

EXPERT REPORT OF JOSEPH JANES

I. Background

As a library educator with expertise in networked information systems such as the Internet, I have been asked to provide expert testimony in this case. I have previously provided expert testimony in one case, *Mainstream Loudoun et al. v. Board of Trustees of the Loudoun County Library*. I am providing my services in this case for free. Some of the exhibits that are attached to this report may be introduced during my testimony. At this time, I do not anticipate using any additional exhibits.

I hold three degrees, all from Syracuse University: an A.B. in Mathematics with a dual major in Information and Library Studies, a Masters in Library Science, and a Ph.D. in Information Transfer. A copy of my curriculum vitae is attached as Exhibit 1. I am currently Assistant Professor in the Information School of the University of Washington. I have been teaching at the graduate level since 1984 in library schools at Syracuse, the University of North Carolina, the State University of New York at Albany, and the Universities of Michigan and Washington. I have taught introductory courses in the areas of online search and retrieval, the use of technologies in library work, statistics and research methods, and reference. I have taught advanced courses or seminars in most of these areas, as well as the development of Internet-based library applications and services, the impacts of technology, and relevance research.

I was also the founder and Director of the Internet Public Library, an online library available to the public at <http://www.ipl.org>. In addition, in the last several years, I have

consulted with the New York Public Library in the building of their new Science, Industry and Business Library, with the Henry Ford Museum in Dearborn, Michigan on Internet applications there and with the Multnomah County Library in Oregon on new reference initiatives. I was a co-founder, with Louis Rosenfeld, of Argus Associates, Inc., an Internet consulting and information architecture firm. I have given presentations, including keynote addresses, at major conferences in the United States, Canada, Great Britain, and Japan, the Netherlands, and have been interviewed on technology/library issues on National Public Radio and the BBC.

My creative and research interests include investigating ways in which the emerging information environment is affecting the practice of librarianship, and how the principles of librarianship can best be used to make this environment easier to use. This includes:

- questions about the use of networked information resources (such as those found on the Internet) to answer ready-reference and more detailed research questions,
- development of network-based services and systems which take advantage of concepts and practices from librarianship,
- thinking about ways in which librarianship can and should evolve to adapt to and take advantage of high-speed computing, high-bandwidth communication and mass interconnection.

I will be providing expert testimony in this case about the degree to which blocking software is blocking access to Internet sites that would be of use or interest in a library context.

II. Methodology

In September of 2001, I received a request from Christopher Hansen of the American

Civil Liberties Union to file an expert report in this case to evaluate whether a sites that had been blocked by various Internet blocking software packages were examples of “overblocking”, that is, were blocked but would be of use, value or interest in a library collection or context. I agreed to participate, and we discussed the nature of the work to be done, including drawing a random sample of those sites and evaluating them individually.

On September 21, 2001, I received a file via electronic from Ben Edelman. That file contained a list of 6775 addresses of Web sites. To provide a high degree of confidence in the results, I chose to draw a sample of 859 of those sites. Sampling theory dictates that a sample of that size would produce a 95% confidence interval of $\pm 2.5\%$ around my estimate of the proportion of sites that were overblocked; that is, in 19 of 20 cases, the true proportion of overblocked sites in the population of 6775 would lie within 2.5% of the proportion found in the sample of 859. See Exhibit 2.¹

I downloaded the list into SPSS, a popular and frequently used statistical software package. In addition to the list of sites, the file included categories under which three of the blocking software packages had blocked each site. I performed simple frequency analyses on the list to determine the proportions of sites blocked by each package in each category. For example, the N2H2 package blocked 915 of the 6775 sites as “Sex”, about 13.5% of the total list.

I used SPSS to draw a simple random sample of 859 of the sites. After drawing a potential sample, I calculated the frequencies of each category for the sample and compared

¹ These web sites are also referenced and archived on CD-R's submitted as Appendices A & B to the Expert Report of Benjamin Edelman.

them to that of the population, trying to get a sample that closely mirrored the population in as many ways as possible. In the sample, 14.3% of the sites were blocked by N2H2 as “Pornography”, and in each case, the sample and population proportions differed by no more than 1%.

To assist in the evaluation of this sample, I recruited, via electronic mail, a number of current and former students at the Information School of the University of Washington. I asked for people with experience, coursework, or background in collection development. A total of 16 people responded and were able to assist; I divided this group into two parts. Those with less direct experience (a group of 11) would evaluate the entire list to identify the most obviously overblocked sites; the second group, those with much more experience in school or public library collection development and reference work, would review the remaining sites to make final decisions.

In the first round, each person evaluated two groups of about 80 sites, and each group of sites was evaluated by two different people. Each group of sites included the following instructions:

Look carefully at each of the Web sites on the list. Please make a notation of any site that appears to meet **any** of the following criteria:

a. Contains information similar to that already found in libraries,

or

b. Contains information a librarian would want in the library if s/he had unlimited funds to purchase information and unlimited shelf space,

or

c. You would be willing to refer a patron (of any age) to the site if the patron appeared at a reference desk seeking information about the subject of the site. For this last criterion, we recognize that you might not refer a young child to a Calculus site just because it would not be useful to that child, but you should ignore

that factor. Informational sites, such as a Calculus site should be noted. A site that is purely erotica should not be noted.

Enter “Yes” in the right-hand column for any site that meets **any** of the above criteria, in your judgment.

Enter “No” for any site that meets **none** of these criteria.

The final sentence in c) above was intended to indicate that erotica sites should be voted “No”; this point was clarified for second-round judges.

Sites that received “Yes” votes from both judges were determined to be of sufficient interest in a library context and removed from further analysis. Sites receiving one or two “No” votes would go to the next round. Requiring two “Yes” votes at this stage is a high standard and raises confidence in these judgments. All judges worked independently and were instructed not to discuss their work with anyone else until all decisions had been made.

One participant in this round was unable to complete his assignment. I decided to exclude the two groups of sites he was to evaluate from further analysis. Since groups of sites were randomly assigned to judges, and the sites were ordered alphabetically, it is unlikely that these two groups of sites differed significantly from the rest of the sample. Excluding these sites leaves a sample size of 699, and this widens the confidence interval for results to $\pm 2.8\%$.

Analysis of the blocking categories of the 160 omitted sites reveals slightly higher proportions for some categories, and slightly lower proportions for others (the N2H2 “Sex” category was 11.3% of these sites), so they do not appear to be significantly different from the sample or population.

III. Results

In the first round of analysis, 243 sites (34.8% of the sample) received “Yes” votes from both judges (See Exhibit 3); 456 sites (65.2%) received one or more “No” votes or could

not be found and were therefore sent forward to the second round of judging. See Exhibit 4.

Some of the judges in the first round had difficulty in deciding what to do with sites that seemed to have a primarily commercial purpose and in most cases voted them as “No”. For the second round of judging, we added a sentence to the judges’ instructions; since most libraries include in their collections the Yellow Pages and other guides or directories of commercial enterprises, and since one could easily imagine people wishing to do research on, say, the ways in which people are selling products or services or otherwise conducting business on the Internet, we instructed judges to vote “Yes” on such sites. The new instructions read:

Look carefully at each of the Web sites on the list. Please make a notation of any site that appears to meet **any** of the following criteria:

a. Contains information similar to that already found in libraries,

or

b. Contains information a librarian would want in the library if s/he had unlimited funds to purchase information and unlimited shelf space,

or

c. You would be willing to refer a patron (of any age) to the site if the patron appeared at a reference desk seeking information about the subject of the site. For this last criterion, we recognize that you might not refer a young child to a Calculus site just because it would not be useful to that child, but you should ignore that factor. Informational sites, such as a Calculus site should be noted. Sites that have a commercial purpose should be included here if they might be of use or interest to someone wishing to buy the product or service or doing research on commercial behavior on the Internet, much as most libraries include the Yellow Pages in their collections. A site that is purely erotica should not be noted.

Enter “Yes” in the right-hand column for any site that meets **any** of the above criteria, in your judgment.

Enter “No” for any site that meets **none** of these criteria

This second round of judging produced the following results: 60 sites could not be found (due to broken links, 404 not found errors, domain for sale messages, etc.), 231 sites were judged “Yes”, and 165 judged “No”. (One site, <http://www.WFTurfClub.com>, was inadvertently not evaluated by the judge; I looked at it and judged it “Yes”.) See Exhibit 5.

Second-round judges had substantially more experience in library reference and collection development work, and so I felt it reasonable to rely on their judgments.

IV. Conclusions

Overall, then, 165 of the 699 sites evaluated were found not to be of any value or use in a library context, 23.6% of the total. 60, or 8.6% could not be found, and therefore 474 of the sites, or 67.8% are examples of overblocking on the part of these blocking packages. We then can be 95% confident that the actual proportion of overblocked sites in the population of 6775 sites is between 65.0% and 70.6%.

Based on this analysis, I conclude that the blocking programs tested block a significant amount of content that would be appropriate in a library setting.

Dated: October 15, 2001

Signed _____

Joseph Janes