Läromedelsförfattarna

SKYDD MOT HÄMTNING AV INNEHÅLL PÅ WEBBPLATS - GUIDE

Läromedelsförfattarna

HUR DU UNDVIKER OTILLBÖRLIGT INHÄMTANDE AV INFORMATION FRÅN DIN WEBBPLATS

Idag finns det ett antal olika sätt att skydda sig från att innehåll otillbörligen hämtas från er webbplats via så kallade ”webcrawlers” (även kända som robotar eller spindlar).
Det första och kanske lättaste du kan göra är att göra det genom att skapa och implementera en ”robot.txt”-fil. Det är en enkel textfil som placeras i rotkatalogen på en webbserver. Dess syfte är att ge instruktioner till webbaserade robotar (som till exempel sökmotorers eller AI-modellers webcrawlers) om vilka delar av webbplatsen som får eller inte får besökas och indexeras. Denna standard kallas för "Robots Exclusion Protocol". Tyvärr är detta endast en form av överenskommelse och inte reglerat av en lag eller fungerar som ett ”hårt” tekniskt skydd.

I detta dokument kan du läsa och förhoppningsvis förstå mer om Robot.txt-filen. Eftersom det inte finns någon lag om webcrawlers så finns det i praktiken ett antal andra skydd eller försvarsmekanismer att implementera för att strikt tekniskt skydda din information. Tyvärr har de flesta av dessa någon form av baksida i form av ökad komplexitet eller minskad användarupplevelse för faktiska användare som du vill ska ta del av innehållet. Exempel på dessa kan vara att lägga informationen bakom en inloggningssida, ta hjälp av en så kallad ”CAPTCHA”-lösning, etablera IP-begränsningar, nyttja dynamisk rendering av innehåll, anpassa http-headers eller andra direktåtkomstbegränsningar.

Utöver dessa finns det även ett antal lagar i Sverige som kan vara relevanta att hänvisa till, även om dessa i sig inte direkt skyddar ditt innehåll för indexering. De vanligaste lagarna att hänvisa till kan vara Dataskyddsförordningen (GDPR), Upphovsrättslagen och Marknadsföringslagen.

Nedan följer en mer detaljerad redovisning om de olika alternativen.

ROBOTS.TXT

robots.txt är en textfil som används av webbplatser för att ge instruktioner om hur olika webcrawlers ska interagera med sidorna på webbplatsen. Denna fil är en del av Robot Exclusion Standard, en konvention som används av webbplatsägare för att styra robotarnas åtkomst till deras webbplatser.

Syfte

Syftet med protokollet är ursprungligen framtaget för att bland annat undvika överbelastning på servrar genom att ge webbplatsägare möjligheten att informera robotar om vilka delar av deras webbplats som inte bör besökas.
Detta har med tiden utvecklats till att fungera som en överenskommelse kring vilken information du som informationsägare vill dela med dig vid olika former av indexering, inklusive användas för träning av AI-modeller.

Hur skapar du denna fil?

Filtyp - Det är en vanlig .txt-fil, ofta kallat textfil, som är en av de enklaste och mest grundläggande filtyperna som finns. Namnet kommer från dess filändelse, ”`.txt`”, som är en standard för textfiler. Det har en universell kompabilitet och kan öppnas och skrivas/redigeras med nästan vilken textredigerare som helst på nästan vilket operativsystem som helst, inklusive Windows, Mac OS, Linux och även på mobila operativsystem
Format - Filen består av regler som definierar åtkomst för olika användaragenter (robotar). Varje regel består av en **User-agent**-rad följt av en eller flera **Disallow** eller **Allow** rader.
User-agent - Specificerar vilken robotregeln gäller. En asterisk (*) används för att representera alla robotar.
Disallow - Ange vilka URL-vägar som är blockerade för den angivna roboten.
Allow - (valfritt) Används för att uttryckligen tillåta åtkomst till delar av webbplatsen, även om en övergripande **Disallow** regel finns.
Crawl-Delay - (valfritt) Ange en fördröjning mellan förfrågningar och servern från en viss robot.

Exempel 1

I detta exempel blockeras åtkomsten till **/privat/** för alla robotar, medan åtkomst till **/offentlig/** tillåts:

*User-agent: **
Disallow: /privat/
Allow: /offentlig/

Exempel 2

I detta exempel blockeras åtkomsten för en specifik user-agent, närmare bestämt OpenAIs GPTBot. Mer information hittar du här: https://platform.openai.com/docs/gptbot

User-agent: GPTBot
Disallow: /

User-agent: ChatGPT-User
Disallow: /

BEGRÄNSNINGAR OCH FÖRBEHÅLL

- Avsaknad av rättslig kraft: robots.txt är en överenskommelse, inte en lag. ”Respektfulla robotar” följer filens instruktioner, men den kan tyvärr även ignoreras av skadlig programvara eller skrapningsverktyg.
- Det är inte att likställa med ett säkerhetsskydd. Därför bör den inte användas för att dölja känslig information eftersom filen enkelt kan ignoreras.

- Skilja på indexering och besökt. Medan robots.txt kan förhindra en robot från att besöka en sida, hindrar den inte sidan från att bli indexerad. Om sidan är länkad från andra platser på internet kan den fortfarande dyka upp i sökresultaten.

SAMMANFATTNING

Kort sammanfattat är användandet av robots.txt ett viktigt verktyg för webbplatsägare för att styra hur deras innehåll crawlas och indexeras av sökmotorer. Korrekt användning kan förbättra en webbplats SEO och laddningsprestanda, medan felaktig användning kan leda till oavsiktlig blockering av värdefullt innehåll eller exponering av känsliga sidor.

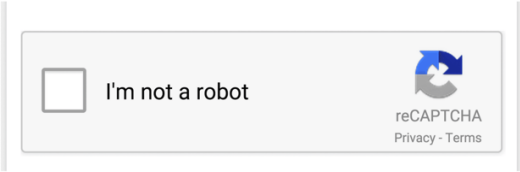
ANDRA MÖJLIGHETER ATT SKYDDA SITT INNEHÅLL

Inloggningsskydd

Avkräva inloggning för åtkomst till vissa delar av webbplatsen. Som det låter är det att lägga innehållet bakom någon form av inloggningsskydd. Det kan ju göras på en mängd olika sätt men är samtidigt mer komplext och kan upplevas som ett stort hinder för tillåtet användande.

CAPTCHA-system

CAPTCHA-system används för att säkerställa att användaren är en människa och inte en robot. Detta kan vara effektivt för att blockera automatiserade skript och robotar som försöker interagera med en webbplats. Exempel på en CAPTCHA kan ses nedan:



IP-Begränsningar

Blockera åtkomst från kända IP-adresser eller IP-intervall som tillhör kända webcrawlers eller botnät. Det finns sidor online som anger IP-adresser på kända webcrawlers, några exempel är: https://udger.com/resources/ip-list ,
https://www.azurespeed.com/Information/AzureIpRanges
Exempelvis, för OpenAI:s webbläsartillägg kommer anrop till webbplatser att göras från följande CIDR-block:
- 23.98.142.176/28
- 40.84.180.224/28
- 13.65.240.240/28

Notera att det kan tillkomma ytterligare adresser eller så kan de ändras med tiden så du får löpande ha kontroll på dessa.

Du kan även begränsa antalet förfrågningar en användare kan göra under en viss tidsperiod för att förhindra överbelastning av servern.

Dynamisk rendering av innehåll

Servera innehåll dynamiskt med hjälp av JavaScript. Vissa crawlers har svårt att tolka JavaScript, vilket kan göra det svårare att skrapa innehållet.

Anpassat innehåll för webcrawlers

Genom att servera enklare, statisk HTML till misstänkta webcrawlers kan du begränsa den information som de enkelt kan skrapa.

Detektion och blockering

Genom att identifiera beteendemönster kan du anpassa hur din webbplats svarar på misstänkt skrapningsaktivitet, eventuellt genom att servera modifierat eller begränsat innehåll.

Anpassa HTTP-headers

HTTP-headers är delar av HTTP-begäran och svar-meddelanden i nätverksprotokollet HTTP. De definierar parametrar för en transaktion. För att skydda sig mot webcrawlers kan du använda följande HTTP-headers nedan.

X-Robots-Tag

Detta är en HTTP-header som fungerar liknande **robots.txt**, men den tillåter finare kontroll eftersom den kan appliceras per sida eller resurs snarare än för hela webbplatsen. Till exempel kan du använda **X-Robots-Tag: noindex** för att förhindra sökmotorer från att indexera en specifik sida.

Strict-Transport-Security (HSTS)

Den här headern tvingar webbläsaren att endast använda säkra anslutningar (HTTPS). Även om detta inte direkt stoppar crawlers, ökar det säkerheten och kan förhindra vissa enklare skrapningsförsök som inte hanterar HTTPS väl.

Content-Security-Policy (CSP)

CSP används för att begränsa vilka resurser som får laddas på en webbsida, vilket kan hjälpa till att förhindra skadliga skript från att köras. Detta påverkar inte direkt webcrawlers, men det hjälper till att säkra din webbplats mot vissa attacker.

Direktåtkomstbegränsningar

Detta handlar om att konfigurera webbservern för att begränsa åtkomst till vissa delar av din webbplats. Vanliga metoder inkluderar bland annat:

Användning av .htaccess-fil (För Apache Web Server)

En **.htaccess**-fil är en konfigurationsfil för användning på webbservrar som kör Apache-programvaran. Den kan användas för att omdefiniera inställningar för mappar och undermappar på servern. Exempel på användningar är:

Begränsa åtkomst efter IP-adress

Du kan konfigurera **.htaccess** för att endast tillåta åtkomst till din webbplats från specifika IP-adresser eller att blockera vissa IP-adresser.

Omdirigera trafik

Du kan omdirigera användare eller bots beroende på deras IP-adress eller användaragentsträng.

Lösenordskydda kataloger

Du kan kräva användarnamn och lösenord för åtkomst till specifika kataloger på din webbplats.

Nginx-konfigurationsfiler

Liknande **.htaccess** för Apache, kan du använda konfigurationsfiler i Nginx för att definiera åtkomstregler, som IP-baserade begränsningar eller grundläggande autentisering.

ICKE-TEKNISKA SKYDD

Utöver olika former av tekniska skydd eller begränsningar bör du även vara tydlig med vad du som innehållsägare vill. Lite tips att använda kan vara:

Dataskyddsförordningen (GDPR)

Om webbskrapningen innebär insamling av personuppgifter, kan GDPR vara tillämplig. GDPR kräver bland annat att insamling av personuppgifter ska ske lagligt, rättvist och transparent.

Marknadsföringslagen

Om en webbcrawler används för att samla information för konkurrensmässiga fördelar, kan det potentiellt ses som en otillbörlig affärsmetod enligt marknadsföringslagen.

Upphovsrättslagen

Upphovsrättslagen kan komma till användning om innehållet som skrapas är skyddat av upphovsrätt.

Automatiskt skydd

Det viktiga att veta är att upphovsrätten är automatisk. Det betyder att ditt originalinnehåll (text, bilder, layout, design, kod etc.) på din webbplats automatiskt skyddas av upphovsrätten från det ögonblick det skapas. Det krävs ingen registrering eller officiell process för att upprätthålla detta skydd.

Vad skyddas?

Upphovsrätten skyddar uttrycket av en idé, inte själva idén. Till exempel, om du skriver en unik artikel eller designar en grafik, skyddas dessa specifika uttryck, men inte den generella idén eller konceptet bakom dem.

Gör ditt upphovsrättsskydd tydligt

Även om upphovsrätten är automatisk, är det bra att tydligt ange ditt upphovsrättsskydd på din webbplats. Här är några steg du kan ta:

Upphovsrättsmeddelande
Inkludera ett upphovsrättsmeddelande på din webbplats. Detta är ofta i formen
”© [År] [Ditt Namn eller Ditt Företags Namn]. Alla rättigheter förbehållna.”
Placera detta meddelande i sidfoten eller på en synlig plats på din webbplats. Har du mer omfattande texter eller verk tillgängliga rekommenderar vi dig att utveckla förbehållet enligt nedan:

”Författaren och ägaren till förekommande texter förbehåller sig alla rättigheter till text- och datautvinning i enlighet med Lagen (1960:729) om upphovsrätt till litterära och konstnärliga verk § 15a. Alla nyttjanden för upplärning av generativa AI-system är strikt förbjudet utan författarens uttryckliga tillåtelse.”

Användarvillkor
Skapa en sida med användarvillkor eller användaravtal där du specificerar hur ditt innehåll får och inte får användas. Detta kan inkludera detaljer om att materialet är skyddat enligt upphovsrättslagen och vad besökare har rätt att göra med ditt innehåll (till exempel läsa, ladda ner för personligt bruk, etc.).

Vattenmärken och Upphovsrättsmärkningar
För fotografier och grafik kan det vara effektivt att använda vattenmärken eller upphovsrättsmärkningar direkt på bilden.