

Crawler AI: registro tecnico di accesso

Configurazione. Da stampare e compilare a mano prima delle prove.

Host e protocollo: _____ Referente: _____

Periodo / fuso orario: _____ Ambiente: _____

Piano proposto: 4 URL pubblici x 4 profili di prova x 2 passaggi = 32 tentativi GET. Esegui una richiesta alla volta e annota ogni orario. Robots.txt e sitemap si controllano a parte. I redirect aggiungono richieste di rete. Con 429 sospendi le prove e verifica i limiti.

Decisioni per funzione. Scrivi consenti / escludi / da decidere. Robots.txt non protegge dati riservati: per questi serve l'autenticazione.

Sistema / funzione	Decisione e perimetro	Responsabile
Googlebot / Ricerca		
Google-Extended / Gemini e fonti		
OAI-SearchBot / ricerca		
GPTBot / addestramento		
ChatGPT-User / richiesta utente		
PerplexityBot / ricerca		
Perplexity-User / richiesta utente		

URL del campione: quattro pagine diverse, pubbliche e autorizzate. Scegli una frase da cercare nell'HTML per riconoscere il contenuto atteso. Per U4, se non hai pagine dipendenti da JavaScript, scegli una struttura diversa dalle altre.

U1 (servizio): _____

Frase: _____

U2 (guida): _____

Frase: _____

U3 (aggiornata di recente): _____

Frase: _____

U4 (JavaScript / altra struttura): _____

Frase: _____

Versione robots.txt / sitemap salvate: _____ Data / ora: _____

Crawler AI: registro tecnico di accesso

Una scheda per URL. Stampa 4 copie di questa pagina per registrare 32 tentativi.

URL U____: _____

Data / fuso: _____ IP della postazione: _____

Profili di prova: CTRL = AuditCrawler/1.0; GOOG = Googlebot/2.1 (+http://www.google.com/bot.html); OAI = OAI-SearchBot/1.4; PPLX = PerplexityBot/1.0. Sono stringhe di prova: non identificano bot autentici. R1 è il primo passaggio, R2 il secondo.

Esito: OK = testo atteso consegnato; KO = blocco o contenuto diverso; NV = prova mancante o non conclusiva. Scrivi l'ora per riga. Salva header e HTML con lo stesso ID, per esempio U1-OAI-R2. Per un timeout annota il codice di uscita di curl.

Profilo	Ora	HTTP / uscita curl	OK KO NV	Frase sì/no	File / ID / note
CTRL-R1					
GOOG-R1					
OAI-R1					
PPLX-R1					
CTRL-R2					
GOOG-R2					
OAI-R2					
PPLX-R2					

Controllo completo: URL finale e redirect, Content-Type, X-Robots-Tag, meta robots, canonical, frase essenziale e link. Anche una pagina di verifica può rispondere con 200. Curl non applica robots.txt: leggi prima le regole. Il solo user-agent di prova non giustifica un'eccezione al firewall (WAF).

Totale per questa scheda: OK ____ + KO ____ + NV ____ = 8.

Crawler AI: registro tecnico di accesso

Richieste autentiche, diagnosi e correzioni. Una copia per il controllo iniziale.

Totale 4 schede: OK ____ + KO ____ + NV ____ = 32 tentativi programmati.

Esempio: supponi di avere 28 OK e 4 KO. Sulla guida U2, OAI riceve 403 in entrambi i passaggi; su U4, PPLX riceve due volte 200 con una verifica anti-bot. Il testo atteso è arrivato in 28 prove su 32 (87,5%): questa quota non misura la visibilità del sito.

Richieste autentiche. Per ciascuna allega i log del firewall o della CDN con data, ora, fuso, IP del visitatore, user-agent, URL, stato HTTP, byte e regola applicata. Chiedi al gestore quale campo riporti l'IP verificato dall'infrastruttura. Collega i log del server tramite ID, se condiviso; altrimenti confronta URL, orario e IP e annota i limiti del riscontro.

Provider / URL / ID della richiesta: _____

Data / ora / fuso: _____ IP del visitatore: _____

Campo del log usato per l'IP: _____

Elenco IP ufficiale o verifica DNS / data / allegato: _____

Identità: verificata / non verificata Stato HTTP / byte: _____

Regola del firewall / riscontro nel server / allegati: _____

Se i log mancano, segna NV per la verifica dell'identità e chiedi al gestore dell'hosting le richieste relative agli URL e agli orari annotati. Se non osservi visite, non puoi concludere che ci sia un blocco.

Intervento: problema e ID delle prove _____

Responsabile: _____ Regola precedente / ripristino: _____

Correzione / data: _____

Nuova prova: ID / data / risultato / allegati _____

Controlli sugli altri profili: _____

Accesso: confermato / incompleto Indicizzazione: _____ Prossimo controllo: _____