Digital Platforms’ Crawling and Use of Publisher Content

December 2023
Crawling & Indexing: Core Concepts

Crawling
Software programs known as crawlers or spiders systematically explore and copy the content from web pages across the internet, following links from page to page to discover and retrieve content.

Indexing
The process of organizing the vast amount of content collected during crawling, categorizing and cataloging web pages' textual, visual, and multimedia elements, so search engines (and other technologies) can quickly retrieve and utilize the information.

Googlebot
Google's web crawler, a sophisticated computer program that follows links on the web, visits web pages, and copies their content for inclusion in Google's search index (and for use in other Google products).
Robots.txt: How Publishers Control Access

**Robots.txt**
A text file placed in a website's root directory that provides instructions to web crawlers, specifying which areas of the site they are allowed or disallowed to crawl and index. It helps website owners control how search engines access and interact with their site's content.

The example to the right gives access to Googlebot but blocks all other crawlers

**Important:** Robots.txt lets publishers control *whether* their content is scraped, not *how* their content is *used*
Publishers Lack Control Over Use of Crawled/Indexed Content

Competitive Market

Publisher Webpage → Crawler → News Universals, Search Results, GAI Training Data

Crawler with Monopoly Power

Publisher Webpage → Crawler → News Universals, Search Results, GAI Training Data
Google’s **Top Stories** box provides a carousel of links to news articles, which often include high-res images and snippets scraped from the publisher.

Clicking the links does take the user to publisher’s website, but when that publisher monetizes through advertising using Google’s dominant ad tech platform, Google extracts up to **30-70%** of the ad revenue for itself.
Google has introduced a **People also ask** section, which displays highlighted excerpts from news articles and other webpages to answer specific questions, obviating the need for users to leave Google’s search results page.
Google’s **About** and **Quick facts** sections are similarly designed to keep users on the search results page.
Even Google’s **Organic** search results again include lengthy snippets and images, discouraging users from clicking through to publishers’ sites.