News Web sites seek more search control

By ANICK JESDANUN, AP Internet Writer

1 hour, 15 minutes ago

NEW YORK - The desire for greater control over how search engines index and display Web sites is driving an effort by leading news organizations and other publishers to revise a 13-year-old technology for restricting access.

Currently, Google Inc., Yahoo Inc. and other top search companies voluntarily respect a Web site's wishes as declared in a text file known as "robots.txt," which a search engine's indexing software, called a crawler, knows to look for on a site.

The formal rules allow a site to block indexing of individual Web pages, specific directories or the entire site, though some search engines have added their own commands.

The new proposal, to be unveiled Thursday by a consortium of publishers at the global headquarters of The Associated Press, seeks to have those extra commands — and more — apply across the board. Sites, for instance, could try to limit how long search engines may retain copies in their indexes, or tell the crawler not to follow any of the links that appear within a Web page.

The current system doesn't give sites "enough flexibility to express our terms and conditions on access and use of content," said Angela Mills Wade, executive director of the European Publishers Council, one of the organizations behind the proposal. "That is not surprising. It was invented in the 1990s and things move on."

Robots.txt was developed in 1994 following concerns that some crawlers were taxing Web sites by visiting them repeatedly or rapidly. Although the system has never been sanctioned by any standards body, major search engines have voluntarily complied.

As search engines expanded to offer services for displaying news and scanning printed books, news organizations and book publishers began to complain.

The proposed extensions, known as Automated Content Access Protocol, partly grew out of those disputes. Leading the ACAP effort were groups representing publishers of newspapers, magazines, online databases, books and journals. The AP is one of dozens of organizations that have joined ACAP.

News publishers complained that Google was posting their news summaries, headlines and photos without permission. Google claimed that "fair use" provisions of copyright laws applied, though it eventually settled a lawsuit with Agence France-Presse and...
agreed to pay the AP without a lawsuit filed. Financial terms haven't been disclosed.

Wade said ACAP could thwart future legal battles and make Web sites more comfortable about putting more material online, including scholarly journals and other items requiring subscriptions.

The new ACAP commands will use the same robots.txt file that search engines now recognize. Web sites can start using them Thursday alongside the existing commands.

Like the current robots.txt, ACAP’s use would be voluntary, so search engines ultimately would have to agree to recognize the new commands. Search engines also could ignore them and leave it to courts to rule on any disputes over fair use.

Google spokeswoman Jessica Powell said the company supports all efforts to bring Web sites and search engines together but needed to evaluate ACAP to ensure it can meet the needs of millions of Web sites — not just those of a single community.

"Before you go and take something entirely on board, you need to make sure it works for everyone," Powell said.

ACAP organizers tested their system with French search engine Exalead Inc. but had only informal discussions with others. Wade said organizers wanted to focus first on getting sites to adopt the system, figuring search engines will follow once a critical mass is reached.

Danny Sullivan, editor in chief of the industry Web site Search Engine Land, said robots.txt "certainly is long overdue for some improvements."

But he questioned whether ACAP would do much to prevent future legal battles. And being an initiative of news publishers, he said, it might lack attributes that blogs, online retailers and other Web sites might need in an updated robots.txt.

Francis Cave, ACAP's technical project manager, said Thursday's plan was only "a first stab. ... We full expect we will need to add to that."

Already contemplated is support for video files, not just text and still images. Cave said online archives such as the British Library and the Internet Archive might also need special commands.