Il ruolo della pseudonimizzazione nel GDPR

Author : Giuseppe Diretto

Date : 1 Marzo 2019

La pseudonimizzazione: questa sconosciuta

Pseudonimizzazione significa, rifacendosi all'etimologia letterale dal greco, *assegnare un nome falso a qualcosa*. Questa semplice ricostruzione comprende tutta la forza che l'utilizzo di uno pseudonimo può avere nella protezione dei dati personali; d'altra parte, l'utilizzo di uno pseudonimo è una pratica applicata comunemente nella nostra interazione con il web: il *nick name* è un esempio classico che dovrebbe servire, falsificando la nostra identità, a proteggerci ma che, spesso, è solo un *nome di battaglia* visto che, nei nostri post, riveliamo molto di più della nostra identità attraverso foto, filmati e così via.

Lo pseudonimo veniva (e viene) usato molto dagli autori e dai giornalisti anche se raramente serviva a proteggerne l'identità. L'ultimo esempio è quello di Elena Ferrante che, a quanto sembra, è lo pseudonimo utilizzato da un misterioso autore di romanzi che qualcuno arriva ad ipotizzare di sesso maschile.

Questo è un tipo di pseudonimo analizzato da uno degli ultimi documenti prodotti dall'European Network and Information Security Agency (ENISA)[i], denominato *"Recommendations on shaping technology according to GDPR provisions - An overview on data pseudonymisation"*[ii]. Il documento dedica, infatti, un paragrafo ai cosiddetti *self?chosen pseudonyms,* concludendo che non possono essere impiegati da titolare o responsabile del trattamento per tutelare i dati personali degli interessati.

A parte questo doveroso inciso, che sgombra il campo da uno dei tanti equivoci che si sono creati da quando è in vigore il Regolamento UE 679/2016 (d'ora in poi GDPR), il documento è una raccolta di elementi utili a comprendere la portata della pseudonimizzazione nella protezione dei dati personali. Come si sa, infatti, la pseudonimizzazione non era presente nel vecchio D.Lgs. 196/2003 (Codice in materia di protezione dei dati personali), nel cui Allegato B – Disciplinare tecnico in materia di misure minime di sicurezza (abrogato dal D.Lgs. 101/2018) si parlava di cifratura dei dati che, come vedremo, è una pratica differente.

La pseudonimizzazione è, quindi, una misura di sicurezza introdotta ufficialmente dal GDPR, che la definisce come:

"Il trattamento dei dati personali in modo tale che i dati personali non possano più essere attribuiti a un interessato specifico senza l'utilizzo di informazioni aggiuntive, a condizione che tali informazioni aggiuntive siano conservate separatamente e soggette a misure tecniche e organizzative intese a garantire che tali dati personali non siano attribuiti a una persona fisica identificata o identificabile"

e che lo stesso GDPR include tra le misure della *privacy by design,* inserendole al primo paragrafo dell'art. 25:

"Tenendo conto dello stato dell'arte e dei costi di attuazione, nonché della natura, dell'ambito di applicazione, del contesto e delle finalità del trattamento, come anche dei rischi aventi probabilità e gravità diverse per i diritti e le libertà delle persone fisiche costituiti dal trattamento, sia al momento di determinare i mezzi del trattamento sia all'atto del trattamento stesso il titolare del trattamento mette in atto misure tecniche e organizzative adeguate, quali la ***pseudonimizzazione****, volte ad attuare in modo efficace i principi di protezione dei dati, quali la minimizzazione, e a integrare nel trattamento le necessarie garanzie al fine di soddisfare i requisiti del presente regolamento e tutelare i diritti degli interessati".*

Appare, dunque, utile riprendere i consigli che l'ENISA ha fornito, cercando di osservarli nella prospettiva di una loro applicazione pratica all'interno delle realtà organizzative.

Pseudonimizzazione, anonimizzazione, crittografazione

Pseudonimizzazione, quindi, vuol dire sostituire i dati identificativi *veri* con dati identificativi *falsi* in maniera che:

- i terzi non possano associare i dati personali ad una persona fisica (interessato);
- il titolare o il responsabile del trattamento possano effettuare la riassociazione quando questo è necessario.

Queste caratteristiche conducono, quindi, a due corollari essenziali:

- il processo di pseudonimizzazione produce, a fronte di un dataset di partenza, due oggetti: il primo è un dataset che, per ogni interessato, contiene lo pseudonimo ed i dati personali che lo riguardano (ma che in nessun modo possono identificarlo) mentre il secondo è un dataset che contiene, sempre per ogni interessato, lo pseudonimo e i dati che ne permettono l'identificazione;
- il secondo dataset deve essere mantenuto separato dal primo, deve essere adeguatamente protetto, deve rimanere nella sola disponibilità del titolare o del responsabile del trattamento e deve essere utilizzato solo quando ciò sia strettamente necessario per le finalità previste.

Per fare un esempio, si supponga di avere un registro scolastico implementato tramite un foglio

elettronico e che contiene:

1. nome e cognome dell’alunno;
2. luogo e data di nascita dell’alunno;
3. indirizzo dell’alunno;
4. nome e cognome del padre;
5. nome e cognome della madre;
6. numero di fratelli dell’alunno;
7. ISEE del nucleo familiare;
8. sport preferito dall’alunno;
9. voti dell’ultimo trimestre.

Quali sono i dati del registro da pseudonimizzare? È intuitivo che **la risposta dipende dal contesto** di riferimento, come confermato anche dal documento ENISA. Certamente i dati del punto 1 sono *direttamente identificativi* ma possono considerarsi *indirettamente identificativi* anche i dati dei punti 2, 3, 4 e 5 se, per esempio, il titolare del trattamento fosse una scuola di un paese con 10.000 abitanti: in contesti così ristretti, infatti, conoscere il nome e il cognome della madre potrebbe facilmente consentire di identificare anche l’alunno. Questo vuol dire che il processo di pseudonimizzazione deve *spezzare* il dataset iniziale in due tronconi, che risulteranno così formati:

dataset1

1. pseudonimo;
2. numero di fratelli dell’alunno;
3. ISEE del nucleo familiare;
4. sport preferito dall’alunno;
5. voti dell’ultimo trimestre.

dataset2

1. pseudonimo;
2. nome e cognome dell’alunno;
3. luogo e data di nascita dell’alunno;
4. indirizzo dell’alunno;
5. nome e cognome del padre;
6. nome e cognome della madre.

È ovvio che il *dataset2*, conformemente al GDPR, costituisce l’*informazione aggiuntiva* per leggere correttamente il *dataset1* e che devono essere “conservate separatamente e soggette a misure tecniche e organizzative intese a garantire che tali dati personali non siano attribuiti a una persona fisica identificata o identificabile”. Ovviamente, le misure tecniche ed organizzative per proteggere le informazioni aggiuntive (cioè il *dataset2*) non potranno più far ricorso alla pseudonimizzazione ma dovranno essere di natura differente.

Per la verità, sempre con riferimento al contesto e per rendere ancora più difficile la questione

(peraltro non affrontata nel documento ENISA), i dati contenuti ai punti 7 ed 8 del registro potrebbero essere dati che identificano indirettamente un soggetto se, statisticamente, risultano ampiamente *fuori range* (tecnicamente *outlier*). Infatti, sempre nel caso di una scuola di un paese con 10.000 abitanti, il fatto che un alunno abbia 12 fratelli è un elemento *fuori range* che lo individua univocamente. Pertanto, risulta opportuno che il processo di pseudonimizzazione sia preceduto da un'analisi statistica accurata (sia per i dati quantitativi sia per quelli qualitativi) affinché siano individuati esattamente i dati che possono identificare gli interessati.

Cosa distingue la pseudonimizzazione dall'anonimizzazione? La possibilità di riassociare i dati personali ad un interessato: in caso di pseudonimizzazione questo è possibile (da parte del titolare o del responsabile facendo ricorso alle *informazioni aggiuntive*) mentre in caso di anonimizzazione questo non è più possibile. Il documento dell'ENISA specifica che i dati anonimizzati non sono più dati personali e non lo saranno mai più (*irreversibilità* del processo), purché l'anonimizzazione sia effettuata correttamente. Sempre per tornare all'esempio, se si vuole rendere anonimo il registro di partenza occorrerà cancellare qualsiasi dato personale che possa identificare, direttamente o indirettamente, l'interessato (ovvero i dati contenuti ai punti 1, 2, 3, 4, 5). Il documento ENISA, peraltro, specifica che una buona anonimizzazione, oltre alla cancellazione dei dati che identificano direttamente o indirettamente l'interessato, dovrebbe riportare, per quanto possibile, gli altri dati a range generici. Per tornare al caso dell'esempio, il numero di fratelli dovrebbe essere rappresentato non più da un numero esatto ma da una collocazione all'interno di intervalli: da 0 a 2, da 3 a 5, oltre 5.

Il rapporto tra pseudonimizzazione e crittografazione è, invece, un po' più difficile da spiegare perché, in alcuni casi, la crittografazione diventa funzionale alla prima fase della pseudonimizzazione (che non abbiamo ancora affrontato) ovvero al calcolo dello pseudonimo. Tuttavia, il documento ENISA riporta la crittografazione al processo che:

- *nasconde* tutti i dati personali, sia quelli che consentono l'identificazione sia gli altri;
- prevede, per natura, il processo inverso, ovvero la decrittografazione;
- si basa su sofisticati algoritmi matematici che trasformano l'insieme di bit contenente i dati originari in un altro insieme di bit utilizzando una chiave (a sua volta, costituita da un insieme di bit). Si tratta, quindi, di un complesso calcolo matematico che, tuttavia, soffre di una vulnerabilità intrinseca: se un *attaccante* viene a conoscenza della chiave e conosce il tipo di calcolo applicato (di solito sono algoritmi standard a livello internazionale) riesce ad accedere all'intero dataset crittografato.

Questo vuol dire che è opportuno **applicare congiuntamente** pseudonimizzazione (prima) e crittografazione (dopo, sulle informazioni aggiuntive).

Tante tecniche, risultati diversi

Una volta chiarito il significato di pseudonimizzazione, occorre soffermarsi sul primo passo per poterlo applicare: il calcolo dello pseudonimo. Il punto di partenza sono i dati da *nascondere*: questo è l'insieme di bit che deve essere *falsificato* ovvero trasformato in un altro insieme di bit. Il documento ENISA presenta alcune alternative per questo passaggio:

- l'**hashing semplice**: si tratta di applicare all'insieme di bit corrispondente ai dati da nascondere (dati che identificano, direttamente o indirettamente, l'interessato) un algoritmo matematico che lo trasforma in un altro insieme di bit che costituisce lo pseudonimo; l'algoritmo di hashing deve essere irreversibile e non dare risultati identici a partire da insiemi di bit differenti; queste caratteristiche sono rispettate da diversi algoritmi noti a livello mondiale - MD5, SHA?1, SHA?2 e SHA?3 - ma ENISA considera solo questi ultimi due sufficientemente solidi;
- l'**hashing con chiave**: si tratta di applicare un algoritmo di hashing che tuttavia tiene conto, nel calcolo, di un'ulteriore variabile scelta da chi effettua la pseudonimizzazione (una chiave numerica);
- l'**hashing *salato* e *pepato***; si tratta di applicare un algoritmo di hashing semplice all'insieme di bit corrispondente ai dati da nascondere che, però, è preventivamente mescolato secondo una certa logica scelta da chi effettua la pseudonimizzazione ed applicata a tutte le istanze da pseudonimizzare (per esempio, scambiare il posto di alcuni bit);
- la **crittografazione** come pseudonimizzazione: si tratta di applicare algoritmi di crittografazione all'insieme di dati che identificano l'interessato per generare lo pseudonimo; naturalmente esistono varie tecniche di crittografazione, principalmente suddivise in crittografazione *simmetrica* (con una sola chiave) e *asimmetrica* (con chiave pubblica e chiave privata);
- la **tokenizzazione,** ovvero l'assegnazione di un numero generato casualmente ad ogni istanza da pseudonimizzare, avendo cura di non riassegnare uno stesso numero già assegnato in precedenza.

Conclusioni

Sebbene la pseudonimizzazione sia uno degli elementi chiave per la protezione dei dati personali, il percorso per un'applicazione concreta e sistematica è lungo e difficile perché richiede un **profondo ripensamento** degli strati software che accedono ai dati.

Come già accennato, la semantica dei dati non è indifferente rispetto all'impiego di questa tecnica e, peraltro, non è indifferente nemmeno il contesto di riferimento. Questo significa che i produttori di software dovranno fornire apposite funzioni che consentano all'utente di configurare, secondo una valutazione dei rischi conforme al GDPR, i dati ai quali applicare la pseudonimizzazione e le modalità con le quali condurla.

Note

[i] https://www.enisa.europa.eu/.

[ii] https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions/at_download/fullReport.

Articolo a cura di **Giuseppe Diretto** e **Francesco Maldera**