

Scraping Reading pages from a site with a program and keeping some of what is on them.

Crawling Following links from page to page the way a search engine does; scraping is what you keep.

Bot Any program that fetches pages without a person clicking.

User agent The name a program gives itself in every request; how a site tells you from a browser.

Rate limit A cap on requests per period; over it the site slows or refuses you.

WAF A firewall in front of a site that blocks requests it judges hostile, bots included.

CDN Servers around the world that serve a site near the visitor; the machine you reach is rarely the operator's.

Bot challenge A test in front of a page to decide whether you are a person; a court has called one a lock.

CAPTCHA A challenge a program is not meant to pass; a careful crawler stops at one.

Login wall Pages that show only after signing in; behind one you are inside the terms.

Terms of service The site's contract with its visitors, whether or not anyone read it.

Browsewrap Terms at a link nobody clicked; they bind only with notice.

Clickwrap Terms you agreed to by ticking a box or making an account; these bind.

Cease-and-desist letter A letter telling you to stop; going on breaks the terms and, in many places, is unauthorised access.

Text and data mining (TDM) opt-out A rightsholder's machine-readable no; in the EU it switches the mining exception off.

Text and data mining Letting a program read works at scale, including to train a model.

Fair use and fair dealing The US doctrine, and its narrower cousins elsewhere, that allow some copying without permission.

Database right A right in the work of compiling a database, in the EU and a few other places, whatever the copyright in its contents.

Technological protection measure A lock on a copyrighted work; defeating it is its own offence.

Authorized access The hinge of every computer-misuse statute, read differently in each.

Personal data Any information about an identifiable person, whether or not they published it.

Data subject The person the personal data is about.

Controller The organisation that decides why and how personal data is processed; a scraper that keeps names is one.

Legitimate interest The basis a scraper usually relies on for personal data, and it comes with a three-part test.

robots.txt A file asking crawlers what not to fetch; evidence everywhere, an opt-out in the EU, a law nowhere.

BODY OF LAW	REACHES YOU THROUGH	THE QUESTION IT ASKS YOU	THE DOCUMENT
Computer misuse and unauthorised access	where the machine you reached belongs	Was something in the way, and did you get around it?	read
Terms of use	wherever the terms say, when they bind you	Did you agree to anything, and did you know the terms were there?	read
Copyright and text-and-data mining	the rightsholder's place, and where the copy is made	Are you copying, and does an exception cover the use?	read
Anti-circumvention	as copyright	Did you defeat a lock on a copyrighted work?	read
Database right	the database maker's place, in the EU and a few others	Did you take a substantial part of a compiled database?	read
Privacy	where the people are, and where you are established	Is there a person in the data, and what is your basis?	read
Unfair competition and hot news	where the competitor is and where the harm lands	Are you competing with the site using its own work?	read
AI law	where the model is placed on the market, and where its output is used	Is a model reading or training on the pages, and whose?	read

The six situations the corpus rates, from open to closed

1. **public unauthenticated.** Fetching pages anyone can view without logging in, the normal open web.
2. **robots disallowed.** Pages the site's robots.txt asks crawlers not to fetch.
3. **behind login.** Content reachable only after signing in to an account.
4. **tos accepted.** Crawling after you have agreed to the site's Terms of Service (e.g. by creating an account).
5. **after technical circumvention.** Getting in by defeating a technical barrier, CAPTCHA, IP block, or rate limit.
6. **after cease and desist.** Continuing to crawl after the site has formally told you to stop.