FILTERING INFORMATION ON THE INTERNET

Look for the labels to decide if unknown software and World Wide Web sites are safe and interesting

by Paul Resnick

The Internet is often called a global village, suggesting a huge but close-knit community that shares common values and experiences. The metaphor is misleading. Many cultures coexist on the Internet and at times clash. In its public spaces, people interact commercially and socially with strangers as well as with acquaintances and friends. The city is a more apt metaphor, with its suggestion of unlimited opportunities and myriad dangers.

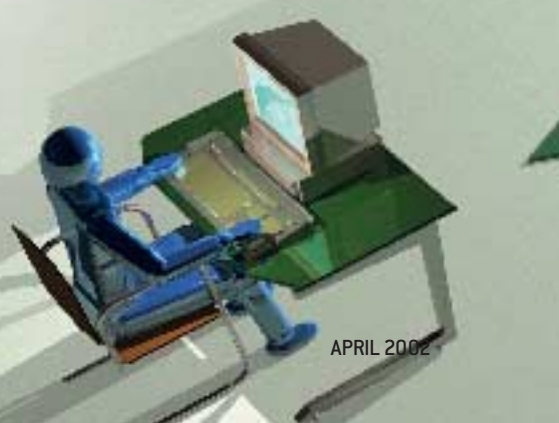
To steer clear of the most obviously offensive, dangerous or just boring neighborhoods, users can employ some mechanical filtering techniques that identify easily definable risks. One technique is to analyze the contents of on-line material. Thus, virus-detection software searches for code fragments that it knows are common in virus programs. Services such as AltaVista and Lycos can either highlight or exclude World Wide Web documents containing particular words. My colleagues and I have been at work on another filtering technique based on electronic labels that can be added to Web sites to describe digital works. These labels can convey characteristics that require human judgment—whether the Web page is funny or offensive—as well as information not readily apparent from the words and graphics, such as the Web site's policies about the use or resale of personal data.

The Massachusetts Institute of Technology's World Wide Web Consortium has developed a set of technical standards called PICS (Platform for Internet Content Selection) so that people can electronically distribute descriptions of digital works in a simple, computer-readable form. Computers can process these labels in the background, automatically shielding users from undesirable material or directing their attention to sites of particular interest. The original impetus for PICS was to allow parents and teachers to screen materials they felt were inappropriate for children using the Net. Rather than censoring what is distributed, as the Communications Decency Act and other legislative initiatives have tried to do, PICS enables users to control what they receive.

What's in a Label?

PICS labels can describe any aspect of a document or a Web site. The first labels identified items that might run afoul of local indecency laws. For example, the Recreational Software Advisory Council (RSAC) adapted its computer-game rating system for the Internet. Each RSACi (the "i" stands for "Internet") label has four numbers, indicating levels of violence, nudity, sex and potentially offensive language. Another organization, SafeSurf, has developed a vocabulary with nine separate scales. Labels can reflect other concerns beyond indecency, however. A privacy vocabulary, for example, could describe Web sites' information practices, such as what personal information they collect and whether they resell it. Similarly, an intellectual-property vocabulary could describe the conditions under which an item could be viewed or reproduced [see "Trusted Systems," by Mark Stefik, page 78]. And various Web-indexing organizations could develop labels that indicate the subject categories or the reliability of information from a site.

Labels could even help protect computers from exposure to viruses. It has become increasingly popular to download small fragments of computer code, bug fixes and even entire applications from Internet sites. People generally trust



FILTERING SYSTEM for the World Wide Web allows individuals to decide for themselves what they want to see. Users specify safety and content requirements (*a*), which label-processing software (*b*) then consults to determine whether to block access to certain pages (marked with a stop sign). Labels can be affixed by the Web site's author (*c*), or a rating agency can store its labels in a separate database (*d*).

that the software they download will not introduce a virus; they could add a margin of safety by checking for labels that vouch for the software's safety. The vocabulary for such labels might indicate which virus checks have been run on the software or the level of confidence in the code's safety.

In the physical world, labels can be attached to the things they describe, or they can be distributed separately. For example, the new cars in an automobile showroom display stickers describing features and prices, but potential customers can also consult independent listings such as consumer-interest magazines. Similarly, PICS labels can be attached or detached. An information provider that wishes to offer descriptions of its own materials can directly embed labels in Web documents or send them along with items retrieved from the Web. Independent third parties can describe materials as well. For instance, the Simon Wiesenthal Center, which tracks the activities of neo-Nazi groups, could publish PICS labels that identify Web pages containing neo-Nazi propaganda. These labels would be stored on a separate server; not everyone who visits the neo-Nazi pages would see the Wiesenthal Center labels, but those who were interested could instruct their software to check automatically for the labels.

Software can be configured not merely to make its users aware of labels but to act on them directly. Several Web software packages, including CyberPatrol and Microsoft's Internet Explorer, already use the PICS standard to control users' access to sites. Such software can make its decisions based on any PICS-compatible vocabulary. A user who plugs in the RSACi vocabulary can set the maximum acceptable levels of language, nudity, sex and violence. A user who plugs in a software-safety vocabulary can decide precisely which virus checks are required.

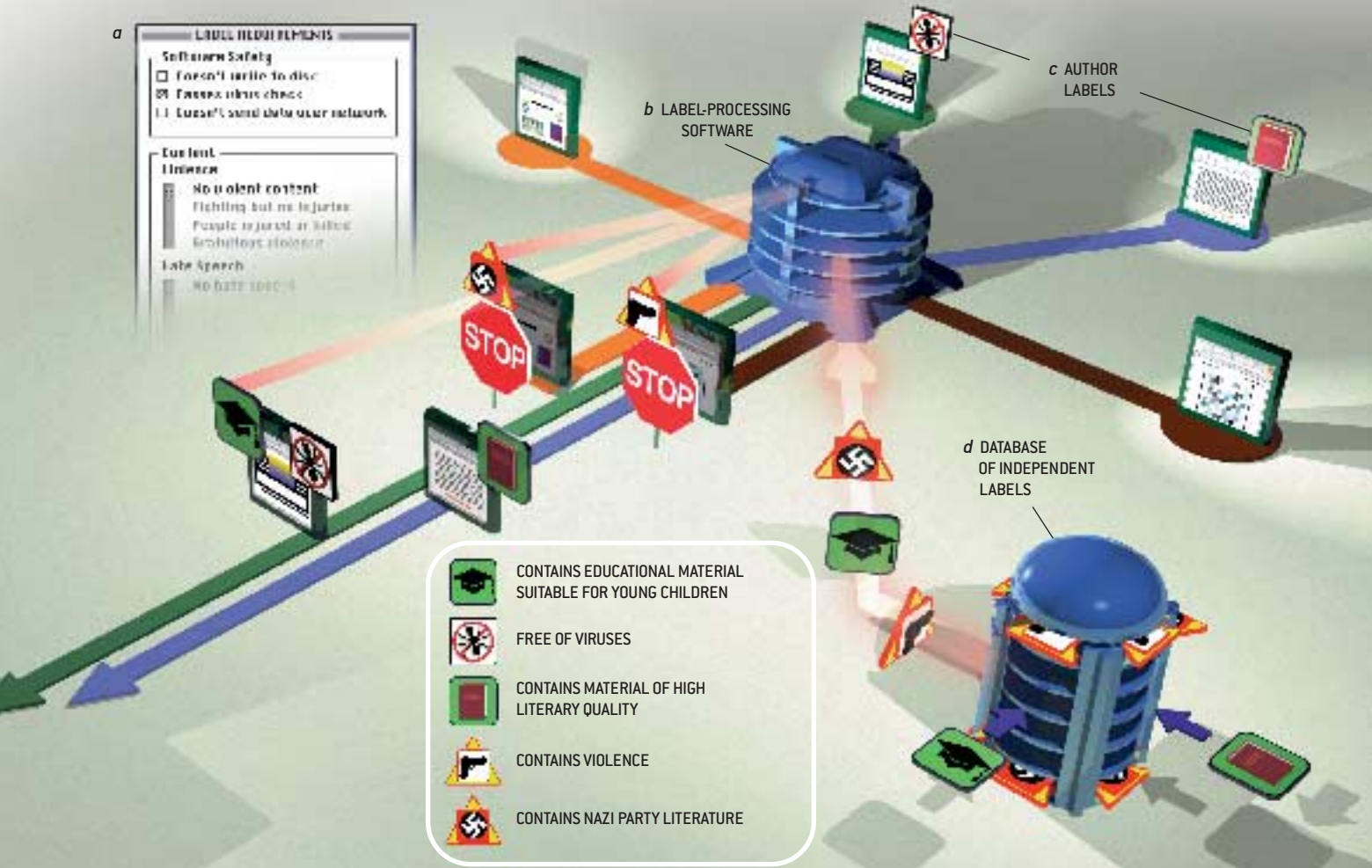
In addition to blocking unwanted materials, label processing can assist in finding desirable materials. If a user expresses a preference for works of high literary quality, a search engine might be able to suggest links to items labeled that way. Or if the user prefers that personal data not be collected or sold, a Web server can offer a version of its service that does not depend on collecting personal information.

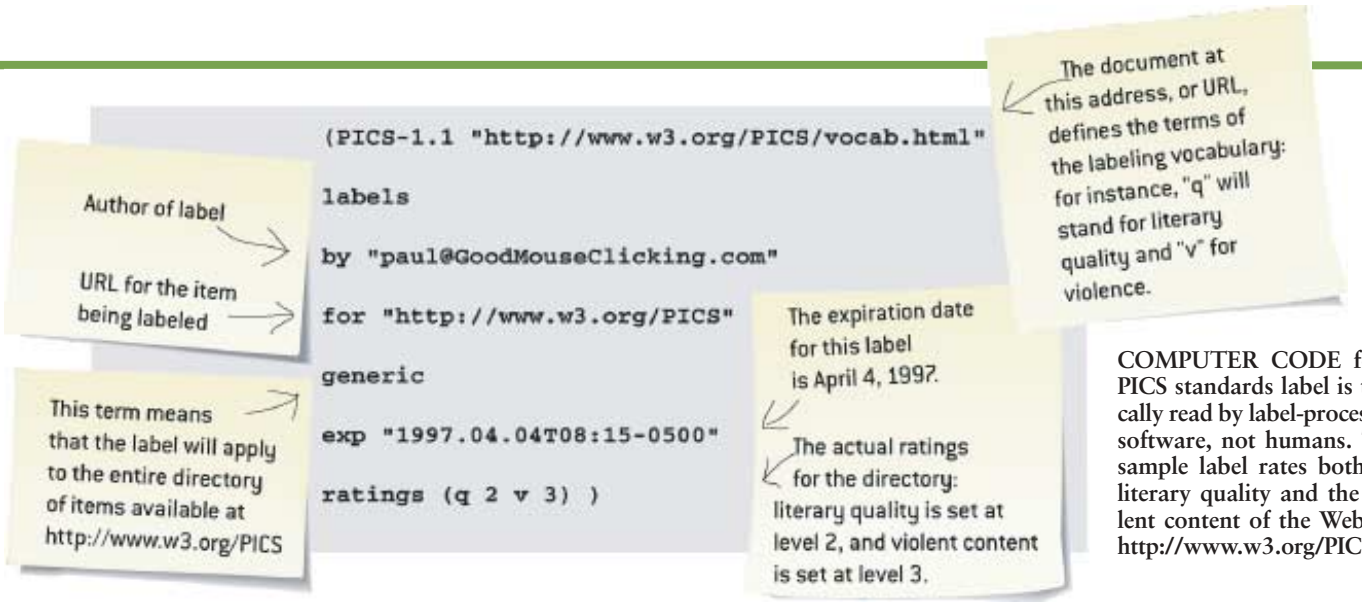
Establishing Trust

Not every label is trustworthy. The creator of a virus can easily distribute a misleading label claiming that the software is safe. Checking for labels merely converts the question of whether to trust a piece of software to one of trusting the labels. One solution is to use cryptographic techniques that can determine whether a document has been changed since its label was created and to ensure that the label really is the work of its purported author.

That solution, however, simply changes the question again, from one of trusting a label to one of trusting the label's author. Alice may trust Bill's labels if she has worked with him for years or if he runs a major software company whose reputation is at stake. Or she might trust an auditing organization of some kind to vouch for Bill.

BRYAN CHRISTIE





JENNIFER C. CHRISTIANSEN

COMPUTER CODE for a PICS standards label is typically read by label-processing software, not humans. This sample label rates both the literary quality and the violent content of the Web site http://www.w3.org/PICS

Of course, some labels address matters of personal taste rather than points of fact. Users may find themselves not trusting certain labels, simply because they disagree with the opinions behind them. To get around this problem, systems such as GroupLens and Firefly recommend books, articles, videos or musical selections based on the ratings of like-minded people. People rate items with which they are familiar, and the software compares those ratings with opinions registered by other users. In making recommendations, the software assigns the highest priority to items approved by people who agreed with the user's evaluations of other materials. People need not know who agreed with them; they can participate anonymously, preserving the privacy of their evaluations and reading habits.

Widespread reliance on labeling raises a number of social concerns. The most obvious are the questions of who decides how to label sites and what labels are acceptable. Ideally, anyone could label a site, and everyone could establish individual filtering rules. But there is a concern that authorities could assign labels to sites or dictate criteria for sites to label themselves. In an example from a different medium, the television industry, under pressure from the U.S. government, has begun to rate its shows for age appropriateness.

Mandatory self-labeling need not lead to censorship, so long as individuals can decide which labels to ignore. But people may not always have this power. Improved individual control removes one rationale for central control but does not prevent its imposition. Singapore and China, for instance, are experimenting with national "firewalls"—combinations of software and hardware that block their citizens' access to certain newsgroups and Web sites.

Another concern is that even without central censorship, any widely adopted vocabulary will encourage people to make lazy decisions that do not reflect their values. Today many parents who may not agree with the criteria used to assign movie ratings still forbid their children to see movies rated PG-13 or R; it is too hard for them to weigh the merits of each movie by themselves.

Labeling organizations must choose vocabularies carefully to match the criteria that most people care about, but even so, no single vocabulary can serve everyone's needs. Labels concerned only with rating the level of sexual content at a site will be of no use to someone concerned about hate speech. And no labeling system is a full substitute for a thorough and thoughtful evaluation: movie reviews in a newspaper can be far more enlightening than any set of predefined codes.

Perhaps most troubling is the suggestion that any labeling system, no matter how well conceived and executed, will tend to stifle noncommercial communication. Labeling requires human time and energy; many sites of limited interest will probably go unlabeled. Because of safety concerns, some people will block access to materials that are unlabeled or whose labels are untrusted. For such people, the Internet will function more like broadcasting, providing access only to sites with sufficient mass-market appeal to merit the cost of labeling.

While lamentable, this problem is an inherent one that is not caused by labeling. In any medium, people tend to avoid the unknown when there are risks involved, and it is far easier to get information about material that is of wide interest than about items that appeal to a small audience.

Although the Net nearly eliminates the technical barriers to communication with strangers, it does not remove the social costs. Labels can reduce those costs, by letting us control when we extend trust to potentially boring or dangerous software or Web sites. The challenge will be to let labels guide our exploration of the global city of the Internet and not limit our travels.

The Author

PAUL RESNICK joined AT&T Labs–Research in 1995 as the founding member of the Public Policy Research group. He is also chairman of the PICS working group of the World Wide Web Consortium. Resnick received his Ph.D. in computer science in 1992 from the Massachusetts Institute of Technology and was an assistant professor at the M.I.T. Sloan School of Management before moving to AT&T.

Further Reading

RATING THE NET. Jonathan Weinberg in *Hastings Communications and Entertainment Law Journal*, Vol. 19; March 1997 (in press). Available on the World Wide Web at http://www.msen.com/~weinberg/rating.htm

RECOMMENDER SYSTEMS. Special section in *Communications of the ACM*, Vol. 40, No. 3; March 1997 (in press).

The Platform for Internet Content Selection home page is available on the World Wide Web at http://www.w3.org/PICS