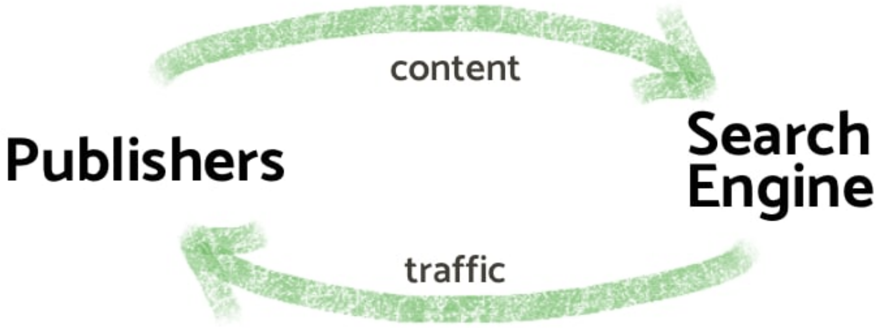
The Context of Scraping Control

While interest in scraping controls have mushroomed with mainstream concern about generative AI and LLMs, it is important to understand that the problem domain is not new. Should a technical solution be designed, it would be unfortunate if it were to solve only a narrow set of concerns that are prevalent today and fail to address pre-existing and, presumable, future issues that are structurally similar. With that in mind, this position paper offers a description of the wider context in which AI-related scraping emerged in the hope of helping inform the discussion. Disclaimer: this paper in no way claims to represent the opinions of the *The New York Times* but it is heavily shaped by my experience there and on solutions I considered while there.

Note that across this paper I use "search" to capture scraping that is used to generate a list of linked results interchangeably with more recent trends for instance in GenAI. I do so because I think it is a mistake to treat these separately.

The Drift of Search Scraping

The *historical* relationship between publishers and search engines was a simple and rather fortuitous mutualistic affair: publishers made their content freely available for search engines to index, which benefits the latter by making them relevant, and in turn the search engine would *only* use that content to index it and make it searchable by its users. That benefitted publishers by driving more traffic to them.

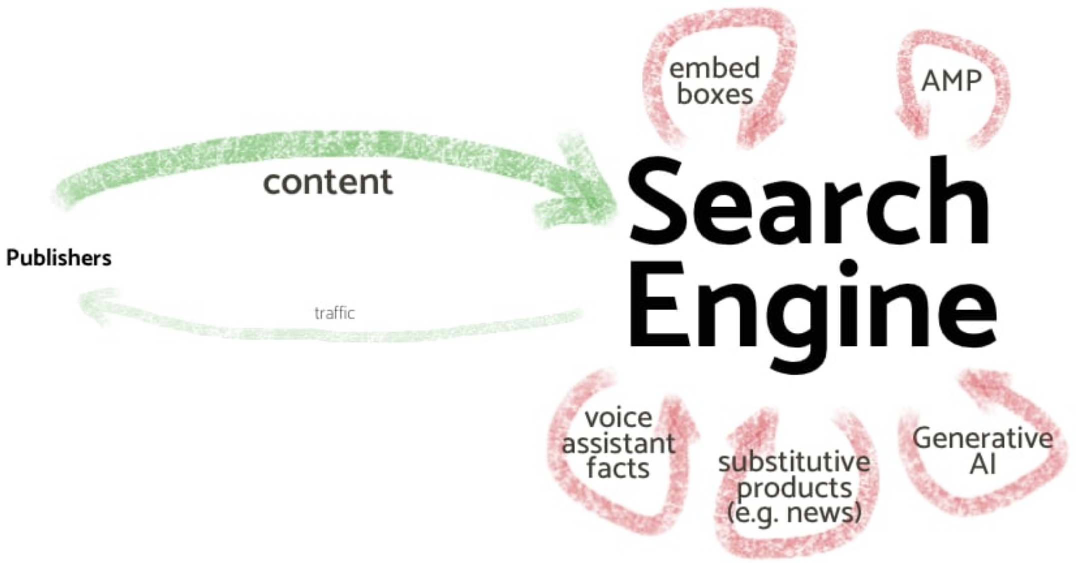


Some of the less savvy publishers complained, but overall this was a good arrangement. This arrangement developed organically in the internet community without deliberate deal-making or the creation of standardised interfaces beyond `robots.txt`.

For this kind of mutualistic symbiosis to persist over time, there needs to be a quick feedback mechanism to punish a defecting party trying to get more out of the relationship than it puts in:

- For publishers, that mechanism is simple: if they don't provide content, they don't get traffic. The incentive to cooperate is strong.
- For search engines, the control mechanism is competition. So long as there's competition between search engines, if one of them starts being extractive with respect to publishers then publishers can simply exclude the engine using `robots.txt`. Sure, they'll lose a little traffic but the search engine suffers from less relevant results (especially as it's likely that many publishers would make the same decision), losing users.

That model held for a while, but as search became increasingly monopolised it started failing. When traffic from a single search engine is your livelihood, you no longer get a say in the relationship: you just do as you're told, no matter how much it hurts you. As Google came to dominate, it started simultaneously doing more with the content than just indexing it, demanding more of publishers, and sending less traffic. In the absence of corrective mechanisms, the relationship shifted from mutualistic to parasitic.



When you take more out of a resource than you put back in, that resource suffers. The effect is not immediate (and therefore doesn't correct behaviour fast enough) but the resource will fail to regenerate, and over time it will become damaged and depleted. That is the sense in which search has become extractive and the reason why the manner in which online content is processed is not sustainable. It's death by a trillion hits: every time search terminates at the search engine (or sooner, as when a voice assistant or LLM replies with a fact extracted from a publisher) instead of going to the publisher's property that's a shaving of a penny less that goes into supporting publishing. Individually, that's imperceptible, but there's on the order of a trillion (10^{12}) searches every year: it adds up.

There is no mechanism ensuring that the publisher $\Longleftrightarrow$ search engine relationship remains mutualistic. In an ideal world, restoring competition would be the answer but that is likely to prove challenging. Browser vendors have a decisive influence over which search engine people use. They get paid through an affiliate marketing scheme that sets up a revenue sharing agreement between search and browsers. While their *intent* may not be monopolistic, they benefit from greater revenue when a search engine is in a position to charge supracompetitive prices. The incentives to voluntarily fix the problem aren't there.

The first take-away from this is that **the problems with AI scraping did not appear ex nihilo, rather they are the continuation of a logic that saw search continuously drifts towards extracting more unreciprocated value and constrained labour from publishers over the past decade.** Previous steps in this direction have included voice assistant responding with facts extracted online, embedded response boxes that provide an answer without a link, AMP, and news aggregators.

A second take-away is that **there is no reason for that drift to stop of its own accord and room for the situation to get worse than it already is with LLMs.** As evidence of what is coming next, 404 Media recently reported that Reddit is now only indexed by Google (presumably based on a direct deal they made), with Reddit's `robots.txt` now `Disallow`ing *all* user agents. A world in which dominant companies pay to secure exclusive access to content (and exclude competition) is here. This is detrimental to innovation in search and AI. If the indexed space fragments, it is also detrimental to users. If the space does *not* fragment because only one dominant player has the means to pay for all big publishers (as is the case with browsers today), then that situation will also be bad for publishers (since they are unlikely to obtain competitive prices in a monopsony).

Solution Space

Scraping controls are unlikely to be able to solve the entire problem of finding an appropriate balance of power between search and publishers, but they should be designed in the context of thinking about a wider solution.

The political dynamics of the situation is simple: we can have a system that works for indices, publishers (understood broadly), and people if and only if it is designed to be *mutualistic* between those who create content and those who index it. That is also the only arrangement in which we can expect to continue to see content published in the open on the internet: the alternatives are trivially unsustainable.

Technology alone will not solve this problem, but we can structure technical solutions so that they work well with (and encourage) good regulatory solutions.

A key component of search power (as weaponised primarily by Google) is tying indexing and/or ranking to the acceptance of search practices that are not mutually beneficial. Classic examples include AMP where the top of the results page is reserved for publishers who do extra free labour and forgo traffic or schema.org where better ranking is given to those who provide structured data that can be reused elsewhere (in a way that benefits neither users nor publishers). **It is therefore useful to create the technical means for publishers to distinguish distinct *purposes* for which their content can be indexed.**

A publisher should therefore be able to convey that they allow scraping for the purpose of indexing and searching used in such ways that they generate traffic, but refuse other purposes such as facts extraction or GenAI. Regulation can then be used to enforce these purposes, to require that purposes be propagated by corpora (e.g. Common Crawl would include allowable purposes per resource), to prevent exclusionary dealing, to specify which purposes are always allowable (for instance for research purposes), to specify if publisher deconsenting is retroactive, and to clarify which purposes might be opt-in only.

Robin Berjon — robin@berjon.com