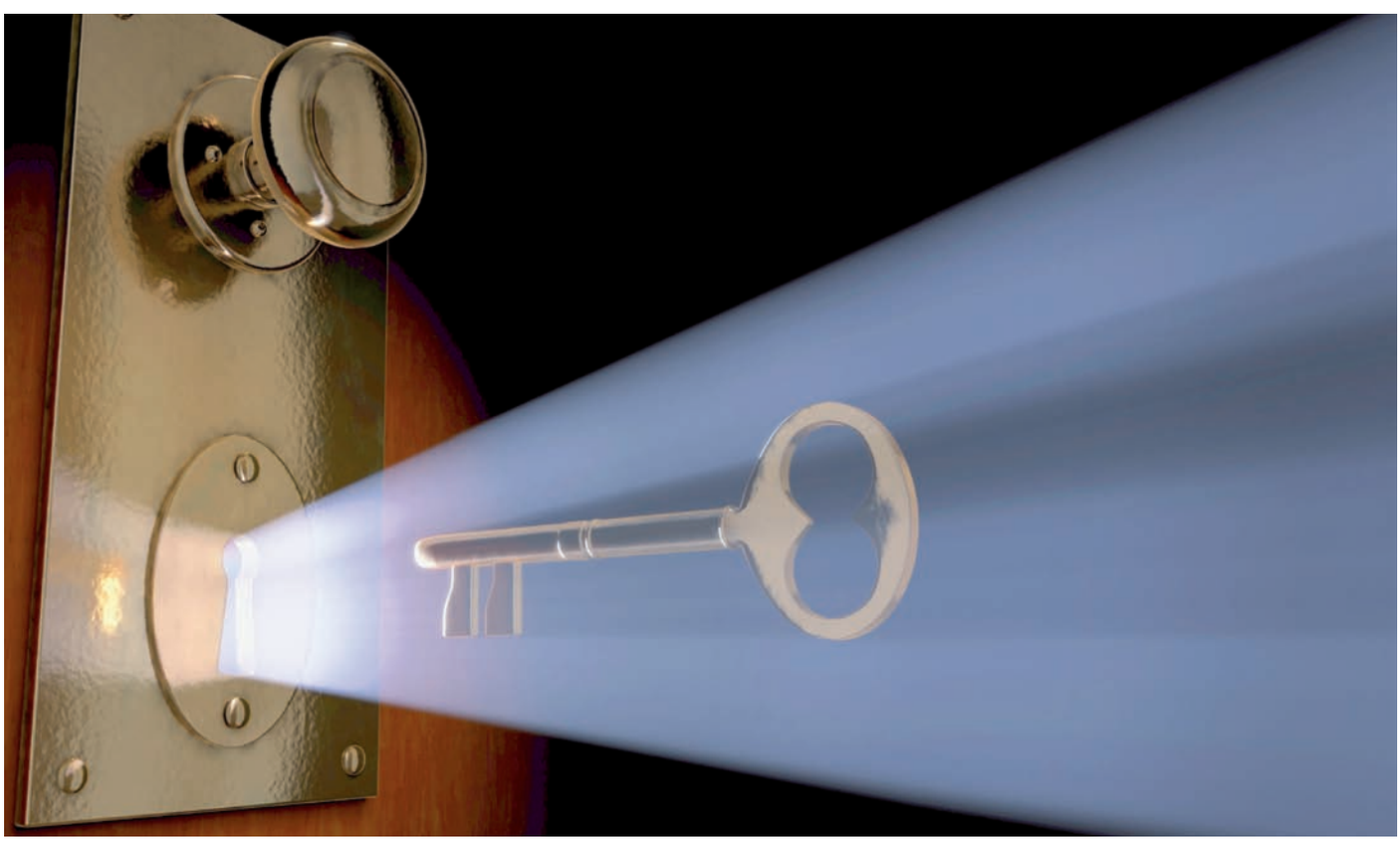
SAFEGUARDING ONLINE CONTENT

The tools behind the ACAP-Robots.txt debate



Don't be put off by the jargon or fooled into thinking the Robots.txt/ACAP (Automated Content Access Protocol) debate is another tiresome squabble amongst programmers. It's true that when you boil it down the whole issue revolves around nothing more than a line of instructions hidden away in a web page, but the implications go right to the heart of the publishing business and its fight for the future.

The publishers and organisations working to forge the ACAP standard will tell you that this is about nothing less than the right to preserve your own publishing lifeblood. Yet a voluble portion of the online community insists that it is just another example of knee-jerk conservatism from publishers failing to grasp the world of wikis and blogosphere.

Two 'simple' tools

Take away the oratory and name-calling and the actual proposal of ACAP is extremely simple. It's about providing increased sophistication ("granularity" in the jargon) for the instructions that set the rules for search engines when they encounter a publisher's content online.

Viewpoint: Media24's Elan Lohmann says ACAP unifies publishers, but the jury is still out on its impact. **Page 18**

Right now the tool that does that, a protocol called robots.txt, is a very simple and some would say blunt tool. It contains the instructions that say whether the content of a site can or cannot be searched. For example, if your robots.txt file contains the line, Disallow: /archive/, then a search engine "bot" (robot) arriving at the site will understand that it can index all the content it finds with the exception of data held in the directory called "archive."

That's about the limit of its subtlety. It can't say how searched content is used. It can't set a time limit for when content goes "out of date" and can no longer be searched. Nor can it do the inverse and set a timeline embargo before which that content can't be searched and shared. All of which would be of great interest to newspaper publishers who want the audience exposure offered by search engine indexing but without the all-or-nothing choice they currently have to make.

Which is where ACAP comes in. The brainchild of a joint initiative of the European Publishers Council, the World Association of Newspapers (WAN) and the International Publishers Association, it is the first draft of a protocol aimed at giving more control to content publishers.

ACAP builds on the role of robots.txt by adding a series of rules and restrictions that can be given to search engines indexing sites. The kind of restrictions being proposed will make instant sense to publishers – for example, instructing search bots precisely when content can be displayed on a third-party site, how long a text sample can be, or when to remove "time-stamped" content that's past its sell-by date.

It could insist that copy remains in the style and format that it was originally laid out in, and it could insist that any third-party use of content taken from the site is attributed, or back-linked so encouraging more visitors to the original. Because these rules can be assigned to any HTML element, they could apply to entire online editions, individual articles, selected paragraphs, or even just to a single line. Better still, ACAP doesn't replace robots.txt; it builds on it so that the old system can happily co-exist with the new for those who don't choose to implement it.

So what's the problem?

All of which sounds sophisticated, sensible and desirable. So why has the proposal met with mixed reaction from online forums and search engines themselves?

Leaving aside a vocal minority who appear to believe that a) all content online must be free, and b) traditional publishers are power-obsessed dinosaurs, there are a number of reasonable objections being made to ACAP as it stands.

Many of these are the kind or relatively small details to be expected of a version 1.0 release of a new protocol and can be expected to be addressed in subsequent developments. For example, the fact that the time specifier – the feature that enables publishers to say how long and when content can be indexed – works in units of days. Some developers consider that to be crude since it doesn't allow for hours/minutes or time zones that are normally accommodated on the web. Another point is that the "maximum length" specification seems to have been designed with text in mind – but ACAP is also intended to cover the indexing of images as thumbnails so a "maximum size" specification would seem to be a sensible addition.

There's also the issue that implementing ACAP is fairly straightforward for publishers who can go ahead and add the code, the proof of which is given by The Times in the U.K., which has already done so with Times Online. The announcement that publishing solutions provider Atex, and certainly others to come, has decided to support ACAP in all future releases of its CMS products should also help establish ACAP as the publishers' standard.

The problem, however, is not the publishers. Nor is it the technical spec. Nor even the rights and wrongs of the approach. The problem is that all of the work on ACAP is for nothing if the search engines don't agree to climb on board and com-

In a nutshell

What does ACAP do?
It gives a search engine instructions on how it can use the content it finds. That might mean date embargos, limits on length, or an instruction to delete material that is "past its sell by date."

How does it work with robots.txt?
It functions alongside robots.txt and adds more sophistication (currently robots.txt only understands "search" or "don't search" as instructions).

What is at stake for publishers?
The chance to control how search engines search, sample, and reuse their content.

Where do search engines stand?
The key ones remain standing back with their arms folded. The question is not whether they can work with ACAP, but more "what's in it for them?"

Francisco Pinto Balsemão, chairman, European Publishers Council; CEO Impresa, Portugal

"I am proud to say that Impresa was one of the first publishers to implement ACAP. ACAP is exactly what has been missing – a useful and more complete tool for publishers who have been working to develop their digital business models. ... Thanks to ACAP, the future of the newspaper and magazine industry will be more secure. We can be sure that our content will be viewed and distributed according to previously agreed terms and conditions."

Bharat Krishna, principal scientist and founder, Google News

"Search engines want to respect publishers' wishes – after all, it is their content. But we aren't mind readers, so it's vital that webmasters tell us how they want their content indexed. This can be done via the Robots Exclusion Protocol (REP), a well-established technical specification that tells search engines which site or parts of a site should not be searchable, and which parts should remain visible in the search results. Using these various technical standards, webmasters can easily give permission to a search engine to crawl and index their content. ... The major 'asks' of ACAP can be met by REP extensions and in fact many already have."

Timothy Balding, CEO, WAN

"How does the ability to express permissions in machine-readable form affect newspapers publishers? Well, this is a bit like asking "how does copyright affect newspaper publishers?" ... On the network, problems of scale make human-mediated permissions impracticable. ACAP offers mechanisms for the expression of permissions in machine-readable form which are particularly well suited to business-to-business relationships in the information supply chain. ACAP does not seek to dictate any particular business model or relationship in the supply chain. Rather it seeks to enable innovative business models for publishers and intermediaries, to the ultimate benefit of all users of information on the network."

ply, and that is the glaring weak spot in the ACAP proposal to date.

Despite some rather vague promises that the major search engines are informally involved with the initiative, it remains the fact that only one relatively minor player (Exalead) has put its name to it. ACAP's proponents point to Exalead as the fourth largest search engine, but unless you live in France the chances are you've never heard of it.

Besides, in this particular race there is no medal for fourth place. Some would argue that with just 12 percent of the market even MSN (Microsoft Network) in third place is a minor player. Realistically, there are only two players that matter because between them they control around 70 percent of the search market and to date Google and Yahoo! appear to be keeping their distance.

There are a few possible reasons for this. On the search engine side, the new permissions metadata could potentially create more work for search engines since ACAP instructions could be inserted into any HTML element, thus requiring more effort to handle.

Plus there is the perception that publishers are dictating to search engines, which are in turn likely to reply by ignoring the new protocol altogether. But realistically, the reluctance of search engines to step forward can be put down to the simplest of all questions; what's in it for them?

ACAP, like robots.txt, is a protocol, not a law. The ageing robots.txt model worked by mutual consent – bots encountering a "no index" message desist from searching that content out of little more than a gentlemen's agreement and unless both sides agree to be bound by it then it becomes redundant. If the search engines don't comply, then the ACAP code is immediately redundant, dead in the water. ACAP promises that it is in the interest of engines and users to implement the protocol because the renewed confidence it gives to publishers over rights would make them more confident about releasing more content to the web. It's a good argument but a hypothetical one if the major search engines don't decide to play ball.

For the moment, that argument hasn't proved compelling enough for Yahoo! and Google. The relative weakness of the appeal to search engines may have contributed to nervous speculation on blogs that ACAP is actually intended not as a protocol but as a first step towards seeking legal compunction to recognise publishers' rights online. This line of thought is probably inspired and fuelled by the successful suing of Google by Belgian publishers last year. In any case, arguments over the moral rights involved are largely irrelevant if the top engines remain unconvinced and it is that, rather than any issues of coding or implementation, that remains the massive hurdle to ACAP's success.

Steve Shipside and Sashi Nair (reader@ifra.com).