another jurisdiction it would cover only that country's own ccTLDs.

Clearly, in-kind penalties can provide a very effective and easily enforceable remedy where jurisdiction would otherwise make enforcement

103. Andrew Orlowski, *Country code chiefs, registrars mull ICANN breakaway*, http://www.theregister.co.uk/2000/11/25/country_code_chiefs_registrars_mull_2 (accessed April 25, 2006); Laurence Helfer and Graeme Dinwoodie, *Designing Non-National Systems: the Case of the Uniform Domain Name Dispute Resolution Policy*, 43 Wm. & Mary L. Rev. 141, 238 (2001); Mueller, *supra* n. 94, at 7-8.

104. This could be done very easily, by amending the root domain name server to indicate that a server other than the one currently in use is responsible for domain names which contain that country code.

105. *See* Joseph Liu, *Legitimacy and Authority in Internet Coordination: A Domain Name Case Study*, 74 Ind. L.J. 587, 593 (1999). U.S. control over the domain name system is accepted internationally because its administration has so far been both unintrusive and effective. Attempting to use the domain name system as a means to enforce its laws in other jurisdictions would be seen as a flagrant abuse, and would likely lead to the establishment of an international organization running a separate root domain name server. This would effectively create two different Internets on the same infrastructure. The universality of the present domain name system is not based on law – it is largely voluntary.

106. Caslon Analytics, *Note on gTLD and 2LD sizes*, http://www.caslon.com.au/dnssizes note.htm (accessed April 25, 2006) (As of January 2003 the .com gTLD was almost 4 times the size of the largest ccTLD(.de), and the .net, .org, .info and .biz domains were each larger than almost all ccTLDs.).

difficult.[107] However, it is important to note that as a centralized rather than decentralized system, DNS is an anomaly on the Internet. Other problems, such as spam cannot be addressed via DNS as it does not have the same centralizing effect in all contexts. In addition, DNS is only effectively centralized with respect to gTLDs; indeed, for countries other than the U.S., DNS is centralized only with regard to each country's ccTLDs, and gTLDs are not regulable at all.[108] Two lessons, then, should be learned from the ACPA. First, where traditional alternatives are ineffective due to the nature of the Internet, in-kind remedies can be a very effective form of relief. Second, DNS, as a centralized point in the architecture of the Internet, can, with some limitations, be used as a way for national governments to regulate the Internet.[109]

D. Cybersquatting and the Uniform Domain Name Dispute Resolution Policy

An alternative approach to cybersquatting is encapsulated in ICANN's Uniform Domain Name Dispute Resolution Policy ("UDRP").[110] ICANN has required all registrars to agree to abide by the UDRP since its adoption in 1999. It allows trademark holders to bring an administrative action before an approved dispute resolution service provider where they allege that an identical or confusingly similar domain name was registered and is being used in bad faith, without the holder having any rights or legitimate interests in the name.[111] Remedies available are limited to cancellation and forfeiture of the domain.[112]

The UDRP has been heavily criticized for perceived procedural biases against defendants and the inconsistency of decisions with the policy itself.[113] In its defense, the UDRP can be seen as nothing more than an extension of the decision making process to include the complaints of interested parties; what ICANN giveth, ICANN taketh away. Furthermore, the rights of both parties to resort to court proceedings are explic-

107. Goldsmith & Wu, *supra* n. 22, at 77-79.

108. Mueller, *supra* n. 94, 225. Cf. Goldsmith and Wu, *supra* n. 22, at 168-71 (stating that '[t]here is, however, a "global law" without which there would be no Internet: the domain name system (DNS).'); Biegel, *supra* n. 43, at 197 (quoting David Post, "[control over the operation of the root server] is quite literally a kind of life-or-death power over the global network itself").

109. Mueller, *supra* n. 94, at 205-08, 219, 221.

110. ICANN, *Uniform Domain Name Dispute Resolution Policy*, http://www.icann.org/udrp/udrp.htm (accessed April 9, 2006) [hereinafter UDRP].

111. *Id.* at Article 4(a).

112. *Id.* at Article 4(i).

113. *See e.g.* Miguel Danielson, *Confusion, Illusion and the Death of Trademark Law in Domain Name Disputes*, 6 J. Tech. L. & Pol'y 3 (2001); Michael Geist, *Fair.com?: An Examination of the Allegations of Systemic Unfairness in the ICANN UDRP*, 27 Brook. J. Int'l L. 903, 935-6 (2002); Mueller, *supra* n. 94, at 193.

itly preserved by the UDRP.[114] Just or not, however, it has been a very significant piece of regulation. ICANN's statistics show that by 2004 there had been almost ten thousand cases involving over fifteen thousand domains, with over two-thirds resulting in cancellations or transfers.[115] The timetable for dispute settlement allows a maximum of 42 days from commencement of proceedings for a decision to be made.[116] In many cases, especially those involving known cybersquatters, the respondent fails to submit a response, further shortening the process.[117]

Like the ACPA, the UDRP utilizes DNS for rapid, effective enforcement. However, the UDRP is not based upon legislation. It is enforced contractually: all registrars providing gTLDs are compelled to apply and enforce it, and all domain name registrants are required to agree to its terms. It is most certainly regulation, in the broad sense in which that term is used here, but it does not come from the usual source (government)[118] and it is a creature of contract rather than legislation. Some concern is justifiable here, as ICANN is seemingly in a position to demand whatever it pleases from domain name owners. Indeed, that is the danger of failure to act legislatively: private authorities or parties with vested interests may take the initiative to establish schemes, in contract or in code, which favor their interests. That concern is mitigated by the oversight of the U.S. government, and the less direct but nonetheless considerable pressure of international scrutiny.

E. Proposing Regulation

Looking to existing regulatory schemes for inspiration, some diverse approaches to problems with similar features to the one at hand provide instruction as to how the regulation of unclean metadata might be undertaken. Spam illustrates the weakness of the conventional enforce-

114. UDRP, *supra* n. 110, at Article 4(k). In any case, it is highly doubtful that ICANN could have taken away the right of a complainant, a non-party to the contract between the registrar and the respondent, to initiate proceedings based on trademark or under the ACPA. It may, however, have been able to require complainants using the UDRP to waive their rights to bring actions in other forums.

115. ICANN, *Statistical Summary of Proceedings Under Uniform Dispute-Resolution Policy*, http://www.icann.org/udrp/proceedings-stat.htm (accessed April 9, 2006).

116. ICANN, *Rules for Uniform Domain Name Dispute Resolution Policy*, http://www.icann.org/udrp/udrp-rules-24oct99.htm (accessed April 9, 2006) (Article 5(a) allows 20 days for the submission of a response, article 6(b) allows 5 days for an appointment of a panel, article 15(b) allows 14 days for a decision, and article 16(a) allows 3 days for communication of the decision to the parties.).

117. Milton Mueller, *Success by Default: A New Profile of Domain Name Trademark Disputes Under ICANN's UDRP*, 14-17, http://dcc.syr.edu/markle/markle-report-final.pdf (accessed April 9, 2006).

118. Although the UDRP could be said to have the tacit support of the U.S. government, given their oversight of ICANN.

ment versus international cooperation approach to problems on the Internet, but shows that legislatures are prepared to regulate on technical matters, even where questions of freedom of speech may be involved. The ACPA and the UDRP highlight the usefulness of in-kind remedies, particularly using the relatively centralized architecture of the domain name system. This provides a means for rapid and effective enforcement, but its limits must be kept in mind. Each jurisdiction controls its own ccTLDs, and while no government outside the U.S. can regulate gTLDs, U.S. control is subject to tacit international consent.

Heeding these lessons, it seems that in-kind regulation is the most attractive option, provided that the effect of extra-jurisdictional actors can be nullified. Although the Internet is frequently referred to as borderless, the natural geographic boundaries of the Internet have been recognized in a number of areas. Sub-networks can be useful in targeting regulation, whether at an internet service provider ("ISP")[119] or a national level.[120] Moreover, ccTLDs carve out national spaces on the Internet, effectively independent of ICANN. IP addresses can be used to determine geographical origin with remarkable accuracy.[121] In less technical ways, language, culture, and user norms also demarcate geographically-rooted portions of the Internet.[122] The difficulty, then, is in matching national jurisdiction to in-kind enforcement in an effective manner. The key is finding, or creating, a virtual border which coincides with the limit of effective powers of enforcement. A scheme based on just such a border is proposed in Part V below.

V. A PROPOSAL – THE CLEAN WEB ACT

The proposal contained in this part is for a piece of legislation targeted at unclean metadata. For want of a catchy name – and distancing the author from the acronyism[123] of recent pieces of U.S. legislation – the Clean Web Act is proposed. The Clean Web Act would not prohibit the use of unclean metadata on the Web. Instead, it would allow authors to guarantee that their metadata is clean on a voluntary basis. By including a guarantee on their page, authors would expose themselves to sanctions if the metadata they provided was in fact not clean. Search engines and other applications would then limit their reliance on

119. Biegel, *supra* n. 43, at 218-19.

120. Goldsmith & Wu, *supra* n. 22, at 3-9, 87-104 (discussing on a French court's effective judgment against Yahoo! and then discussing China's censorship efforts based on the structure of its carefully architected national network).

121. *Id.* at 58-62.

122. *Id.* at 50-53.

123. Acronyism, n. the flagrant abuse of acronyms to garner support, particularly popular in the United States around the turn of the century. See the USA PATRIOT Act, the CAN-SPAM Act, the TEACH Act, the PROTECT Act and the proposed US SAFE WEB Act.

metadata to those sites which guarantee the accuracy of their metadata and are in a jurisdiction in which that guarantee is credible. Enforcement would be primarily via DNS, and would be backed by civil penalties.

A. Unclean Metadata

Any number of reasonable definitions of unclean metadata may be given. Unclean metadata might be defined, as it is in this paper, as metadata which is false, inaccurate or misleading. Alternatively, the criteria applied to deceptive subject headings in e-mail used in CAN-SPAM could be adapted for use here: actual knowledge, or knowledge fairly implied on the basis of objective circumstances, that *the metadata* would be likely to mislead a *person or computer*, acting reasonably under the circumstances, about a material fact regarding the contents or subject matter of the *Web page*.[124] The primary danger in this definition is excessive narrowness. For instance, the simple term 'false' may be not be sufficient to cover all undesirable metadata. Whatever definition is adopted, a jurisprudence based on analogous areas, such as trade practices, is sure to develop to delimit the boundary between clean and unclean metadata more exactly.

B. The Guarantee

The use of a guarantee makes participation voluntary. This voluntariness is a very appealing characteristic because it places only the most minor restriction on what can be done on the Web. The best way in which to allow authors to guarantee the cleanness of the metadata contained in their Web page is to embed the guarantee itself as metadata.[125] This could be read by search engines and browsers alike, and would not interfere with the functionality of the Web at all.[126] If the guarantee was present, search engines, or indeed, any application which has a use for the metadata,[127] could treat the metadata as authoritative and trustwor-

124. 15 U.S.C. § 7706 (a)(2) (2006 (emphasis added).

125. For example, as a HTML tag such as <meta name="CleanWebAct"value="guarantee"/>.

126. World Wide Web Consortium, *User Agent Conformance*, http://www.w3.org/TR/xhtml1/#uaconf (accessed April 26, 2006) (The HTML standard requires that tags which are not understood be ignored. Many Web pages already contain metadata which is not relevant to Web browsers, such as information on versioning and authorship used by the author to manage the content, and tags which cannot be understood by all browsers, such as frames).

127. For example, browsers might use the metadata to annotate the user's browsing history, to allow the user's browsing history to be searched or to organise documents which have been viewed together in conceptual groups (a kind of automatic bookmarking).

thy.[128] Penalties would be imposed upon those presenting guarantees but not complying with them.

Moreover, assuming for the moment that enforceable penalties are available, search engines can limit their trust of guarantees to those sites which are in jurisdictions which effectively enforce compliance with such guarantees. For example, if the United States and Australia had enacted this kind of legislation, and were effectively enforcing compliance with it, search engines utilizing metadata would limit their searches to Web pages which fulfilled two conditions: they contained the guarantee of cleanness and they were in the '.us' or '.au' ccTLDs. The decision on the reliability of any particular jurisdiction would be made by the search engine or application in question.[129] This allows the legislation to be effective whether enacted in one jurisdiction or all, and provides an impetus for jurisdictions to enact and enforce the system in order that sites within their ccTLDs might be included in metadata-based searches.

C. PENALTIES

DNS provides the most attractive enforcement mechanism for this scheme. If a Web site is non-compliant, that is, it provides a guarantee but has unclean metadata, its domain name would be forfeited, effectively removing it from the Web.[130] This is a simple, low cost, and effective penalty available to all countries against Web sites with domain names in their ccTLDs. Furthermore, it is exercisable against all Web sites, whether the author or owner is in the regulating jurisdiction or not.

It is arguable that the benefit gained from having a Web site listed highly in the results of popular searches even temporarily would be worth the sacrifice of a domain name. A domain name might be bought for the sole purpose of attracting a burst of traffic in order to redirect users to a valid site or display advertising. The best solution to this problem is to limit the registration of domain names in each ccTLD to those with a presence in the jurisdiction.[131] This is the case in some ccTLDs

128. S Bellovin, *RFC 3514: The Security Flag in the IPv4 Header*, http://www.ietf.org/rfc/rfc3514.txt (accessed April 19, 2006) (This may remind Internet-savvy readers of the famous 'evil bit' RFC, requiring that malicious communications identify themselves in order that they may be efficiently intercepted by security devices).

129. It would seem likely that authorities, either formal or informal, would develop rating systems and make pronouncements on the levels of compliance in various jurisdictions.

130. The Web site would still be accessible via its IP address, but that has little bearing on the effectiveness of the sanction.

131. This might seem quite a restrictive condition, but it is hardly unfair. The very purpose of ccTLDs is to allow Web sites to identify themselves with a particular geographic area. Those who wish to obtain a domain name in a particular country but do not have the

already.[132] With this limitation in place, civil or even criminal penalties could be used to further deter blatant or repeat violations.[133]

D. ENFORCEMENT

Responsibility for enforcement is best placed in the hands of a government agency. Existing agencies, especially those charged with monitoring deceptive advertising and other unfair trade practices, may be well-placed to take on such a role.[134] A government agency would collect information on breaches of the legislation from a wide variety of sources in order to act against offenders. Users of search engines could be provided with simple tools for providing feedback where they believe a page's metadata is not accurate, based on its lack of relevance to their search.[135] Where a sufficient number of reports were made, search engines could automatically forward information to the agency. Many companies watch their search rankings very closely, and would be able to provide information on violations to the agency. Search engines already appear to have tools for monitoring this kind of abuse.[136] The agency could run a Web crawling robot which compares page content to

---

required presence could use agents to register the appropriate domains, indemnifying them against penalties. Some countries may choose to expand registration to parties in other countries with whom they have reciprocal enforcement agreements.

132. *See e.g.* auDA, *Current auDA Published Policies*, http://www.auda.com.au/policies/ (accessed April 12, 2006) (Australia, for instance, already has such a restriction for all of its ccTLDs; Domain name licences may only be allocated to a registrant who is Australian, as defined under the eligibility and allocation rules for each 2LD, [such as .com.au or .edu.au].."); CIRA, *Policy Development Process,* http://www.cira.ca/en/cat_Registrar.html (accessed April 12, 2006) (for Canada's similar provisions); NeuStar.Inc., *The usTLD Nexus Requirements*, http://www.neustar.us/policies/docs/ustld_nexus_requirements.pdf (accessed April 12, 2006) (Although not strictly a ccTLD, the .us gTLD has a nexus requirement which requires that the registrant be a person resident in the U.S., a U.S. entity or organization or a foreign entity or organization which has a 'bona fide' presence in the U.S.).

133. Anonymity would not appear to be a great difficulty. If the contact information in the registry proves inaccurate – and it shouldn't given that registrars need to verify that the registrant has sufficient connection to the ccTLD in question – forfeiture would be automatic and investigation aimed at imposing more serious penalties could proceed based on the means of payment for the domain, the content of the Web site and the location of the host computer.

134. *See e.g.* 15 U.S.C. § 45(a)(1) (2206) (Yhe FTC is perfectly placed to carry out such a role, based on the mandate which states, "[u]nfair methods of competition in or affecting commerce, and unfair or deceptive acts or practices in or affecting commerce, are hereby declared unlawful.").

135. Feedback could be given to the search engine via a simple link named "irrelevant" displayed beside each link in a search engine's results, or directly to the agency by way of a button in the browser pressed with the offending page loaded.

136. BBC News, *BMW Given Google "Death Penalty"*, http://news.bbc.co.uk/1/hi/technology/4685750.stm (accessed April 26, 2006); Tom Espiner, *Google Blacklists BMW.de*, http://news.zdnet.com/2100-1009_22-6035412.html (accessed April 26, 2006).

metadata in order to identify potential breaches.[137]

E. Variations and Extensions

One interesting variation is the use of software for the generation of keywords. One piece of metadata would be reserved for the output of software which analyzes the Web page. That metadata would then be assessed as clean if it matched the output of the software, allowing the automation of compliance monitoring. This could be used in addition to human-authored metadata, and it is easily extended to allow a number of pieces of software to be used.[138]

However, this introduces a number of problems. First, this would encourage Web site owners to design Web pages to produce a certain set of metadata, recreating the technological arms race of software writers versus Web page authors, which has led to the problem at hand. Second, each piece of software would have to be optimized for a particular type of page, academic papers or news articles for instance. Otherwise, there is no reason to believe that a variation of the scheme with software-produced metadata could be any more effective than existing search engines.[139]

This is only one example of how a platform for clean metadata could be built upon by custom applications. Allowing the reservation of certain metadata tags for particular purposes – in the manner of port numbers – would encourage standardization and interoperability and allow innovation.[140]

---

137. *See* Whatis?.com, *Look It Up*, *Crawler*, http://whatis.techtarget.com/whome/0, 289825,sid9,00.html (accessed oct. 15, 2007) ("Crawler is a program that visits Web sites and reads their pages and other information in order to create entries for a search engine index.").

138. *See* Lessig, *supra* n. 34, at 100-108. For example, if the FTC were to supply a metadata generation tool, it would reserve the use of a tag such as <meta name= "autowords"value="FTC">traffic monitoring enforcement speed limits</meta>. Other software providers could reserve other values, such as <meta name="autowords" value= "Google">. . .</meta>. This would allow a competitive market for the use of such software to emerge, resulting in a market-selected standard. Lessig's comments about the benefits of open code, especially in terms of transparency, bear consideration here. A positive and not unlikely result is the emergence of a cooperatively developed standard tool from the technical community.

139. However, the metadata would be distributed rather than centralized, realizing the benefits discussed previously. *Supra*, pt. II (discussing the impact of claen metadata in the search space).

140. *Infra*, pt. VI (discussing the probabilities of success).

VI. CHALLENGES AND ASSESSMENT

A. GLOBAL TOP LEVEL DOMAINS AND HOSTS WITHOUT DOMAIN NAMES

The scheme outlined above is based solely upon ccTLDs, excluding both gTLDs and hosts without domain names from offering guarantees of clean metadata. With respect to gTLDs, this problem can be dealt with in at least four ways. First, as highlighted by the ACPA, the U.S. can regulate this namespace. In order to be effective, however, the U.S. would have to limit registrations to persons within its jurisdiction. This would be highly contentious as gTLDs are seen as global addresses. Second, gTLDs could be abolished, leaving only ccTLDs. This too would be controversial in light of the value presently assigned to such domains. Thirdly, gTLDs could simply be excluded from the system, leaving sites with those types of domain names unable to offer guarantees of clean metadata. However, that would prevent participation by a very large and important set of Web sites.[141]

Finally, gTLDs could be included in the system by way of representations from sites located in regulated ccTLDs. If a site with a domain name in a gTLD wished to give a metadata guarantee, it would have to obtain a domain name in a ccTLD. This is not a difficult requirement to comply with, as a single Web site may have multiple domain names. The site's guarantee would then include a list of sites to which it wishes to extend its guarantee of clean metadata – sites for which it takes responsibility.[142] While the simplest scenario is offering a guarantee for the same site accessed via its gTLD name, guarantees could be extended to mirrors in unregulated ccTLDs or even completely unrelated sites. Hosts without domain names could be included in the same manner. Allowing sites to warrant the cleanness of other sites' metadata allows the system to be extended to gTLDs and unregulated ccTLDs without compromising the global nature of the former or the independence of the latter.

B. ICANN IS AN INAPPROPRIATE REGULATORY TOOL

There have been strong arguments made against the use of ICANN as a policy-based regulator. These arguments focus on ICANN's lack of democratic legitimacy as well as international suspicion, and they are persuasive. ICANN is a technical body; its mandate, and indeed its continued existence, is premised upon its role being restricted to ensuring that DNS functions. It is therefore a strength of the proposed scheme

141. Caslon Analytics, *supra* n. 108 (Almost two-thirds of all existing domain names were gTLDs in 2003. Their commercial significance gives them even greater importance.).

142. In addition, the site receiving the guarantee would have to include a reference to the site giving the guarantee so that those viewing the receiving site can go and verify the existence of the guarantee.

that it does not rely upon ICANN for enforcement. Each jurisdiction is capable of regulating its own ccTLDs by imposing requirements on whatever entity runs its registry. This is not a scheme in which centralization of power in the root server is used to create an international autocracy. It is a scheme in which national governments exercise their power to regulate their own jurisdictions in the most effective manner possible.

C. Third Party Speech

Many Web sites do not have control over all of the pages which are published on sites using their domain name. Examples include blog hosting sites and free Web hosts.[143] If a client of the Web site owner offered a guarantee on their Web page, the site owner, as the owner of the domain name, would be liable for any inaccuracy. This problem can be solved by employing an existing mechanism. At present, search engines check for a file named 'robots.txt' on each site which they index.[144] It is found in a standard location, at the root of the site.[145] This file contains instructions from the site operator indicating which parts of the site are to be indexed and which are not. A similar scheme, or an expansion of the information in robots.txt, could be introduced for the meta-guarantee to indicate that guarantees are not reliable for certain parts of the site. This would prevent third parties from rendering the Web site owners who host their Web pages liable for failure to comply with a guarantee.

D. Non-participation and Non-compliance

This kind of scheme exhibits network effects: "the value of [the] system. . .to its users tends to increase as other users adopt the same system."[146] As such, it needs to reach a critical mass of use in order to survive, but after that critical mass has been reached, the network effects encourage universal adoption.[147] For this scheme, it might be said that there is little incentive in exposing oneself to the risk of sanctions if search engines are not utilizing the metadata; but search engines will not invest in the development necessary to rely on the metadata until

143. Interactive sites, such as those on which users can post comments, do not suffer from the same problem. Comments are published in the body of a Web page (the part which contains the content of the page), whereas the guarantee would be located in the head section of the page (separate from and preceding the body), meaning that users cannot insert a guarantee in their comments.

144. *See* The Web Robots Pages, *A Standard for Robot Exclusion*, http://www.robotstxt.org/wc/norobots.html (accessed April 26, 2006) (In fact, this practice is not encapsulated in an informal standard, although major search engines generally respect it.).

145. *See e.g.* http://www.google.com/robots.txt (accessed April 26, 2006).

146. Milton Mueller, *supra* n. 94, at 52.

147. *Id.* at 53.

guarantees are sufficiently widespread to make them useful. In the early stages of adoption, a certain number of parties need to invest in the technology either altruistically or with a view to the long-term benefits. Often an outside impetus such as a subsidy is necessary in order to reach the critical mass.[148]

Failure to reach critical mass is a risk in the implementation of this scheme, but it is not a major obstacle. First, there are no competing standards in this case. The choice is between non-participation and participation, and legitimate Web publishers would perceive the benefit of improved findability that is gained through participation. Second, compliance is a simple matter for most Web publishers, who would not have unclean metadata on their pages in the first place. Third, the discouraging effect of the risk of sanction can be minimized by taking a light-handed approach to enforcement initially. Fourth, the investment required to utilize clean metadata is almost trivial. A search engine relying on clean metadata could be implemented by a group of computer science students without difficulty.[149] Fifth, adoption can begin in sub-communities such as academic circles, reducing the scale of operation of the search. With a reduced scale, the cost of search would be lower, and each sub-community would have its own, much lower critical mass. Sixth, if attaining critical mass proves difficult in spite of all these factors, the government could mandate use of the scheme and provision of a search engine by its various agencies. The cost of compliance with such a requirement would be insignificant as government agencies would presumedly not use unclean metadata.

Note, however, that these factors would only become relevant if the legislation was not warmly received initially. If there was doubt as to the usefulness of clean metadata or the potential efficacy of the scheme, Web site owners might take a wait and see approach. However, if the Web community believed in the benefits of clean metadata and the capacity of this proposal to guarantee it, the achievement of critical mass would be a foregone conclusion. The mere proposal of this kind of legislation would generate comment and controversy amongst technologists and Web developers. Its likelihood of success would have been debated at length before its formal commencement. Critical mass would then be a question of popular support based upon belief in the likely efficacy of the legislation, defaulting to gradual adoption based upon individual benefit if popular commitment is initially absent.

148. *Id.* at 83-85.

149. Obviously there are degrees of sophistication, and the development of a fully fledged search engine would require a more significant investment. However, given the success of information technology icons such as Google, it is hard to believe that either entrepreneurial developers or funding would be difficult to find.

Assuming that a critical mass is achievable, the problem switches from lack of participation to overwhelming non-compliant participation. A culture of non-compliance is avoided by making the scheme voluntary, but the credibility of the scheme could be undermined by a large number of Web sites blatantly abusing the system. The key here is clearly effective and persistent enforcement. Whereas a light-handed approach might be applied to those who appear to have inadvertently breached the guarantee in its early days, a much stricter approach ought to be taken to those who blatantly abuse the guarantee after it has reached a critical mass. Also, in the search domain, the area where clean metadata would be most immediately useful, targeted searches will exclude sites which use too much metadata. For example, a researcher looking for academic articles and trying to avoid commercial Web sites might require that the metadata not include the word "shop." A publisher who has misleadingly included the word "shop" will then be excluded from this search. This discourages over-indulgence in the use of metadata, as metadata can be used to exclude as well as include.

E. The Difficulty of Judgment

Undeniably, there is potential for difficult cases in assessing the cleanness of metadata. However, this is not a legitimate reason to reject the whole scheme. Both the judiciary and the executive make difficult judgments on a daily basis. A wide variety of laws allow for circumstances of dubious legality. Our courts have centuries of experience in resolving exactly these types of issues. Moreover, they do it transparently and accountably, subject to the checks and balances of public power. They can clarify and refine the law, taking into account the various interests which emerge through conflict. This is a far superior alternative to leaving these judgments in the hands of private companies, operating behind closed doors and with no binding rules.

Furthermore, when considering the use of clean metadata as a platform, many of its applications will not require hard judgments as to accuracy. Only where the metadata is determined by human interpretation of content will there be the possibility of contention; where metadata is used in a way that is not dependent on human judgment, there is no room for debate. There is no possibility of difficult cases where metadata is generated by machine analysis, rule or formula.[150]

150. A good example is the hotel search engine described below under "H. Probabilities of Success." In that case the metadata in question is a summary of hotel room features, prices and availability. The accuracy of the metadata is determinable trivially and without room for controversy.

F. Free Speech and Discrimination

While free speech advocates might be placated by the voluntary nature of this scheme, Internet libertarians might argue that it divides information into classes: trustworthy and untrustworthy. Those who are unable to give a guarantee, whether because they do not own the domain name on which they publish or because they cannot risk the sanctions, are relegated to relative obscurity.[151] That argument might be extended to suggest that metadata guarantees could be a tool for censorship or propaganda. Information agreeable to the government is given trusted status; dissent must remain untrustworthy, lest the publisher be pursued by the government.

In response, one might point to the fact that the proposed scheme is not centralized in any way. There is no authority which has to be satisfied before a guarantee can be given. Even in enforcement, authority is distributed internationally, rather than being concentrated in one institution or agency. A page which is shut down in one jurisdiction can be posted by another site owner in another jurisdiction without fear of being pursued by the same authority. Jurisdictions which regulate metadata for political ends will quickly cease to be trusted by users. Where a government is prepared to use metadata regulation to suppress views, the effect is likely to be trivial in comparison to the impact of the direct censorship which that government is engaged in. Then as now, the results of a search for "Tiananmen Square" on a Chinese search engine would not be expected to be impartial. This leads on to a more complete answer to this criticism.

Trust already operates as an informal filter on the Web. Users assess information based on the nature of its source. News from the New York Times is considered more authoritative than the ramblings of a soapboxing citizen. Credit card information is given to Dell Corporation far less reticently than it is given to Joe's Computer Shop, no fixed address. Web users link authenticity and credibility to accountability. The New York Times is accountable; Joe's computer shop is not. To a Web user in the U.S., a U.S. company might be more accountable than a Nigerian one. The proposed scheme simply formalizes trust derived from accountability in a way which allows it to be utilized by search engines. Indeed, it is more egalitarian than simply relying on known brand names or institutions, as the meta-guarantee of Joe's Computer Shop is

151. *Infra*, pt. VI (discussing howe the extensibility of the Web is not effected). This argument is flawed to the extent that the existing mechanisms of the Web – current search engines, for example – are not compromised in any way by the proposed scheme. The reference to 'relative obscurity' is somewhat misleading – the sites might be less findable in some respects than those offering a metadata guarantee, but they would still be as findable as they are presently.

equivalent to the meta-guarantee of Dell Corporation. In addition, trust and authority are closely tied to findability,[152] and democratized search democratizes findability.

G. Extensibility: The Web is not Effected

One highly favorable characteristic of the proposed scheme is its extensibility, in both of the senses described above. First, it does not limit the functionality of the Web in any way. Just as the cost of compliance to individual parties is nil, the cost to the public in terms of loss of facility or potential in the Web is nil. This is a powerful counter-argument to claims of infringement upon free speech and undue restrictiveness. The market is free to vote with its feet. Failure costs the public nothing, and will surely provide insights into future regulation of the Internet.[153] The Web is not being upgraded to a new model, more desirable to government or commerce; rather an extension is being offered – one might say a new layer is being added[154] – leaving users free to make a choice whether to utilize it or not.

The proposal at hand seeks to bring clean metadata to the Web by imbuing a particular artifact in code – the guarantee expressed in metadata – with legal significance. Although the effect is through code, and code is generally considered to be the architecture of the Internet,[155] giving legal significance to a statement in code in this way is not a change to the architecture of the Internet. While almost all of the architecture of the Internet is code, not all code is architecture. The architecture of the Internet is made up of layers: at the bottom is hardware, controlled by code; at the top is code, running on hardware. But on top of this architecture, there can be more code still, not defining the space but rather acting as the medium of interaction with it. Here, code is relevant not as architecture but as a manifestation of the actions of persons in the regulated space. Code is a way to make a promise, and law regulates by giving the promise consequences.

Regulation which does not require restrictive changes to the architecture enables choice. That is the point made in the introduction to this paper – this scheme is about helping users of the Web find the Web they are looking for. The Web is infinitely multi-faceted, and the scheme proposed above does nothing to limit that. The proposed scheme simply adds a new facet. Better said, it adds an infinite set of new facets, novel be-

152. Morville, *supra* n. 2, at 157.

153. There is, of course, a cost to the administration in the lost efforts setting up the appropriate regulatory agencies and other executive modifications necessary to implement the laws; but laws change frequently enough that this is little more than the cost of government in the normal course.

154. Wu, *supra* n. 32, at 1189-92.

155. Lessig, *supra* n. 34, at 85-99.

cause they can operate upon clean metadata. This may be more useful in some fields than in others. Considering search alone, some searches might be better made using existing tools. However, some will benefit from separate traditional and metadata-based searches, and some will utilize tools which combine traditional and metadata-based approaches. Certain areas might be best searched using only metadata-based tools.[156]

This segues neatly into the second type of extensibility discussed above, the applicability of the scheme itself for different purposes. The manner in which clean metadata would be provided is amenable to use for a variety of purposes, as a platform for any number of uses besides search. A government agency could publish a standard for metadata used in a particular niche market, relying on the established system of guarantees to prosecute those who abuse the standard by offering inaccurate metadata. Better yet, any private or public body could do the same, without needing formal powers of proclamation or regulation, as abusing a published standard would clearly constitute providing unclean metadata. If further regulation is needed, the scheme serves as a model for other regulatory efforts.

H. Probabilities of Success

With virtually no cost of failure and a considerable potential dividend, the only rational argument against implementing the proposed scheme is the belief that its failure is a foregone conclusion. Against that, one can do little more than to point to the detailed examination of potential obstacles undertaken above. All that is needed is political will. In light of successful campaigns for legislation against spam,[157] that is certainly achievable. Indeed, given the willingness to regulate evidenced by anti-spam legislation, there is reason to believe that the political capital realizable from a successful exercise of regulatory power on the Web will be enough to inspire legislators.[158] However, in case a little more illustration is needed, below is an example of the manner in which the proposed scheme could become established and useful, assuming its enactment by a national legislature.

First, academics in specialist fields begin to annotate their Web sites using metadata to enable colleagues and students to better locate relevant research. Adopting the Dublin Core schema, information scientists

156. *See* Morville, *supra* n. 2, at 48-54, 139-41 (discussing different measures of effectiveness in information retrieval and addressing the layering of technologies, combining advantages, and allowing user choice, as "the genius of the AND").

157. *See supra* n. 84 and accompanying text (Successful in the sense that legislation was enacted, not that spam was stopped, or even reduced.).

158. *See supra* n. 88 and accompanying text (Especially considering that the problem can be cast in terms of pornography.).

and librarians lead the way, realizing that actions speak louder than words (even perfectly classified words). A small team of undergraduate information science students develop a simple search engine using Boolean queries to index these resources.[159] Other small research groups follow suit, creating pockets of guaranteed clean metadata in academic corners of the Web accessed using custom search tools.[160]

A graduate researcher proposes to the National Hoteliers' Association, say, that all its members be encouraged to use a metadata schema which he has created to describe hotel rooms.[161] The association can then provide an authoritative search tool for finding hotel rooms, saving members the costs of advertising on other search engines. Smaller hotels, invisible on regular search engines, flock to the idea. Although the NHA declines to fund the search engine – established interests being what they are – the researcher believes in the idea enough to establish Hoogle. Similar operations pop up in other niches, gradually but inevitably bringing guaranteed clean metadata to the commercial Web.

As momentum builds, regular search companies begin to take notice of this new force. They begin to allow users to include metadata as an influence on search results, and establish alternative tools for metadata-based searching. Clean metadata has gone mainstream, and the result is a wave of fraudulent Web sites using deceptive metadata to attract traffic. The government agency in charge responds with a wave of domain name forfeitures, prompting a number of court cases (although the majority of offenders do not contest their penalties). The predictability of enforcement is improved as parties are made more aware of the limits of the law, bringing formerly hesitant Web site authors into the fold. Seeing the success of the scheme in this jurisdiction, other countries enact similar legislation, seeking to foster local Web-related innovation. Guaranteed clean metadata becomes part of the global Web infrastructure.

---

159. The technology required is proven and well-known; it simply needs to be applied to the problem at hand. More than likely, an existing, freely available piece of software could be adapted to serve such a purpose.

160. *See* Dublin Core Metadata Initiative, *People involved in the Dublin Core Metadata Initiative*, http://dublincore.org/about/participants/ (accessed April 17, 2006) (listing the kinds of parties already involved in metadata-related activities). Given that educational institutions are trying to foster the growth of knowledge, that effective search is so important in research, and that academics tend to be early adopters of new technology, it is highly likely that this kind of scheme would be embraced at universities. Universities (the institutions themselves) and academics are not parties with an incentive to publish unclean metadata. Some already publish clean metadata. Government bodies are in a similar position.

161. For example, <meta name="NHA:stars"value="4"/><meta name="NHA:price" value="300"/>. Any number of alternative implementations exist, such as nesting data within a single meta tag.

VII. CONCLUSION

It seems clear that the establishment of a system of regulation to ensure clean metadata on the Web is both feasible and worthwhile. The proposal explored above accounts for the challenges of Internet regulation by tying the scheme to national top-level domains, known as ccTLDs. Doing so secures its enforceability, giving credibility to Web site owners' guarantees as to the cleanness of their metadata. This allows the realization of the myriad benefits of clean metadata within national jurisdictions that implement and enforce such legislation. Better search – faster and more democratic – is an important benefit, but it may merely be the beginning. The true potential of clean metadata could come as a platform for innovation, for visions as grand as the Semantic Web.

By finding a way to regulate behaviors on the Web without altering its architecture, we can introduce the benefits of regulation – the stability and predictability valued by individuals and commerce alike – without sacrificing the flexibility and freedom that have made the Web what it is. We can extend the Web, allowing users to find the Web that they are looking for. And while legislation bringing clean metadata to the Web could be an important step in bringing the Web to its full potential, it could be more significant still as the start of a habit of regulation of the Internet which is explicit, direct, facilitative, application-specific, creative, and informed.