GDPR: conservazione e pseudonimizzazione

L’articolo 32 del Regolamento (UE) 2016/679 del Parlamento Europeo e del Consiglio, più comunemente noto con l’acronimo GDPR, indica misure di sicurezza tecnico e organizzative che titolari e responsabili del trattamento, tenendo conto dello stato dell’arte e dei costi di attuazione, nonché della natura, dell’oggetto, del contesto e delle finalità del trattamento, del rischio per i diritti e le libertà delle persone fisiche, se del caso, mettono in atto per garantire un livello di sicurezza *adeguato* al rischio cui è esposto il trattamento dei dati personali.

L’indicazione non è prescrittiva e dev’essere valutata, coerentemente con il sopra citato articolo, in considerazione delle specifiche attività di trattamento. Nell’analisi delle richiamate misure tecniche emerge la complessità dovuta dall’adozione di meccanismi di pseudonimizzazione o de-identificazione, i quali non possono essere assoluti e devono considerare lo specifico contesto operativo delle attività di trattamento. L’utilizzo delle suddette misure tecniche, infatti, implica anche l’introduzione di ulteriori complessi meccanismi da attivarsi qualora l’informazione protetta debba successivamente essere ulteriormente trattata.

Risulta necessario ricercare e sviluppare soluzioni tecnologiche che non banalizzino l’applicazione di norme e regolamenti cogenti, finalizzati, come nel caso del GDPR, alla tutela delle persone fisiche con riguardo ai loro dati personali.

Dal punto di vista normativo, a livello ISO, sono in corso attività di normazione tese alla pubblicazione di un “Privacy enhancing data de-identification framework”. La de-identificazione è un mezzo per facilitare l’uso dei dati

personali, come definiti dall'articolo 4 del GDPR, in modo che non si possa identificare un individuo, o un gruppo di individui, o compromettere i suoi diritti e libertà con riguardo ai suoi dati personali. L'uso appropriato di tecniche di de-identificazione può sostenere la conformità con i requisiti normativi e i principi di "protezione dei dati".

La de-identificazione richiede, almeno, una valutazione delle informazioni aggiuntive a disposizione di un individuo, o di un gruppo, che può venire a conoscenza di dati personali in modo inappropriato, e come queste possono essere combinate per rivelare o scoprire dati personali.

Al fine di sapere quali informazioni aggiuntive saranno disponibili a un generico attaccante e conoscere dunque la possibilità di attacchi e motivazioni per la reidentificazione, la de-identificazione richiede sia una valutazione del contesto e delle circostanze in cui i dati personali saranno resi disponibili ai destinatari, sia una valutazione dei dati stessi al fine di determinare come le informazioni aggiuntive disponibili a un generico avversario possono essere utilizzate per rivelare o scoprire i dati personali.

Le norme in corso di definizione hanno lo scopo di fornire indicazioni per governare l'uso appropriato delle tecniche di de-identificazione dei dati descritte nella ISO/IEC 20889[1] che migliora la terminologia di de-identificazione dei dati e la classificazione delle tecniche. Questo quadro di de-identificazione può essere applicato in qualsiasi punto del ciclo di vita dei dati: dalla progettazione dei mezzi di raccolta dei dati, al riutilizzo interno, alla messa a disposizione dei dati a partner esterni oppure, anche, l'archiviazione. I destinatari dei dati possono quindi essere interni o esterni al custode dei dati che sta implementando procedure e pratiche di de-identificazione. In generale si possono identificare tre scenari operativi:

- l’utilizzo, ed il riuso, di dati personali, implica che il custode mantenga la supervisione sui dati de-identificati mentre li rende disponibili a un dipartimento interno o a un gruppo funzionale;
- la condivisione esterna implica che il custode mantenga la supervisione sui dati de-identificati mentre li rende disponibili a un destinatario di dati esterno (ad esempio, attraverso un portale web o un’applicazione in cloud);
- il rilascio esterno implica che il custode trasferisca la supervisione sui dati de-identificati a un destinatario di dati esterno.

In ognuno di questi tre scenari, il processo di de-identificazione stesso può essere trasferito a una terza parte, separata dal custode o dal destinatario.

Nel seguito sono investigate due possibili soluzioni tecnologiche di pseudonimizzazione compatibili con le “Linee Guida sulla formazione, gestione e conservazione dei documenti informatici” emesse da AgID il 9 settembre 2020 ed attuative dal 1° gennaio 2022, data a partire dalla quale è abrogato il previgente DPCM 3 dicembre 2013 di pari argomento.

La trasmissione del documento informatico ai fini della sua conservazione è affiancata dalla generazione dei c.d. metadati di conservazione, che caratterizzano il documento informatico anche da un punto di vista archivistico, andando a descrivere un insieme di proprietà che permettono la corretta tenuta del documento informatico nel tempo.

Del documento informatico, oltre a informazioni, come ad esempio la sua impronta digitale, o *hash*, che consente di verificare nel tempo l’integrità del documento stesso, i citati metadati si arricchiscono, in questa nuova definizione, di ulteriori dati come le informazioni relative agli stakeholder del documento quali l’autore, il mittente, il destinatario, l’assegnatario, l’operatore o altri ancora,

definiti dal nodo *Soggetti*.
Questi ruoli possono contenere dati personali, come definiti all'art. 4 del GDPR, che, per loro natura, non possono essere sottoposti a mascheramento senza vanificare il loro inserimento fra i metadati stessi. Queste informazioni sono esemplificazione di criticità a cui le organizzazioni, impegnate nella conservazione dei documenti informatici, devono far fronte con soluzioni tecnologiche sempre più evolute.
La creazione del Pacchetto di Versamento, la sua successiva trasformazione in pacchetto di Archiviazione e, infine, Pacchetto di Distribuzione, deve quindi considerare sia il rispetto dei vincoli normativi sia l'esigenza di garantire l'esecuzione dei diritti riconosciuti agli interessati, previsti negli artt.li dal 15 al 21 del GDPR, come il Diritto di accesso dell'Interessato (Art.15) o il Diritto alla cancellazione (Art. 17).
All'interno di ogni fase del processo di conservazione, tutti i dati personali, in transito o a riposo, devono essere opportunamente protetti giungendo anche al mascherato e alla loro de-identificazione al fine di garantirne la riservatezza, in un processo end-to-end.

Tale processo deve, dunque, mantenere la caratteristica di piena reversibilità, affinchè sia sempre possibile risalire al dato originale, ma rispettare i vincoli definiti dallo schema illustrato nell'Allegato 5 "I Metadati". Dal momento che i campi destinati a ricevere dati personali sono definiti come di tipo *string* (alfanumerico), il loro valore è potenzialmente di indefinita lunghezza almeno fino a quando non dev'essere memorizzato all'interno di basi dati o supporti tecnologici vincolati da restrizioni dimensionali.

Due sono le tecnologie di pseudonimizzazione applicabili a questo contesto:

- il Format-Preserving Encryption, e
- l'adozione di algoritmi crittografici a chiave

asimmetrica.

Il **Format-Preserving Encryption (FPE)**, è un algoritmo basato sull'implementazione dell'algoritmo AES a chiave simmetrica e di una funzione di permutazione pseudo-casuale atta a mantenere inalterato il formato del dato.

Nella sua applicazione è prevista la definizione preliminare di un dizionario descrittivo del dominio dei caratteri del testo da cifrare, il cui output è l'applicazione della funzione di permutazione pseudo-casuale che, attingendo dallo stesso dizionario, produce un testo cifrato della medesima lunghezza.
Questo accorgimento permette di integrarsi all'interno di servizi preesistenti o laddove il dato sia sottoposto a validazioni che limitano l'utilizzo dei caratteri a sottoinsiemi predefiniti come il formato delle carte di credito, i numeri telefonici, generalità, importi etc. La principale vulnerabilità di tale algoritmo risiede nella lunghezza del dato da trattare: più questo è di lunghezza minima, più risulta facile individuare la chiave di cifratura.

L'adozione di algoritmi crittografici a chiave asimmetrica, come ad esempio RSA o ECC, già ampiamente implementati nei comuni linguaggi di programmazione, che tuttavia generano un output di dimensioni più elevate rispetto all'input fornito. Questa tecnica è quindi da utilizzarsi in contesti in cui il dato cifrato non sia soggetto a validazioni restrittive o laddove l'informazione non sia memorizzata all'interno di campi limitati eccessivamente in lunghezza.

In entrambe le soluzioni non dev'essere trascurato l'effort computazionale di decifratura *a run-time* laddove il dato debba essere ricercato all'interno di una base dati: processo che può complicarsi laddove si vogliano applicare filtri di ricerca più complessi come il "contiene", "inizia/finisce con…" eccetera.

Quando l’aggiunta di un ulteriore strato di sicurezza non costituisca ostacolo implementativo, può essere adoperata una tecnica ancora più robusta e votata al totale controllo sui dati da parte dell’Interessato, nota come ***zero-knowledge encryption***. In base a questa tipologia di soluzione, adoperata già da alcuni provider di posta elettronica, l’Interessato è l’unico detentore della chiave privata per decifrare il dato, mentre la cifratura dei dati avviene prima che l’informazione sia trasmessa al Titolare (o Responsabile) del trattamento.
Una volta che l’informazione è stata mascherata, solo l’Interessato può applicare l’operazione di decifratura, a patto che la chiave privata sia custodita correttamente e mai persa.

Va da sé che il punto di forza dell’approccio *zero-knowledge* è anche quello di sua massima debolezza: laddove l’Interessato perdesse la chiave privata o la *passphrase* per recuperarla, nessuno sarebbe potrebbe risalire ai dati cifrati, paradossalmente rischiando di andare a violare il principio di integrità e disponibilità dei dati personali, normato dallo stesso GDPR e da altre norme come, ad esempio, la ISO/IEC 27001.

La scelta di una delle soluzioni precedentemente descritte comporta inevitabilmente da parte degli attori coinvolti l’utilizzo di accorgimenti tecnologici per la gestione e la generazione delle chiavi di cifratura/decifratura.
L’utilizzo della tecnica di mascheramento FPE ha come principale vulnerabilità la presenza di una singola chiave per la cifratura e decifratura dei dati, poiché si basa sulla condivisione di un “segreto”; il suo trattamento deve tenere conto quindi della possibilità da parte di chiunque in possesso di tale quantità di sicurezza di applicare la funzione di decifratura per risalire al dato in originale.

La tecnica a chiave asimmetrica, invece, introduce per

progettazione, l'utilizzo di due chiavi, una pubblica ed una privata, responsabili rispettivamente della cifratura e decifratura dei dati: un procedimento computazionalmente più sicuro e che sacrifica l'aspetto funzionale nell'integrazione in processi *legacy*.

Un approccio ***zero-knowledge encryption (ZKE),*** certamente più articolato nella sua strutturazione ed implementazione, pone l'interessato al centro di un più ampio processo di sicurezza, tanto da diventarne elemento imprescindibile. La sua implementazione permette di raggiungere traguardi di sicurezza per nulla banali e scontati, andando oltre la ormai, quasi, scontata cifratura del dato a riposo e riducendo notevolmente anche le conseguenze derivanti da un indesiderato cyberattack al repository dei dati. In questa tipologia di eventi, infatti, l'attaccante spesso guadagnato accesso a tutta la struttura informatica del suo bersaglio, vanificando dunque, ogni meccanismo di protezione. Con un processo ZKE ben congeniato, l'attaccante non avrebbe alcun vantaggio nel minacciare la diffusione di dati che nessuno al di fuori dell'interessato, soggetto diverso dal custode dei dati, è in grado di decifrare, oppure la cifratura di un dato già cifrato, poiché anche la chiave di decifrazione non sta nelle disponibilità del gestore del repository dei dati.

Note

[1] ISO/IEC 20889:2018 – Privacy enhancing data de-identification terminology and classification of techniques

Articolo a cura di **Davide Leonardi** e **Luciano Quartarone**



Luciano Quartarone, in oltre 20 anni nel mondo dell'ICT, ha sviluppato significative esperienze in diversi settori del mondo ICT, concentrandosi su PKI, firme elettroniche e sistemi di conservazione. Ha curato diversi importanti progetti di system integration, specializzandosi sui temi della sicurezza

delle informazioni, dell’analisi del rischio e della conformità a norme nazionali e internazionali. È auditor ISO/IEC 27001, di cui è anche trainer certificato presso il PECB e Senior Lead Implementer presso lo stesso ente. Membro del consiglio direttivo UNINFO, partecipa alla Commissione UNINFO UNI/CT 510 – Sicurezza e Commissione UNINFO UNI/CT 522 – eBusiness e Servizi Finanziari. Ha collaborato alla redazione di diverse norme nazionali in ambito sicurezza delle informazioni, long-term data preservation, distruzione dei supporti dati, definizione di profili professionali in ambito sicurezza delle informazioni e protezione dei dati personali. Da dicembre 2018 è CISO e Data Protection Director per il gruppo Archiva

https://www.ictsecuritymagazine.com/articoli/gdpr-conservazione-e-pseudonimizzazione/