WEB SCRAPING E DATA PROTECTION

17 dicembre 2024

AUTORI

Aurora Agostini
Partner

Giulietta Minucci
Counsel

Jessica Giussani
Associate

Giovanni Lombardi
Associate

Il fenomeno del *web scraping* - tecnica che consente la raccolta massiva e sistematica di dati accessibili online mediante l'utilizzo di software automatizzati - sta sollevando, nell'ultimo periodo, questioni di particolare rilevanza sotto il profilo della protezione dei dati personali: tale pratica, infatti, benché ampiamente utilizzata per alimentare modelli di intelligenza artificiale e per finalità di *business intelligence*, pone rischi concreti e significativi per la tutela della riservatezza dei dati personali, con particolare riguardo alle categorie particolari di dati di cui all'art. 9 del Regolamento Europeo 2016/679 ("GDPR").

Web scraping: nel mirino delle Autorità Garanti per la Privacy

Sedici Autorità per la Protezione dei Dati Personali provenienti da diverse parti del mondo, in collaborazione con alcune delle maggiori aziende di *social media* a livello globale, hanno lavorato insieme per raccogliere osservazioni e definire linee guida sulle pratiche e istruzioni riguardanti il *data scraping* e la tutela della *privacy*. Questo sforzo congiunto ha portato alla pubblicazione di una Dichiarazione Congiunta, che rappresenta la continuazione del percorso avviato nell'estate del 2023. Il documento, oltre a delineare principi generali, fornisce indicazioni operative e strumenti pratici specificamente calibrati sulle esigenze delle piccole e medie imprese (PMI), al fine di supportarle nell'implementazione di adeguate misure di prevenzione e contrasto all'estrazione massiva di dati.

Partendo dalla definizione di *web scraping*, l'articolo esamina la pronuncia del Garante Privacy italiano e delle autorità internazionali in materia di *web scraping*, fornendo indicazioni pratiche sia per chi utilizza tecniche di *scraping* sia per i gestori di siti web che intendono proteggere i propri contenuti da estrazioni non autorizzate.

LE LINEE GUIDA DEL GARANTE PRIVACY ITALIANO

Il Garante per la protezione dei dati personali Italiano ha recentemente pubblicato una nota informativa mediante la quale ha fornito puntuali indicazioni su come difendere i dati personali pubblicati *online* da soggetti pubblici e privati dalla c.d. tecnica del *web* scraping, ossia dall'estrazione sistematica di dati da siti *web* con l'ausilio di *bot.* Le indicazioni fornite dal Garante Privacy Italiano, seppur di natura non obbligatorie, rappresentano un utile strumento per i Titolari del Trattamento che vogliono tutelare le proprie informazioni, nonché per quegli operatori che utilizzano tecniche di *web scraping* come modello di *business* o come strumento per l'implementazione del proprio modello di *business.*

Nelle Linee Guida, il Garante definisce il *web scraping* come attività di raccolta, memorizzazione e conservazione massiva e indiscriminata di dati, inclusi quelli personali. Detta tecnica, come già chiarito in un precedente articolo, non è in linea di principio illegale, a patto che i dati oggetto delle attività di *scraping* siano liberamente accessibili sui siti e siano usati per scopi statistici o di monitoraggio dei contenuti. Considerando l'aumento delle pratiche di *web scraping* non autorizzate, il Garante fornisce una serie di misure idonee a contrastare, o quantomeno a mitigare, il *web scraping* non autorizzato e segnatamente:

- la creazione di aree riservate mediante le quali i dati sono accessibili solo previa registrazione, così da consentire di sottrarre i dati alla disponibilità indiscriminata, limitando o quantomeno riducendo le attività di *web scraping*:

- l'inserimento di clausole *ad hoc* nei termini d'uso del servizio al fine di vietare – quale misura preventiva di natura giuridica – l'uso di tecniche di *scraping*;

- il monitoraggio del traffico di rete, ed in particolare il monitoraggio delle richieste HTTP ricevute, al fine di identificare flussi di dati anomali ed implementando tecniche quali il "*rate limiting*";

- l'implementazione di limitazioni ai *bot*, mediante l'inserimento di verifiche CAPTCHA, modifiche periodiche del *markup* HTML, incorporazione dei contenuti in oggetti multimediali

Il provvedimento del Garante rappresenta senz'altro un passo importante per la protezione dei dati personali nel contesto del *web scraping* e conferma l'importanza di circoscrivere questo fenomeno con una particolare attenzione alla normativa privacy applicabile.

LE LINEE GUIDA DEL GARANTE PRIVACY OLANDESE

L'interesse per il *web scraping* e le implicazioni lato *privacy* non è di interesse solo per l'Autorità Italiana. Difatti, già a maggio 2021, l'Autorità Garante della Privacy Olandese ha preso una posizione sul *web scraping,* pubblicando delle linee guida per gli operatori. In particolare, il documento dell'Autorità Olandese esplora il fenomeno del *web scraping*, analizzando le implicazioni legali e i rischi per la privacy in conformità con il GDPR.

Il documento dell'Autorità olandese evidenzia come il *web scraping*, frequentemente impiegato per finalità commerciali o di sviluppo tecnologico, comporti necessariamente la raccolta di ingenti quantità di dati personali, ivi incluse categorie particolari di dati, con conseguenti potenziali pregiudizi per i diritti e le libertà fondamentali degli interessati. In particolare, l'Autorità sottolinea come, ai fini della conformità al GDPR, il trattamento dei dati raccolti mediante tecniche di *scraping* debba necessariamente trovare fondamento in un'idonea base giuridica - quale, ad esempio, il legittimo interesse del titolare del trattamento - la cui applicabilità deve essere oggetto di attenta valutazione mediante un test di bilanciamento con i diritti fondamentali dei soggetti interessati.

Di particolare rilievo risultano le indicazioni fornite dall'Autorità in merito all'implementazione del principio di trasparenza, mediante l'adozione di misure quali la predisposizione di informative *privacy* complete ed esaustive e la tempestiva comunicazione agli interessati. È cruciale limitare il trattamento dei dati trattati e rispettare il principio di minimizzazione che richiede che i dati personali trattati siano adeguati, pertinenti e limitati a quanto necessario rispetto alle finalità per cui vengono raccolti e utilizzati. Tra le altre misure suggerite, si annoverano la pseudonimizzazione, la rimozione immediata dei dati non necessari e il rispetto di *standard* tecnici come il protocollo robots.txt.

L'Autorità evidenzia altresì i limiti del *web scraping*, illustrando come, in molte situazioni, esso possa violare i principi del GDPR, soprattutto se utilizzato senza il consenso degli interessati o per scopi non giustificabili, come la profilazione non autorizzata o il monitoraggio delle attività *online*.

Per garantire la conformità normativa, è essenziale effettuare una valutazione preventiva dell'impatto sulla protezione dei dati (DPIA), specialmente in contesti di utilizzo su larga scala o in ambiti che implicano dati particolari.

DICHIARAZIONE CONGIUNTA SUL DATA SCRAPING

Come visto, il fenomeno del data scraping è indubbiamente una questione centrale nel dibattito sulla protezione dei dati personali e, sebbene questa pratica possa essere impiegata legittimamente in determinati contesti, il suo utilizzo senza autorizzazione può dar luogo a violazioni significative della normativa privacy. In tale conteso, le Autorità Garanti per la protezione dei dati di 16 Paesi hanno reso nota una nuova dichiarazione congiunta sul data scraping e la tutela della privacy, successiva alla dichiarazione congiunta del 2023, in cui veniva già sottolineata l'urgenza di proteggere le piattaforme online dall'estrazione non

autorizzata, invitando le imprese a implementare controlli robusti per bloccare attività illecite.

L'iniziativa si inserisce nel più ampio contesto di un crescente interesse normativo, a livello globale, volto alla regolamentazione e mitigazione dei rischi connessi all'estrazione automatizzata di dati personali dalle piattaforme digitali. La dichiarazione, elaborata con il contributo delle Autorità Garanti di diverse giurisdizioni (tra cui Canada, Regno Unito, Repubblica Popolare Cinese, Confederazione Elvetica e Regno di Norvegia) e con la partecipazione attiva dei principali operatori del settore dei social media (tra cui Meta, LinkedIn e X), fornisce indicazioni operative di particolare rilevanza in materia di tutela da iniziative di *web scraping* non autorizzate.

La dichiarazione, da un lato, evidenzia innanzitutto come le corporate debbano rispettare le normative sulla protezione dei dati personali quando utilizzano le informazioni estratte mediante tecniche di scraping per sviluppare modelli di intelligenza artificiale o per scopi commerciali e, dall'altro, suggerisce l'adozione di misure di sicurezza dinamiche e *multileval* al fine di prevenire scraping non autorizzato.

Tra le misure suggerite, vi è (i) l'uso di captcha, (ii) di sistemi di blocco degli indirizzi IP (come peraltro suggerito anche dal Garante Privacy Italiano), nonché (iii) l'utilizzo di strumenti basati sull'intelligenza artificiale per identificare e prevenire comportamenti anomali e (iv) la progettazione di piattaforme/siti web con meccanismi intrinseci che ostacolino l'estrazione automatica dei dati.

La dichiarazione congiunta rappresenta un importante passo avanti nella collaborazione internazionale per affrontare le sfide poste dal *web scraping*. Le aziende operanti nel settore tecnologico devono adottare un approccio proattivo per garantire la conformità normativa, proteggendo i dati personali degli utenti e implementando misure di sicurezza avanzate. La cooperazione tra autorità garanti per la protezione dei dati personali e il settore privato rimane essenziale per mitigare i rischi associati a queste pratiche e promuovere un ecosistema digitale più sicuro e trasparente.

CONCLUSIONI

In conclusione, le recenti iniziative delle autorità garanti per la protezione dei dati personali dimostrano come il fenomeno del *web scraping* richieda una risposta coordinata da parte delle imprese e dei regolatori. La strategia raccomandata non si limita alla reazione a violazioni già avvenute, ma prevede un approccio proattivo e multilivello, che combina innovazione tecnologica, *policy* aziendali e *compliance* normativa.

AUTORI

Aurora Agostini

Partner

Giulietta Minucci

Counsel

Jessica Giussani

Associate

Giovanni Lombardi

Associate