NOPREVIEW - The missing X-Robots-Tag

Posted on 7 August, 2007

Google provides previews of non-HTML resources listed on their SERPs:

These "view as text" and "view as HTML" links are pretty useful when you for example want to scan a PDF document before you clutter your machine's RAM with 30 megs of useless digital rights management (aka Adobe Reader). You can view contents even when the corresponding application is not installed, Google's transformed previews should not stuff your maiden box with unwanted malware, etcetera. However, under some circumstances it would make sound sense to have a NOPREVIEW X-Robots-Tag, but unfortunately Google forgot to introduce it yet.

Google is rightfully proud of their capability to transform various file formats to readable HTML or plain text: Adobe Portable Document Format (pdf), Adobe PostScript (ps), Lotus 1-2-3 (wk1, wk2, wk3, wk4, wk5, wki, wks, wku), Lotus WordPro (lwp), MacWrite (mw), Microsoft Excel (xls), Microsoft PowerPoint (ppt), Microsoft Word (doc), Microsoft Works (wks, wps, wdb), Microsoft Write (wri), Rich Text Format (rtf), Shockwave Flash (swf), of course Text (ans, txt) plus a couple of "unrecognized" file types like XML. New formats are added from time to time.

According to Adam Lasnik currently there is no way for Webmasters to tell Google not to include the "View as HTML" option. You can try to fool Google's converters by messing up the non-HTML resource in a way that a sane parser can't interpret it. Actually, when you search a few minutes you'll find e.g. PDF files without the preview links on Google's SERPs. I wouldn't consider this attempt a bullet-proof nor future-proof tactic though, because Google is pretty intent on improving their conversion/interpretation process.

I like the previews not only because sometimes they allow me to read documents behind a login screen. That's a loophole Google should close as soon as possible. When for example PDF documents or Excel sheets are crawlable but not viewable for searchers (at least not with the second click) that's plain annoying both for the site as well as for the search engine user.

With HTML documents the Webmaster can apply a NOARCHIVE crawler directive to prevent non paying visitors from lurking via Google's cached page copies. Thanks to the newish REP header tags one can do that with non-HTML resources too, but neither NOARCHIVE nor NOSNIPPET etch away the "view-as HTML" link.

<speculation>Is the lack of a NOPREVIEW crawler directive just an oversight, or is it stuck in the pipeline because Google is working on supplemental components and concepts? Google's yet inconsistent handling of subscription content comes to mind as an ideal playground for such a robots directive in combination with a policy change.</speculation>

Anyways, there is a need for a NOPREVIEW robots tag, so why not implement it now? Thanks in advance.

URL: http://sebastians-pamphlets.com/nopreview-the-missing-x-robots-tag/