COMPLIANCESME

Response to the public consultation on Guidelines 03/2026 on web scraping in the context of generative AI · 11 October 2026 · compliancesme.com

Reading a page to answer a question is a different act from copying it into a training corpus. The guidelines should let a publisher welcome the first and refuse the second.

To the European Data Protection Board,

ComplianceSME publishes compliance documentation for small and medium companies trading across the European Union and the United Kingdom, and the Global Fuel Report, a weekly report on the state of the world's fuel supply. We write as the owner of a website that gets scraped. The draft guidelines speak to the companies doing the scraping. Nobody in them speaks for us.

Our position is simple and we have built our website around it. AI search tools are welcome to read our pages, quote them with attribution, and send the reader to us. Nobody is welcome to copy our content into a training dataset. We publish machine-readable signals saying exactly that: our robots file admits named AI crawlers, our terms reserve our text and data mining rights under Article 4(3) of Directive (EU) 2019/790, and a reservation file at a standard address states that training use is refused.

The draft gives that position no legal effect. We ask the Board for six changes.

1. **Permission for one purpose is refusal of the others.** Paragraphs 37 and 64 and Examples 4 and 5 treat a website as either protected by technical barriers or open to everything. Example 4 even reads openness as raising the expectation that pages will be used "to develop AI models". So a publisher who lets AI search in strengthens the legal case of the company harvesting its pages for training. Please state plainly: a machine-readable signal that permits one purpose refuses the purposes it does not name.

2. **A clear written refusal must count on its own.** Paragraph 37 recognises opposition to scraping only through technical measures such as logins and CAPTCHAs. Those punish every human reader, and a small publisher should not have to wall its site to earn legal recognition of a refusal it has already published where every scraper can read it. Please state that a machine-readable refusal counts without a barrier behind it, and that a scraper who reads a clear refusal and carries on anyway has that held against it in the balancing test under paragraph 53.

3. **Name the signals.** The draft cites "ai.txt files" twice without defining them. It does not say which signals a scraper must read, in what order they rank, or how often they must be re-checked. A publisher cannot state a refusal reliably when nobody has said where a refusal lives. Please name the minimum set of signals, state that a reservation under Article 4(3) weighs in the balancing test, and set a rule for how fresh a scraper's reading of the signals must be.

4. **Crawlers must say who they are.** A website can only treat a search crawler differently from a training harvester if crawlers identify themselves at the moment of the request. The draft asks for disclosure only afterwards, in the controller's public notice under paragraph 30. Please state that operating an unidentified or rotating crawler weighs against the scraper. And please correct paragraph 66(c): as

written, a site that serves a declared crawler a purpose-appropriate view of its own pages is treated the same as a site that deceives. The deceiving site and the honest publisher serving a declared crawler are not the same thing.

5. **If the content can be licensed, scraping it is not necessary.** Paragraph 51 asks whether a less intrusive way of pursuing the purpose exists, and paragraph 52 lists some. It misses the plainest one: where the publisher offers the same content under licence, taking the licence is equally effective, intrudes on nobody whose data sits in those pages, and keeps the source accountable for accuracy. Please name licensed access from the source as an alternative the necessity test must consider.

6. **If you send scrapers to us, make them honour our terms.** Paragraph 42 tells scrapers to prefer sources that are "official, trustworthy, and maintained" for the sake of accuracy. We are such a source. The draft asks nothing of the scraper in return: no duty to honour the source's stated conditions, no attribution, no seat for the source in the balancing test. Please pair the recommendation with that expectation.

One distinction runs through all six points: reading a page to answer a question and sending the reader to the source is a different act from copying that page into a training corpus. A publisher should be able to welcome the first and refuse the second, in signals your guidelines recognise. We would welcome the chance to expand on any of these points.

Adam White
ComplianceSME
compliancesme.com
adam@compliancesme.com