Filtering the Internet:

A Best Practices Model

by members of
the Information Society Project
at Yale Law School

http://www.law.yale.edu/infosociety

J.M. Balkin
Beth Simone Noveck
Kermit Roosevelt

September 15, 1999

1. Introduction

The rapid growth of the Internet during the 1990's has led inevitably to repeated calls for regulation

of Internet content. The most widespread justification for regulation has been that the Internet contains material that is harmful to children because of its bad language, sexual explicitness or violent content.

Calls for protecting children nevertheless beg the question whether regulation is best achieved through government sanctions or through a combination of industry self-regulation and individual choice. We believe that industry self-regulation and facilitation of end-user choice through technology provide a better solution. Technological solutions like filters allow concerned parents and other Internet users to avoid harmful or objectionable material without violating the freedom of expression of Internet publishers. This paper discusses the pros and cons of various filtering systems and proposes the best practices for Internet self-regulation using filtering technology.

Many nations responded to fears about the Internet with legislation that tried to punish those who provided disfavored content. America's Communications Decency Act ("CDA") is perhaps the most well-known example among many others.[1] The U.S. Supreme Court struck down the CDA as unconstitutional under the freedom of speech guarantees of the First Amendment to the U.S. Constitution, but Congress (and many states) promptly passed new legislation.

Obviously, some governments may seek to control Internet content for reasons other than the protection of children. Some governments and some politicians may wish to control the viewing and reading choices of adults explicitly; others may attempt to control adults through the guise of regulation designed to protect children. In general, we think that regulations designed for these purposes are illegitimate interferences with freedom of thought and expression. Obviously we cannot prevent governments from engaging in antidemocratic practices or politicians from engaging in demagogic appeals. Rather, our arguments are designed to show why governments sincerely interested in protecting children should prefer self-regulatory solutions. Moreover, nothing in this report should be understood as opposing legal regulation of child pornography. Child pornography is unlikely to be dealt with effectively by filtering solutions. Filtering solutions generally require rating of sites. However, child pornographers are unlikely to draw attention to themselves by permitting rating or engaging in self-rating. The issue before us is the best way to guarantee adults access to content they have a right to view while allowing parents to shield children. We think that direct legal prohibitions on Internet content are a poor solution.

There are important and obvious problems with trying to impose such legal restrictions. National sovereigns usually legislate territorially, but the Internet crosses national boundaries effortlessly. A general international treaty on Internet content is highly unlikely, given the wide cultural diversity of the planet and the need for near universal participation in order to prevent technological end-runs. As a result, governments attempting to control Internet content through law have engaged in half-way measures: Each government has attempted to impose its own penalties on whatever content providers it can lay hands on.

---

[1] In a well-publicized case, a German court convicted the managing director of Compuserve Germany of assisting in the dissemination of pornographic material on the basis of the presence of such material in a Compuserve newsgroup. *See* Cyber-Rights & Cyber-Liberties 1999. The *Somm* decision imposes liability on a gatekeeper entity (Compuserve) rather than the actual content providers. Imposing gatekeeper liability may be more effective than trying to reach the actual content providers, but it is even less desirable from a free speech perspective.

This practice combines unpredictable sanctions with near-total ineffectiveness because many content providers simply cannot be reached by many territorial governments. Worse still, this haphazard arrangement can produce arbitrary and inconsistent patterns of prosecution that chill freedom of thought and expression without actually protecting children.

Some governments have considered requiring website operators to install age verification systems as a way of safeguarding children. Generally speaking, age verification systems require either the use of a credit card number or an adult verification code which can be purchased using a credit card. The credit card number or the age verification code is then typed in each time the user visits a site with content deemed inappropriate for children. The definition of what is inappropriate for children is determined by statute or administrative regulation. If a site contains such material, website operators are subject to fines or criminal sanctions if they fail to install effective age verification schemes.

Obviously, many of the same problems of territorial limitation apply to these proposals: much Internet content will come from beyond the territorial boundaries of governments requiring age verification, and each government will have different criteria about what is inappropriate for children. But age verification systems have other serious drawbacks: They are both problematic and ineffective as a general strategy for keeping children from inappropriate content. First, age verification strongly hinders Internet surfing by adults. Most adult users do not wish to have to insert an age verification identification every time they visit a new site, and they are even more loath to type in their credit card number. As a result, age verification will deter adults as well as children from visiting web sites, particularly web sites that adults have every right to visit.

Second, maintaining an age verification system is an expensive and complicated proposition for nonprofit organizations and for individuals who are not selling goods and services but simply wish to communicate with others. Age verification works best for commercial pornographers, because they already seek to sell adult customers goods and services, which are usually paid for by credit card. But age verification is a genuine hardship for an organization like the National Abortion Rights League, or for the average individual who seeks to put content on the web that might contain material inappropriate for children-- for example, strong language or discussions of human sexuality. (We should add that adult verification requirements have virtually no effect on child pornographers, who are largely underground in any case.)

Third, most adult verification systems require only the use of a credit card to prove age. There is no guarantee that the person who holds the credit card actually is an adult. Moreover, there is a vibrant business on the Web for persons bartering false adult verification identifications and passwords.

At best we think adult verification schemes have a place in commercial pornography sites-- and we note that many commercial pornography sites already include them-- but we think that those sites can also adequately be dealt with through the use of filtering systems.

For these reasons, we believe that industry self-regulation and technological facilitation of end-user choice offer a better solution to the problem of harmful Internet content. Industry self-regulation does not mean mere forbearance by content providers; it also involves putting technological solutions in the hands of end-users to filter out content they do not wish to receive. Filtering through software specifications is the most widely supported technological method of blocking Internet content. Web sites and other forms of Internet content receive ratings which can be read by a filter located in the end-user's browser or other

software interface. In this report, we focus mainly on rating and filtering of content delivered on the World Wide Web; however the filtering solutions we recommend can be adapted to other forms of Internet content, including chat rooms and Usenet newsgroups.

A self-regulatory solution has two distinct but equally important components: the use of technology to facilitate choice by end users, and location of that choice at a decentralized, or local level, rather than at a national or global level. Obviously whatever technological solutions individuals employ can also be employed by governments. Hence governments may be tempted to impose filtering centrally or hierarchically, upstream from their citizens. In this way governments may try to use filtering technology to block all content that these governments deem inappropriate for their citizens to view.

The Information Society Project strongly opposes such hierarchical filtering by governments. We emphasize that a self-regulatory solution must include both technological filtering and decentralization. Conceivably, governments may be more successful in blocking content through technology than through law, especially if their citizens do not know that their access to the Internet is being filtered without their approval. However, from the standpoint of freedom of thought and expression, state-imposed hierarchical filtering is no better than legal punishments and may in fact be much worse. The purpose of this study is to show how a decentralized technological solution that places choice in end users about what and whether to filter can protect children and other citizens from being exposed to harmful or offensive content.

Because there is simply too much information on the Internet, filtering of some form will be inevitable. Market forces will demand and produce increasingly technologically proficient filters for different purposes. The most familiar of these are search engines, which grow in power and discrimination with each passing day. Thus, there is no need to ask whether filtering and filtering technologies should be implemented. They already are. The appropriate question is what kinds of filtering systems are desirable and for what purposes. Cyberspace is plastic and can be shaped into a wide variety of architectures. The choice of filters plays an important role in the architecture of cyberspace. Thus the choice we face in choosing filters is a choice about what kind of architecture cyberspace will have. This is an important and weighty decision, because architecture regulates human conduct and shapes human thought (Lessig 1998, Boyle 1997). Technology in practice is not neutral in its effects or in the values that it promotes or hinders. Moreover, filtering and rating systems will surely be used for purposes other than protecting children from harm. In this way they may have significant and unintended effects on the evolution of culture (Balkin, 1996).

One argument against the development of filtering software is that it will be employed by governments or large private entities without the consent of end-users to censor private expression upstream. However, as noted above, we think that market forces will inevitably produce filtering systems of increasing proficiency. Moreover, opposing the development of filtering systems for end-users may have even worse consequences. Unless governments are assured that decentralized filtering will produce an acceptable degree of self-regulation, they will inevitably turn to legal sanctions that will stifle freedom of expression more directly. Thus, we should not avoid designing effective end-user filters because of a fear that they may someday be adapted by governments or other powerful entities for unscrupulous purposes. The fight against hierarchical filtering by governments is best carried on through political and legal pressure rather than by simply opposing technological development. Moreover, such criticisms overlook the fact that technology can also be used to fight hierarchical filtering, by developing effective methods to prevent or discourage governments from filtering Internet content upstream.

In order to avoid the dangers of upstream filtering by governments who fail to respond to political and legal pressure, we think it best to design the architecture of the filtering system to frustrate such attempts before they occur. Thus, we propose a system of *encrypted ratings*. Under this system, all ratings sent from web pages will automatically be encrypted with a weak (three to eight digit) encryption scheme. This encryption will take virtually no time to decode in the end-user's browser. However, a government's proxy server will have to decode each and every one of these ratings in order to decide whether the relevant Web pages should be blocked. The cumulative effort required to do this will be prohibitive: If more than a small number of people try to use the Internet at once, decoding even weak encryption of all of these ratings will be sufficient to overload any proxy server. In short, the point of weak encryption of ratings is to use the bottleneck feature of upstream filtering against itself. If weak encryption of ratings is built into the filtering system, governments will be unable to filter upstream without their proxy servers grinding to a halt. As a result, all filtering will have to be performed at the end-user's browser. Thus, by designing the architecture correctly, we can actually promote free speech values. We can safeguard decentralization and end-user autonomy, and turn the tools of censorship against would-be censors.[2]

In the following discussions, we emphasize end-user autonomy as a central concern. Nevertheless, the expression "end-user" is necessarily a term of art. For example, if the goal of filtering is to protect children, their parents or guardians are properly the relevant "end-users," even though the children use the computer.[3]

---

[2]Even without encryption, end-users can take other steps to defeat upstream filtering. A second solution involves what we might call "trojan ratings." It involves sending targeted false ratings to the IP addresses of countries using hierarchical filtering, while sending true ratings to requests from all other IP addresses. The point of such “trojan ratings” is to create an credible threat against oppressive governments. If a government uses hierarchical filtering, private parties can threaten to sabotage it by targeting its proxy servers. It is not necessary for everyone to engage in this practice as long as a sufficient number of rated sites announce publicly that they will do so. While the first solution-- encrypted ratings-- can be analogized to a law or mandate that is built into the architecture of the system, the second solution is a form of "civil disobedience" that individuals can employ based on the architecture of the system.

[3]We cannot hope to offer a general discussion of the problem of who is the "end-user" in this report, nor do we think it necessary, given that our basic concern is interactions between parents and children. We do suggest, however, that the decision by a third party that person may not use a computer to access certain content from the Internet demands some sort of justification. The burden should be on the filterer to justify the denial of another persons's access. The most plausible justifications for restricting access are that the third party owns the computer or that the third party has a relation of legitimate authority over the user. For privately-owned computers, the brute fact of ownership may often be a good enough reason, although a relation of legitimate authority will often also be present. Thus parents may restrict their children's use of their computers. (We also assume that private individuals should be able to restrict use of their computers by friends and social guests, even without a relation of legitimate authority.) Private employers may restrict employees' use based on ownership of the computer and legitimate relations of authority and workplace control. Nevertheless, restricting employee access may involve technological surveillance and invasions of

The aim of this report is to set out a best practices model for Internet filtering systems. The values that this model seeks to promote are the following:

End-user autonomy. End-users, rather than intermediaries such as Internet Service Providers ("ISPs") or nation-states, should decide whether and how to filter Internet access. As noted above, we oppose upstream or hierarchical filtering by governments. Equally important, meaningful choice by end-users requires a variety of filtering options that reflect different cultural values and ideologies. We hope to promote this variety by creating a market for filtering options and by carefully limiting the intellectual property rights in filtering systems that might allow third parties to restrain the market's operation. Finally, the system must feature a user-friendly interface that encourages actual use of its features and makes choice a real possibility for the vast majority of end-users. No system can truly promote autonomy unless most people can operate it fairly easily.

Protection of freedom of thought and expression for content providers. The point of using filtering systems instead of legal prohibitions on content is to allow individuals to publish what they want on the Internet while allowing end-users to filter out things they do not want. A good filtering system will respect the ideological diversity of content providers as well as end users. It will not block pages whose content is unrelated to the criteria used for filtering, and it will not attempt to block pages because they are critical of the filtering system being employed. As a default rule, the system should not block unrated sites unless the end-user specifically requests this option. As we will discuss in more detail below, our preferred solution relies heavily on rating by content providers themselves, which is usually called "first-party" rating. First-party rating is not only more practical, it also places important decisions in the hands of content providers, for example; whether to rate the entire site, individual pages, or individual elements within a site, and, perhaps equally importantly, whether to rate at all.

Protecting freedom of expression is and must be a serious issue in filtering design. To understand why this is so, we must understand which speakers the burden of Internet filtering falls most heavily upon. When most people speak of protecting children from inappropriate content, they normally have in mind child pornography and commercial pornography. As we have pointed out previously, filtering systems are

---

privacy. These are separate questions that must not be overlooked.

Free speech values suggest that the government should be treated differently than private employers. We do not think, as a general matter, that governments should be able to impose mandatory filtering on government-owned computers if those computers are made generally accessible to the public. Nevertheless, certain important contexts, like computers used for educational purposes in public schools, may involve a relationship of legitimate educational authority that justifies filtering. By contrast, a general governmental mandate that citizens use filters or particular filtering settings on their own computers restricts the use to which private parties can put their own computers. This constitutes a genuine, and in our view, unacceptable threat to freedom of speech and conscience. Finally, we do not support the notion that governments may force parents to place filters on their computers in order to protect children; it should be up to parents, in the first instance, to decide what their children should be exposed to.

not particularly well suited to dealing with child pornography. Because mere possession of child pornography is illegal in many countries, child pornographers do not wish to draw attention to themselves. They are unlikely to rate their sites as containing child pornography or to stay in one place long enough to be part of list of rated sites by third parties. They are best dealt with through law enforcement and other devices of industry self-regulation like telephone hot lines. The same is true for people who use e-mail or chat rooms to entice children. They are unlikely to rate themselves as pedophiles.

Commercial pornographers, on the other hand, are in the business of making money. Hence they usually require a credit card or some other form of adult verification in order to enter a site. Obviously, these sites may have free "come-ons" designed to entice visitors. However, because their target audience is adults, it is not surprising that many commercial sites have signed up to be rated as adult sites by various forms of filtering software. Commercial pornography sites do not resist being identified as "adult". Being "adult" is part of their advertising. As a result, we think that commercial pornographers are the most likely to fit easily into our proposed filtering system.

It turns out then, that the burden of filtering is likely to fall most heavily on those persons who fit into neither of these categories-- the vast majority of Internet speakers and publishers who are not in the commercial pornography business and who do not trade in child pornography. This includes political activists who discuss issues like human sexuality, abortion, rape, and drug use, take controversial and unpopular stands, as well as authors of artistic work with erotic themes. These speakers simply want to be able to reach audiences that are willing to listen to them. We should design a filtering system that ensures that they are able to do so.

Ideological Diversity and Flexibility. A good filtering system must be flexible enough to permit development of many different instantiations that reflect the wide cultural and ideological diversity on the planet. As values change over time, the system must be flexible enough to accommodate these changes.

Capacity for organic growth. Ideally, filtering devices should be both forward and backward compatible. New filtering software should be able to use specifications and read ratings currently in place. Older software should be easily adaptable to read ratings based on new specifications as they evolve.

Transparency. End-users must know when access has been blocked and why. For example, when access is blocked, filtering software should not simply display a generic error message such as "Error. Document not found." The software should instead explain that information has been blocked and give the reason for the denial of access, stating the basic criteria for the rating employed, for example: "Access to this site has been denied because the requested data contains content rated above 2 on the RSACi violence scale." The software should also reveal where the filtering has occurred—at the end-user's browser, or upstream, by a corporate intranet, an ISP, or a proxy-server.

Transparency means different things for end-users, for people who assign ratings to web sites, and for programmers who create implementations of a common filtering specification from which end-users may choose. For end-users, transparency means having enough information to make reasonable choices about which filtering system to use. Thus, the end user must know the basic categories upon which filtering occurs. If the system involves scalar values that measure intensity, such as 0 to 4, the end user must be able

to ascertain what these different values mean in practice. For persons who rate web sites-- including people who self-rate-- transparency means that the substantive meaning of different ratings is easily understandable and publicly available. For the programmers who create different ratings systems, transparency means that information about all aspects of the specification is fully public so that programmers can create different implementations. Moreover, it means that all aspects of the different implementations are public so that others can examine and criticize their work. Because end-users may lack programming expertise, what is transparent to programmers is not necessarily transparent to end-users or even to persons who rate web sites; however, it is sufficient that anyone with programming skills can gain access to information about ratings specifications if they so desire.

Open Source. As the discussion of transparency suggests, we advocate a basic filtering system with a set of specifications that can be implemented by anyone. The specifications will be public and free; no organization will hold intellectual property rights that allow it to constrain the use of the system. Network effects will probably lead to a certain degree of standardization, and we recommend that a non-profit organization be created to oversee the development of an initial basic vocabulary of content descriptors. But the content of the system should be determined in large part by the preferences of end-users, website operators and public interest organizations; it should not be imposed from the top down.

Privacy. Some approaches to filtering-- notably, certain implementations of the PICS-based system discussed below-- raise troubling privacy issues, because the process of filtering generates and places in the hands of third parties a list of material requested by end-users. We must not allow filtering systems to erode end-user privacy.

Compatibility between different rating systems. Filtering software must be able to accommodate different rating systems individually and in conjunction. Filtering software must allow different ratings systems to "talk to each other" and be applied seriatim or in combination. End-users who can turn to a large number of different rating systems have greater choice than those who must rely on a small number of systems. But end-users who can use different systems *together* have the greatest degree of freedom in constructing a filter to suit their particular needs.

These values are related, and each contributes to the central value of end-user autonomy. For example, transparency promotes autonomy because end-users who do not know that material has been blocked can neither readily evaluate the performance of a filter nor intelligently choose a different one. An open source approach also promotes autonomy, by giving end-users a voice in the development of the system. We can thus put our basic normative goal very simply: end-users should have a choice about whether and what to filter.

Filtering systems that attempt to offer a meaningful choice confront the same basic problem. For end-users to have real autonomy, they must be able to choose among a wide variety of ideologies and values embodied in different filtering systems. They must be able to choose a system that blocks what they want it to block and permits what they want it to permit.

The problem is who will do the rating. Generally speaking rating is either done by content

providers, called "first-party rating," or by separate organizations, called "third-party rating." Web site owners who rate their own web pages are first-party raters; organizations like the Anti-Defamation League or companies like Cybersitter or NetNanny who rate sites are third-party raters.

The problem with third-party ratings is that there is too much material. Third-party rating is too expensive and time consuming. Nor can end-users rely on first-party rating by website operators who rate their sites according to ideological criteria. Even if content providers were willing to self-rate according to these criteria, there is no guarantee that they would do so accurately. In short, first-party rating according to ideological standards is too unreliable. The task, then, is to construct a system that accommodates a diversity of ideologies while remaining inexpensive enough to be feasible and reliable enough to be useful.

Such a system must both describe Internet content and evaluate the description. It must, for example, both express the fact that a Web page contains depictions of violence and attach ideological significance to that fact by restricting or permitting access. The problem is how to achieve both tasks. Third parties can bring a consistent ideological perspective to bear in evaluation. But they cannot possibly describe all the content on the Internet. We might ask first parties to describe their content, but they cannot consistently evaluate it. Perhaps even more important, description and evaluation are not mutually exclusive categories. Descriptions of content are not entirely value-free, especially the content that most ratings systems would be interested in. Descriptions of nudity, language, violence and sexuality inevitably involve some normative or ideological evaluation.

Our solution relies on a division of labor between first and third parties. We ask first parties (website operators) to describe their content, but in terms that are likely to lead to convergent practices. In other words, we are less concerned with whether the descriptions are value-free (an impossible goal in any case) than with whether most first parties will apply them in roughly the same way. The goal is not ideological neutrality but predictable convergence in behavior. One might call these descriptions "objective" but a more accurate term would be "intersubjectively convergent."

We then ask third parties to produce "templates" that combine and rank combinations of these content descriptors in ways that match their ideological preferences. Simply put, a template takes the raw materials of content description and decides which combinations are better and which are worse with respect to a given value system. Thus, we do not ask third parties to be ideologically neutral-- indeed, we specifically ask them to rank certain types of content based on their values about what is good and bad, and what is more or less harmful to children. The goal of third parties in the system is to set up basic standards of evaluation that will be applied to the convergent descriptions of first parties. Because the basic task of third parties is to set up ratings templates, they do not have to rate every site, although they are free to rate particular sites individually and add those ratings to the mix.

Because ratings templates will be relatively simple and easy to set up, we expect many different organizations will be willing to create them. Our goal is to make it possible for an organization to set up a new template with only a day or two of work. Moreover, because the templates will be publicly available, organizations can model their efforts on previous templates, making the costs of template creation even smaller. Finally, because all templates will be based on a common language, end-users (or other organizations) can mix and match them to produce custom templates suitable to their ideological tastes.

This approach allocates to each group the tasks that they are most able and most willing to perform.

Before explaining our recommendations in more detail, we will first discuss the history of different filtering systems and their pros and cons. Then we will discuss the technological devices out of which our system will be constructed.

2. Filtering Systems

2.1. *First Generation Filters*

Filtering software is not new. Early approaches to filtering have tended to take two primary forms. One form screens documents before allowing access. If the screening detects forbidden words such as "breast," "sex" or "homosexual," access is denied. This technique deals with the twin problems of description and evaluation by performing a mechanical and very crude evaluation of all content requested. It judges content based on the presence or absence of forbidden terms. Unsurprisingly, text-based screening is very bad at recognizing changes in context: for example, text-based filters may block discussions of breast cancer and legal debates over homosexual marriage because they contain the words "breast" and "homosexual." The University of Kansas, having installed the text-screening Surfwatch program, found it had cut off access to its own Archie R. Dykes Medical Library. One way to mitigate the harshness of the automatic blocking rule is to delete the offending terms, but this leads to bizarre results, converting sentences such as "Traditionalists oppose homosexual marriage" to "Traditionalists oppose marriage" (Weinberg 1997:460).

Text-based blocking software is still being refined, and more sophisticated approaches attempt to capture contextual subtleties by considering factors such as repetition and proximity of forbidden terms. But the refinement seems unlikely to produce a satisfactory system. Context is simply too complex for mechanical evaluation, and, more important, mechanical evaluation of pictures is still technologically quite distant.

The second major approach relies on third-party organizations to evaluate all content by individual inspection. Evaluators generate lists of acceptable and unacceptable sites; software then either restricts access to the unacceptable sites ("blacklisting") or allows access to only the acceptable ones ("whitelisting"). CyberPatrol, for example, offers both a "CyberNOT Block List" of sites deemed unsuitable for children and a "CyberYES List" of approved sites. Users can configure the software either to exclude the blacklist or to allow access only to the whitelist.

We have already noted that relying entirely on third-party rating is impractical given the volume of Internet content. Invariably both blacklists and whitelists are underinclusive. Blacklists miss some sites that should be blocked; whitelists omit some sites that do not contain harmful content. Indeed, the mere existence of both blacklists and whitelists acknowledges this failing, for a perfectly inclusive blacklist would be identical to a perfectly inclusive whitelist.

Even manufacturers of blocking software understand their inadequacies in covering the Internet; blacklisting software programs such as CyberPatrol are now usually compatible with PICS-based filtering, which will be described momentarily. But these programs are not only ineffective; they have other, equally serious failings. They also offend several of the values discussed above as a necessary consequence of their technological structure. First, blacklists are usually not transparent. Some blacklisting programs

reveal to users that access has been blocked. But they need not do so. Moreover, once a site is blocked, the program does not have to offer any explanation, except that the program's raters found the site unsuitable based on their criteria. There is no requirement that these criteria be made public. Among existing blacklist programs, CyberPatrol offers some degree of transparency. It evaluates sites along fifteen different dimensions, from "violence/profanity" to "alcohol & tobacco." In effect, it contains fifteen different blacklists; end-users may activate as few or as many as they choose. CyberPatrol allows some information about why a site has been blocked, because it can disclose on which blacklist the site fell. Solid Oak Software's CyberSitter, by contrast, maintains a single undifferentiated blacklist (Dobeus 1998:633).

Lack of transparency is a serious problem for blacklisting software. Users of blacklist filters tend to be parents concerned about restricting their children's access to inappropriate material. But consensus at that level of generality about what should be blocked (material "unsuitable for children") does not translate into consensus about which sites belong on the blacklist. Even when more specific criteria are listed (e.g., "foul language") the basis for decision is not readily available. Decisions about individual sites are inevitably ideological, and blacklists force parents to accept the undisclosed ideologies of the third-party raters. Thus, blacklists tend to be not only underinclusive but also overinclusive; they block sites to which end-users do not object.

The list of sites blocked by common blacklisting programs is troubling: Jonathan Weinberg notes that CyberPatrol blocked the Electronic Frontier Foundation's censorship archive, the Animal Rights Resource Site, the League for Programming Freedom (a group opposing software patents), and Usenet newsgroups including alt.feminism, soc.feminism, and alt.support.fat-acceptance. CyberSitter blocks the National Organization of Women website, the Penal Lexicon (a British site covering prisons and penal affairs), and Web pages that criticize its blocking decisions (Weinberg 1997:461-62).

Compounding the problems created by the lack of transparency is the second defect of blacklisting programs: they are generally not open source. Their blacklists are typically treated as proprietary information, withheld from end-users and protected by intellectual property rights. End-users have no voice in determining the content of the blacklist, and are in fact severely restricted in their ability even to learn which sites are blocked. (When the Netly News, a subsidiary of Time-Warner Pathfinder, created a search engine designed to allow end users to find out which sites CyberSitter blocked, Cybersitter retaliated by blocking the more than 150,000 web pages on pathfinder.com.) (Wagner 1999:762-63, Weinberg 1997:462).

Withholding the list of blocked sites is not simply an arbitrary affront to transparency; it is a central feature of most blacklisting software. Blacklists employ lists of offending sites that are deliberately kept secret through technological devices and by legal protections like trade secret law. Third-party rating is labor-intensive and consequently expensive. The makers of blacklisting software sell their products to recoup these expenses and turn a profit. Because the value of blacklisting software consists primarily in its database of unacceptable sites, revealing a company's database of offending sites would undermine the market value of the product. Competitors could free-ride on the hard work done by the company's employees in locating and rating thousands of websites.[4]

---

[4]Another argument is that transparency would give children a ready-made list of inappropriate sites. Indeed, most filtering systems can be inverted and used to seek out objectionable material. This concern

Neither text screening nor blacklisting, then, is a viable way to construct a filtering system. Text screening is too crude. Blacklisting relies entirely on third parties; consequently it is expensive, underinclusive, and nontransparent. It gives end-users very little choice in the ideological content of the filtering system. The failings of text-screening and blacklisting filters show that first-party rating must be an essential component of any filtering system. In the next section, we look at various software solutions that take the first steps towards integrating the work of first- and third-party raters.

2.2. *Filtering through general software specifications: PICS, PICSRules, and RDF.*

2.2.1. PICS

The Platform for Internet Content Selection, ("PICS") is a set of software specifications for filtering systems created by the World Wide Web Consortium ("W3C"). PICS allows the creation of labels that can be associated with individual Internet addresses (Universal Resource Locators, or URLs). Labels can also be associated with an IP address --the identifying address of a computer connected to the Web-- in which case they apply to every document retrieved from that computer. The basic idea behind PICS is to create a standard for metadata: information about information. For example, if data consists in a picture on a page, metadata might include the statement that there is a certain type of picture on a page. A PICS label is essentially a statement about what data resides on the page at a certain URL. The statement generally takes the form of an assertion that the data has certain properties, for example, that it is a picture, that it contains guns, that it is violent, that it has been rated by a certain organization, and so on.

Different ratings systems will use different systems of labels; the PICS format simply outlines the basic format that all such systems must use so that they can be read by PICS compatible filters. Thus, strictly speaking, PICS itself is not a rating system; it merely provides common specifications for creating labels that are part of a rating system. The most well-known content rating system using the PICS software specification is RSACi, which was originally created and maintained by the Recreational Software Advisory Council on the Internet. (The RSACi organization has recently been merged into a larger organization, the Internet Content Rating Alliance (ICRA), which is now entrusted with maintenance and development of the RSACi rating system.) RSACi is only one possible implementation of the PICS software specification. It is not identical with PICS. We will discuss the advantages and disadvantages of the RSACi ratings system later in this report.

In the kinds of filtering systems we are concerned with, the properties used to describe data will be the categories and scales associated with particular rating systems. Thus a typical PICS label might say that the data at worldwarII.com/dday/omaha.jpeg (a fictitious URL, intended to designate a photograph

---

seems relatively insubstantial however, given that the settings of filtering systems are supposed to be accessible to parents but not children; blacklists could similarly be protected by passwords known only to parents. More significantly, finding pornography on the Internet is anything but difficult. Simply searching for "PICS" (to say nothing of "sex") on Lycos will turn up a host of pornographic sites, and search engines devoted to pornography, like http://www.sexhound.com, also exist.

of the Omaha beach landing) contains content rated 3 on the Violence category of RSACi, an Internet ratings system. A label may contain more than one assertion. The same label might also say that the document is a picture, that it is historical, or that it shows real-life combat. For labels to be useful end-users must have filtering software that recognizes the properties "picture," "historical," and "real-life combat."

Programmers can associate PICS labels with entire websites, pages within those sites, or particular items on particular pages. A label associated with the hypothetical worldwarII.com website would apply to every document retrieved from worldwarII.com. If the documents differ significantly—say, a text description of the Lend-Lease program and a photo of Allied troops storming Omaha beach— there are advantages to create different labels for each individual document. However, the more labels added, the longer it takes to rate a site and to filter the site when it is visited.

Using a PICS-based rating system to filter content involves several different tasks, and it is important to recognize that each of them can be performed by different parties (Resnick 1999). First, one must establish a rating system. A rating system requires the development of a standard vocabulary and categories for labels. By "vocabulary" we mean any description of content. By "category" we mean organization or grouping of vocabulary elements along a particular axis. For example, in the RSACi rating system "mild expletives," "moderate expletives or profanity," and "strong language, obscene gestures, or hate speech" are elements of the RSACi vocabulary. Note that the last two elements are *disjunctive*, that is, the rating applies if *any* part is satisfied. As we will describe below in more detail, disjunctive vocabulary elements create special problems for a ratings system.

RSACi organizes these vocabulary elements into a single "language" category, whose range of permissible values ("scale") runs from zero to four. Each scalar value from zero to four is associated with a different vocabulary element. Thus, "mild expletives" are assigned a scalar value of 1, "moderate expletives or profanity" receive a value of 2, and "strong language, obscene gestures, or hate speech" have a value of 3. (Because this category is disjunctive, a site receives a 3 rating if it contains either a swear word, an obscene gesture, or a racial epithet.) A browser reading RSACi labels, with a "language" filtering setting of <3, will permit sites containing the first two groups of vocabulary elements but block those containing the third. A set of categories offered as a unit is a "rating system" or a "template" (W3C 1999d).

A ratings system can have many different categories each containing many different vocabulary elements. For example, RSACi's rating system contains the categories "Nudity," "Sex," "Violence," and "Language." In RSACi each category has a five-point scale, running from zero to four, associated with different vocabulary elements.

As noted, PICS allows the creation of many kinds of ratings systems. For example, a different rating system, concerned with the medium of presentation, might have the categories "Picture," "Text," "Video," and "Audio." Each of these categories would probably have only two values on its scale: 1 (for yes) and 0 (for no). The vocabulary elements would thus be "Picture" and "No Picture," "Text" and "No Text," etc. Still another rating system, modeled on the Motion Picture Association of America (MPAA) ratings, might estimate the maturity required to view the rated material. It would have only one category, "Minimum Recommended Age." The scale within this category might run from 1 to 18, in which case the vocabulary elements would be "Suitable for Age 1," "Suitable for Age 2 and Under," "Suitable for Age 3 and Under," and so on. Or it might be a coarser five-point scale like the MPAA, with vocabulary elements "Suitable for All Ages," "Suitable for Pre-Teens," "Suitable for Young Teens," "Suitable for Older Teens,"

and "Adult."

As these examples should suggest, vocabulary elements may range from relatively "objective" descriptions that can be reliably coded by a wide variety of raters (e.g., "Picture" and "No Picture") to relatively "subjective" descriptions that are likely to produce considerable disagreement in application (e.g., "Suitable for Young Teens"). Choosing between the reliability of vocabulary elements that will be coded similarly by most people and the ideological richness of more subjective vocabulary elements is one of the basic dilemmas for filtering systems. This dilemma is related to the choice between first- and third-party rating; if first-party rating is to be reliable, vocabulary elements must be "objective," not in the sense that they are value-free, but in the sense that they will produce convergence in rating because most people will apply them in the same way.

Because current filtering systems generally focus on protection of children, most rating systems use similar categories such as sexual or violent content, indecent language, and promotion of drug use. They vary considerably in the "objectivity" (expected convergence in rating) of their vocabulary elements. RSACi's elements are designed to promote behavioral convergence in rating, while SafeSurf's are more subjective and controversial. SafeSurf, unlike RSACi, tries to reflect common moral judgments (presumably those in the United States) about factors that make particular content more or less objectionable. Thus, for example, SafeSurf's "Nudity" category distinguishes between "Dictionary, encyclopedic, news, medical references" (value 3); "Classic works of art presented in public museums for family view" (value 4); "Artistically presented with full frontal nudity" (value 6); and "Erotic frontal nudity" (value 7). Hence the SafeSurf "Nudity" category may produce divergence in coding in many different ways. For example, people in different cultures or with different values may disagree about what is "artistic," and "erotic," as well as what is "classic."

The developers of PICS hope for a wide variety of rating systems in order to maximize end-user choice (W3C 1999f). Moreover, through a further development, PICSRules, they have made it possible to employ different rating systems in conjunction. We agree that the homogeneity of existing rating systems is troubling. If third-party rating systems were cheaper to produce, greater diversity would be more likely. Hence one of the aims of our proposed system is to reduce the cost of creation.

Second, someone must rate sites by assigning labels to sites. To rate a site is to assign labels to content. A typical label for a website might look like this: "Nudity = 4; Language = 3; Violence =5." Again, rating can be done by first-party content providers or third parties. W3C expects both first and third parties to be active in the rating of content, and we will also recommend a mixture of first- and third-party ratings. While we believe that a mixture of first- and third-party rating is preferable from a policy perspective, an equally important reason is that third parties will simply be unable to rate the entire Internet.

Third, labels must be distributed to those who request them for filtering. PICS-based filtering may be performed either by the end-user's browser, or by an entity further upstream-- for example, a search engine, a proxy-server, or an Internet Service Provider ("ISP"). Upstream filtering—especially if not disclosed to or consented to by end-users— threatens free speech values. We thus recommend that filtering be performed at the end-user level. If filtering is performed upstream, the software should report this fact. It should also indicate that access has been denied and explain why.

Rating labels can be distributed either by website operators ("first-party distribution") or by other organizations ("third-party distribution"). First-party distribution probably the easiest and least costly

method. Programmers can insert rating labels as a header in the language of a web page. Third-party distribution is more complicated, because the end-user has to fetch the rating labels associated with a particular web page from some other place on the Internet. W3C contemplates the existence of "label bureaus" which will store third-party labels and provide them to browsers (W3C 1999b). This creates a real danger of bottlenecks, as millions of web surfers request labels from a modest number of label bureaus. One can also store third-party labels in end-users' browsers. However, given the vast and growing number of websites, this approach is not feasible unless third parties rate only a very small fraction of websites. It is unlikely, in any event, that third parties will have the resources to keep up with the rapidly expanding Internet, and the system we discuss below contemplates a limited role for third-party raters.

Fourth, ratings will not work unless someone writes filtering software that can read the labels. The software required is actually quite minimal; it need be able only to read labels and follow instructions about restricting access based on what the rating labels say. We expect that in the future most browsers will incorporate filtering software, and indeed the most recent versions of Internet Explorer and Netscape Communicator are PICS-compatible.

Fifth, someone-- either the end-user or another person-- must choose among the available filtering settings permitted by the software. In short, someone must operate the filtering software to decide which ratings are acceptable and which are not. Once again, in general, we believe that this choice should be made at the end-user level. In other words, filtering settings should be specified in browsers, rather than upstream. This reduces the danger that filtering will be done without the end-user's knowledge or consent. This does not mean that third-parties should not develop and distribute their own rating systems, only that the end-users should be free to choose which rating systems to employ.

Sixth, and finally, filtering software must be installed and run. Generally, we also recommend that this take place at the end-user level, rather than upstream, in the interest of promoting end-user autonomy.

The PICS specification was designed to be flexible and to accommodate many different kinds of ratings systems. It has been enhanced by new specifications, including PICSRules, which we discuss next.

2.2.2. PICSRules

PICSRules, like PICS, is a development of W3C. It is a language for writing filtering rules that allow or block access to URLs based on PICS labels that describe those URLs (W3C 1999c). In general, PICSRules operate by specifying the organizations whose labels are to be consulted, and then articulating policies for applying those labels, specified according to their categories. For example, a PICSRule may instruct filtering software to look up the RSACi label for a particular URL. Then it may give instructions to use RSACi's violence and language categories in certain ways while ignoring its sex and nudity categories. In slightly greater detail: the "serviceinfo" clause of a PICSRule specifies the organizations whose labels are to be consulted and gives the URL of a label bureau from which to retrieve the labels. It also controls whether or not the rule will use labels embedded in the requested document. "Policy" clauses determine whether a URL will be accepted or rejected. They may direct the filtering software to accept or reject based on information coded in the URL itself or the scalar values of any categories of the labels specified in the serviceinfo clause. Once a policy clause instructs the software to accept or reject the URL, it does so regardless of later clauses. Thus earlier policy clauses have priority. A set of clauses defining

preferences will be referred to as a "profile."

The importance of PICSRules lies in its ability to coordinate various rating systems through multiple policy clauses. A PICSRule could, for example, give the following instructions:

> (a)Three particular URLs are to be accepted and four blocked. (These instructions use "AcceptByURL" and "RejectByURL" policy clauses.)
>
> (b) Any URL rated by the ArtFriends service as having artistic content>2 is to be accepted. Any URL rated by People for the American Way as having political content>3 is to be accepted. (These instructions use "AcceptIf" policy clauses.)
>
> (c) URLs not accepted under the first two clauses are to be rejected unless RSACi rates them as having violence<3. (This instruction uses a "RejectUnless" policy clause; it will reject all URLs rated higher than two on the RSACi violence category, and also all URLs not rated on that category.)
>
> (d) URLs not disposed of by the first three rules are to be accepted. (This instruction uses the policy clause "AcceptIf 'otherwise'".)

This particular rule is broadly similar to filtering using the RSACi violence category alone and blocking unrated URLs, but it allows greater precision in articulating what is to be blocked and thus greater user choice in determining the ideology embodied in the filter settings. That is, it reflects the judgment that artistic or political content, as determined by ArtFriends and People for the American Way, is sufficient to redeem material rated as violent by RSACi standards. By employing rating systems developed by different organizations, PICSRules would allow a basic rating system to be refined and overridden by the avowedly value-laden judgments of organizations trusted by users.

2.2.3. RDF

The Resource Description Framework ("RDF") is yet another framework for describing and exchanging metadata. As noted in the discussion of PICS, metadata, involves statements or descriptions about properties of Web pages or individual documents on those pages. RDF provides a standardized and machine-readable syntax for making these statements and descriptions (Bray 1999, Cowan, 1999, Flynn 1999, W3C 1999a, W3C 1999e).

A thorough technical analysis of RDF beyond the scope of this report. In essence, RDF is quite similar to PICS; it can do anything PICS can. Both specifications allow the creation of statements about Internet documents. As a sample statement, consider the proposition that worldwarII.com/dday/omaha.jpeg is a picture containing real-life historical violence. We have seen already that this proposition can be expressed in a PICS label, which can be attached to the document and read by a PICS-compatible browser. RDF permits parties to do precisely the same thing.

The syntax of RDF allows for the attribution of "properties" (specific characteristics with scalar values) to "resources" (anything corresponding to a URL). (We should note that the value of a property can also be a "literal"-- i.e., any string of characters-- but these do not play a significant role in filtering systems.) A specific resource together with a named property plus the value of that property for that

resource is an RDF "statement." These three elements of a statement are called, respectively, the "subject," the "predicate," and the "object." A single RDF statement can attribute several properties, just as a single PICS label can express several different ratings. A description of worldwarII.com/dday/omaha.jpeg might consist of a single statement attributing four properties—here, "picture," "real-life," "historical," and "violence." For the properties "picture," "real-life," and "historical" the scale would probably have only the values 1 and 0, corresponding to "Yes" and "No." The property "violence" might have a broader range of values—it might, for example, be the RSACi Violence category, in which case its scalar values would range from 0 to 4, and worldwarII.com/dday/omaha.jpeg would receive a 3. Website operators labeling their own documents need not use the RSACi properties; they would be free to use whatever properties they chose, and even those they invented. However, because the utility of metadata depends on the existence of a common vocabulary of properties, one can expect some degree of standardization in properties.

RDF differs from PICS in that it provides a more general treatment of metadata, and W3C plans to reformulate a future version of PICS as an application of RDF. W3C also aims to produce a conversion algorithm from PICS 1.1 to RDF. Presumably some form of RDF-based system will eventually supersede PICS-based filtering. The recommendations we offer here are consistent with either PICSRules or RDF treatments of metadata. For purposes of this discussion we will speak of PICS (including PICSRules) as the software basis of our proposed system, with the understanding that the system we propose can also be adapted to and supported by RDF.

3. Content Analysis of Internet Ratings Systems

3.1. *Ideological Effects in the Construction of Rating Systems.*

In this section, we examine the potential ideological biases of ratings systems that attempt to facilitate end-user choice. No ratings system can or should be value-free, but one of the goals of Internet self-regulation should be a filtering technology that is compatible with a wide variety of cultures, ideologies, and value systems. If there are many different rating systems in operation, the ideological slant of each is not a major concern, because end-users can simply pick the one that best matches their preferences. However, if network effects prevent a diversity of ratings systems, or cause ratings systems to converge toward a single model or a small group of models, then some end-users will either be forced to employ a system that does not adequately respect their values or face the choice of using no filtering system at all.

In fact, the market for rating systems does have very significant network effects. It is unlikely that website operators will be willing to rate their sites according to a large number of different or incompatible systems. As a larger number of websites rate according to a particular system, its value to end-users will increase. Similarly, as more end-users filter based on that system, the incentive for other website operators to self-rate according to that system will increase. This is especially so if end-users who filter block unrated sites. If network effects operate in a predictable way, one or two ratings systems will eventually emerge as a de facto standard.

We should distinguish between two different kinds of network effects. One of these is not troublesome; the other is. The first network effect is that ratings systems compatible with PICS and

PICS/Rules may become a de facto standard. This means that there will be convergence toward an underlying software protocol but not toward a particular set of ratings categories, vocabulary elements or scalar values. Even if PICS becomes the basic software substrate for ratings systems, there can be many different PICS-compatible systems reflecting many different ideologies and value systems. Indeed, this was the original purpose behind the PICS specification. Moreover, because of the way that PICSRules are constructed, end-users can mix and match different ratings systems as long as they are all based on the PICS standard. This kind of network effect is not at all troublesome. In fact, it tends to promote diversity of rating and filtering systems.

The second kind of network effect occurs when a particular *substantive system* of rating becomes the de facto standard. For example, currently the RSACi system appears to have the largest base of rated sites. It has a fixed set of categories, vocabulary elements, and scalar orderings. Conceivably, network effects might also make RSACi the de facto standard for substantive ratings systems. We think this would be unfortunate, not because RSACi is a particularly bad ratings system, but because of inherent limitations in all ratings systems that code content as more or less appropriate for end-users. Simply put, all such ratings systems face a series of difficult tradeoffs, and all such systems reflect particular ideological biases. This is true even if the ratings system strives, as RSACi does, for value neutrality and objectivity. No single ratings system can adequately serve all end-users on a culturally diverse planet. If network effects produce convergence on one system this will simply impose a particular ideological bias on all.

Value choices and ideological biases can enter into ratings systems directly or indirectly. They enter directly when the designers consciously attempt to promote a particular ideology or set of values. For example, a ratings system that coded Internet content for being "consistent with Christian scriptural values" or "inconsistent with Christian scriptural values" would be rather deliberately promoting a particular religious agenda. Nevertheless, value choices also enter into the process indirectly through tradeoffs that are inherent in any ratings system. To see why, we must consider the various constraints that will affect any such system.

1. *Reliability and Convergence in Rating*. If a ratings system relies on more than one person to rate content, it is important that most people rate according to the system in roughly the same way. For example, a system that relies on first-party rating must ensure that content providers around the world use the ratings criteria consistently. Sometimes this is expressed by saying that the categories and vocabulary in the ratings system must be "objective." However, there are at least two different senses of the word "objective." One means simply that the terms are likely to produce behavioral convergence: most people in most situations will apply them in roughly the same way. The other meaning of "objective" is "value-neutral" or "ideology free." There is no guarantee that any ratings system can achieve this. The choices made by ratings systems always involve values, and those values are always contestable at some level. Hence, ratings systems at best can strive for objectivity-as-reliability.

Nevertheless, the more that a system strives for objectivity-as-reliability, the less likely it will be to adequately respond to and reflect people's differing ideologies and value systems. After all, such a system is designed to produce convergence despite ideological diversity. A reliable system will also be less likely to take contextual differences into account, for example, deciding whether expression is "artistic" or violence is "justified." As a result, a reliable system may be ideologically slanted toward an arbitrary or idiosyncratic set of values rather than towards "neutral" or "objective" ones.

2. *Coarseness*. All ratings systems feature differing degrees of coarseness. By coarseness is meant

the number of categories into which the system divides the world (Balkin 1996). The more coarse a system, the more undifferentiated its judgments. A system that lumps homosexual and heterosexual intercourse, kissing and handholding into a single category of "sexual conduct" is more coarse than a system that distinguishes between kissing, handholding and intercourse, or between heterosexual and homosexual sexual conduct. A system that distinguishes racial epithets from sexual vulgarities is less coarse than one that treats both as instances of "bad language." The MPAA movie ratings system in the United States is quite coarse because language, nudity, sexual situations, and violence all contribute in the determination of whether a movie is rated G, PG, PG-13, R, or NC-17. Thus, a very violent movie, a movie with few well-chosen expletives, and a movie that features a nude love making scene all may be rated in "R" in the United States (meaning that no one under 17 is to be admitted without parent or guardian). The RSACi ratings system is much less coarse than the MPAA system because it has four separate categories each of which features five scalar values. Coarseness has ideological effects because different ratings system imposes a more differentiated or less differentiated vision of what is harmful or inappropriate for children. Deciding what categories count and how much will fall within them is often politically controversial.

3. *Equivalency.* Coarseness is the question of how many categories and degrees of ratings there are in the system. Equivalency is the question of what things the system treats as falling within the same category (e.g., "bad language") or category level (e.g., level 3 in the language category)(Balkin 1996). Thus, a system that places profanity, obscene gestures, and racial epithets in the same category level states that the three are just as bad or just as harmful for purposes of the ratings system. A system that treats homosexual kissing and heterosexual intercourse as "explicit sex" treats the two as equally inappropriate for children to view. Two systems can be equally coarse in that they have the same number of categories but they can make different things equivalent. For example, suppose system one has three categories, (1) profanity and hate speech; (2) nudity and (3) violence, while system two's categories are (1) profanity and nudity; (2) hate speech; and (3) violence. The two systems are equally coarse, but hate speech is treated as equivalent to different things. All ratings systems make certain things equivalent as soon as they create categories and establish category levels. But these decisions are some of the most ideologically controversial. For example, the decision to place hate speech in the same category with profanity is not ideologically neutral, nor is the decision to treat heterosexual and homosexual sexual activity as equivalent (or, for that matter, to distinguish between the two).

4. *Scalar Ordering*. Many ratings systems have degrees of intensity or harmfulness within categories. As noted previously, the RSACi system has a language category in which level 1 is "mild expletives," level 2 is "moderate expletives or profanity," and level 3 is "strong language, obscene gestures, or hate speech." Level 3 has important equivalency effects: it makes hate speech equivalent to strong language and obscene gestures. Equally important, it treats hate speech as more troublesome than "moderate expletives" or "moderate profanity." Scalar ordering refers to situations when two or more elements placed in a category are not treated as equivalent but are ranked in order. Obviously there is more than one way to do this, and the choices are not value free. For example, the decision to rank hate speech as worse than one kind of profanity or gesture but not as bad as another is not politically neutral. A particularly remarkable example of the political choices inherent in scalar ordering is RSACi's decision to rank sports violence at level 0, lower than any other form of violence (and equivalent to "no violence").

To see how these considerations interact in practice, we will discuss them in the context of what

is perhaps the most widely used content rating system, RSACi. Because of the inevitable ideological effects of ratings systems, we believe that it is a fool's errand to settle upon or to accept a single content ratings system for first-party rating. Rather, the best solution is to create a flexible platform on which many different forms of first- and third- party rating can interact. This system will try to capture many of the real advantages of RSACi without the disadvantages that necessarily arise from a unitary content rating system.

3.2. *The RSACi Content Rating System*

The RSACi system is a set of categories, vocabulary, and scalar values built on top of the PICS specification. The original mission of the Recreational Software Advisory Counsel (RSAC) was to rate video games for violent content, bad language, sex, and nudity. In 1996 RSAC turned its sights on the Internet, and RSACi was born. The RSACi rating system is an adaptation of the video game rating system. The system has four categories, each comprising five sets of vocabulary elements. The categories are Sex, Nudity, Language, and Violence, and the vocabulary elements for each range from "None" through progressively stronger examples.[5] RSACi ratings are generated by completing a questionnaire which asks about the presence of various vocabulary elements. The value for each category is determined by the highest valued vocabulary element present.[6] Like other PICS labeling systems, RSACi may be used to rate entire sites, particular pages, or particular documents on a page.

RSACi aspires to a substantial degree of "objectivity" in its vocabulary elements. What RSACi means by this is what we have called objectivity-as-reliability. The RSACi ratings are generated by a series of yes or no questions about the presence of types of defined content. Because the definitions are explicit and turn on relatively uncontested concepts ("humans injured or killed with small amounts of blood" rather than "illegitimate violence against righteous people") RSACi expects that different people will tend to give the same answers about the same content. The RSACi system is objective in the sense that it does not require "subjective" judgments about, for example, legitimate violence or artistic nudity. This objectivity is obviously essential if self-rating is to be reliable.

In order to achieve reliability, therefore, the RSACi system thus tends deliberately to exclude contextual factors from the vocabulary elements. The SafeSurf rating system, in contrast, includes "Technical Reference," "Non-Graphic-Artistic," and "Graphic" (in ascending order of intensity) as vocabulary elements for categories such as "Profanity," "Violence," and "Nudity" (W3C 1999d). A

---

[5]Of course, what constitutes a "stronger" example is a value-laden question, as is the determination of which vocabulary elements fall within a category. An ideal filtering system would allow end-users some freedom in defining their own categories, and we attempt to implement this ideal below.

[6]The number of questions exceeds the number of vocabulary elements because RSACi's vocabulary elements are disjunctive. "Extreme hate speech" or "crude, vulgar language" both suffice to assign a "language" value of 4. Similarly, "wanton and gratuitous violence," "torture," and "rape" all suffice for "violence" value 4. We doubt the wisdom of disjunctive vocabulary elements. Deciding whether "extreme hate speech" and "crude, vulgar language" are equivalent is more appropriately left to end-users.

depiction's technical or artistic nature is treated as a mitigating factor, reducing the depiction's assigned scalar value. RSACi's vocabulary elements make no distinction between technical and non-technical or artistic and non-artistic portrayals of nudity and violence. Nor do they distinguish between pictorial and textual portrayals.

The fact that RSACi strives for objectivity-as-reliability does not mean that it achieves another sort of objectivity: a rating system that has no ideological slant or bias. The RSACi system falls short of value-neutrality in several respects. RSACi's decision not to consider mitigating factors such as artistic or technical content might seem to avoid ideology: sexual or violent content will be coded for what it is, rather than how it is presented. But the rejection of the idea of mitigating factors is itself ideologically fraught; the perspective that equates Michelangelo's David, a Playboy centerfold, and a medical textbook because all feature frontal nudity is not neutral but embodies a particular (and idiosyncratic) ideological stance.

Second, and perhaps more serious, RSACi's vocabulary elements are constructed in disjunctive fashion. For example, "hate speech," "strong language," and "obscene gestures" all qualify for level 3 on the language category; thus, the level 3 vocabulary element of this category is effectively "hate speech, strong language, or obscene gestures." What this means, of course, is that the RSACi system treats "hate speech." "strong language," and "obscene gestures" as equivalent—indeed, identical. This is hardly an ideologically neutral decision. Whether we think that hate speech, strong language, and obscene gestures should be treated identically by the ratings system will depend on how we feel about the "harms" each may produce. The fear that children will acquire coarse habits of speech or conduct is surely distinct from the fear that they will acquire racial bias. And regardless of the qualitative difference, many Americans would probably find RSACi's categories of "obscene gestures" (e.g., extending the middle finger) and "strong language" (e.g., "dick") significantly less offensive than "hate speech," which includes the paradigmatically inflammatory "nigger."

The reason that the RSACi system seems to hold out the promise of objectivity as value-neutrality is its focus on the seemingly objective category of "harm." The RSACi rating system is explicitly oriented towards filtering out material that is harmful to children. The original rating system was developed by "a team of academics, psychologists, and educators" in response to the threat of congressional legislation on violence in computer games. RSACi describes its system as "based on the work of Donald F. Roberts of Stanford University, who has studied the effects of media on children for nearly 20 years" (Microsys.com 1999, W3C 1999d). The inclusion of different types of content as disjunctive components of vocabulary elements presumably reflects a judgment that these are equally harmful to children. This sort of judgment obviously depends upon the premise that "harm to children" is a concept that can be measured scientifically, so that the decision that "hate speech," "obscene gestures," and "strong language" are equivalent is not a value choice but an "objective" scientific truth verified by empirical evidence. Unfortunately, this is not the case.

In any event, RSACi's appeal to scientific data about "harm" does not explain the construction of its actual rating system. The appeal to science falls short in two ways. First, the RSACi vocabulary elements are too coarse to accommodate the social science data. The studies on which RSACi relies—and indeed the statements of Roberts himself—argue that the harm various content inflicts on children depends on context (Roberts 1999). For example, violence that is punished has different effects from violence that is rewarded; similarly, violence portrayed as a regrettable but necessary response to aggression has different

effects from unprovoked violence. Seeking reliability, RSACi largely eliminates contextual factors from its vocabulary elements. But in so doing it eliminates the very factors that determine how harmful content is.

The second failing is perhaps even more important. Social science data cannot provide an objective metric for the measurement of harm for the simple reason that "harmful to children" is a contested category. What counts as harm is an inherently ideological question. Presumably most people would agree that material that incites children to unjustified and unprovoked violence is harmful. If incitement to violence were the only danger, scientific objectivity might be possible. Nevertheless, we still face the difficulty that not all violence is necessarily harmful or socially undesirable. Police officers are required to exercise violence to preserve social order; soldiers are required to fight to defend their country, and physical force may be necessary (even if regrettable) in self-defense or to protect innocent third parties. But even if the terms "unjustified" and "unprovoked" could be given objective definitions, the Internet raises concerns far broader than incitement to violence. The danger of unfiltered access is that children will be exposed to the "wrong" values. Which values are wrong is obviously not a question to be settled by social science. Some parents may think their children benefited by exposure to material suggesting that homo- and heterosexuality are orientations of equal dignity; others may think such material quite harmful.[7]

3.3. *A Critique of RSACi as a Unitary System*

---

[7] A final feature of the RSACi system is unrelated to content, but no less important. It is RSACi's response to the problem of first-party mislabeling and forgery. The problem is serious, for an RSACi tag—or any PICS label—is easy to forge. A website operator could simply write an RSACi tag without registering with RSAC or completing its questionnaire, or she could give false answers to the RSACi questions. The issue is significant because the legal rights and technological powers that third-party rating organizations may wield against first-party website operators will shape the legal and economic environment in which filtering takes place. Careful allocation of these rights and powers is essential to ensure that a filtering system does not collapse under the weight of high transaction costs or collective action problems.

RSACi is pursuing both legal and technological solutions to the problem of inaccurate self-rating. On the legal side, RSACi requires website operators to register their sites and enter into a contract before they are permitted to license RSACi labels (which RSACi refers to as "service marks"). The "Terms and Conditions" of use require website operators to acknowledge that RSACi "owns" its labels and "has established significant rights and valuable good will therein." *See* RSACi 1999. They must also agree to RSACi audits and, if misrating is found, RSACi may terminate the license agreement or employ "corrective labeling, consumer and press advisories and postings on appropriate Web Sites." *Id.*

RSACi's technical solutions include the incorporation of digital signatures into its labels to make forgery more difficult and the development of a web crawler that will visit RSACi tags and compare them to the database of registered users to detect unauthorized use. A full discussion of the appropriate allocation of rights and powers is beyond the scope of this report. However, our general conclusion is that third-party rating organizations should have fairly broad technological powers but fairly narrow legal rights.

Our discussion of blacklisting programs demonstrated why first-party rating must form an essential component of a successful filtering system. The RSACi rating system takes the opposite approach from blacklists: it relies almost exclusively on first-party rating according to a single substantive system. We now explain why its defects will make reliance on a unitary system for first parties inadequate as well.

RSACi was conceived as a complete and self-sufficient filtering system. It is superior both to first generation blacklisting systems and to the MPAA (Motion Picture Association of America) system that simply rates in terms of suitability for particular ages. It is more transparent, more flexible, and more "objective" in the sense of producing convergent and reliable coding. URLs are blocked not because their content has been deemed unsuitable but because they contain specific content that the end-user has decided not to receive.

Many of RSACi's flaws stem from its historical origins in video-game ratings. RSACi was not originally intended as a multipurpose filtering system, facilitating the choice of end-users with widely differing ideologies, desires, and purposes. Instead, its stated purpose is simply to protect children from "harmful content" as that term would be understood by a largely American audience. This narrow goal produces a system that is too coarse to be useful to end-users with divergent ideologies and values. In this sense, the RSACi system is ideologically too weak: it focuses on only a few ideological concerns and does nothing with respect to many others.

On the other hand, in many other respects the RSACi system is also ideologically too strong. It imposes an ideology that many end-users may not share. The fact that RSACi takes the goal of protecting children as a given creates the illusion that its "objectivity" extends beyond reliability to value-neutrality. The claim, made with varying degrees of explicitness by RSACi's proponents, is that the rating system encodes no contestable ideology but simply reflects and enforces the results of scientific studies about harm to children. However, this claim (or more accurately, this assumption) is simply not true. As we noted previously, RSACi is much too coarse to fully accommodate the scientific data on which it relies. The experimental literature has reached the unsurprising conclusion that the effect on viewers of depictions of violence depends heavily on context (Brown 1999). Donald Roberts, RSACi's in-house expert, identified nine contextual factors: 1) the nature of the perpetrator; 2) the nature of the victim; 3) the reason for the violence (whether it is justified or unjustified); 4) the presence of weapons; 5) the extent of the violence; 6) the realism of the depiction; 7) whether the violence is rewarded or punished; 8) the consequences of the violence as indicated by harm or pain cues; and 9) whether humor is involved (Roberts 1999).

RSACi's violence vocabulary elements do not accommodate all of these contextual factors. They do not take into account the presence of weapons or humor, the reason for the violence, whether it is rewarded or punished, or whether violence is presented as desirable or regrettable.[8] The result is a system that treats as identical depictions that are very different. Roberts comments, "It does not take a scientific background to sense that the consequences to viewers of a film like *Schindler's List* … are quite different than the consequences of a film like *Natural Born Killers*." Yet oddly enough, the RSACi system would

[8]Surprisingly, even RSAC's video game rating system is more sensitive to context; it differentiates between violent depictions based on whether the target is threatening or non-threatening, and whether the violence is rewarded or not. *See* Roberts 1999.

rate *Schindler's List* and *Natural Born Killers* identically: each contains wanton, gratuitous violence, and therefore each merits the highest rating (4). Similarly, Jonathan Wallace complains that the RSACi system would require him to rate "An Auschwitz Alphabet," a report on the Holocaust containing descriptions of violence done to inmates' sexual organs, as equivalent to "the Hot Nude Women site" (Weinberg 1997:462-63).

The problem of context is important in another way. Avoiding the "harm to children" that comes from exposure to violent content is not a value-free goal. Even if exposure to certain kinds of violence produces aggressive behavior by children, it is by no means clear that all aggression is equally troublesome. Some forms of aggression-- like standing up for one's rights, or being willing to protect other people who are being harmed, bullied, or attacked -- may actually be socially valuable. It is by no means clear that children are harmed if they learn these lessons. Although some parents might be happy if their children were raised to be pacifist in virtually all circumstances, others might disagree. The real question is whether exposure to violent content creates socially undesirable aggression, and that is a question that the RSACi system does not purport to answer.

The problem of equivalencies pervades the RSACi system. By using disjunctive vocabulary elements, RSACi equates many types of content that many end-users may find quite different. Which types of content should be treated as equivalent is a political choice. RSACi's apparent belief that material harmful to children is both the universal target of filtering systems and a widely-agreed upon concept leads it to construct a system that is both too narrow and ideologically fraught. RSACi's set of categories is tailored to a narrow, secularized and Eurocentric notion of harm to children. Blasphemy, sexism, and homophobia, for example, are captured, if at all, by the vocabulary elements of "profanity" (treating something regarded as sacred with irreverence) and "hate speech" (devaluing a person or group on the basis of race, ethnicity, religion, nationality, gender, sexual orientation, or disability). But people disagree- often quite strongly-- on what is sacred, and they also disagree with equal vehemence about whether content devalues a group or simply reflects basic truths and foundational values. For example, last year U.S. Senate Majority Leader Trent Lott was quoted as describing homosexuals as sinners who are in need of treatment like kleptomaniacs. If Senator Lott's views were reprinted on a website, some parents would insist that it be filtered at a fairly high level because it demeans homosexuals, while others would find it entirely innocuous or at most a poorly expressed declaration of worthy religious sentiments.

Similarly, each RSACi category contains contestable value choices either in its construction of vocabulary elements or in its scalar hierarchy of better and worse vocabulary elements within each category. The Violence category, for example, treats "sports violence" the same as no violence; each gets a rating of zero. Yet the influence of sports violence on children is an area of increasing concern. Many parents might be concerned that a professional wrestling website will tempt their children to practice hammerlocks and piledrivers.

The Violence category also equates "rape" with "wanton, gratuitous violence." It does not distinguish between news stories, educational sites, and endorsements of rape or violence. As a result, news stories about the use of rape as a tool of war, date rape awareness sites, and pornography sites catering to "rape fantasies" are treated as identical.

The Sex and Nudity categories do not distinguish between homosexual and heterosexual content. Some parents may think that two men kissing is equivalent to a man and a woman kissing and therefore

should receive the same rating. On the other hand, other parents may view the former as being as sexually explicit as a man and a woman having intercourse.

The Language category equates "crude language," which includes the standard obscenities, with "extreme hate speech," which includes advocacy of violence against racial, ethnic, and religious groups. Parents may well not want their children to encounter the word "fuck." Nevertheless, their reasons may be quite different than the reasons they do not want their children viewing a website urging genocide as a means of achieving racial purity. Some parents may find the "N" word much more troubling than the "F" word; some parents may think genocidal advocacy and the spread of prejudices and stereotypes to be much more harmful to the next generation than the spread of coarse language.

A good filtering system will be ideologically capacious even if it cannot be strictly neutral. Even if it inevitably makes value choices through its choices of category it should allow users to select from among a variety of ideologies. By contrast, RSACi's attempts at neutrality by trying as far as possible to *eliminate* ideology. Nudity is nudity, regardless of whether it is artistic or educational. Violence is violence, regardless of whether it is endorsed or deplored. And criticism of groups is coded identically regardless of whether the defamed group is Catholics, atheists, or Satanists. But that is not the elimination of ideology; it is the imposition of an ideology—and a very peculiar one at that, a wooden and formalistic neutrality likely to undermine the ideological concerns of many different kinds of end-users.

These criticisms of RSACi should be understood in context, for RSACi is a significant advance on previous efforts at content filtering. Some of its defects are correctable. For example, RSACi could create separate "Sex" categories for homosexual and heterosexual content. But more important, many of these defects are inherent in any filtering system that relies exclusively on first-party ratings. To be useful, first-party ratings must be reliable. To be reliable, they must produce convergence in rating by different individuals. And if they are convergent in this way, they will be ideologically thin. They will not, by themselves, allow end-users to filter according to many different kinds of ideological preferences.

What the analysis of RSACi shows, then, is that a successful filtering system will require contributions from both first- and third-party raters. In the next section, we offer a new proposal that borrows the best features of RSACi while avoiding many of its disadvantages.

4. A Proposal for a Ratings System: The Layer Cake Model

The central problem we face is to design a system that relies on both first and third parties, that is easy for raters to rate and end users to use, that is flexible, that can accommodate many different value systems and ideologies, and can grow over time. The solution we propose aims to achieve all of these goals.

4.1. *The Basic Model*

Our proposal is to distribute the work of rating and filtering between first parties and third parties. One can think of the filtering system as a three layer cake that sits on a plate. The "plate" is the software specification, which includes PICS, PICSRules, and (eventually) RDF. The first layer of the cake is a basic

vocabulary that will be used by first parties in rating their sites. This vocabulary will consist of between thirty to sixty basic terms and expressions. The actual number will depend on a balance of considerations, including comprehensiveness and ease of use. Too many terms will be difficult for first parties to code. Too few will be insufficient to do the work of description.

For purposes of comparison, RSACi currently contains four categories of five levels each, or twenty basic vocabulary terms. This might suggest that RSACi contains only twenty vocabulary elements. In fact, several of the RSACi levels contain multiple terms, for example, level 3 of language contains "strong language, obscene gestures, or hate speech." This conjunction of terms is an important source of the ideological effects of RSACi's coarseness and equivalency. When the vocabulary elements in RSACi are fully separated, the number is closer to forty, so that an upper limit of sixty elements is not unreasonable.

Unlike the RSACi system, first parties will not code their sites with scalar numbers. Instead, they will simply list all applicable vocabulary elements as a header in their webpages. They may choose to rate their entire website, individual pages within the website, or individual elements within web pages. We expect that most first-party labeling will rate the entire site to save time. But first parties will always have the option to offer separate ratings for pages and individual elements because they want specific parts of their site to be accessible to more people. The choice to differentiate parts of the site properly lies in the hands of the individual content provider, the party that wants to be heard and visited by others, and who therefore can best decide how much effort he or she wishes to invest.

The vocabulary elements should be chosen to be "objective" in the sense of objectivity-as-reliability describe above. In other words, they should involve descriptors that most first parties will apply in roughly the same way. Objectivity-as-reliability will make the task of first-party rating easier and well as more predictable. The more controversial the descriptor, the more time first parties may spend agonizing over the right approach.

The second layer of the cake consists of ratings templates. They are created by third parties. Third parties will take the forty to sixty vocabulary elements and arrange them into categories and scalar orders. Different templates will have different degrees of coarseness. They will make different things equivalent, and they will rate elements in different scalar orders. Organizations that create templates do not have to include all of the vocabulary elements; they will include only those that are relevant to their particular ideological concerns. For example, assuming that the basic vocabulary elements are similar to those in RSACi, one template might make hate speech equivalent to strong profanity ("four letter words" in English) and set both at level 3. A second template might differentiate the two, and place hate speech at a higher level than strong language. A third might do exactly the opposite. A fourth template might treat hate speech as an entirely separate category from four letter words. A fifth might choose not to filter for hate speech at all, but focus only on profanity.

By separating the vocabulary elements from the construction of templates, we can better allow third parties to reflect their value systems while still preserving the reliability of ratings by first-party raters. We should not pretend that the choice of basic vocabulary elements has no ideological overtones. It clearly does. Any set of vocabulary elements at the first level will affect and restrict the kinds of templates that can be created at the second level. For example, if there is no vocabulary element for blasphemy, then third parties who wish to filter for blasphemy will have no codings from first parties to work with. In addition, every vocabulary element contains its own elements of coarseness and equivalency. A vocabulary element

like "strong profanity" makes some expressions equivalent to others and does not permit further differentiation. Nevertheless, the goal is to allow a sufficiently diverse array of vocabulary elements at layer one to facilitate a wide variety of choices of template construction at level two. No system is perfect in all respects, but we think this approach is an improvement on a unitary system.

The basic point is that by combining a basic vocabulary at level one with flexibility at level two we can achieve much greater diversity and provide more end-user choice than in a unitary system. At first it might appear that a common vocabulary at layer one will produce an unacceptable limitation of ideological diversity. In fact, precisely the opposite is true. If ratings systems existed with multiple and inconsistent vocabularies, network effects would soon reduce the number of ratings systems to a very small number. Because of the Internet's enormous size, any Internet ratings system must rely on the good graces of first-party raters. Most of these first parties are unlikely to code their sites for more than one ratings system. As a result, those first parties that seek to be rated will over time tend to converge on a smaller and smaller number of ratings systems, until at last there may be only one or two practical alternatives. This "winner take all" phenomenon is the predictable result of network effects. On the other hand, if first parties can be assured that all templates will be compatible with a fixed set of vocabulary elements, no template will necessarily dominate all of the others, because each template is, in principle, compatible with and combinable with all of the others. Ideological diversity thus is enhanced by having a common language at layer one, much in the same way that other common standards enhance creativity and diversity. It is the common set of standards at lower levels, in the "plate" (i.e, the software specification) and in the first layer of the cake (i.e., the vocabulary) that makes possible greater differentiation and diversity at higher levels.

Diversity can only be achieved if there are a wide variety of templates to choose from. However, because third parties do not have to rate individual sites in order to create templates, their costs are greatly reduced. Indeed, the point is to make templates sufficiently easy to create that many non-profit and ideologically driven organizations will want to create their own templates, which they will give away for general distribution. Thus, we do not expect that there will be a market for templates. Rather, we expect that many nonprofit organizations with ideological agendas will have an interest in making templates widely available to anyone who wants them. If template construction were expensive, the lack of a market would be a drawback. However, the system is designed to be inexpensive and easy to use. Indeed, we believe that an organization can probably create a new template in an afternoon, or at the most a few days. Moreover, because the content of these templates will be public and in the public domain rather than secret and proprietary, any organization can model their own template on the work of previous organizations. In contrast to the first generation of filters, free riding on previous templates is positively encouraged. This should greatly speed up the process of template creation.

Note, moreover, that, at the second layer of the cake, the end-user is free to combine different ratings templates from different organizations. For example, imagine two templates, one from the Moral Majority that rates content based on suitability for age, and one from People for the American Way which rates according to language, nudity, sexual content and violence. An end-user (or yet another third party) might want to block out all content that either the Moral Majority or People for the American Way deem objectionable for children, only content that both templates deem objectionable, or content that the Moral Majority template filters except for content that People for the American Way rates as less than two on its language scale. Indeed, third parties might act as arbitrageurs between different ideological preferences of

different organizations and attempt to offer end-users the best of both (or all) possible ideological worlds.

The combination of the first and second layers of the cake will achieve reliability in self-rating and accommodate a wide diversity of ideological concerns. Nevertheless, some might still object that a common vocabulary will not filter out everything that should be filtered and that a reliable first-party system will be insufficiently sensitive to context. The third layer of our cake is designed to address these problems by further refining the system.

The third layer of the cake is a set of third-party ratings of individual sites. Three types of third-party ratings are possible. First, existing third-party blacklist systems like those involved in CyberPatrol can be conjoined with the results of filtering from ratings templates. Although this solution is technologically feasible, we do not recommend it because of the transparency problems inherent in first-generation blacklist software.

Second, third parties can create their own PICS-compatible ratings systems that can work together with or be superimposed on top of the layer two templates. For example, a group might create a ratings system that included contextual and ideological categories like "artistic" or "racist" or "insensitive to women." However, these ratings systems would be different from templates. The group would have to create the categories and vocabulary and also rate a large number of sites according to those categories and vocabulary. These ratings could then be combined with existing templates.

For example, parents concerned about their children's exposure to intolerance (as defined by certain trusted groups), pornography (but not artistic or educational material), and violence (but not in historical material) could opt for the following profile: Use the People for the American Way Layer Two template, and add the following modifications for Layer Three: Any URL rated above 2 on the NAACP's "Racist" category, The National Organization for Women's "Sexist" category, or the Anti-Defamation League's "Anti-Semitic" category will be rejected. Any URL rated above 2 on the National Endowment for the Art's "Artistic" category, the National Council of Teachers' "Educational" category, or the American Historical Association's "Historical" category will be accepted. (All of these categories refer to hypothetical additional ratings systems imposed at level three). Note that using these third parties' ratings systems will require that end users fetch rating labels from sites operated by these third parties or from a centralized label bureau. However, the most efficient storage method would be to put the ratings in the browser, periodically updated by downloads. This vision of integrated first- and third-party rating systems can provide great ideological richness and flexibility to layers one and two. However, it is costly for third parties, because they have to rate individual sites in addition to coming up with their own PICS-compatible ratings system.

A third possible solution is for organizations to create redemptive lists, i.e., lists of sites that, because of their context, should not be blocked even though their content falls within particular "objective" descriptor categories like nudity, profanity or racist epithets. Redemptive lists are a kind of whitelist. Because redemptive lists indicate sites that shouldn't be blocked, the collective action problems of constructing such lists tend to work in the opposite direction from blacklists: Individual website owners have strong incentives to identify themselves to sympathetic third-party organizations as sites that should not be blocked. This saves these organizations time in identifying and rating sites. Moreover, because organizations have ideological reasons to make sure that certain sites are not blocked, they have incentives to make their redemptive lists of sites available to other organizations with similar ideologies. This will result in a pooling of resources to create redemptive lists.

In short, layer three consists of traditional blacklist filters, ancillary ratings systems, and redemptive lists that can be combined with the results of filtering using layer two templates. Thus, an end user could combine the Moral Majority's layer two template with a layer three redemptive list from an organization with very different agendas, for example those sites that People for the American Way has listed as acceptable for young adults on the basis of language and sexual situations. In this way the system can achieve considerable flexibility as well as ideological richness. End-users need not form these combinations by themselves. In the system we propose there is plenty of opportunity for interested organizations to act as ratings "arbitrageurs": They can produce combinations of ratings systems and offer them as a package to the general public.

4.2. *Complicating the Model: Adding Contextual Judgments to the Layer One Vocabulary*

Still further refinements of this system are possible. One of the biggest problems in designing a ratings system is accommodating judgments of context. For example, many people think that artistic nudity is preferable to nudity that is merely erotic and designed to titillate. The problem is that terms like "artistic" and "erotic" may not produce reliable convergence in first-party ratings. Different kinds of violence may be more harmful to children; variations in context may also be quite important. For example, violence by evil people that is suitably punished, or violence that appears as part of news reports may be less harmful than unpunished or so-called "gratuitous" violence.

One way of refining the system is to include contextual operators ("artistic," "news reporting", "cartoon" etc.) in the first layer vocabulary and then simply leave it up to template constructors whether to make use of these operators. Remember that a template does not have to include all vocabulary elements; it can be as coarse or as comprehensive as the template constructor wishes. Thus, some templates may use some of these contextual operators on the grounds that the tradeoff between reliability and contextual judgment is worthwhile, while other templates will strike a different balance and avoid them. Nevertheless, we note that the more contextual elements are added to the basic vocabulary in layer one, the larger the vocabulary becomes and the more time-consuming and burdensome for first-party raters. Therefore, we do not recommend adding a large number of these operators. Instead, we think the burden of deciding what is appropriate for children because it is "artistic" is best left to whitelists operating at Level Three. As we have noted previously, a whitelist system can sometimes be more finely attuned to ideologically sensitive judgments, and it can draw on the collaborative efforts of many different organizations.

Nevertheless, we do think that one can enhance the precision and ideological flexibility of the system at Layer One by adding a small number of contextual descriptors that can describe context in relatively reliable ways. For example, if contextual operators might include distinct descriptors for cartoon violence and sports violence.[9] Further, the vocabulary could distinguish between media: the system could

[9]It might be desirable to incorporate other factors that the sociological literature finds relevant to the effect of violence on viewers, such as whether the violence is rewarded or punished, and what the reason for the violence is. These sorts of contextual distinctions require value judgments that are inconsistent with the reliability required of first-party labels—whether violence is "legitimate" is obviously a value-laden determination.

offer distinct descriptors for text, audio, pictures, and real-time chat. These sorts of contextual descriptions would not sacrifice very much in terms of reliability.[10]

The demand for contextual operators is perhaps greatest in the case of news. News organizations regularly describe violence, and so it is understandable that they wish to be rated differently from non-news sites that contain violent depictions. We certainly do not rule out an experimental use of a content descriptor for news organizations. Nevertheless, there is some danger that non-news organizations will attempt to use this operator to prevent being filtered, and there will be inevitable controversies about what is or is not news. For these reasons, we think a better solution is to rely on whitelists for news organizations at Layer Three. Because the question of "what is news" is ideologically contested, we think it is better for it to be assigned to third party raters, who can draw up lists of news organizations. Even if the advisory board ultimately decides to include a contextual operator for news organizations in Layer One, we think that a whitelist system provides an invaluable form of insurance in case first-party ratings about news do not prove to be reliable.

4.3. *Open Source and The Creation of a Ratings Board*

The task of constructing a preliminary set of content descriptors will require some degree of centralization and standardization at the beginning. There must be an initial standard set of vocabulary elements that first-party raters can use. We recommend the creation of a board of people for this purpose who have both interest and expertise in the problem of content filtering. We do not think that the board should be for-profit, nor should it be under the auspices or control of any business organization. Ideally, in addition to experts on civil liberties and Internet policy, the board should include social scientists who can

---

[10]Still another way of increasing precision would be to offer different types of content descriptors that could be conjoined in a label when the label is generated. These types might include, for example, "media," "content," and "modifier." "Media" descriptors specify the medium of the content: text, audio, picture, chat etc. "Content" descriptors specify categories of content, such as violence, sexual content, language etc. "Modifier" descriptors specify content more precisely or more contextually; they might include such values as "political," "cartoon," "news," "medical," and "homosexual." First-party raters could rate for each type of descriptor; the types would then be conjoined to produce labels such as "text: violence (political)" or "picture: sex (homosexual) (medical)." Use of these conjoining types will work only if rating is done at the level of individual documents; otherwise, a web page that contains both political cartoons and pictures of violence might be labeled as "picture: violence (cartoon) (political)." RDF permits the labeling of individual documents, and that degree of precision seems desirable; it allows, for example, parents to decide that their children may read news accounts of violence but not see accompanying photographs. The problem with this solution is that many site operators will be unwilling to spend the time to rate pages, text, or photographs individually. Thus, although we offer this solution as a theoretical possibility, we are unsure whether it can be implemented practically.

advise about what kinds of content are more and less harmful to children. However, developers must recognize that "harmful to children" is a contested concept, and they must be aware that filtering will be used for purposes beyond protecting children. The members of the board must also be sensitive to cultural differences; in particular, content descriptors should not track only those concepts salient to Western cultures and Western preoccupations. Finally, the board should also strive to create easy-to-understand guides and questionnaires that reduce the number of questions and complications necessary for first parties in different places around the globe to produce acceptable ratings.

We recommend that elaboration and refinement of the rating system should be conducted on an open source model. The set of content descriptors will be part of the public domain, available for use by anyone at no charge.[11] New content descriptors can be proposed by anyone. The board in charge of constructing the preliminary set should periodically decide whether new descriptors will be approved as part of the standard vocabulary. Board approval, however, would be only an endorsement, neither necessary nor sufficient for practical success. Content descriptors that are not popular will simply remain unused; conversely, descriptors that are not approved by the board can nonetheless become widely used. What descriptors become used is simply a matter of choice for the first-party raters and the third parties constructing templates. As in other open source situations, there will inevitably be feedback between the board and installed base of users and raters. Thus we think that the board will be responsive to features repeatedly demanded by persons who use the ratings system every day.

We thus envision that the board will not only oversee the development of the first generation of vocabulary elements, but also act as a clearinghouse for the open source process that will result in successive updates. We do not think that "official" updates of the set of vocabulary elements should occur very often. One revision every two years is more than enough, primarily because it will require end-users to update their software and first-party raters to re-rate their websites. The board should assist in promulgating updates; it can also encourage users to download new software and first parties to update their websites based on changes to the first layer vocabulary.

4.4. *End-user Interfaces: How to Ensure Ease of Use*

Ease of use for end-users is an important consideration for any filtering system. Ease of use not only enhances the central value of end-user autonomy; it also helps ensure that people use the filtering system.

Ease of use is not inconsistent with a system that is both flexible and powerful. The proposal we describe features several different layers and many possible options for innovation. But it is important to distinguish between the complexity of the *filtering system* and the complexity of the *user interface*. A car

---

[11]The reliance on first-party rating and the lack of intellectual property protection for content descriptors obviously raises the question of misrating. Misrating should be handled by technological means. Organizations whose templates are compromised by misrated URLs can maintain downloadable lists of corrected ratings for those URLs; browsers would give these ratings priority over imbedded descriptors.

is an extremely complex piece of machinery, but its user interface is designed to make it easy to drive. A cake can be baked from many ingredients, but this does not make it difficult to eat.

Software companies spend millions of dollars a year to make their user interfaces easy to use despite the complexity of the underlying software engines. We see no reason why this learning cannot be adapted to filtering, which, in many ways, involves a much less complicated piece of software.

We do not purport to design the actual user interface in this report: the interface will be integrated into the end-user's browser, and so the actual implementation is a job for professional software engineers. However, we do offer the following recommendations about how to enhance ease of use and thus promote end-user autonomy.

First, when the end user purchases the computer, the dealer can offer to set up the filtering system at the point of purchase, just as dealers currently offer to install many other pieces of software.

Second, the filtering system should be accompanied by a step-by-step "wizard," akin to the devices Microsoft currently employs for its software suite. When the end-user boots up the computer for the first time, the wizard can ask "Do you want to protect your children from harmful content now?" and take the end-user step-by-step through the process. The end-user can choose to install a filter at that point or delay the process until later.

Third, access to filters should be easy to find and readily available to end-users whenever they are using their browsers. Buttons indicating access to the filter setting should be prominently displayed on the browser's main window, rather than hidden several layers down in the browser menu. The end-user should have the opportunity to turn filtering on or off with at most a few clicks of the mouse (and the typing of a password, in the case of turning off the filter).

Fourth, when adults are surfing the Net, they may not wish to be filtered. It is also important to remind them of the fact that they are being filtered. Therefore, when the browser blocks a site, it should not only tell the end-user that the site has been blocked, and why the site has been blocked, but also contain a small check box. Checking this box will allow the end-user to turn off filtering for the remainder of the session (i.e., while the browser is running) by typing in a password. The browser should also prominently display a button or toolbar that controls filtering settings so that the end-user can adjust filtering even if he or she does not wish to turn all filtering off. (Of course, changing filtering settings will also require the use of a password.).

Fifth, adjusting filtering ratings should be done with buttons and slides that allow settings to be changed with a few clicks of the mouse. If filtering templates are associated with particular languages or interest groups, they can be represented with icons (like national flags or other symbols) that make it easier for the end-user to identify their source and purpose.

Sixth, filtering wizards should make it easy for end-users to download new templates from the Internet and updated versions of older templates (including whitelists and other third-party ratings). End-users should also be able to quickly and easily create their own blacklists and whitelists by adding particular web sites they encounter to lists of approved or disapproved sites.

An easy-to-use interface is perfectly compatible with a system that allows third parties considerable leeway in designing filtering templates and providing independent ratings and white lists. It is largely a question of good software design. The end-user does not need to know how many options are available to the template designer-- the end user only needs to know how to operate the software interface before

him or her. End users do not need to know the details of the system and its complexities any more than they need to know how an engine works to drive a car or they need to know how to bake a cake in order to eat it.

End-users are not the only parties who need an easy-to-use interface. First party raters will also need help in rating their sites. Our purposed system contemplates that the Layer One vocabulary will contain between thirty to sixty basic terms. These terms should be organized and accompanied by questionnaires that help take the first-party rater through the system. Creating these questionnaires and testing them for ease of use can and should be an important task of the advisory board that designs the Layer One vocabulary. We also think that a software program or "wizard" could be designed to assist in this process. Finally, the system should be integrated into web authoring tools, so that authors and designers can easily integrate the system into their web content.

5. Conclusion

In sum, our proposed system consists of three layers placed on top of a software specification:

1. Layer One: A basic vocabulary for first-party raters.

2. Layer Two: A series of templates constructed by third-party raters that combine and rank these vocabulary elements in many different ways. In addition, multiple templates can be combined and added to refine the filter.

3. Layer Three: An assortment of blacklist filters, ancillary ratings systems, and redemptive lists maintained third-party raters, that can be combined and added to the results of layers one and two.

The proof of the pudding, we recognize, will be in the construction of Layer One. Putting together the initial set of content descriptors will require important tradeoffs between precision, reliability, and increased workload for first-party raters. We do not think, however, that the tradeoffs are unacceptable. And we think that the result will better serve the interests of content providers and end-users alike than a unitary system of content rating and filtering.

The Layer Cake Model

Additional third party ratings can be superimposed on the results of Layer Two to increase flexibility and ideological diversity.	Examples: • Blacklists: Lists of specific sites judged unsuitable for children. • Whitelists: Lists of specific sites judged suitable for children (e.g. news sites, educational sites, etc.). • Any PICS (or RDF) compatible filtering system. • Combinations of features of any Layer Two rating templates	**Layer 3**
Rating templates are created by third parties. They arrange the basic vocabulary into different categories and scalar orders within categories (e.g., from 1 to 5 in the category of "Nudity."). Templates are open source and can build on each other. Templates can be designed for different countries and cultures. They reflect the different ideologies and value systems of different third parties.	Examples of possible rating templates: • Third Party *A* creates a template based on *categories* with *scalar orderings*: Language (1 through 5) Nudity (1 through 5) Sex (1 through 5) Violence (1 through 5) The end user selects how high a level he or she wants for each category. This is an *adjustable rating template.* • Third Party *B* creates a template based on judgments about what is appropriate for children of a particular *age*: Age 0-6 Age 6-9 Age 10-13 Age 14-17 Age 18 and Over This template is also adjustable. • Third Party *C* creates a template based on a particular set of values, which is not adjustable.	**Layer 2**

A set of **basic vocabulary elements** that first parties will use to rate their own sites. These elements are chosen and defined through detailed questionnaires so that most first parties will employ them in roughly the same way. Descriptions of content that lack this degree of convergence by first party raters do not appear in Layer One.	Examples of basic vocabulary: • Descriptions of *content*: Possible examples: "no nudity," "full frontal nudity," "strong expletives," "extreme violence." (N.B. Each of these content labels will be further specified and defined through a questionnaire.) • Descriptions of *context*: Possible examples: medical treatise, medical advice, sports programming, shopping • Descriptions of *information requested*: Possible examples: asks for credit card number, asks for personal information (name, address, home telephone number), asks for household income. • Descriptions of *media*: Possible examples: picture, text, streaming video	**Layer 1**
A **basic software specification** for labeling content	PICS, PICS Rules, or RDF	**Plate**

Jack M. Balkin is Knight Professor of Constitutional Law and the First Amendment, Yale Law School, and Director, The Information Society Project.
Beth Simone Noveck is Director of International Programs, The Information Society Project.
Kermit Roosevelt is Resident Fellow, The Information Society Project.

Bibliography

American Civil Liberties Union (ACLU) (1999): Censorship in a Box, http://www.aclu.org/issues/cyber/box.html (visited April 24, 1999).

The Associated Press (1999): Sports Violence Seen as Dangerous Influence on Kids, http://interactive.cfra.com/1998/06/03/38469.html (visited May 17, 1999).

Balkin, J.M. (1996): Media Filters, The V-Chip, and the Foundations of Broadcast Regulation, *Duke Law Journal* 45:1133.

Boyle, James (1997): Foucault in Cyberspace: Surveillance, Sovereignty, and Hardwired Censors, *Univ. Cin. Law Review* 66:177.

Bray, Tim (1999): RDF and Metadata, http://xml.com/xml/pub/98/06/rdf.html (visited April 24, 1999).

Brown, Jason (1999): *Children and Television Violence*, http://www.compusmart.ab.ca/jbrown/Violence.html (visited May 17, 1999)

Center for Democracy and Technology (CDT)(1999): Internet Family Empowerment White Paper: How Filtering Tools Enable Responsible Parents to Protect Their Children Online http://www.cdt.org/speech/empower.html (visited April 24, 1999).

Communications Decency Act, 47 U.S.C. Sections 223 et seq.

Cowan, John (1999): RDF Made Easy, http://www.ccil.org/~cowan/XML/RDF-made-easy.html (visited April 24, 1999).

Cyber-Rights & Cyber-Liberties (1999): Felix Somm Decision in English, http://www.cyber-rights.org/isps/somm-dec.htm (visited April 24, 1999).

Dobeus, Jonathan (1998): Rating Internet Content and the Spectre of Government Regulation, *John*

Marshall Journal of Computer & Information Law, 16:625.

Electronic Frontier Foundation (1999): Policy on Public Interest Principles for Online Filtration, Ratings and Labeling Systems,http://www.eff.org/policies/filtration_policy.html (visited May 17, 1999).

Filtering Facts (1999): How Filters Really Work, http://www.filteringfacts.org/howfilt.htm (visited April 24, 1999).

Flynn, John (1999): Frequently Asked Questions about the Extensible Markup Language, http://www.ucc.ie/xml/ (visited May 17, 1999).

Ledingham, Jane (1999): The Effects of Media Violence on Children, http://www.media-awareness.ca/eng/med/home/resource/famvlnc.htm (visited May 17, 1999).

Lessig, Lawrence (1998): What Things Regulate Speech: CDA 2.0 vs. Filtering, *Jurimetrics* 38:629.

Mediascope (1999): National Television Violence Study, http://www.media-awareness.ca/eng/med/home/resource/ ntvs.htm (visited April 24, 1999).

Microsys.com (1999): Recreational Software Advisory Council Launches Objective, Content-Labeling Advisory System for Internet, http://www.microsys.com/Profiles/RSAC_1.HTM (visited April 24, 1999).

Raymond, Eric S. (1999): The Cathedral and the Bazaar, http://www.tuxedo.org/~esr/writings/cathedral-bazaar/cathedral-bazaar.html (visited May 17, 1999).

Roberts, Donald F. (1999): Media Content Rating Systems: Informational Advisories or Judgmental Restrictions?, http://www.rsac.org.fra_content.asp?onIndex=36 (visited April 24, 1999).

RSACi, RSACi Terms and Conditions, http://rsac.org/content/register (visited May 17, 1999).

Reno v. ACLU, 117 S.Ct. 2329 (1997).

Resnick, Paul (1999): PICS Censorship, & Intellectual Freedom FAQ, http://www.si.umich.edu/~presnick/pics/intfree/FAQ.htm (visited April 22, 1999).

W3C (1999a): Extensible Markup Language (XML), http://www.w3.org/XML/#faq (visited May 17, 1999).

W3C (1999b): PICS Label Distribution Label Syntax and Communication Protocols, http://www.w3.org/TR/REC-PICS-labels (visited April 24, 1999).

W3C (1999c): PICSRules 1.1, http://www.w3.org/TR/REC-PICSRules (visited April 24, 1999).

W3C (1999d): Rating Services and Rating Systems (and Their Machine Readable Descriptions), http://www.w3.org/TR/REC-PICS-services (visited April 24, 1999).
W3C (1999e): Resource Description Framework (RDF) Model and Syntax Specification, http://www.w3.org/TR/REC-rdf-syntax/ (visited April 24, 1999).

W3C (1999f): Statement on the Intent and Use of PICS: Using PICS Well, http://www.w3.org/TR/NOTE-PICS-Statement (visited April 24, 1999).

Wagner, R. Polk (1999) Filters and the First Amendment, *Minnesota Law Review* 83:755.

Weinberg, Jonathan (1999): Rating the Net, *Hastings Comm/Ent Law Journal* 19:453 (available at http://www/msen/com/~weinberg/rating.htm).

Working Party (1999), Report, http://www2.echo.lu/legal/en/internet/wpen.html (visited April 24, 1999).